Masterarbeit

im Studiengang
Data Science

Identifikation des Anwendungspotenzials Maschinellen Lernens in der allgemeinmedizinischen Versorgungsforschung

Strukturierung und prototypische Implementierung anhand eines Anwendungsbeispiels

vorgelegt von

Michael Anton Paulitsch

an der Hochschule der Medien Stuttgart am 16. Mai 2022

zur Erlangung des akademischen Grads eines

Master of Science

Erstprüfer:	Prof. Dr. Peer Küppers
Zweitprüfer:	Prof. Dr. Johannes Hartig

Ehrenwörtliche Erklärung

Hiermit versichere ich, Michael Anton Paulitsch, ehrenwörtlich, dass ich die vorliegende Masterarbeit mit dem Titel: „Identifikation des Anwendungspotenzials Maschinellen Lernens in der allgemeinmedizinischen Versorgungsforschung“ selbstständig und ohne fremde Hilfe verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen wurden, sind in jedem Fall unter Angabe der Quelle kenntlich gemacht. Die Arbeit ist noch nicht veröffentlicht oder in anderer Form als Prüfungsleistung vorgelegt worden.

Ich habe die Bedeutung der ehrenwörtlichen Versicherung und die prüfungsrechtlichen Folgen (§ 26 Abs. 2 Bachelor-SPO (6 Semester), § 24 Abs. 2 Bachelor-SPO (7 Semester), § 23 Abs. 2 Master-SPO (3 Semester) bzw. § 19 Abs. 2 Master-SPO (4 Semester und berufsbegleitend) der HdM) einer unrichtigen oder unvollständigen ehrenwörtlichen Versicherung zur Kenntnis genommen.

Offenbach am Main, 16.05.2022,

Kurzfassung

Anwendungen Maschinellen Lernens werden als Forschungsmethode in der Domäne der allgemeinmedizinischen Versorgungsforschung so gut wie nicht genutzt, während deskriptiv- und inferenzstatistische sowie qualitative Auswertungsverfahren dominieren. Dementsprechend liegt das Ziel dieser Masterarbeit zum einen in der Identifikation potenzieller Anwendungsgebiete Maschinellen Lernens in dieser Domäne und zum anderen auf Basis dieser identifizierten Anwendungsgebiete in der prototypischen Entwicklung, Anwendung und Evaluation eines für die allgemeinmedizinische Versorgungsforschung repräsentativen Anwendungsfalls.

Für die Identifikation dieser Anwendungsgebiete wurden eine systematische Literaturrecherche in medizinischen Fachdatenbanken, ein Screening allgemeinmedizinischer Fachzeitschriften sowie Interviews mit Expert:innen der allgemeinmedizinischen Forschung durchgeführt. Hieraus abgeleitete Anwendungsgebiete für Maschinelles Lernen können unter anderem in *Bedarfsermittlung von Patient:innen*, *Patientensicherheit* und *Health Literacy* gruppiert werden.

Aus dem Anwendungsgebiet der *Health Literacy* wurde für diese Arbeit ein prototypischer Anwendungsfall mit folgender beispielhafter Fragestellung abgeleitet: Welche Faktoren sind mit Falschangaben durch Patient:innen bei einer ärztlichen Konsultation assoziiert? In diesem Fall sind die Falschangaben auf die Nennung nicht ärztlich vergebener Diagnosen (*Overreporting)* sowie die fehlende Nennung ärztlich vergebener Diagnosen (*Underreporting*) bezogen.

Auf Basis eines entsprechenden Datensatzes wurde für die Beantwortung der Fragestellung eine inferenzstatistische Methode mit einer Methode Maschinellen Lernens theoriegetrieben aufeinander aufbauend verbunden: Anhand der Inferenzstatistik wurde geprüft, welche Faktoren mit einem *Over-* oder *Underreporting* als Outcome-Variable statistisch signifikant assoziiert sind und anhand Maschinellen Lernens, ob anhand dieser vorab identifizierten Faktoren ein *Over-* oder *Underreporting* von Patient:innen vorhergesagt werden kann. Hierbei wurden mehrere relevante Variablen identifiziert (wie das Ausmaß der *physischen Lebensqualität*), die allerdings nicht ausreichten um die Outcome-Variable angemessen vorherzusagen. Dies kann unter anderem auf die mangelnde Erhebung potenziell relevanter Variablen (wie *Health Literacy*) oder methodischer Probleme (ungleiche Häufigkeitsverteilung in der Outcome-Variable) zurückzuführen sein.

Zusammenfassend kann geschlussfolgert werden, dass Maschinelles Lernen als weitere theoriegetriebene Forschungsmethode ergänzend genutzt werden kann: Während beispielsweise inferenzstatistische Verfahren überprüfen, mit welchen Variablen die Ausprägungen einer Outcome-Variable assoziiert oder kausal verknüpft sind, können Verfahren Maschinellen Lernen testen, ob diese Variablen korrekt Ausprägungen einer Outcome-Variable vorhersagen können. Darüber hinaus könnten datengetriebene Methoden Maschinellen Lernens bei Verfügbarkeit großer Datenmengen Hinweise für die Verbesserung oder auch Entwicklung wissenschaftlicher Theorien liefern.

Abstract

Applications of machine learning are hardly used as a research method in the research of general medical health services, while descriptive and inferential statistics as well as qualitative evaluation methods dominate. Accordingly, the aim of this master's thesis is, on the one hand, to identify potential areas of application for machine learning in this domain and, on the other hand, based on these identified areas of application, to develop, apply and evaluate a prototype as a use case that is representative of general medical health services research.

To identify these areas of application, a systematic literature search in medical specialist databases, a screening of journals in general medicine and interviews with experts in general medicine were carried out. Derived areas of application for machine learning can for example be grouped into *assessment of patients' needs*, *patient safety* and *health literacy*.

For this work, a prototypical use case with the following exemplary question was derived from the field of application of *health literacy*: Which factors are associated with incorrect information provided by patients during a medical consultation? This incorrect information relates to the naming of diagnoses not made by a doctor (*overreporting*) and the failure to name diagnoses made by a doctor *(underreporting*).

On the basis of a data set, an inferential statistical method was combined with a machine learning method in a theory-driven way to answer the question: the inferential statistics were used to test which factors are statistically significant associated with over- or underreporting as an outcome variable and testing by machine learning whether these previously identified factors are able to predict over- or under-reporting in patients. Several relevant variables were identified (such as the extent of *physical quality of life*), but these were not sufficient to predict the outcome variable adequately. For example, this may be due to the lack of collection of potentially relevant variables (such as *health literacy*) or methodological problems (unbalanced distribution of values in the outcome variable).

In summary, it can be concluded that machine learning can be used as a further theory-driven research method: While inferential statistical methods, for example, check which variables are associated or causally linked with the levels of an outcome variable, machine learning methods can test whether these variables can predict levels of an outcome variable correctly. Further, data-driven methods of machine learning could provide indications for the improvement or development of scientific theories if large amounts of data are available.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Formelverzeichnis

1 Einleitung

Die Anwendung von Methoden Maschinellen Lernens hat sich über die Zeit hinweg in zahlreichen Domänen verbreitet[1&2]: beispielsweise sowohl in der Betrugserkennung in Form der Klassifizierung von Spam-Emails[3], im Transportwesen in Form der autonomen Steuerung von Fahrzeugen[4] als auch in prominenten Anwendungsbeispielen wie der automatischen Gesichtserkennung anhand von Künstlicher Intelligenz gestützter Kameras[5].

Eine breite Anwendung findet Maschinelles Lernen auch in den Wissenschaftsdomänen der *Gesundheitsversorgung* („Health Care“)[6&7]: Eine dieser Domänen stellt die Disziplin der Humanmedizin dar. Als prominente erfolgreiche Anwendungsbeispiele in dieser Domäne können die Klassifizierung von Hautkrebs anhand *Tiefer Neuronaler Netze*[8] oder die Vorhersage des Verlaufs einer Prädiabetes-Vorstufe zu Typ 2-Diabetes auf Basis elektronisch gespeicherter Routinedaten[9] genannt werden.

Im Laufe der vergangenen Jahre wurden in vielen humanmedizinischen Fachdomänen die dort jeweilige Anwendung Maschinellen Lernens in Form von Reviews und Übersichtsarbeiten reflektiert beziehungsweise diskutiert: zum Beispiel in der Augenheilkunde[10], der Gastroenterologie[11], der Kardiologie[12], der Neurochirurgie[13], der Nephrologie[14], der Neurologie[15] oder der Urologie[16]. Im Kontrast hierzu zeigte eine für diese Masterarbeit durchgeführte systematische Literaturrecherche keine entsprechende wissenschaftliche Übersichtsarbeit für die Fachdomäne der Allgemeinmedizin[17].

Somit zeigt sich, dass für die Allgemeinmedizin und den ihr zugehörigen Forschungsdomänen (wie die Versorgungsforschung und die Lehrforschung) das Potenzial Maschinellen Lernens weder systematisch aufbereitet und reflektiert, noch etwaige entsprechende Anwendungsgebiete aufgezeigt wurden. Dies steht im Kontrast zu der in diesem Abschnitt beschriebenen zunehmenden allgemeinen Bedeutung Maschinellen Lernens, insbesondere in der Humanmedizin und ihrer entsprechenden Fachdomänen.

[1] (Vgl. Chahar und Kaur 2020).
[2] (Vgl. Ray 2019).
[3] (vgl. Bhowmick und Hazarika 2018).
[4] (H. Nguyen et al. 2018).
[5] (Hassan und Abdulazeez 2021).
[6] (Doupe, Faghmous und Basu 2019).
[7] (Topol und Verghese 2019).
[8] (Esteva et al. 2017).
[9] (Anderson et al. 2015).
[10] (Armstrong und Lorch 2020).
[11] (A. Adadi, S. Adadi und Berrada 2019).
[12] (Dudchenko, Ganzinger und Kopanitsa 2020).
[13] (Donepudi 2020).
[14] (Lemley 2019).
[15] (Valliani, Ranti und Oermann 2019).
[16] (Salem et al. 2021).
[17] Die Literaturrecherche wird detailliert in Abschnitt 3.1 erläutert

Dementsprechend besteht das Haupziel dieser Masterarbeit in der Identifikation des Anwendungspotenzials Maschinellen Lernens als Forschungsmethode in der Domäne der allgemeinmedizinischen Versorgungsforschung. Der Fokus auf die Versorgungsforschung liegt darin begründet, dass der Aspekt der Versorgung den zentralen Inhalt allgemeinmedizinischer Tätigkeit darstellt (siehe Abschnitt 2.1).

In dieser Arbeit soll dargelegt werden, ob und in welchen Anwendungsgebieten der allgemeinmedizinischen Versorgungsforschung Methoden Maschinellen Lernens zur Lösung von Problem- beziehungsweise Fragestellungen herangezogen werden und dabei herkömmliche Foschungsmethoden (wie unter anderem die Inferenzstatistik) ergänzen oder sogar ersetzen könnten. Die daraus resultierenden Erkenntnisse sollen im Anschluss an einem konkreten Anwendungsfall durch eine prototypische Umsetzung exemplarisch dargestellt und beschrieben werden.

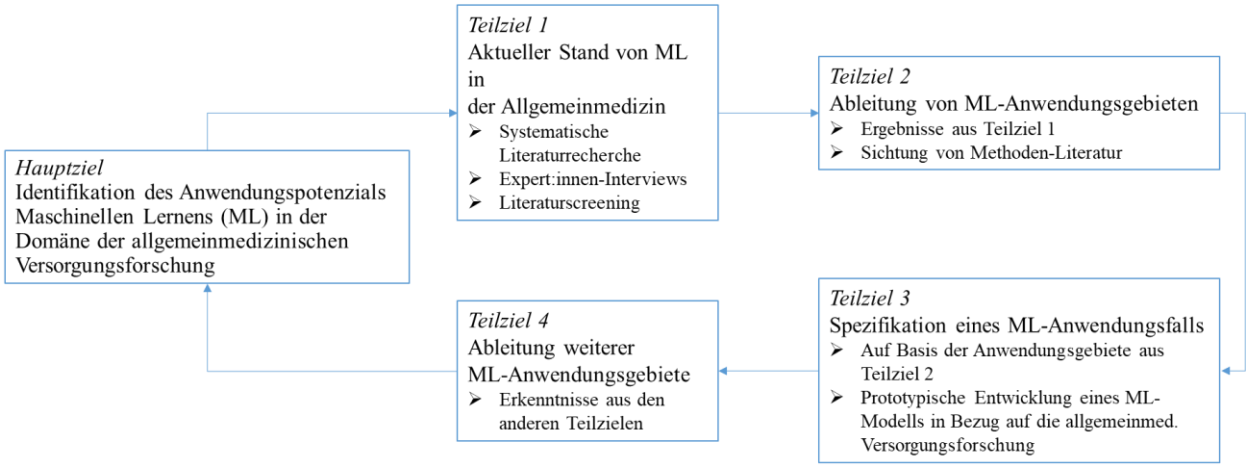

Abbildung 1: Haupt- und Teilziele dieser Masterarbeit und ihre Umsetzung

Dieses Hauptziel wird in vier Teilziele untergliedert (Abbildung 1), deren jeweilige Umsetzung im Folgenden erläutert wird:

Teilziel 1: Identifikation des aktuellen Stands des Einsatzes von Methoden Maschinellen Lernens in der Domäne, um den Wissenstand, die Anwendungsverbreitung als auch den jeweiligen Anwendungszweck systematisch zu erfassen. Dementsprechend wurde dieses Teilziel anhand von drei Schritten umgesetzt:

- Eine systematische Literaturrecherche zur Identifikation wissenschaftlicher Publikationen, die Maschinelles Lernen als Forschungsmethode in der allgemeinmedizinischen Versorgungsforschung eingesetzt haben.
- Ein Literaturscreening allgemeinmedizinischer Fachzeitschriften zur Identifikation der bislang in der Allgemeinmedizin grundsätzlich und maßgeblich eingesetzten Forschungsmethoden.

- Die Durchführung von Interviews mit Expert:innen der allgemeinmedizinischen Forschung zur Identifikation des aktuellen Stands Maschinellen Lernens in der allgemeinmedizinsichen Forschungspraxis.

Teilziel 2: Ableitung domänenspezifischer Anwendungsgebiete Maschinellen Lernens sowie eines repräsentativen Anwendungsfalls anhand folgender Schritte:

- Auf Basis der in der systematischen Litereraturrecherche als relevant identifizierten Publikationen
- Auf Basis von Aussagen Expert:innen in den durchgeführten Interviews.

Teilziel 3: Spezifikation, Durchführung und Evaluation eines für die Domäne repräsentativen Anwendungsfalls für den Einsatz Maschinellen Lernens:

- Suche nach einem geeigneten Datensatz für den ausgewählten Anwendungsfall
- Auswahl einer angemessenen Methode Maschinellen Lernens

Teilziel 4: Ableitung weitere domänenspezifischer Anwendungsgebiete für die Methoden Maschinellen Lernens:

- auf Basis der Systematisierungs- und Ableitungserkenntnisse der Teilziele 1 und 2
- der prototypischen Ergebnisse aus Teilziel 3

2 Grundlagen und Forschungsmethoden der Allgemeinmedizin

In diesem Kapitel werden die für das Erreichen der Ziele dieser Masterarbeit benötigten Konzepte und methodischen Ansätze beschrieben: Dazu zählen die Beschreibung der Domäne der allgemeinmedizinischen Versorgungsforschung sowie die Darstellung Maschinellen Lernens und weiterer Forschungsmethoden.

2.1 Die Domäne der allgemeinmedizinischen Versorgungsforschung

Allgemeinmedizinische Versorgungsforschung kann als ein Bestandteil der übergeordneten Domäne der Gesundheitsversorgung angesehen werden: Die Domäne der Gesundheitsversorgung umfasst die Vorsorge, Behandlung und das Management von Erkrankungen sowie die Aufrechterhaltung des mentalen und physischen Wohlbefindens mit Hilfe von Dienstleistungen medizinischer und mit diesen assoziierten Berufsgruppen[18]. Diese die Gesundheitsversorgung bereitstellenden Berufsgruppen entstammen unter anderem den Domänen der Psychologie, Pharmazie, Physiotherapie, Zahn- sowie der Humanmedizin.

Die Humanmedizin befasst sich im weitesten Sinne mit der Diagnostik, Prophylaxe und Therapie körperlicher und seelischer Erkrankungen des Menschen[19]. Hierbei bezeichnet Humanmedizin sowohl die Wissenschaft menschlicher Krankheiten als auch die klinisch-praktische Anwendung und kann nach Uexküll und Wesiack als die Summe aller „Regeln, Programme oder Rezepte“ verstanden werden, “die Menschen die Möglichkeit eröffnen, anderen Menschen zu helfen, ihre Gesundheit wiederzugewinnen und zu erhalten“[20]. Die Humanmedizin gliedert sich in zahlreiche internistische (wie die Kardiologie oder Gastroenterologie), chirurgische (wie die Herz- oder Abdominalchirurgie) und weitere (unter anderem Neurologie oder Pädiatrie) Fachdomänen.

Die in dieser Arbeit behandelte Fachdomäne der Allgemeinmedizin umfasst „die Grundversorgung aller Patienten mit körperlichen und seelischen Gesundheitsstörungen in der Notfall-, Akut- und Langzeitversorgung sowie wesentliche Bereiche der Prävention und Rehabilitation“[21]. Diese Grundversorgung von Patient:innen wird auch als Primärversorgung bezeichnet („Primary Care“): Sie gilt als erste Stufe der Krankheitsversorgung und Prävention für den Großteil der Bevölkerung[22] und

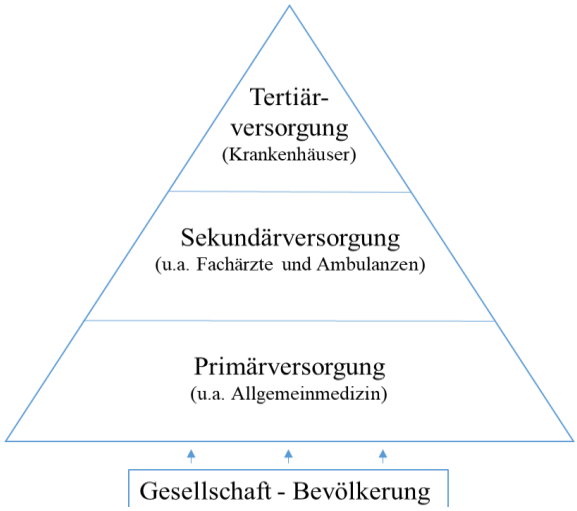

Abbildung 2: Drei Stufen medizinischer Versorgung (eigene Abbildung in Anlehnung an die fh Gesundheitsberufe OÖ[22])

[18] (Vgl. Harcourt 2008).
[19] (Vgl. DocCheck Medical Services GmbH 2022b).
[20] (Uexküll und Wesiack 1988, 607).
[21] (Deutsche Gesellschaft für Allgemeinmedizin und Familienmedizin e. V. 2002).
[22] (Vgl. fh Gesundheitsberufe OÖ).

wird neben allgemeinmedizinischen und internistischen Hausärzt:innen sowie von der Notfallmedizin übernommen. Von der Grundversorgung abzugrenzen ist die Sekundär- und Tertiärversorgung, bei der bei ersterer spezialisierte Fachärzte die ambulante Versorgung übernehmen und bei letzterer die stationäre Aufnahme in Krankhäuser einzuordnen ist (siehe Abbildung 2).

Dem Grund- beziehungsweise Primärversorgungsauftrag entsprechend liegt ein wesentlicher Teil der allgemeinmedizinischen Wissenschaft in der Versorgungsforschung: Versorgungsforschung wird allgemein als „ein multidisziplinärer Ansatz zur Erforschung der Umsetzung wissenschaftlicher Erkenntnisse in die Praxis der Gesundheitsversorgung hinsichtlich ihrer Wirkung auf Qualität und Effizienz in individueller und sozioökonomischer Perspektive“ definiert[23]. Zusammenfassend betrachtet kann Versorgungsforschung als an den Behandlungsauswirkungen orientiert beschrieben werden, um eine gute Gesundheitsversorgung zu gewährleisten[24].

Die Versorgungsforschung kann von anderen Forschungstypen wie der Grundlagen-, der epidemiologischen oder der klinischen Forschung abgegrenzt werden:

- Die Grundlagenforschung sucht nach Erkrankungsursachen[25] anhand biochemischer, genetischer und physiologischer Untersuchungen und betreibt Studien zu Arzneimittel- und Materialeigenschaften[26] ohne dabei einen direkten Anwendungsbezug zu beinhalten[27].
- Die epidemiologische Forschung untersucht zum einen die Ausbreitung gesundheitsbezogener Zustände (beispielsweise onkologischer Erkrankungen) und Ereignisse (wie kardiovaskuläre oder thromboembolische Ereignisse) in bestimmten Populationen, sowie zum anderen die mit diesen assoziierten Faktoren (wie genetische und umweltbedingte Risikofaktoren) und untersucht zum anderen die aus diesen Untersuchungsergebnissen resultierenden Anwendungen (wie Präventionsmaßnahmen)[28].
- Zu klinischer Forschung zählen unter anderem die Untersuchung der Auswirkungen konkreter, eher von der gesamten Versorgung isoliert zu betrachtender Therapien, wie die Gabe von Medikamenten oder die Anwendung chirurgischer Operationstechniken.

Eine typische Aufgabe der allgemeinmedizinischen Versorgungsforschung besteht entsprechend der Untersuchung von Behandlungsauswirkungen in der Analyse und Vorhersage eines Outcomes (wie beispielsweise die Krankheitslast chronisch kranker Patienten mit einer rezidivierenden depressiven Störung) und, inwiefern dieser aufgrund einer bestimmten allgemeinmedizinischen Versorgungsstrategie (zum Beispiel in Form von *Case Management*) verändert werden kann[29] (siehe

[23] (Schrappe et al. 2005, 1).
[24] (Vgl. Kuhlmey 2011, 918).
[25] (Kuhlmey 2011, 918).
[26] (Vgl. Röhrig et al. 2009, 263).
[27] (Vgl. DocCheck Medical Services GmbH 2022a).
[28] (Vgl. Porta 2014, 95).
[29] (Vgl. Gensichen et al. 2009).

Abbildung 3). Viele dieser Studien werden längsschnittlich durchgeführt, um das Ausmaß dieser Krankheitslast über den zeitlichen Verlauf hinweg zu untersuchen.

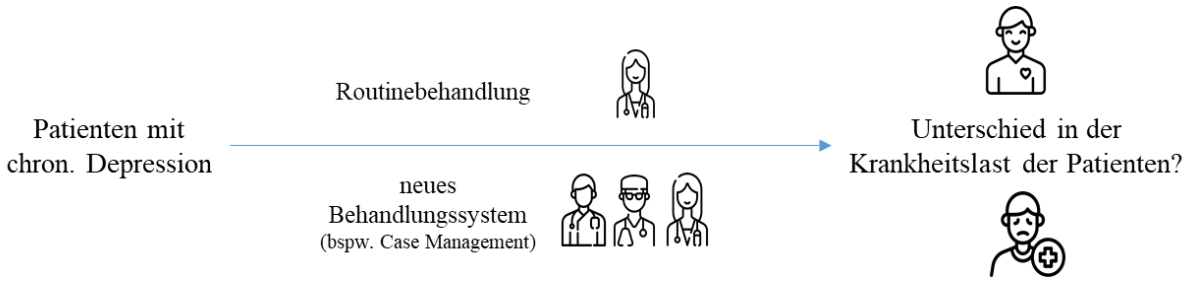

Abbildung 3: Ein repräsentatives Studiendesign in der allgemeinmedizinischen Versorgungsforschung

Um in diesen Studien interpretierbare Ergebnisse zu erhalten, müssen die allgemeinmedizinische Versorgungsforschung als auch alle anderen wissenschaftliche Domänen Forschungsmethoden nutzen, denn, wie unter anderem von Döring und Bortz formuliert[30], sei keine Wissenschaft ohne Forschungsmethoden möglich.

[30] (Vgl. Döring und Bortz 2016, 4).

2.2 Forschungsmethoden

Gemäß Eid[31] umfasse der wissenschaftliche Methodenbegriff „alle Mittel und Wege, die dem Erkenntnisgewinn und der praktischen Anwendung wissenschaftlicher Erkenntnisse dienen“. Unter diese Mittel und Wege können eine große Menge quantitativer als auch qualitativer Methoden fallen, die in den folgenden Abschnitten beschrieben und systematisiert dargestellt werden sollen (Überblick in Abbildung 4).

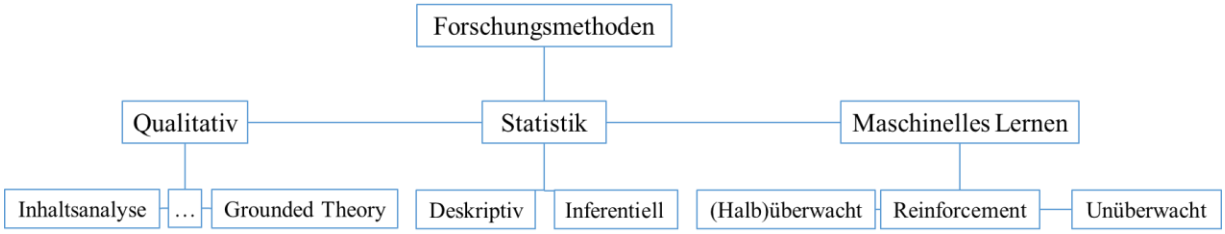

Abbildung 4: Gliederung von Forschungsmethoden

Allgemein kann die Anwendung von Forschungsmethoden in theorie- und datengetrieben unterteilt werden. Nach Maass et al. generiere theoriegetriebene Forschung Wissen auf Basis der Identifikation von Konstrukten und der unter ihnen bestehenden Beziehungen, wobei diese Konstrukte und ihre Beziehungen Abstraktionen spezifischer Phänomene[32] darstellen würden[33]: Vom Ablauf theoriegetriebener Forschung her würden zunächst Hypothesen entwickelt, diese anhand von gesammelten Daten analysiert und auf Basis der Ergebnisse theoretische Schlüsse gezogen. Bei theoriegetriebener Forschung würden zumeist geringe Datenmengen vorliegen.

Datengetriebene Forschung dagegen generiere Wissen anhand der explorativen Identifikation von Mustern unter anderem anhand der Analyse vorhandener Daten. Diese Muster-Identifikation basiere auf der Auswahl und Anwendung analytischer Methoden bei häufig sehr große Datenmengen.

Verschiedene Arten von Forschungsmethoden können je nach Anwendung theorie- oder datengetrieben eingesetzt werden, wie zunächst im folgenden Abschnitt anhand von Maschinellem Lernen gezeigt wird.

2.2.1 Grundlagen Maschinelles Lernens

Allgemein formuliert kann Maschinelles Lernen als die Wissenschaft beschrieben werden, Computern die Fähigkeit zu geben, selbstständig auf Basis von Daten zu lernen (zusammengefasst anhand von [34, 35 & 36]). Eine differenziertere Definition nach Mitchell[37] lautet, dass beim Maschinellen Lernen “a computer program is said to learn from experience *E* with respect to some task *T* and some performance measure

[31] (Eid, Gollwitzer und Schmitt 2017, 35).
[32] Bspw. ersichtlich an der Wandelbarkeit des Gesundheits- und Krankheitsbegriffs (vgl. Franke 2012).
[33] (Vgl. Maass et al. 2018, 1254–55)
[34] (Vgl. Géron 2019, Kap. 1).
[35] (Vgl. C. N. Nguyen und Zeigermann 2021, 3).
[36] (Vgl. Samuel 1959).
[37] (Mitchell 1997, 3–4).

P, if its performance on *T*, as measured by *P*, improves with experience *E*". Mitchell führt zur Veranschaulichung dieser Definition ein Beispiel an, wie ein Computer das Erkennen von Handschriften erlernt:

- *Task:* Das Erkennen und Klassifizieren von handgeschrieben Wörtern
- *Experience:* Eine Datenbank handgeschriebener Worte, die bereits klassifiziert wurden.
- *Performance Measure:* Die relative Anzahl korrekt klassifizierter Wörter.

Die Aufgabe des Computerprogramms bestünde demnach in dem Erkennen beziehungsweise dem Klassifizieren ihm noch nicht präsentierter handgeschriebener Wörter auf Basis des Lernens anhand einer Datenbank, in der Wörter bereits vorab klassifiziert wurden. Inwieweit das Computerprogramm diese Aufgabe erfüllen kann, wird anhand des Ausmaßes der korrekt klassifizierten Wörter erfasst.

Wie in diesem Beispiel gezeigt, kann Maschinelles Lernen zur Klassifikation (der Zuordnung beziehungsweise der Vorhersage kategorialer Daten in vorab gegebene Klassen) angewandt werden. Weitere Anwendungsmöglichkeiten bestehen in der Regression (eine Zuordnung beziehungsweise Vorhersage metrischer Daten) sowie zum Clustering (der Suche nach noch nicht gegebenen Klassen).[38]

Maschinelles Lernen gilt als ein Subtyp Künstlicher Intelligenz und wird unter anderem für die Entwicklung von Algorithmen intelligenter autonomer Systeme eingesetzt[39] und kann als Teil des Fachs der Informatik angesehen werden.

Allerdings grenzt beispielsweise De-yu[40] Maschinelles Lernen von traditioneller Programmierung in folgender Form ab (siehe Abbildung 5): Bei der traditionellen Programmierung würden einer Maschine Daten (zum Beispiel Daten eines Bewerbers um einen Kredit) und vordefinierte Regeln in Form eines Programmes (unter welchen Bedingungen zu erwarten ist, ob ein Kredit zurückgezahlt würde) eingespeist. Daraufhin gäbe die Maschine aus, ob von einer Kreditrückzahlung des Bewerbers auszugehen sei. Beim Maschinellen Lernen dagegen würden historische Daten vergangener Bewerber als auch ein entsprechender Output (ob von einem in der Vergangenheit liegenden Bewerber ein Kredit zurückgezahlt wurde) in eine Maschine eingegeben, die gemeinsam ein sogenanntes Vorhersage-Modell in Form eines Programms ergäben, das automatisch auf Basis der Daten eines neuen Bewerbers angäbe, wie wahrscheinlich die Kreditrückzahlung sei.

[38] (Vgl. Géron 2019, Kap. 2 & 3).
[39] (Vgl. Plaue 2021, 3).
[40] (Vgl. De-yu 2021).

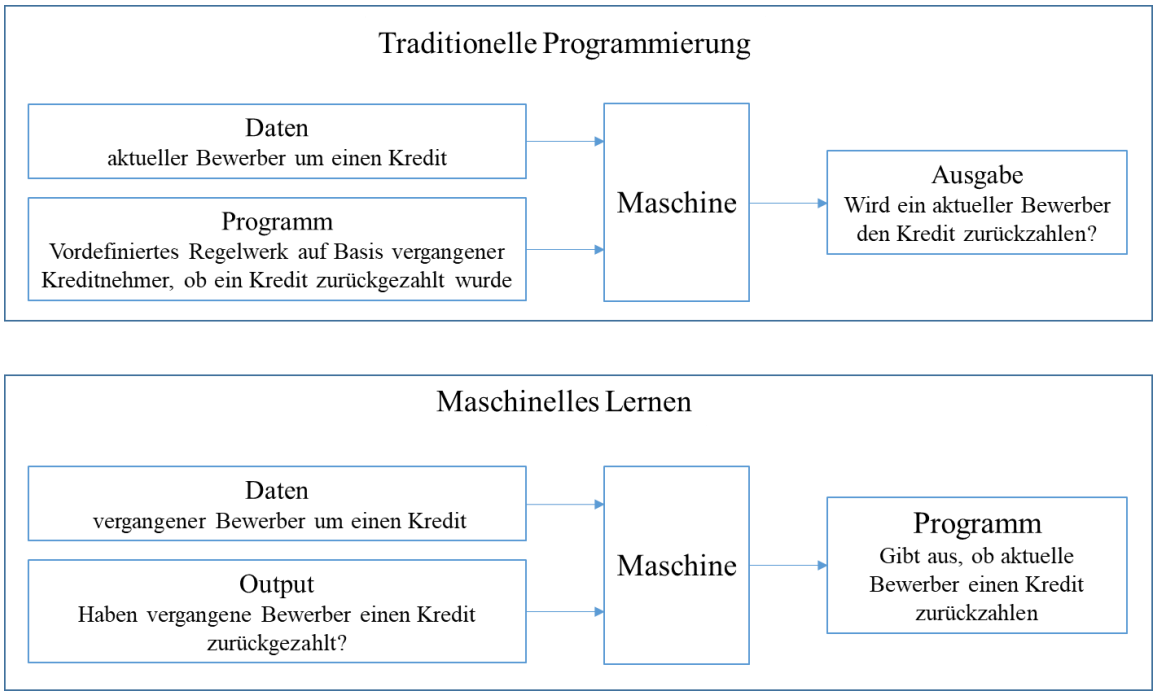

Abbildung 5: Der Unterschied Maschinellen Lernens zur traditionellen Programmierung (eigene Abbildung in Anlehnung an De-yu[41])

Neben diesen grundsätzlichen Eigenschaften Maschinellen Lernens, können dessen Anwendungen hinsichtlich verschiedener Forschungsansätze und Anwendungssysteme unterteilt werden. Diese Differenzierung wird in den folgenden Abschnitten beschrieben.

2.2.1.1 Datengetriebenes Maschinelles Lernen

Maschinelles Lernen kann in datengetriebener Form durchgeführt werden: Hierbei werden Muster nur auf Basis von Daten erkannt und weder Experten-Wissen noch Intuition fließt in diesen Prozess mit ein[42]. Nach Montáns et al. würden datengetriebene Methoden auf Basis von Rohdaten realer Beobachtungen einen unverzerrten impliziten Zugang zu Lernerfahrungen[43] ermöglichen[44]. Dies basiere nach Brunton und Kutz auf enorm großen und wachsenden Datenmengen, die durch günstige Sensoren, dem Anstieg der Rechenleistung sowie nahezu unbegrenzter Speicherkapazität und Transfer-Fähigkeiten verfügbar würden[45].

Als ein Beispiel für datengetriebene Forschung auf Basis Maschinellen Lernens in einer humanmedizinischen Domäne kann die Vorhersage des Verlaufs einer Prädiabetes-Vorstufe zu Typ 2-Diabetes bei Patient:innen angeführt werden[46]: Die entsprechenden Analysen wurden auf Basis eines Satzes elektronisch gespeicherter Gesundheitsdaten durchgeführt. Bei der Auswahl der für die Analysen

[41] (Vgl. De-yu 2021).
[42] (Vgl. Brunton und Kutz 2019, 4).
[43] „Lernen“ im Sinne Maschinellen Lernens
[44] (Vgl. Montáns et al. 2019, 846).
[45] (Vgl. Brunton und Kutz 2019, 9).
[46] (Vgl. Anderson et al. 2015).

zu berücksichtigenden Variablen wurde explizit auf vorher aufgestellte Hypothesen verzichtet und somit nicht auf vorhergehende Publikationen sowie Vorannahmen zurückgegriffen.

2.2.1.2 Theoriegetriebenes Maschinelles Lernen

Bei dem theoriegetriebenen Ansatz werden in Modelle Maschinellen Lernens Vorannahmen und Theorien integriert[47]. Dadurch könne nach Cox et al. ein Modell Muster in den Daten erkennen ohne eine anerkannte Theorie über das untersuchte Phänomen zu verletzen[48]. Nach Manhart würde hierbei versucht, eine Theorie in ein lauffähiges („runable") Programm zu transformieren[49]. Als konkretes Beispiel könne bei *Künstlichen Neuronalen Netzen* das Modell theoriegeleitet spezifiziert werden, indem der Domäne entsprechende Attribute ausgewählt und der Theorie entsprechend Knoten miteinander verbunden werden[50]. Somit könnten mit bekannten wissenschaftlichen Theorien übereinstimmende Modelle vor dem Lernen scheinbarer sowie verzerrter Muster geschützt werden und diese seien somit interpretier- als auch generalisierbar. Diese sei insbesondere bei Problemstellungen wichtig, die mit hohen Risiken wie im Gesundheitswesen einhergingen.[51]

Eine dementsprechend repräsentative theoriegetriebene Anwendung besteht in der Forschung zur Prävention von Suiziden: Nach Cox et al. [52] wäre ein sensitives (hoher Anteil korrekt positiver Klassifikation) und spezifisches (hoher Anteil korrekt negativer Klassifikation) Modell zu Vorhersagen von Suizidversuchen ein großer Erfolg. Aber solch ein Modell sei aufgrund von oft vorliegender mangelnder Interpretierbarkeit für die klinische Forschung nutzlos, wenn es auf allen möglichen Daten beruhe. Klinische Forschung suche nach Gründen für einen Suizid beziehungsweise dem Verstehen der unter Phänomenen liegenden Mechanismen und basiere auf Replizierbarkeit. Ein rein auf Daten basierendes Modell könne somit nicht an Kliniken zur Suizid-Prävention vermittelt werden.

Als ein theoriegetriebenes Beispiel Maschinellen Lernens in der Domäne der Suizid-Prävention kann eine Studie zur Vorhersage zukünftiger Suizidversuche herangeführt werden[53]: In dieser wurde auf die Ergebnisse vorangegangener Studien aufgebaut und eine auf diesen beruhende Vorauswahl potenziell relevanter Variablen (wie dem Vorhandensein einer Psychose oder rezidivierenden depressiven Episoden) getroffen. Dies hatte laut der Autoren zu einem Modell in Richtung einer exakten und skalierbaren Suizid-Risiko-Ermittlung geführt.

Als Vorteile des theoriegetriebenen Ansatzes führt Manhart unter anderem folgende Aspekte auf [54]: die Formalisierbarkeit eines Modells, die Möglichkeit zur Durchführung von Experimenten und bessere

[47] (Vgl. Cox et al. 2020, 2).
[48] (Vgl. Adombi, Adoubi Vincent De Paul, Chesnaux und Boucher 2021, 2683).
[49] (Vgl. Manhart 1996, 423).
[50] (Vgl. Karpatne et al. 2017, 2322).
[51] (Vgl. Karpatne et al. 2017, 2320–21).
[52] (Vgl. Cox et al. 2020, 3).
[53] (Vgl. Walsh, Ribeiro und Franklin 2017, 22071).
[54] (Vgl. Manhart 1996, 423).

Einsichten in die Funktionsweisen eines Modells, die die Undurchsichtigkeit von Computer-Modellen reduziere.

2.2.1.3 Interpretable Machine Learning und Performance

Wie aus dem vorangegangenen Abschnitt ersichtlich, werden theoriegetriebene Modelle Maschinellen Lernens mit dem Begriff der *Interpretierbarkeit* verbunden, die der Undurchsichtigkeit von Modellen entgegenwirken soll. In diesem Abschnitt wird dargelegt, was unter Interpretierbarkeit zu verstehen ist.

Die Angabe einer konkreten Definition *Interpretierbaren Maschinellen Lernens* (*Interpretable Machine Learning*) gestaltet sich als schwierig, da es nach Murdoch et al.[55] ein umfassendes aber auch wenig definiertes Feld sei. Sie selbst beschreiben dieses Feld als die Extraktion relevanten Wissens aus einem Modell Maschinellen Lernens, auf Basis der innerhalb der Daten bestehenden und durch das Modell gelernten Beziehungen.

Nach Masis[56] gehe *Interpretierbares Maschinellen Lernens* der Frage nach, inwiefern man einem Modell trauen und die Bedeutung eines Modell-Algorithmus darlegen könne („to explain the meaning of it"). Das Thema der Interpretierbarkeit sei relevant, da Anwendungen Maschinellen Lernens von Menschen programmiert würden und diese Anwendungen auf Basis eines Designs handelten und dies zu schwerwiegenden Konsequenzen führen könne: Als Beispiele hierfür seien in der Vergangenheit liegende und auf Modellen Maschinellen Lernens basierende ungerechtfertigte Verweigerungen von Begnadigungen im Strafvollzug, die fälschliche Entlassung von Schwerkriminellen sowie Aussagen, in Wahrheit schwer belastete Luft sei gefahrlos zu atmen, anzuführen. Entsprechende zu solchen Problemen führende Fehler und Schwachstellen in Modellen könnten anhand interpretierbarem Maschinellen Lernens korrigiert werden.

Masis differenziert *Interpretierbares Maschinelles Lernen* in drei Teilbereiche: Der *Fairness* (Ausschluss von Voreingenommenheit in Sinne eines *Discernible Bias*), der *Accountability* (ob die Verantwortung für Vorhersagen auf Etwas oder Jemanden verlässlich zurückgeführt werden können) und der *Transparenz* (ob erklärt werden kann, auf welche Weise und warum Vorhersagen zustande kamen).

Zu dem Teilbereich der Transparenz zählt Masis unter anderen die Aspekte der *Interpretierbarkeit* (*Interpretability*[57]) und der *Erklärbarkeit* (*Explainability*):

- Ein hohes Ausmaß an *Modell-Interpretierbarkeit* bedeute somit, dass die Schlüsse („inference") eines Modells auf eine von Menschen interpretierbare Weise beschrieben werden können: warum der Input eines Modells einen bestimmten Output hervorbringe.

[55] (Vgl. Murdoch et al. 2019, 22071).
[56] (Vgl. Masis 2021, Kap. 1).
[57] Masis unterscheidet zwischen dem Überbegriff des *Interpretable Machine Learnings* und dem untergeordneten Aspekt der *Interpretability*

- *Erklärbarkeit* umfasse alle Aspekte, die auch die Interpretierbarkeit betreffe. Es beinhalte zusätzlich die Transparenz im Sinne von für Menschen zugänglichen Erklärungen der inneren Prozesse eines Modells („inner workings“) und des Trainingsprozesses.

Interpretierbarkeit kann auf der globalen als auch der lokalen Ebene betrachtet werden[58]: Globale Interpretation beinhaltet das Verständnis des Einflusses der Input-Attribute auf das gesamte Modell. Um diesen Einfluss zu erfassen, gibt es verschiedene Methoden, zu denen nach Masis die *Feature Importance* zu den wichtigsten zähle. *Feature Importance* gibt den relativen Einfluss jedes Attributs auf eine Vorhersage eines Outcomes über den gesamten Datensatz hinweg gemittelt als ein einzelnes Gewicht in Form eines Wertes an[59]. Dieses wird aus der Erhöhung des Vorhersagefehlers bei Weglassen des jeweiligen Attributes abgeleitet. Lokale Interpretation dagegen betrifft das Darlegen des Einflusses der Input-Attribute auf individuelle Vorhersagen (also beispielsweise einer einzelnen Person).

Anwendungen Maschinellen Lernens können je nach Ausmaß der genannten Teilbereiche des *Interpretierbaren Maschinellen Lernens* in *Black* oder *White Box-Modelle* unterteilt werden. *Black Box-Modellen* fehle es nach Masis[60] an Transparenz, denn es sei nur der jeweilige In- und Output, aber nicht das Zustandekommen des Outputs aus dem Input beobachtbar: Auf konkrete Modelle bezogen kann man an *Künstlichen Neuronalen Netzen* als auch *Support Vector Machines* erkennen, dass sie schwierig zu erklärende komplexe mathematische Funktionen beinhalten[61]. Rudin bezeichnet *Tiefe Neuronale Netze* als *Black Box-Modelle* erster Ordnung, da sie hochgradig durch sich selbst definiert („highly recursive“) seien[62].

Transparente oder annähernd transparente Modelle werden dagegen als *White Box-Modelle* bezeichnet: Sie gelten als intrinsisch interpretierbar und ihre Funktionen seien näher an der menschlichen Sprache (hierzu zählen Methoden wie *Decision Trees*).[63]

Nach Lundberg et al. sollten Modelle unter anderem für Anwendungen in der Medizin sowohl *interpretierbar* als auch „accurate“ sein[64]. *Accuracy* ist ein Maß der *Performance* einer Anwendung Maschinellen Lernens: Die Performance gibt die Güte beziehungsweise Genauigkeit („correctness“) einer Vorhersage wider und diese Genauigkeit wird in *Performance*-Kennzahlen wie *Accuracy*, *Precison*, *Recall*, *F-Score* oder *Specificity* ausgedrückt[65]. Welcher Kennzahl-Typ als Maß dieser Güte genutzt wird, richtet sich danach, ob ein *Classifier* (ein Algorithmus zu Klassifikation diskreter Outcomes) oder ein *Regressor* (ein Algorithmus zur Vorhersage metrischer Outcomes) genutzt wird.

[58] (Vgl. Masis 2021, Kap. 2).
[59] (Vgl. Géron 2019, Kap. 7).
[60] (Vgl. Masis 2021, Kap. 1).
[61] (Vgl. Loyola-Gonzalez 2019, 154097).
[62] (Vgl. Rudin 2019, 207).
[63] (Vgl. Loyola-Gonzalez 2019, 154107).
[64] (Vgl. Lundberg et al. 2020, 57).
[65] (Vgl. Sokolova und Lapalme 2009).

Das Ausmaß dieser Güte kann neben der Datenqualität, der Verarbeitbarkeit der Datenmenge (in Form von Kosten für die Berechnungen „computation cost“ und Zeitaufwand) auch durch das Ausmaß der *Interpretierbarkeit* bedingt sein[66].

In verschiedenen Domänen können unterschiedliche *Performance*-Kennzahlen angemessen beziehungsweise wichtig seien. Es sei daher möglich, dass bestimmte Kennzahlen bei Algorithmen Maschinellen Lernens gut funktionieren, aber bei anderen suboptimal. Daher müssten Algorithmen über eine Bandbreite von *Performance*-Kennzahlen evaluiert werden.[67]

Solch eine Evaluation kann anhand sogenannter *Benchmark*-Datensätze standardisiert durchgeführt werden: An diesen kann beispielsweise festgestellt werden, ob eine Methode Maschinellen Lernens tatsächlich vorhandene Muster erkennt. Weiterhin können auch verschiedene Methoden hinsichtlich ihrer Stärken und Schwächen verglichen werden. [68]

Als ein Beispiel solch eines *Benchmark*-Datensatzes kann *Caltech-101* angeführt werden, an dem das Erkennen von Bildern durch Modelle getestet wird: In einer Übersichtsarbeit von Angelov et al.[69] wurden verschiedene Studien zu diesem *Benchmark*-Datensatz zusammengefasst und die *Performance* verschiedener Modelle Maschinellen Lernens der *Interpretierbarkeit* gegenübergestellt. Aus diesen Ergebnissen kann für diesen Datensatz geschlossen werden, dass *Black Box-Modelle* wie *Tiefe Künstliche Neuronale Netze* und *Support Vector Machines* zwar mit einer höheren *Performance* einhergehen und gleichzeitig mit einer niedrigen *Interpretierbarkeit* assoziiert sind. *Naïve Bayes* dagegen zeigt eine deutlich geringere Performance, aber eine höhere Interpretierbarkeit.

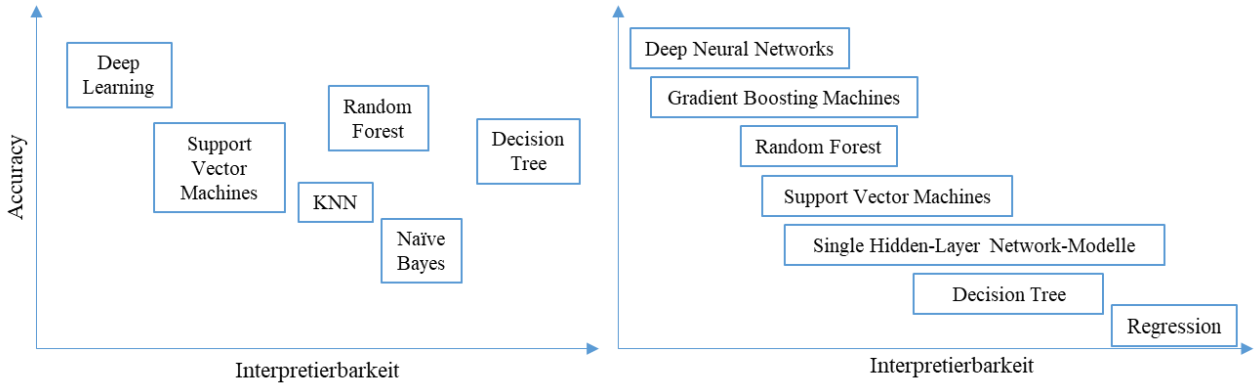

Abbildung 6: Performance im Verhältnis zur Interpretierbarkeit am Benchmark-Datensatz Caltech-101 (eigene Abbildung in Anlehnung an Angelov et al.[70]) und datensatzunabhängiger Darstellungen (eigene Abbildung in Anlehnung an Zhang[71])

[66] (Vgl. Gupta et al. 2021).
[67] (Vgl. Caruana und Niculescu-Mizil 2006).
[68] (Vgl. Olson et al. 2017, 1).
[69] (Vgl. Angelov et al. 2021).
[70] (Angelov et al. 2021).
[71] (Vgl. Zhang, o. J.).

Diese Ergebnisse stützen auch allgemeine datensatzübergreifende Zusammenfassungen von Methoden Maschinellen Lernens, die *Performance* und *Interpretierbarkeit* gegenüberstellen[72]: Wie aus diesen verallgemeinerten Ergebnissen ersichtlich, geht eine höhere Performance tendenziell mit einer niedrigeren *Interpretierbarkeit* einher (siehe Abbildung 6).

So kann geschlussfolgert werden, dass wenn eine hohe *Performance* in einer Anwendung als prioritär gegenüber der *Interpretierbarkeit* anzusehen ist, beispielsweise sogenannte *Deep Learning-Modelle* angewandt werden können[73]. Was unter diesen Modellen zu verstehen ist, wird im nächsten Abschnitt dargelegt.

2.2.1.4 Wide Learning- und Deep Learning-Modelle

Eine Unterscheidung unter anderem *Decision Trees* und *Künstliche Neuronale Netze* betreffend liegt in *Wide Learning*- und *Deep Learning-Modellen*[74&75]: Ein *Wide-Modell* bildet lineare Beziehungen ab und besitzt eine umfangreiche Menge miteinander vernetzter („cross-linked“) Features[76], die direkt mit einer Outcome-Variable in Beziehung gesetzt werden. In Abbildung 7 ist links ein Beispiel für solch ein Modell dargestellt.

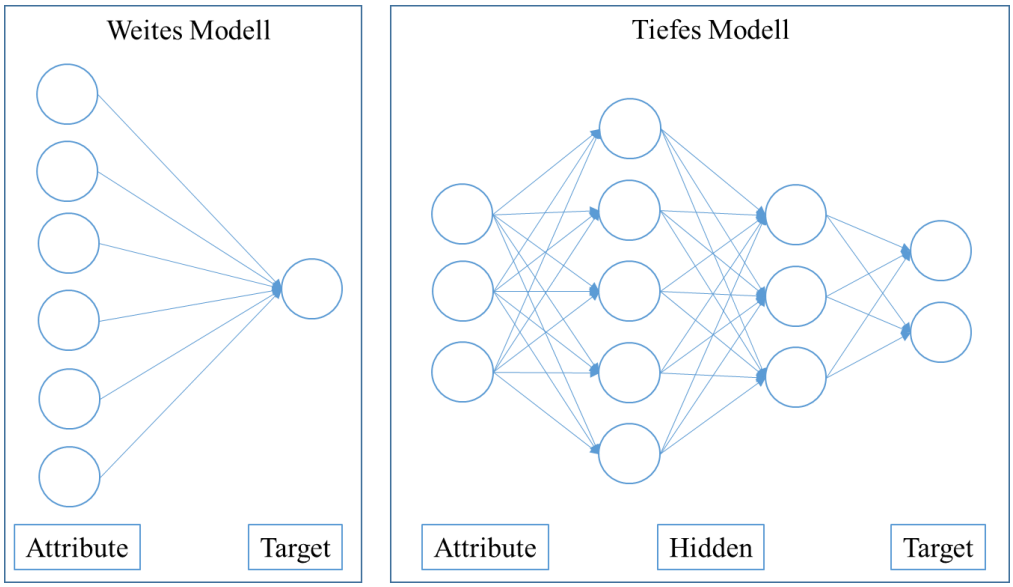

Abbildung 7: Visualisierung des Unterschieds zwischen Wide- und Deep-Modellen am Beispiel Künstlicher Neuronaler Netze

Deep Learning-Modelle bestehen aus einer Reihe von Schichten[77]. Da diese aufeinander aufbauen, ergibt sich, wenn sie graphisch abgebildet werden, eine Tiefe, durch die der Name *Deep Learning*

[72] (Vgl. Zhang, o. J.).
[73] (Vgl. Koleva 2020).
[74] (Vgl. Cheng 2016).
[75] (Vgl. Gour 2019).
[76] eine genaue Anzahl von Features wird nicht definiert, ab dem von einem Wide-Modell gesprochen wird
[77] (Vgl. Glassner 2021, Part 1).

entstanden sein soll[78]. Entsprechend fassen Patterson und Gibson *Deep Learning* derart zusammen[79]: *Deep Learning* befasse sich mit einem *Künstlichen Neuronalen Netz*, dass aus mehr als zwei Schichten bestehe. Diese Netze bestünden entweder aus einer von vier fundamentalen Netzwerk-Architekturen (wie *Convolutional* oder *Recurrent Neural Networks*) oder aus einer Variation der genannten Architekturen (wie ein *Hybrid Convolutional and Recurrent Neural Network*) und besäßen mehr Neurone als andere Netze (wie einem *Multilayer Feed-Forward Network*). Die Schichten seien in komplexerer Weise miteinander verbunden und benötigten eine "cambrian explosion" von Rechenleistung, um Modelle zu trainieren.

2.2.1.5 Systeme Maschinellen Lernens

In den vorangegangenen Abschnitten wurden grundlegende Eigenschaften Maschinellen Lernens beschrieben. Im folgenden Abschnitt sollen nun konkrete Methoden Maschinellen Lernens dargestellt werden. Diese können jeweils unterschiedliche Eigenschaften besitzen, die sie als ein bestimmtes *System* charakterisieren, wie beispielsweise ein *Online Model-Based Supervised Learning-System*. Was dies bedeutet wird in diesem Abschnitt dargestellt.

Überwachtes Maschinelles Lernen (Supervised Machine Learning): Dient der *Klassifikation* einer Zielvariable (*Target*) anhand von *Attributen* (mit einem zugeordneten Wert auch *Feature* genannt) und basiert auf der statistischen Auswertung eines Trainingsdatensatzes in Form einer Stichprobe mit bereits bekannter Kategorienzuordnung (*Labels*)[80&81]. Für diesen Ansatz bekannte Methoden sind zum Beispiel *Künstliche Neuronale Netze*, *Decision Trees*, *Support Vector Machines*, *k-Nearest Neighbors* sowie *lineare* und *logistische Regressionsanalysen*.

Beispiel: Eine bekannte Anwendung überwachten Lernens ist der Email-Spam-Filter. Zunächst werden Emails als *erwünscht* oder *unerwünscht* gekennzeichnet (*Labels* einer *Target*-Variable) und jeweils mit den *Features* mehrerer *Attribute* in Beziehung gesetzt (Betreff vorhanden oder nicht vorhanden, im Mail-Text „Viagra", Empfängerfeld an „Undisclosed recipients" gerichtet und so weiter). Neue Emails werden anhand dieser Informationen wiederum in *unerwünscht* oder *erwünscht* klassifiziert.[82]

Unüberwachtes Maschinelles Lernen (Unsupervised Machine Learning): Dient dem *Clustering*, ebenfalls einer Art *Klassifikation* allerdings ohne eine *Target*-Variable mit entsprechenden *Labels*. Die zu Grunde liegende Struktur des Datensatzes wird anhand der *Features* der *Attribute* erschlossen: Zum einen kann ein Datensatz hierbei anhand ähnlicher *Features* in *Klassen* (die dann den erlernten *Labels* entsprechen) eingeteilt werden. Zum anderen können diese Verfahren einer Dimensionsreduktion dienen, anhand derer hoch miteinander korrelierende *Features* zusammengefasst werden. Bekannte

[78] (Vgl. Goodfellow, Bengio und Courville 2016, 1–2).
[79] (Vgl. Patterson und Gibson 2017, Kap. 3).
[80] (Vgl. Plaue 2021, 189–94).
[81] (Vgl. Géron 2019, Kap. 1).
[82] (zum Beispiel vgl.Christina, Karpagavalli und Suganya 2010).

Methoden für das *Clustering* sind unter anderem *K-Means* sowie *DBSCAN* und für die Dimensionsreduktion die *Hauptkomponentenanalyse*.[83&84]

Beispiel: Klassifikation von Konversationen auf Twitter zur Analyse von Fehlinformationen hinsichtlich des therapeutischen Einsatzes des Antimalaria-Medikaments Hydroxychloroquin bei einer Covid-19-Infektion. Hierfür wurden über eine Millionen Tweets auf Twitter gesammelt und diese anhand einer Methode *unüberwachten Lernens* hinsichtlich ihres Inhalts in vier Klassen *geclustert*: In „Gerüchte" („rumours"), „Prävention und Therapie", „Maßnahmen von Regierungsseite" sowie „Verschwörung"[85].

Halb-überwachtes Lernen (semi-supervised Machine Learning): Wird eingesetzt, wenn nur ein Teil der Daten mit *Labels* versehen ist und der Trainingsdatensatz somit aus einer Kombination von Daten mit und ohne *Labels* besteht[86]. Dabei werden einander ähnliche Daten anhand *unüberwachten Lernens* in vorhandene *Labels* gruppiert (*pseudo-labeling*). Im Anschluss werden diese Einordnungen anhand von *überwachtem Maschinellem Lernen* überprüft. Die meisten Methoden zum halb-überwachten Lernen bestehen daher aus Kombinationen *unüberwachtem* und *überwachtem Lernens*[87]: Zum Beispiel basieren *Deep Belief Networks* auf unüberwachten *Restricted Boltzmann Machines*.

Beispiel: In Rahmen einer Studie zur Anwendung eines semi-überwachten Ansatzes wurde die Klassifikation einer Auswahl von Texten des Internetauftritts von BBC-Sport überprüft. Zehn Prozent der verwendeten Artikel erhielten ein *Label* („athletics", „cricket", „football", „rugby" und „tennis"). Anhand der *Kohonen Self Organizing Map*-Methode wurden die anderen Artikel mit Label versehen und im Anschluss diese Klassifikation anhand von Methoden überwachtem Lernens überprüft[88].

Reinforcement Learning: Lernen wird hierbei bei einem sogenannten *autonomen Agenten* durch Interaktion mit der Umwelt erzeugt. In dieser Interaktion wird die Verhaltensanpassung an der Maximierung eines spezifischen Signals der Umwelt (*Belohnung* versus *Bestrafung*) orientiert[89]. Dieser Ansatz soll in dieser Arbeit allerdings nicht weiter ausgeführt werden, da er nicht in das Schema der Anwendung von Forschungsmethoden passt.

[83] (Vgl. Géron 2019, Kap. 3).
[84] (Vgl. Plaue 2021, 255–57).
[85] (Vgl. Mackey et al. 2021).
[86] (Vgl. van Engelen und Hoos 2020, 374).
[87] (Vgl. Géron 2019, Kap. 1).
[88] (Vgl. Barman und Chowdhury 2020).
[89] (Vgl. Géron 2019, Kap. 1).

Zusätzlich kann die Art des Lernens der Systeme *überwachten Lernens* anhand folgender Eigenschaften charakterisiert werden[90]:

- *Online versus Batch-Learning:* Hiermit wird die Art des Trainings des Modells charakterisiert. Beim Batch-Learning wird das Modell außerhalb der Anwendung trainiert und das Training wird immer wiederholt, wenn neue Daten vorliegen. Beim Online-Learning dagegen werden neue Datensätze einzeln oder in Mini-Batches gleich in die Anwendung integriert.
- *Instance Based- versus Model Based-Learning:* Bei Modellen, die ein *Instance Based-Learning* betreiben wird ein neuer Datenpunkt mit denen im Trainingsdatensatz verglichen und anhand der Ähnlichkeit zu diesen eine Vorhersage beispielsweise in Hinblick auf eine *Klassifikation* vorgenommen. Eine entsprechende Methode ist *k-nearest Neighbours*. Beim *Model based-Learning* dagegen werden anhand des Trainingsdatensatzes Parameter geschätzt. Eine Zuordnung eines neuen Datenpunktes bei einer Klassifikation wird anhand dieser Parameter vorgenommen. *Support Vector Machines* nutzen beispielsweise diese Trainingsmethode.

[90] (Géron 2019, Kap. 1).

2.2.2 Klassische Forschungsmethoden der allgemeinmedizinischen Versorgungsforschung

Im vorangegangenen Abschnitt wurde Maschinelles Lernen als eine Forschungsmethode beschrieben. In der Forschung werden viele weitere Methoden eingesetzt, die in ihrer Entstehung zeitlich deutlich weiter zurückliegen[91]: Hierzu zählen insbesondere die deskriptive und inferentielle Statistik, aber auch qualitative Verfahren zur Auswertung von Interviews, Beobachtungen oder Gruppendiskussionen.

2.2.2.1 Statistische Methoden

Statistik kann allgemein als Lehre von Methoden zum Umgang mit quantitativen Informationen beschrieben werden[92]. Eine konkretere Definition nach Diaz-Bone[93] lautet, dass Statistik das Wissenschaftsgebiet sei, in dem Methoden und Techniken zur Analyse numerischer Daten entwickelt würden. Zusätzlich sei Statistik auch eine angewandte Wissenschaft, anhand derer ihre Methoden und Techniken in den empirischen Wissenschaften bei der Datenanalyse zum Einsatz kämen. Statistik kann in deskriptive als auch inferentielle Statistik aufgegliedert werden.

Deskriptive Statistik: Deskriptive Statistik umfasst alle Methoden, anhand derer empirische Daten zusammenfassend dargestellt und beschrieben werden können und zu diesen Methoden zählen Grafiken, Tabellen und Kennwerte[94]: Als Kennwerte gelten Maße zur zentralen Tendenz (wie der Modus, Median oder das Arithmetische Mittel), zur Streuung (insbesondere Varianz oder Range) aber auch Prozent- und Häufigkeitsangaben. Als Graphiken zur Beschreibung von Datenverteilungen werden unter anderem Box-Plots oder Balken- und Histogramme genutzt. Deskriptive Statistik kann sowohl zur explorativen Datenanalyse (Daten anhand von Darstellungen und Berechnungen nach Mustern und Zusammenhängen untersuchen[95]) eingesetzt werden, als auch zur Beantwortung wissenschaftlicher Fragestellungen.

Inferenzstatistik: In der Inferenzstatistik (auch schließende Statistik genannt) werden auf Basis von Stichprobendaten anhand der Überprüfung von Hypothesen und der Schätzung von Populationsparametern induktiv allgemeingültige Aussagen formuliert[96]. Was dies bedeutet, wird im folgenden Abschnitt erläutert[97]:

- *Schließen:* Nach Bandyopadhyay und Forster sei die einzig bedeutsame Aufgabe der Statistik, reliable Schlüsse zu ziehen. Dementsprechend formuliert und testet die Inferenzstatistik Aussagen über Grundgesamtheiten (Populationen) auf Basis von Stichproben.

[91] (Vgl. Lehmann 2011).
[92] (Vgl. Rinne 2008, 1).
[93] (Vgl. Diaz-Bone 2019, 12)
[94] (Vgl. Schäfer 2010, 131).
[95] (Vgl. Schäfer 2010, 99).
[96] (Vgl. Bortz 2005, 85).
[97] (Vgl. Bandyopadhyay und Forster 2011, 2).

- *Induktiv:* Nach der induktiv geleiteten Statistik werden aus einer Menge beobachteter Daten Schlüsse auf eine unbeobachtete Menge von Daten gezogen, allerdings ohne eine Sicherheit der Gültigkeit dieser Schlüsse.

Diese induktiven Schlüsse können sowohl theoriegetrieben als auch datengetrieben gezogen werden: Einerseits können theoretische Annahmen in Form von Hypothesen anhand von Stichproben hinsichtlich ihrer wahrscheinlichen Gültigkeit in der Population hin geprüft beziehungsweise getestet werden. Andererseits können ebenfalls auf Basis von Stichproben Mittelwerte oder Varianzen in Bezug auf ihre Gültigkeit als Populationsparameter hin geschätzt werden, ohne dabei theoriegetriebene Annahmen zu berücksichtigen.

Der Begriff der datengetriebenen Inferenzstatistik wurde in der Literatur bislang nicht systematisch aufbereitet. Allerdings kann aus der in den vorangegangenen Abschnitten vorgenommenen Beschreibung datengetriebener Forschung auf bestimmte inferenzstatistische Anwendungen geschlossen werden: Wie bereits wiedergegeben generiere nach Maass et al.[98] datengetriebene Forschung Wissen anhand der Identifikation von Mustern durch die Analyse vorhandener Daten. Das Vorgehen in Form von *Stepwise-Prozeduren* bei inferenzstatistischen Regressionsanalysen ähnelt dieser Beschreibung: Johnsson charakterisiert diese Prozeduren derart, dass deren Zweck reine Deskription, Vorhersage oder das Aufdecken von Beziehungen von Variablen sei. Das Hauptziel beinhalte, relevante Regressoren aus einer Anzahl von potenziell relevanten auszusortieren [99]. Dies bedeutet, dass bei diesem Vorgehen nur getestet wird, ob auf einen geschätzten Mittelwert, eine Varianz oder einem Zusammenhangsmaß in der Population als statistisch signifikant geschlossen werden kann.

Unter theoriegetriebener Inferenzstatistik kann verstanden werden, dass Vorannahmen beziehungsweise Theorien anhand statistischer Hypothesen auf Basis einer Stichprobe hinsichtlich ihrer Gültigkeit in der Population getestet werden können: Beispielsweise kann die aus Vorerfahrung und Vorannahmen abgeleitete Hypothese getestet werden, nach der bei einer Covid-19-Infektion Männer im Kontrast zu Frauen mit höherer Wahrscheinlichkeit einen schwereren Krankheitsverlauf erleiden. Dies kann anhand einer Zufallsstichprobe hinsichtlich der relativen Anzahl einer Einweisung von Männern auf eine Intensivstation und auftretender Todesfälle gemessen werden. Aus dieser Stichprobe können anhand inferenzstatistischer Methoden Populationsparameter geschätzt werden, die in diesem Fall den Zusammenhang zwischen der biologischen Geschlechtszugehörigkeit mit einem schweren Krankheitsverlauf ausdrückt: Laut einer Metanalyse besitzen auf Basis verschiedener Stichproben Männer ein statistisch signifikantes 2.84mal höheres Risiko hinsichtlich einer Einweisung auf eine Intensivstation und ein 1.29mal höheres Risiko zu versterben als Frauen[100]. Hierbei wurde die

[98] (Vgl. Maass et al. 2018, 1254).
[99] (Vgl. Johnsson 1992, 21).
[100] (Vgl. Peckham et al. 2020).

theoretisch abgeleitete Hypothese in Form statistischer Parameter anhand einer Stichprobe hinsichtlich ihrer Gültigkeit in der Population bestätigt.

Um solch induktive Schlüsse anhand inferenzstatistischer Forschungsmethoden ziehen zu können, herrsche laut Bandyopadhyay und Forster[101] eine vielfältige Debatte unter anderem über die Wahl eines *inferenzstatistischen Paradigmas*. Hierbei unterscheiden sie vier sich nicht alle gegenseitig ausschließende Paradigmen[102]: Dem *Klassischen* bzw. *Fehler-*, dem *Bayesianischen*, dem *Likelihood-* und dem *AIC-* Paradigma. Diese Paradigmen schließen sich in einer einzelnen Anwendung nicht gegenseitig aus, sondern können sich auch ergänzen.

Somit können zum Beispiel regressionsanalytische Verfahren auf Basis des *Klassischen* oder des *Bayesianischen Paradigmas* durchgeführt werden und bei beiden Paradigmen Modelle anhand des *AICs* und verwandter Indizes gegeneinander getestet werden.

Auf diesen Paradigmen basieren eine größere Menge an Forschungsmethoden, anhand derer man auf Unterschiede (Mittelwertsvergleiche anhand von *t-* und *U*-Tests sowie *Varianzanalysen*) oder auf Zusammenhänge (*Korrelations- und Regressionsanalysen* oder *Strukturgleichungsmodelle*) in der Population testen kann.

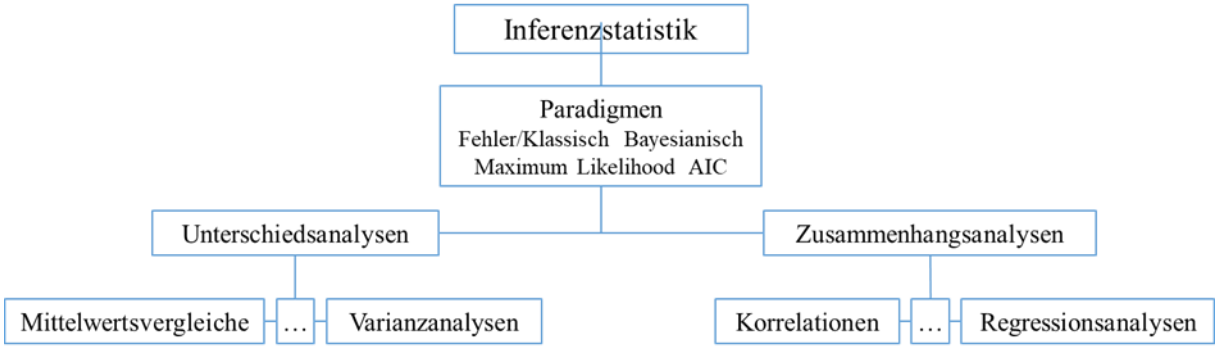

Abbildung 8: Gliederung inferenzstatistischer Methoden

Eine Übersicht inferenzstatischer Methoden wird in Abbildung 8 dargestellt und die Paradigmen im Folgenden kurz skizziert:

- *Fehler-Statistik (auch Klassische oder Frequentistische Statistik):* Unter diesem Paradigma werden alle Methoden zusammengefasst, die Fehlerwahrscheinlichkeiten nutzen. Diese Fehlerwahrscheinlichkeiten basieren auf relativen Fehlerhäufigkeiten bei wiederholten Stichprobenerhebungen, deren Nichtberücksichtigung zu der fälschlichen Akzeptanz beziehungsweise Verwerfung von Hypothesen führt. Das Paradigma der *Fehler-Statistik* kann

[101] (Vgl. Bandyopadhyay und Forster 2011, 2).
[102] allerdings wird von ihnen nicht definiert, was sie unter einem Paradigma verstehen, aber sie scheinen nicht den Wissenschafts-Paradigmen nach Thomas Kuhn zu entsprechen (vgl. Anand, Larson und J. T. Mahoney 2020).

selbst in mehrere Ansätze unterteilt werden[103], von denen als die beiden bekanntesten das *Null Hypothesis-Testing* (*Signifikanztest*) nach Ronald A. Fisher[104,105 & 106] und die *Decision-Theory* (*Hypothesentest*) nach Jerzy Neyman und Egon S. Pearson[107,108 & 109] zu nennen sind.

Trotz gewichtiger Unterschiede[110] ist beiden Ansätzen das Aufstellen und Testen beziehungsweise Verwerfen von Hypothesen anhand eines *p*-Wertes gemein: Bei einem statistischen Test dieses Paradigmas wird ein quantitativ zu interpretierender Wert berechnet (nimmt Werte zwischen 0% und 100% an), der als die bedingte kumulative Wahrscheinlichkeit der beobachteten (oder eines extremeren) empirischen Evidenz *E* (zum Beispiel ein *t*-Wert) gilt, unter der Annahme, die Nullhypothese H_0 sei wahr: $P(E|H_0)$.

- *Bayesianische Statistik:* Im Kontrast zur *Fehler-Statistik* werden in diesem Paradigma bedingte Wahrscheinlichkeitsaussagen *P* in Bezug auf Hypothesen *H* relativ zur empirischen Evidenz *E* getroffen: P(H|E). Hierbei kombiniert man Vorannahmen über die Gültigkeit der zu testenden Hypothesen (in einer sogenannten A-Priori-Wahrscheinlichkeit) mit empirischen Stichproben in einer A-posteriori-Wahrscheinlichkeit werden. Somit soll es zu einer reliableren Aussage über die Gültigkeit der Ergebnisse kommen[111].
- *Likelihood-Paradigma*: Basiert auf dem *Law of Likelihood*: Dieses Paradigma begründet sich darauf, eine Hypothese einer anderen als überlegen anzusehen, wenn die entsprechenden Daten eine hierfür statistische Evidenz bieten. Anhand von *Likelihood Rati*o-Maßen werde das Ausmaß der besseren Unterstützung einer Hypothese gemessen.
- *AIC-Paradigma:* Dieses bezieht sich auf die Auswahl eines passenden Modells. Das Kriterium gibt das Ausmaß an, in dem ein Modell in der Lage ist, die *Accuracy* von Vorhersagen zu maximieren und das eigentliche Signal vom Rauschen zu unterscheiden.

Qualitative Auswertungsmethoden: Verfahren dieser Gruppe von Forschungsmethoden werten qualitative Daten aus. Dabei handelt es sich um nicht-numerische Daten (verbale beziehungsweise textuelle oder visuelle). Diese Daten werden interpretierend (hermeneutisch) ausgewertet. Zu diesen Methoden zählen unter anderem die *Qualitative Inhaltsanalyse*, die *Grounded-Theory-Methodologie*, die *Narrative Analyse* und aber auch spezifische Verfahren wie qualitative Analysen von Videomaterial und Kinderzeichnungen.[112]

[103] (Vgl. Lehmann 2011).
[104] (Vgl. Gigerenzer 2004).
[105] (Vgl. Stang und Kowall 2020).
[106] (Vgl. Pernet 2015).
[107] (Vgl. Gigerenzer 2004).
[108] (Vgl. Stang und Kowall 2020).
[109] (Vgl. Pernet 2015).
[110] (Vgl. Gigerenzer 2004).
[111] (Vgl. Döring und Bortz 2016, 615–16).
[112] (Vgl. Döring und Bortz 2016, 599–601).

Eine Unterteilung in theorie- und datengetriebene Methoden ist schwer vorzunehmen: Nach Döring und Bortz seien die Methoden stark datengesteuert ausgerichtet, folgten allerdings einem hypothesen- und theoriebildenden Erkenntnisinteresse[113].

Eine differenzierte Perspektive stellen Kuckartz und Rädiker am Beispiel der *qualitativen Inhaltsanalyse* vor[114]: Die in diesem Verfahren vorgesehene Kategorienbildung könne deduktiv (vor der Auswertung) anhand einer Theorie festgelegt werden. Ein anderer Weg wäre, die Kategorien induktiv (anhand der gewonnen Daten) zu erstellen. Allerdings könne die induktive und auch deduktive Kategorienbildung unabhängig von Theorien gestaltet und sich auch am Alltagswissen oder subjektiven Erfahrungen orientieren und somit für die theoriengenerierende als auch beschreibende Datenanalyse verwendet werden.

2.2.3 Unterschiede der verschiedenen Forschungsmethoden

In den vorangegangenen Abschnitten wurden die Grundlagen verschiedener Forschungsmethoden beschrieben. In diesem Abschnitt sollen nun die Unterschiede hinsichtlich der Anwendungsmöglichkeiten in der Forschung erläutert werden: Doupe et al.[115] beschreiben die Anwendungsmöglichkeiten Maschinellen Lernens dahingehend, dessen Methoden zu Vorhersagen zu nutzen, wen oder was ein bestimmtes Outcome beträfe (zum Beispiel in Bezug auf das Ausmaß von Gesundheitskosten). „Causal Research“ ziele dagegen auf die Identifikation der die Outcomes verursachenden Faktoren ab (wie die Anzahl chronischer Erkrankungen).

Nach Bzdok et al.[116] besäße solch eine Vorhersage anhand Maschinellen Lernens keine Erfordernis, die ihr zugrunde liegenden Mechanismen zu verstehen, während Inferenzstatistik ein mathematisches Modell der Entstehung des Outcomes erschaffe, um entweder Hypothesen zu überprüfen oder ein System formal zu verstehen. Als erläuterndes Beispiel aus der Biologie wird von den Autoren angeführt, dass Methoden Maschinellen Lernens Personen mit einer Erkrankung identifizieren könnten, während statistische Methoden untersuchten, welche biologischen Prozesse mit der Dysregulation eines Gens bei der entsprechenden Erkrankung assoziiert seien. Allerdings fügen die Autoren hinzu, dass die Grenze zwischen Maschinellem Lernen und Inferenzstatistik schwammig sei und sowohl Methoden der Statistik als auch des Maschinellen Lernens für Vorhersagen genutzt werden könnten[117]. Zusätzlich zeigte sich in dem Abschnitt über *interpretierbares Maschinelles Lernen* (Abschnitt 2.2.1.3), dass auch hierbei mit einer Outcome-Variable relevant assoziierte Faktoren identifiziert werden können.

[113] (Vgl. Döring und Bortz 2016, 599).
[114] (Vgl. Kuckartz und Rädiker 2022, 70–103).
[115] (Vgl. Doupe, Faghmous und Basu 2019, 808).
[116] (Vgl. Bzdok, Altman und Krzywinski 2018, 233).
[117] (Vgl. Bzdok, Altman und Krzywinski 2018, 233–34).

Eine Abgrenzung der qualitativen Auswertungsverfahren wird aufgrund ihrer Komplexität, ihrer Anwendungsmöglichkeiten und dem begrenzten Rahmen dieser Studie nicht vorgenommen.

2.3 Zusammenfassung

Die in dieser Arbeit behandelte Domäne der allgemeinmedizinischen Versorgungsforschung beschreibt eine humanmedizinische Disziplin, die die wissenschaftliche Untersuchung der Grundversorgung von Patient:innen behandelt. Sie ist von anderen Feldern wie der Grundlagen-, epidemiologischen oder klinischen Forschung abzugrenzen, in denen Themen wie die Krankheitsentstehung oder pharmakologischen Wirkungen analysiert werden.

Die in diesem Kapitel dargestellten Forschungsmethoden können weitgehend anhand von zwei Eigenschaften beschrieben werden: Zum einen, ob sie theorie- oder datengetrieben genutzt werden und zum anderen, ob sie dem Maschinellen Lernen, der Deskriptiv- oder Inferenzstatistik oder qualitativen Auswertungsverfahren angehören.

Maschinelles Lernen kann tendenziell eher für Vorhersagen eines Outcomes verwendet werden, während Inferenzstatistik für eine stichprobenbasierte Schätzung von Populationsparametern eingesetzt wird, um das Zustandekommen eines Outcomes anhand anderer Variablen erklären zu können. Allerdings muss beachtet werden, dass diese Grenzen zwischen den Methoden als fließend anzusehen sind.

Auf Basis der in den vorangegangenen Abschnitten beschriebenen Begriffe wird im Folgenden nun versucht, das Hauptziel dieser Masterarbeit in Form der Identifikation des Anwendungspotenzials Maschinellen Lernens in der allgemeinmedizinischen Versorgungsforschung darzulegen. Dieses Hauptziel wurde in vier Teilziele untergliedert, deren jeweilige Umsetzung aufeinander aufbauend beschrieben werden.

3 Maschinelles Lernen in der allgemeinmedizinischen Versorgungsforschung

In diesem Teilziel wurde versucht festzustellen, in welchem Ausmaß Maschinelles Lernen in der allgemeinmedizinischen Versorgungsforschung verbreitet ist und angewendet wird. Es wurden drei Methoden genutzt, um diese Bestandsaufnahme zu erarbeiten:

1. Eine systematische Literaturrecherche in medizinischen Datenbanken, um die bisher veröffentlichten wissenschaftlichen Arbeiten in der allgemeinmedizinischen Versorgungsforschung zu identifizieren, die Maschinelles Lernen verwendet haben.
2. Ein Screening allgemeinmedizinischer Fachzeitschriften, um einen Überblick über die in dieser Domäne derzeit maßgeblich eingesetzten Forschungsmethoden zu erhalten.
3. Durchführung von Interviews mit Expert:innen der allgemeinmedizinischen Forschung, ob und in welcher Weise theoretisches Grundlagenwissen, praktische Anwendungserfahrungen sowie Überlegungen zum Einsatz des Maschinellen Lernens verbreitet sind.

Im Folgenden werden die Ergebnisse der drei Methoden zusammen mit der entsprechenden Vorgehensweise dargestellt.

3.1 Identifikation von Anwendungen Maschinellen Lernens in wissenschaftlichen Publikationen der allgemeinmedizinischen Versorgungsforschung

Das Ziel dieser Literaturrecherche bestand in der Erfassung aller relevanter Publikationen der allgemeinmedizinischen Versorgungsforschung, in denen Maschinelles Lernen als Forschungsmethode genutzt wurde. Die hier eingesetzte Suchstrategie orientiert sich an dem von Droste standardisierten Ablauf [118 & 119]: Dieser gliedert sich in acht Schritte, beginnend bei der Formulierung der Fragestellung bis hin zur Evaluation der eingesetzten Suchstrategie. Der formale Ablauf als auch seine Entsprechung für die Literaturrecherche wird in Abbildung 9 dargestellt und in den darauffolgenden Abschnitten beschrieben.

1. Recherchierbare Fragestellung: Zunächst wird eine recherchierbare Fragestellung erarbeitet. Für dieses Teilziel lautet sie wie folgt: „Wie viele und welche Studien nutzten in der Domäne der allgemeinmedizinischen Versorgungsforschung Maschinelles lernen als Forschungsmethode?“

2. Konzepterstellung: In diesem Schritt wurde zunächst berücksichtigt, dass die zu identifizierenden Studien einen bestimmten Forschungstyp (der Versorgungsforschung und nicht epidemiologische oder Grundlagenforschung) in einer konkreten humanmedizinischen Fachdomäne (der Allgemeinmedizin und keine hausärztliche Innere Medizin) betreffen, in dem eine bestimmte Forschungsmethode

[118] (Droste 2008).
[119] (Droste und Dintsios 2011).

eingesetzt wurde (Maschinelles Lernen und keine statistischen Verfahren). Nur eine Kombination der drei Konzepte gilt als relevanter Treffer.

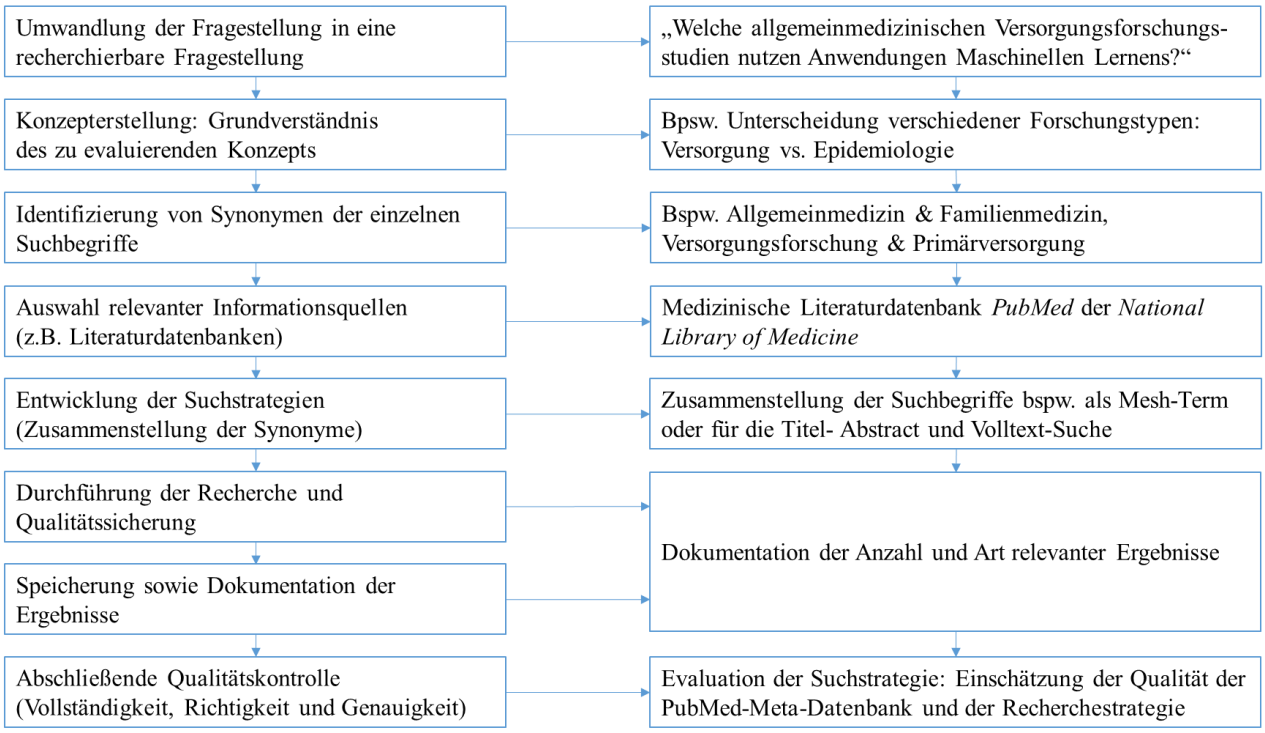

Abbildung 9: Formaler Ablauf der Literaturrecherche in acht Schritten links und die entsprechende Umsetzung in dieser Arbeit rechts

3. Identifizierung von Synonymen: Allgemeinmedizin kann auch durch andere medizinische Begriffe benannt werden, wie *Familienmedizin*. Versorgung und Versorgungsforschung betrifft auch die *Primärversorgung*, und dieser Begriff sollte daher auch in der Literaturrecherche verwandt werden.

4. Auswahl relevanter Informationsquellen: Für diese Literaturrecherche wurde die Meta-Datenbank *PubMed*[120] ausgewählt. Durch die dort eingegebenen Suchanfragen wird unter anderem auf die medizinische Fachdatenbank *MEDLINE* sowie dem *PubMed Central-Archiv* zugegriffen. Die Suche erfolgt anhand Englischer Begriffe.

5. Entwicklung der Suchstrategie: Die unter Punkt 3 identifizierten Synonyme wurden - wenn vorhanden - als Mesh-Terms (Schlagworte für die Katalogisierung von Artikeln in *PubMed*) eingegeben oder als Suchbegriffe für alle Felder wie Titel, Abstract, Volltext und weitere genutzt, um ein breites Feld von Publikationen erfassen zu können. Die Suche wurde auf bis Ende 2021 begrenzt. Der konkrete Suchbefehl wird in Abbildung 10 dargestellt.

[120] (National Center for Biotechnology Information 1988).

```
((“General Practice”[Mesh]) OR (“Primary Health Care”[Mesh]) OR (“Ambulatory Care”[Mesh]) OR “primary care”[All Fields]) OR (“family practi*”[All fields]))
AND (“Machine Learning”[Mesh])
AND (2011:2021[pdat])
```

Abbildung 10: Der Suchbefehl für diese Literaturrecherche

6. und 7. Durchführung der Recherche und Dokumentation der Ergebnisse: Auf Basis dieser Suchstrategie ergaben sich 428 Publikationen. Hierbei zeigte nur eine einzelne einen Bezug zur Versorgungsforschung in der Allgemeinmedizin: Das Ziel dieser gefundenen Studie bestand in der Ermittlung des Vorhersage-Potenzials eines asynchronen Teleconsulting-Services, der in Katalonien seit 2015 zur Kommunikation zwischen Allgemeinmedizinern und Patient:innen besteht[121].

Die anderen gefundenen Publikationen lassen sich anderen Domänen zuordnen und großteilig hinsichtlich des Anwendungsziels Maschinellen Lernens gruppieren: Zur Identifikation eines Risikos für eine Erkrankung oder eines Ereignisses (wie einem Schlaganfall), als Unterstützungstool für die Diagnostik oder im Rahmen von Versorgungsforschung in anderen nicht direkt mit der Allgemeinmedizin vergleichbaren Domänen:

1. Risikoidentifikation für bestimmte Erkrankungen auf Basis Maschinellen Lernens anhand von Daten aus allgemeinmedizinischen Praxen. Diese Risikoidentifikation fällt somit in den Bereich der epidemiologischen Forschung und nicht der Versorgungsforschung:
    - Die Ermittlung des Risikos eines Schlaganfalls bei Patient:innen in Zusammenhang mit einem bestimmten Typ oraler Antikoagulation, die in allgemeinmedizinischen Praxen behandelt wurden[122]
    - Die Vorhersage des Lungenfunktionsverlustes von Patient:innen mit COPD, deren Daten unter anderem am *UK Royal College of General Practitioners* gesammelt wurden[123]
2. Diagnostische Unterstützung: Eine weitere Gruppe von Publikationen betrifft die Entwicklung von Anwendungen, die unter anderem Allgemeinmediziner:innen bei der Diagnosestellung Unterstützung geben können. Dies betrifft somit eher die Klinische Forschung, als die Versorgungsforschung und betrifft nicht nur die Allgemeinmedizin.
    - Die Risiko-Einschätzung hinsichtlich pigmentierter Hautläsionen per Smartphone[124]
    - Das frühzeitige Erkennen von Demenz bei Patienten anhand elektronischer Patientendaten[125]
    - Diagnostik von Patienten mit der Parkinson-Krankheit anhand ihrer Sprache[126]

[121] (Vgl. López Seguí et al. 2020).
[122] (Vgl. Kostev et al. 2021).
[123] (Vgl. Nikolaou et al. 2021).
[124] (Vgl. Chin et al. 2020).
[125] (Vgl. Ford et al. 2019).
[126] (Vgl. Carrón et al. 2021).

3. Versorgung: Es wurden mehrere Studien gefunden, die das Thema Versorgungsforschung betrafen, sich allerdings nicht auf allgemeinmedizinische Praxen beziehen, sondern eher spezifische Probleme anderer Versorgungssysteme betrachteten:
    - Die Verlässlichkeit von Patient:innen hinsichtlich der Wahrnehmung von Arztbesuchen im US-amerikanischen Gesundheitsversorgungssystem für Armee-Veteranen (*Veterans Affairs*) wahrzunehmen[127]
    - Das Risiko einer kurzzeitigen Wiedervorstellung von Patient:innen in einer Notaufnahme nach vorheriger Entlassung[128]
    - Die Identifikation von Patient:innen, die in US-amerikanischen *Ambulatory Care Centers*[129] unpünktlich zu Terminen erscheinen könnten[130]
    - Die Identifikation von Medikations-Fehlern in medizinischen Zentren in den USA[131]

Somit kann festgehalten werden, dass bis auf eine Ausnahme, in der Domäne der allgemeinmedizinischen Versorgungsforschung noch keine Studien mit Forschungsmethoden aus dem Bereich des Maschinellen Lernens durchgeführt wurden.

8. Qualitätskontrolle: Als ein Qualitätsmerkmal der durchgeführten Literaturrecherche und der aufgeführten Ergebnisse kann folgende Studie aus dem Jahr 2019 herangeführt werden: Es wird in ihr bestätigt, dass eine Qualitätskontrolle durch Förderung und Wartung beider Datenbanken (*Pubmed Central-Archiv* und *Medline*) auf die eine *Pubmed*-Suche zugreift, gesichert sei[132].

3.2 Überblick und Verbreitung von Forschungsmethoden in aktuellen Studien der allgemeinmedizinischen Versorgungsforschung

Aufgrund der Ergebnisse der systematischen Literaturrecherche kann festgehalten werden, dass Maschinelles Lernen so gut wie keine Anwendung als Forschungsmethode in der Domäne der allgemeinmedizinischen Versorgungsforschung gefunden hat.

Dieser Abschnitt beschäftigt sich daher mit der Frage, welche Forschungsmethoden statt Maschinellem Lernen bislang in dieser Domäne Anwendung gefunden haben. Auf Basis der Antworten auf diese Frage kann erfasst werden, welchen Forschungsmethoden Maschinellem Lernen gegenüberzustellen ist, um dadurch dessen Potenzial als Ergänzung oder Ablösung herkömmlich angewandter Methoden ableiten zu können. Dementsprechend wurde ein Literaturscreening durchgeführt, um explorativ einen Überblick zu erhalten, wie häufig welche Art von Forschungsmethoden in der gesamten Domäne der

[127] (Vgl. Wong, Schuttner und Reddy 2020).
[128] (Vgl. Chmiel et al. 2021).
[129] Eher als Zentrum mit verschiedenen Facharztrichtungen zu verstehen
[130] (Vgl. Srinivas 2020).
[131] (Vgl. Schiff et al. 2017).
[132] (Vgl. Ossom Williamson und Minter 2019).

allgemeinmedizinischen Forschung eingesetzt wurden. Dabei wird der Fokus auf die Allgemeinmedizin grundsätzlich gelegt und nicht nur auf die Versorgungsforschung.

Hierfür wurden die Publikationen von drei allgemeinmedizinischen Fachzeitschriften betrachtet: Das *European Journal of General Practice (EJGP)*[133], das *British Journal of General Practice (BJGP)*[134] sowie die *Zeitschrift für Allgemeinmedizin (ZfA)*[135]. Auf diese Weise wurden die maßgeblichen Publikationen auf europäischer als auch deutscher Ebene abgedeckt und auch die Nähe zum deutschen Gesundheitssystem bewahrt. Daher wurden keine US-amerikanischen Fachzeitschriften in das Screening mit aufgenommen. Es wurden die Jahre 2020 und 2021 in das Screening eingeschlossen, da sie als repräsentativ für aktuelle Entwicklungen angesehen werden können. Alle Originalartikel (im *BJGP* als „Research Article" bezeichnet) der Jahre 2020 und 2021 wurden hinsichtlich der für die Beantwortung der in dem Artikel jeweiligen Fragestellung genutzten Methode durchsucht und kategorisiert.

Tabelle 1: Anzahl der gescreenten Artikel je Zeitschrift und Jahrgang

Zeitschrift	*ZfA*		*BJGP*		*EJGP*		*Summe*
Jahr	2020	2021	2020	2021	2020	2021	
Anzahl der Artikel	39	30	102	111	17	32	331

In der Summe wurden über alle Zeitschriften hinweg 331 Originalartikel betrachtet (siehe Tabelle 1). Hierbei zeigte sich, dass die verwendeten Forschungsmethoden maßgeblich in deskriptiv- und inferenzstatistische Verfahren (zumeist verschiedene Formen *linearer Modelle* und *logistischer Regressionsanalysen*) sowie qualitative Auswertungsverfahren (maßgeblich in Form von Auswertungen von Interviews und Gruppendiskussionen) unterteilt werden können. Unter den als *Sonstige* klassifizierte Verfahren fielen die Anfertigung systematischer Reviews beziehungsweise Literaturrecherchen oder Lehrveranstaltungsentwicklungen (Tabelle 2).

Tabelle 2: Verwendete Forschungsmethoden je Jahr und Zeitschrift

	Deskriptive Statistik		*Inferenzstatistik*		*Qualitative Auswertung*		*Sonstige Verfahren*	
	2020	*2021*	*2020*	*2021*	*2020*	*2021*	*2020*	*2021*
EJGP	9	14	11	18	5	13	0	0
BFGP	42	33	50	54	33	46	10	7
ZFA	26	23	12	4	13	8	3	2
Gesamt	77	63	73	76	51	63	13	9

Die Anzahl der gezählten Methoden übersteigt die Anzahl der gescreenten Publikationen, da in einigen für die Beantwortung von Fragestellungen verschiedene Ansätze verwandt wurden: Zum einen

[133] (WONCA Europe, o. J.).
[134] (Royal College of General Practitioners, o. J.).
[135] (Deutsche Gesellschaft für Allgemeinmedizin und Familienmedizin e. V., o. J.).

Verknüpfungen aus deskriptiv- und inferenzstatistischen Verfahren oder sogenannte *Mixed-Methods*[136], die qualitative und quantitative Forschungsmethoden in einem Untersuchungsdesign kombinieren.

3.3 Interviews mit Expert:innen der allgemeinmedizinischen Forschung

Aufgrund der Ergebnisses des Literaturscreenings und der Literaturrecherche, die die Anwendung Maschinellen Lernens als Ausnahme zeigten, wurden Interviews[137&138] mit Expert:innen der allgemeinmedizinischen Forschung geführt, um den aktuellen Stand Maschinellen Lernens in der Allgemeinmedizin konkreter zu erfassen.

Anhand von Expert:innen-Interviews sollte die Forschungsfrage beantwortet werden, ob und in welchem Ausmaß theoretisches Grundlagenwissen und praktische Anwendungserfahrungen sowie Überlegungen zum potenziellen Einsatz Maschinellen Lernens in der allgemeinmedizinischen Forschung (allgemein und die Versorgungsforschung betreffend) verbreitet sind.

Nach Bogner et al.[139] würden Expert:innen-Interviews den Zweck verfolgen, Aussagen von Personen zu sammeln, die ein spezifisches Praxis- oder Erfahrungswissens hinsichtlich eines klar begrenzbaren Problemkreises (in diesem Fall die Allgemeinmedizin) besäßen. Anhand dessen hätten diese Personen die Möglichkeit, anhand ihrer Deutungen das konkrete Handlungsfeld sinnhaft und handlungsleitend für Andere zu strukturieren. Auf das Feld der Allgemeinmedizin bezogen kann dies bedeuten, dass diese Expert:innen Forschungs- und Projektschwerpunkte setzen und somit den Ein- oder Ausschluss Maschinellen Lernens inklusive des entsprechenden Diskurses zumindest maßgeblich mitbestimmen können.

3.3.1 Interview-Entwicklung und Konzeption der Durchführung und Auswertung

Die Interviewpartner:innen wurden auf Basis der für die interessierende Domäne entsprechenden Expertise als auch ihrer Durchsetzungs- und Entscheidungskompetenz ausgewählt. Dementsprechend wurden als Interviewpartner Instituts- und Arbeitsbereichsleiter:innen, wissenschaftliche Mitarbeiter:innen als auch praktisch tätige Allgemeinmediziner:innen ausgewählt.

Für diese Interviews wurde ein spezifischer Interviewleitfaden erstellt[140] und die Interviews vor Ort oder über ein Videokonferenzsystem durchgeführt.

Als Interviewtyp wurde das *informative Expert:innen-Interview*[141] gewählt: Hiernach werden Informationen über den interessierenden Untersuchungsbereich in Form von technischem oder Prozesswissen gesammelt, das im weiteren Verlauf ausgewertet wird. Für die Auswertung von

[136] (Vgl. Kelle 2014).
[137] (Vgl. Bogner, Littig und Menz 2014).
[138] (Vgl. Wassermann 2015).
[139] (Vgl. Bogner, Littig und Menz 2014, 13).
[140] Siehe Anhang A
[141] (Vgl. Bogner, Littig und Menz 2014, 23).

Interviews, die auf eine Informationsgewinnung abzielen, sei nach Bogner et al.[142] die *qualitative Inhaltsanalyse* das Auswertungsverfahren der Wahl. Kurz zusammengefasst kann unter *qualitativer Inhaltsanalyse* die systematische und methodisch kontrollierte wissenschaftliche Analyse von Texten, Bildern, Filmen und anderer Kommunikationsinhalte verstanden werden[143]. Da das Ziel der Interviews in dieser Arbeit in der Erlangung praktisch verwertbaren Wissens liegt und nicht in der Theoriebildung, ist die von *Kuckartz* geprägte Variante einer *qualitative Inhaltsanalyse* geeignet[144].

Die grundsätzliche zentrale Herangehensweise *qualitativer Inhaltsanalysen* liegt in der Kategorienbildung anhand derer das gesamte Datenmaterial codiert wird, das zur Beantwortung der Forschungsfragen relevant ist[145]:

- Als Kategorie wird das Ergebnis einer Klassifizierung von Einheiten verstanden, wobei diese Einheiten unter anderem Prozesse, Aussagen, Diskurse oder Ideen sein können[146]. Die Gesamtheit aller Kategorien wird als *Kategoriensystem* bezeichnet[147].
- Unter Codierung wird das Herstellen einer Verbindung zwischen einer Textstelle wie aus einem Interview und einer Kategorie verstanden. Die entsprechende Textstelle wird als codiertes Segment bezeichnet[148].

Die Analyse der Interviewinhalte erfolgt inhaltlich strukturierend[149]: Hierbei können die Kategorien während der Interview-Auswertung gebildet werden (induktives Vorgehen) oder schon vorab festgelegt werden (deduktives Vorgehen). Für die hier vorliegende Fragestellung wurde ein deduktives Vorgehen gewählt und die Kategorien anhand der Leitfaden-Fragen erstellt, da diese Fragen die maßgeblich interessierenden Inhalte abbilden. Nach Kuckartz und Rädiker verläuft wie in Abbildung 11 dargestellt eine inhaltlich strukturierende qualitative Inhaltsanalyse in sechs Phasen[150]:

[142] (Vgl. Bogner, Littig und Menz 2014, 72).
[143] (Vgl. Kuckartz und Rädiker 2022, 39).
[144] (Vgl. Kuckartz und Rädiker 2022, 52).
[145] (Vgl. Kuckartz und Rädiker 2022, 39).
[146] (Vgl. Kuckartz und Rädiker 2022, 53).
[147] (Vgl. Kuckartz und Rädiker 2022, 61–63).
[148] (Vgl. Kuckartz und Rädiker 2022, 67).
[149] (Vgl. Kuckartz und Rädiker 2022, 104, 129-130).
[150] (Vgl. Kuckartz und Rädiker 2022, 132–56).

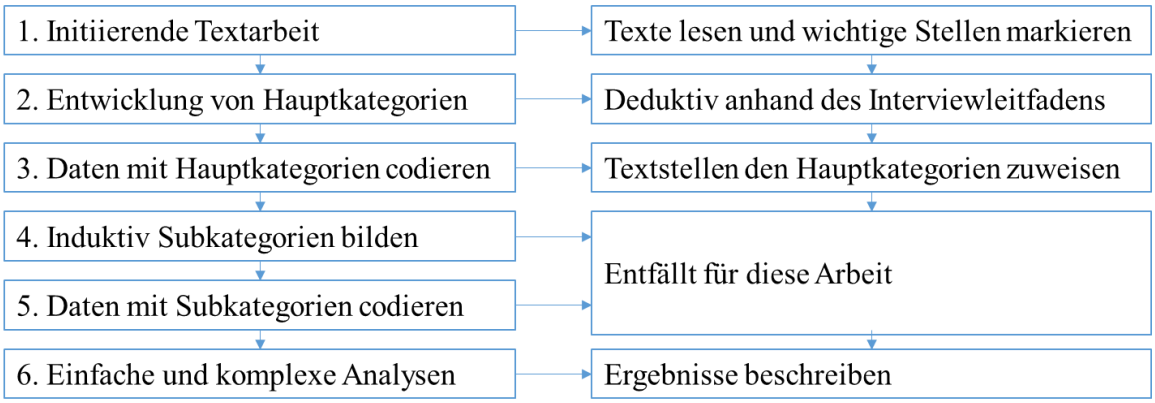

Abbildung 11: Ablauf einer inhaltlich strukturierenden qualitativen Inhaltsanalyse nach Kuckartz[143]

3.3.2 Ergebnisse

Mit sechs Expert:innen wurden Interviews geführt und die Ergebnisse werden hinsichtlich ihrer theoretischen Kenntnisse, ihrer Erfahrungen im nichtberuflichen Alltag, sowie in beruflicher Hinsicht im Folgenden dargestellt: Die Ergebnisse zeigen[151], dass die Expert:innen ein allgemeines und auch teilweise vertieftes Verständnis bezüglich Maschinellem Lernen besitzen. Maschinelles Lernen wird mit „Künstlicher Intelligenz“, „Datenauswertung“, „Automatisierung“ und „Verbesserung auf Basis zusätzlicher Daten“ assoziiert.

In ihrem nichtberuflichen Alltag sind sich über die Hälfte der Gesprächsteilnehmer:innen über die Präsenz maschinellem Lernen bewusst: bei Internetsuchmaschinen wie Google, bei Routenplanern, Applikationen im eigenen Tesla sowie der Steuerung der Energieversorgung in Smart Homes.

In der konkreten allgemeinmedizinischen Forschungs- und klinischen Praxis wurden keinerlei Erfahrungen in eigenen Publikationen als Haupt- oder Koautor:in oder anhand spezifischer Anwendungen gemacht. Zwei Befragte hatten in ihrer beruflichen Laufbahn unter anderem bei Publikationsprojekten mitgearbeitet, in denen beispielsweise das Volumen bestimmter Gehirnareale auf Basis Maschinellen Lernens berechnet wurde oder waren als Koautorin an nicht-allgemeinmedizinischen Publikationen beteiligt, in denen Maschinelles Lernen angewandt wurde.

Die in den Interviews angegebenen potenziellen Anwendungen Maschinellen Lernens in der Allgemeinmedizin lassen sich in zwei Gruppen unterteilen: Als Anwendung in der klinischen Praxis und als Forschungsmethode für konkrete Fragestellungen in der Versorgungsforschung:

Als Anwendungen in der klinischen Praxis wurden folgende Tools genannt:

- Als Hilfestellung zur Vorhersage schwerer Verläufe bei Patient:innen mit diffusen Symptomen
- Als Unterstützung zu Diagnosestellung für Ärzt:innen mit wenig Erfahrung

[151] Die Kernaussagen der einzelnen Interviews werden in Anhang B zusammengefasst

- Zur Vorhersage des Krankschreibungsbedarfs von Patient:innen, um diese bei schweren Erkrankungen nicht zusätzlich mit regelmäßigen Praxis-Besuchen zu belasten
- Vorhersage, welche Patient:innen zu spezialisierten Fachärzt:innen überwiesen werden sollten
- Als Hilfestellung in der Verwaltung in der Lehre und Weiterbildung am Fachbereich, beispielsweise in Form von automatisierter Terminfindung oder Empfehlung weiterer Inhalte auf Online-Auftritten.

Für die Versorgungsforschung wurden folgende Vorschläge genannt:

- Grundsätzlich wurde angemerkt, dass große Datenmengen explorativ genutzt werden könnten.
- Als konkreter Anwendungsfall in der Versorgungsforschung wurde die Identifikation von Patient:innen beschrieben, die bei Konsultation von Allgemeinmediziner:innen keine korrekten Angaben hinsichtlich ihrer Gesundheitsangaben kommunizieren (beispielsweise die an sie von anderen Ärzt:innen vergebenen Diagnosen).

Aus den Interview-Ergebnissen lässt sich zusammenfassen, dass ein Grundverständnis Maschinellen Lernens bei den Expert:innen gegeben ist, einige Ideen für konkrete Anwendungen in der klinischen Praxis vorhanden sind, aber für die konkrete Forschung noch kaum konkretere Planungen angedacht wurden.

3.3.3 Limitationen

Für diese Interviews wurde kein Pretest zur Leitfaden-Entwicklung und Interview-Umsetzung durchgeführt, anhand dessen die Funktionalität des Interviewleitfadens als auch die Umsetzung aus Sicht der Interviewten reflektiert würde[152]. Eine auf Basis eines Pretests gezieltere Befragung hätte durch Übungseffekte bei der Interviewdurchführung und Rückmeldungen der Interviewten präzisere Antworten ergeben, beispielsweise hätte gezielter nach Vorstellungen in den verschiedenen Forschungsdomänen (allgemein und konkret nach Versorgungsforschung) gefragt werden können.

3.4 Zusammenfassung des aktuellen Stands Maschinellen Lernens in der Domäne

Das Teilziel 1 bestand aus der Bestandsaufnahme, ob und in welcher Weise Maschinelles Lernen in der allgemeinmedizinischen Versorgungsforschung angewandt wird (Stand: Ende 2021).

Hierbei zeigte sich anhand einer systematischen Literaturrecherche, dass in wissenschaftlichen Publikationen Maschinelles Lernen bislang keine breite Anwendung als Forschungsmethode in der allgemeinmedizinischen Versorgungsforschung gefunden hat. Nur eine einzelne Publikation mit einer Anwendung Maschinellen Lernens lässt sich auf diese Domäne zurückführen. Die in allgemeinmedizinischen Fachzeitschriften grundsätzlich vorherrschenden Forschungsmethoden lassen sich auf Basis eines Literaturscreenings maßgeblich in inferenz-, deskriptivstatistische und qualitative Auswertungsverfahren verorten. Aus Sicht von Expert:innen allgemeinmedizinischer Forschung sind

[152] (Vgl. Bogner, Littig und Menz 2014, 34).

Methoden Maschinellen Lernens zwar bekannt, aber noch keine Anwendungen in der Domäne erfolgt oder geplant.

4 Ableitung von Anwendungsgebieten und eines Anwendungsfalls Maschinellen Lernens in der allgemeinmedizinischen Versorgungforschung

In diesem Abschnitt werden die aus der Erfassung des aktuellen Stands in Kapitel 3 resultierenden Ergebnisse systematisiert: Zum einen in allgemeine potenzielle Anwendungsgebiete Maschinellen Lernens in der Domäne, zum anderen wird ein konkreter Anwendungsfall aus einem der Gebiete abgeleitet, anhand dessen Maschinelles Lernen als Forschungsmethode beispielhaft eingesetzt werden kann.

4.1 Ableitung von Anwendungsgebieten

Aus der Literaturrecherche ergab sich ein einzelner konkreter Anwendungsfall Maschinellen Lernens in der allgemeinmedizinischen Versorgungsforschung in Form einer Studie aus Katalonien[153]: Hierbei wurden die Inhalte asynchroner Textnachrichten von Patient:innen an Allgemeinmediziner:innen unter anderem in das *Vorhandensein eines erhöhten Versorgungsbedarfs* oder der *Vermeidung eines ‚Vor Ort'-Termins* klassifiziert. Somit kann ein allgemein formuliertes Anwendungsgebiet die *Ermittlung des Versorgungsbedarfs* aus Perspektive des Patienten darstellen.

Weitere Anwendungsfälle aus der Literaturrecherche entstammen allerdings aus Anwendungsgebieten anderer Grundversorgungsdomänen (*Veteran Care*, Notaufnahmen oder ambulante Versorgungszentren) und beziehen sich auf Aspekte der Pünktlichkeit oder der Verlässlichkeit von Patient:innen. Diese Aspekte sind allerdings auch für die Allgemeinmedizin relevant und können unter dem Anwendungsgebiet der *Adhärenz und Compliance* zusammengefasst werden. Ein weiterer Anwendungsfall einer anderen Grundversorgungsdomäne umfasst das Thema der *Identifikation von Medikationsfehlern*, das ebenso für die Allgemeinmedizin bedeutsam ist und innerhalb des Anwendungsgebiets der *Patientensicherheit* verortet werden kann.

Auf Basis der Interviews wurden insbesondere Anwendungen zur Unterstützung zur Diagnosestellung aufgezählt, unter anderem zur Identifikation potenziell schwerer Erkrankungen auf Basis von zunächst diffuser Symptome wie Abgeschlagenheit oder zur Unterstützung noch weniger erfahrener Ärzt:innen im Alltag der klinischen Praxis. Hinsichtlich der Versorgungsforschung wurde die Identifikation von Patient:innen-Gruppen vorgeschlagen, die falsche Angaben in der Kommunikation mit medizinischem Fachpersonal machen, insbesondere in Bezug auf die Kommunikation von Diagnosen durch die Patient:innen. Solch ein Anwendungsfall kann unter dem Gebiet der *Gesundheitskompetenz* (*Health Literacy*) eingeordnet werden, die den kompetenten Umgang mit Gesundheitsinformationen miteinschließt.

[153] (Vgl. López Seguí et al. 2020).

Die eben aufgeführten Anwendungsfälle mit allgemeinmedizinischem Versorgungsforschungsbezug werden in Abbildung 12 systematisch mit den zugehörigen allgemeinen Anwendungsgebieten dargestellt.

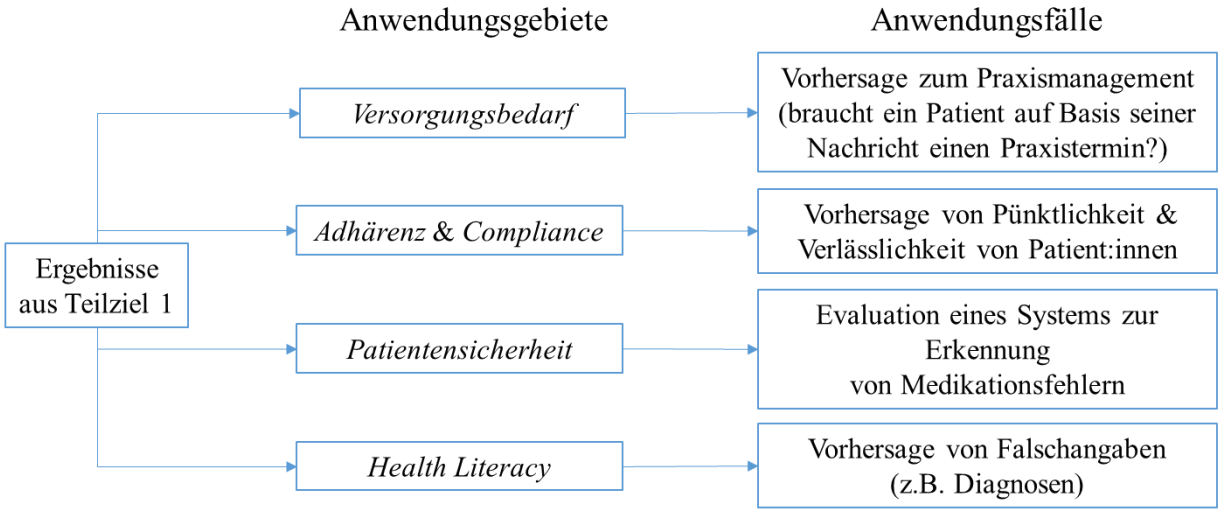

Abbildung 12: Aus Teilziel 1 abgeleitete Anwendungsgebiete für die Versorgungsforschung

4.2 Ableitung eines Anwendungsfalls

Als exemplarisches Anwendungsgebiet wird für dieses Teilziel das der *Health Literacy* aufgegriffen. *Health Literacy* (Gesundheitskompetenz) kann allgemein formuliert bei Menschen als gegeben angesehen werden, wenn sie in der Lage sind, sich um ihre Gesundheit zu kümmern („manage their health“)[154] und beinhaltet unter anderem Wissen über Gesundheit und Gesundheitsversorgung, sowie die Fähigkeit, Gesundheitsinformationen zu finden, zu verstehen, zu interpretieren und zu kommunizieren[155].

Der letztgenannte Aspekt beinhaltet auch die Kommunikation mit medizinischem Personal. Zu diesem zählen auch Allgemeinmediziner:innen und nach einem Rollenbild allgemeinmedizinischer Ärzt:innen im Gesundheitswesen würde dieser eine zentrale Position als verantwortliche/r Koordinator:in einnehmen und den Überblick über die Gesamtversorgung erhalten[156]. Somit würden bei ihm/ihr als Koordinator:in alle relevanten Informationen über Patient:innen zusammenlaufen, zu denen auch von allen Fachrichtungen gestellte Diagnosen zählen. Da es allerdings beispielsweise in Deutschland keine zentrale Datenbank mit gespeicherten Diagnosen der Patient:innen gibt, auf die Ärzt:innen Zugriff haben, sind diese unter anderem auch auf die Selbstauskunft von Patient:innen angewiesen, um alle relevanten Informationen zu erhalten. Allerdings zeigt sich, dass manche Patienten vergebene Diagnosen in einem Gespräch mit Ärzten nicht nennen (wird als *Underreporting* bezeichnet) oder auch

[154] (Vgl. Keleher und Hagger 2007, 24).
[155] (Vgl. Irving Rootman et al., 16).
[156] (Vgl. Deutsche Gesellschaft für Allgemeinmedizin und Familienmedizin e.V. 2012, 9).

Diagnosen nennen, die nicht vergeben wurden (*Overreporting*). Eine Übereinstimmung zwischen offiziell vergebenen Diagnosen und Patientenaussage wird als *Agreement* bezeichnet.

Diese Selbstauskünfte von Patient:innen hinsichtlich des eigenen Gesundheits- und Risikostatus ist in verschiedener Hinsicht relevant, da diese zum einen auf die Versorgung bezogen für eine angemessene Therapie und Prävention korrekt sein sollten. Zum anderen ist dies auch für epidemiologische Studien bedeutsam, in denen Erkrankungs-Prävalenzen (Erkrankungsrate in einer Population zu einem bestimmten Zeitpunkt) geschätzt werden.

Somit kann *Health Literacy* hinsichtlich verschiedener Aspekte als ein grundsätzliches Problem in der Gesundheitsversorgung angesehen werden. Eine nicht den Tatsachen entsprechende Nennung von Diagnosen ist ein in Studien regelmäßig auftretendes Problem. Allerdings zeigten sich in diesen Studien inkonsistente Ergebnisse, da die betrachteten Krankheitsgruppen jeweils in unterschiedlichem Ausmaß Falschangaben zeigten: Beispielsweise zeigte sich in einer Studie ein *Underreporting* zu eigentlich diagnostiziertem Herzversagen, Diabetes[157] und in einer anderen Studie ebenfalls zu Bluthochdruck, aber mit gleichzeitigem *Agreement* zu Diabetes[158], sowie ein geringeres *Agreement* bei Herz- , Lungen- und Gefäßerkrankungen[159]. Diese Auswahl an Studien zeigt, dass diese Selbstauskünfte von Patient:innen ein relevantes Problem darstellen, aber zumeist in älteren Studien behandelt wurden. Somit ist eine beispielhafte Implementierung hinsichtlich solch eines Anwendungsfalls lohnenswert.

Eine Fragestellung innerhalb dieses Problemfelds kann sich auf einen Versuch der Erklärung des Zustandekommens der Falschangaben durch Patient:innen beziehen (wie durch den Bildungsstand, psychische Beeinträchtigungen oder dem Alter von Patient:innen), sowie der auf diesen Ergebnissen basierenden Identifikation von Patient:innen, die potenziell Falschangaben äußern könnten. Die konkrete Aufstellung als auch Beantwortung solch einer Fragestellung durch einen entsprechenden Anwendungsfall wird im nächsten Abschnitt dargestellt.

[157] (Vgl. Okura et al. 2004).
[158] (Vgl. Goldman 2003).
[159] (Vgl. Malik et al. 2011).

5 Prototypische Umsetzung einer beispielhaften Studie als repräsentativer Anwendungsfall Maschinellen Lernens in der allgemeinmedizinischen Versorgungsforschung

In diesem Abschnitt soll der repräsentative Anwendungsfall aus einem der abgeleiteten Anwendungsgebiete entwickelt und prototypisch anhand einer Methode des Maschinellen Lernens angewandt sowie evaluiert werden.

5.1 Fragestellung der Studie

Die übergeordnete Fragestellung in dieser beispielhaften Studie wird wie folgt formuliert:

> *Mit welchen Faktoren sind diagnosebezogene Falschangaben (Over- und Underreporting) von Patient:innen hinsichtlich sie selbst betreffender medizinischer Diagnosen assoziiert?*

Der Zweck dieses Anwendungsfalls kann im Sinne der Entwicklung theoretischen Wissens verortet werden, da es sich bei der allgemeinmedizinischen Versorgungsforschung um eine Wissenschaftsdomäne handelt. Dementsprechend wurde ein theoriegetriebenes Vorgehen gewählt. Nach Klaus sei theoretisches Wissen ein wesentliches Element der Struktur von Wissenschaft und eine Theorie würde durch die Entwicklung eines Systems empirischer Erkenntnisse entstehen[160]. Daraus abgeleitet entspricht ein reiner und allumfassender datengetriebener Ansatz nicht dem Zweck der Entwicklung theoretischen Wissens.

Auch in der Domäne der allgemeinmedizinischen Versorgungsforschung basiert die Erarbeitung theoretischen Wissens auf empirischen Erkenntnissen: Diese empirischen Erkenntnisse wurden in dieser Domäne wie in Abschnitt 3.2 bislang durch deskriptiv- und inferenzstatistische als auch qualitative Forschungsmethoden gewonnen. In dieser Arbeit geht es nun um eine Veranschaulichung, wie Maschinelles Lernen als eine weitere Forschungsmethode empirische Erkenntnisse in dieser Domäne beitragen kann.

Wie in Abschnitt 2.2.3 dargelegt, können inferenzstatistische Verfahren eingesetzt werden, um zu erklären, durch welche Ausprägungen welcher Variablen ein Outcome zustande kommt. Maschinelles Lernen dagegen eignet sich eher zu einer Vorhersage eines bestimmten Outcomes auf Basis von Variablen. Somit können beide Forschungsmethoden jeweils empirische Erkenntnisse hinsichtlich der Erarbeitung theoretischen Wissens liefern: Zum einen, in welcher Weise bestimmte Phänomene aufgrund welcher Variablen auftreten (anhand von Inferenzstatistik) und zum anderen, in welchem Ausmaß diese Variablen ausreichen, das entsprechende Phänomen vorherzusagen (durch Maschinelles Lernen). Übertragen auf den hier vorliegenden Anwendungsfall können inferenzstatistische Verfahren testen, ob und in welchem Ausmaß bestimmte Variablen (wie *Alter*, *Bildung* oder bezogen auf bestimmte Erkrankungen) mit *Over*- und *Underreporting* zusammenhängen, und Maschinelles Lernen

[160] (Vgl. Klaus 1976, 1310).

kann überprüfen, ob diese durch die inferenzstatistische Testung bestätigten Variablen in der Lage sind, Patient:innen korrekt zu klassifizieren (in die, die korrekte Angaben machen und die, die keine korrekten Angaben machen). Somit kann Maschinelles Lernen eine weitere Facette empirischer Erkenntnisse liefern, und zwar, wie präzise die Ausprägungen einer interessierenden Variable vorhergesagt werden kann.

Auf Basis dieser Überlegungen wird folgender Ablauf für eine theoriegetriebene Arbeit in der Domäne der allgemeinmedizinischen Versorgungsforschung vorgeschlagen:

1. Zunächst werden auf Basis vorangegangener empirischer Erkenntnisse und theoretischer Überlegungen Variablen zusammengetragen, die mit der Art des *Reportings* der Patienten (*Agreement* oder *Over-* und *Underreporting*) in Zusammenhang stehen könnten. Hierfür wird ein entsprechender Datensatz gesucht, der die entsprechenden Variablen beinhaltet.
2. Auf Basis eines inferenzstatistischen Vorgehens wird untersucht, ob und in welchem Ausmaß diese vorab ausgewählten Variablen mit dem *Reporting* der Patient:innen in Beziehung stehen.
3. Anhand einer Methode Maschinellen Lernens wird darauf aufbauend überprüft, inwieweit die aus der inferenzstatistischen Analyse als relevant eingestuften Variablen zu einer Vorhersage der Art des *Reportings* der Patient:innen geeignet sind.

Auf diese Weise können die mit einer besseren *Vorhersage-Performance* assoziierten Black Box-Modelle (wie *Künstliche Neuronale Netze* oder *Gradient Boosting-Methoden*) mit auf hohe Interpretierbarkeit ausgerichteten Verfahren wie die der Inferenzstatistik (wie *logistische Regressionsanalysen*) verbunden werden.

5.2 Datengrundlage und deskriptive Statistik

Um solche Modelle zu überprüfen, müssen zunächst entsprechende Daten gesammelt werden. *Over-* und *Underreporting* als auch *Agreement* können durch einen Vergleich der in Selbstauskunft von Patient:innen kommunizierten Diagnosen mit offiziell vergebenen Diagnosen als Variablen operationalisiert werden.

Beispielsweise wurde im Rahmen der niederländischen SMILE-Studie (die prospektive Kohortenstudie *Study of Medical Information and Lifestyles in Eindhoven*)[161] in einem Teilprojekt solch eine Überprüfung vorgenommen[162]: Hierfür wurden im Rahmen der Studie Patient:innen nach dem aktuellen oder früheren Vorhandensein einer Diagnose in jeweils einer von 14 Gruppen chronischer Erkrankungen gefragt (unter anderem nach Lungen-, Herz- oder Darmerkrankungen oder Diabetes und Epilepsie). Diese Angaben wurden mit gespeicherten Gesundheitsdaten der Patient:innen verglichen und somit eine Aussage hinsichtlich *Agreement*, *Over-* oder *Underreporting* bei jeder der 14 Krankheitsgruppen ermöglicht. In den Niederlanden werden in Form eines Registers alle offiziell an Patient:innen

[161] (Vgl. van den Akker et al. 2008).
[162] (Vgl. van den Akker et al. 2015).

vergebene Diagnosen zentral gespeichert. Neben der diagnosebezogenen Selbstauskunft und der gespeicherten Diagnosen wurden in dieser Studie auch soziodemographische (wie *Alter* und *Bildungstand)* und klinische Daten gesammelt. Die Befragungen der Patient:innen wurden im Mai 2010 beendet und die gespeicherten Gesundheitsdaten im November 2010 abgerufen.

Ein Teil der Variablen dieses Datensatzes der SMILE-Studie wurde für diese Masterarbeit zur Verfügung gestellt: Der Datensatz beinhaltet die Daten von 2.893 Patient:innen, die in 145 allgemeinmedizinischen Praxen behandelt wurden. Hieraus ergibt sich eine sogenannte Multilevel-Struktur, die in Abbildung 13 dargestellt ist: Level 3 repräsentiert die höchste Ebene der allgemeinmedizinischen Praxen. In jeder Praxis ist eine Gruppe der rekrutierten Patient:innen (Level 2) und jeder Patient weist ein *Agreement*, *Over-* oder *Underreporting* auf jeder der 14 abgefragten Krankheitsgruppen (die dem Level 1 entsprechen) auf.

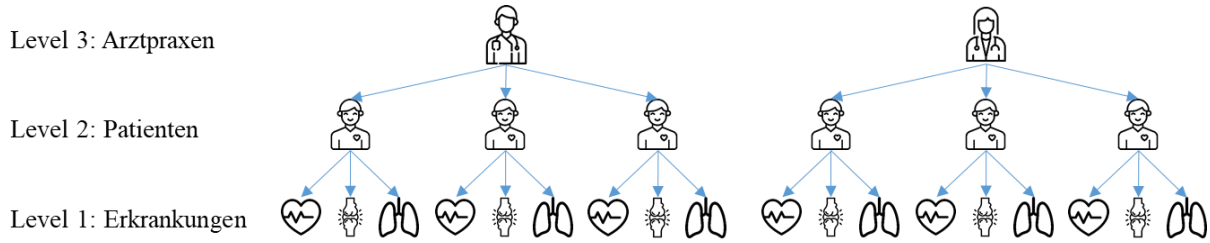

Abbildung 13: Hierarchische Struktur der Daten

Die Stichprobengröße als auch die Anzahl von Variablen kann als repräsentativ für Studien in der allgemeinmedizinischen Versorgungsforschung angesehen werden (als Beispiele können die *PRoMPT*[163]- oder die *PICANT*[164]-Studien betrachtet werden). Für solche Studien wird zumeist eine Stichprobengröße von mehreren Hundert oder höchstens mehreren Tausend Teilnehmern rekrutiert und eine begrenzte Auswahl hinsichtlich der für die jeweiligen Fragestellungen spezifisch zu erhebenden Variablen getroffen.

Die für diesen Anwendungsfall ausgewählten Variablen werden wie beschrieben theoriegetrieben ausgewählt: Hierfür werden vorab die für das Modell einzuschließenden Variablen auf Basis von Ergebnissen vorangegangener Fachliteratur als auch anhand theoretischer Vorüberlegungen ausgewählt. In Tabelle 3 werden die entsprechenden in die Modelle einzuschließenden Variablen inklusive des vorangegangenen empirischen Befunds dargestellt.

[163] (Vgl. Gensichen et al. 2009).
[164] (Vgl. Mertens et al. 2019).

Tabelle 3: Eingeschlossene Variablen zur Vorhersage des Reportings der Patient:innen

Variable	*Operationalisierung*	*Skalenniveau*	*Quelle*
Krankheitsgruppen	*Screening questionnaire of the Dutch Association of General Practitioners*	Nominal (14 chronische Erkrankungen)	Z.B: *Underreporting* zu gegebenem Herzversagen und Diabetes[165] oder *Agreement* zu Diabetes[166]
Geschlecht	Abfrage in Fragebogen	Binär (männlich/weiblich)	Frauen zeigen mehr *Agreement* bei allen abgefragten Erkrankungen (verschiedene Herzkreislauf-erkrankungen, Schlaganfall und Diabetes) [167], Frauen zeigen mehr *Agreement* bei Bluthochdruck[168]
Alter	Abfrage in Fragebogen (Geburtsdatum minus Tag der Erhebung)	Metrisch (Patient:innen ab einem Alter von 55 Jahren in Studie eingeschlossen)	Desto jünger desto mehr *Agreement* bei allen abgefragten Erkrankungen (verschiedene Herzkreislauf-erkrankungen, Schlaganfall und Diabetes) [169] Je älter, desto mehr *Overreporting* bei Schlag-anfall und *Underreporting* bei Arthritis[170]
Körperliche Funktionsfähigkeit	Körperliche Funktionsfähigkeit (Subskala des *SF-36*)	Metrisch (*0 – 100*, höherer Wert bedeutet höhere Lebensqualität)	Beeinträchtigung bei allen abgefragten Erkrankungen mit *Overreporting* assoziiert (Lungen-, Herz- und onkologische Erkrankungen, periphere Artheriosklerose, Schlaganfall, Diabetes und Arthritis)[171]

Fortsetzung der Tabelle auf Seite 41

[165] (Vgl. Okura et al. 2004).
[166] (Vgl. Goldman 2003).
[167] (Vgl. Okura et al. 2004).
[168] (Vgl. Merkin et al. 2007).
[169] (Vgl. Okura et al. 2004).
[170] (Vgl. Kriegsman et al. 1996).
[171] (Vgl. Kriegsman et al. 1996).

Fortsetzung von Tabelle 3:

Psychisches Wohlbefinden	Subskala des *SF-36*	Metrisch (*0 - 100*, höher Wert bedeutet höhere Lebensqualität)	Erhöhtes *Underreporting* bei Gelenkserkrankungen[172]
Angsterkrankung	Subskala der *Hospital Anxiety and Depression Rating Scale (HADS)*	Binär (Vorliegen oder Nicht-Vorliegen einer Angsterkrankung)	Bspw. das Vorliegen einer Angststörung mit *Overreporting* bei Diabetes assoziiert[173]
Depression	Subskala der *Hospital Anxiety and Depression Rating Scale (HADS)*	Binär (Vorliegen oder Nicht-Vorliegen einer Angsterkrankung)	Das Vorliegen einer Depression mit *Underreporting* bei Schlaganfall-bezogenen Beschwerden [174]
Krankheitszahl	Anzahl aller in dem elektronischen Register eingetragenen chronischen Erkrankungen	Metrisch	Bei keinen zusätzlichen chronischen Erkrankungen eher *Agreement*[175]
Bildung	Abfrage der Schul- und Ausbildungsabschlüsse	Ordinal (niedriger, mittlerer und hoher Bildungsabschluss)	Ab 12 Bildungsjahren mehr *Agreement*[176]

Zusätzlich wurde auch berücksichtigt, inwieweit die Arztpraxen und die Patient:innen über die Krankheitsgruppen hinweg systematisch mit dem Zustandekommen von *Over-* und *Underreporting* sowie *Agreement* assoziiert sind:

- Beispielsweise, ob Patient:innen unabhängig von der Erkrankung eine Neigung haben, gestellte Diagnosen nicht zu nennen
- Oder ob in Arztpraxen gehäuft Diagnosen von Patient:innen genannt, nicht genannt oder mit den gespeicherten Daten übereinstimmen.

Wenn methodisch möglich, können somit Patient:innen und Arztpraxen als Variable in Analysen miteingeschlossen werden.

[172] (Vgl. van den Akker et al. 2015).
[173] (Vgl. van den Akker et al. 2015).
[174] (Vgl. van den Akker et al. 2015).
[175](Vgl. Okura et al. 2004).
[176] (Vgl. Okura et al. 2004).

Auf Basis dieser vorangegangen Befunde könnte beispielsweise als zu prüfende theoretische Annahme behauptet werden, dass unter anderem ein höherer Bildungsgrad, weibliches Geschlecht als auch ein jüngeres Alter bei Patient:innen tendenziell mit einer korrekten Angabe von vergebenen Diagnosen einherginge. Männliches Geschlecht und höhere körperliche Beeinträchtigungen gingen dagegen mit einer erhöhten Wahrscheinlichkeit einher, nicht vergebene Diagnosen beziehungsweise nicht vergebene Diagnosen anzugeben.

Auf Basis weiterer theoretischer Vorüberlegungen könnten andere Variablen, wie beispielweise ein Maß für die bei Patient:innen gegebene *Health Literacy-Kompetenz* relevant sein. Diese wurde bislang nicht in vorangegangene Analysen eingeschlossen, allerdings auch für diesen in dieser Arbeit genutzten Datensatz nicht erhoben.

Die für diese Arbeit erhobenen Daten werden im Folgenden deskriptiv beschrieben. Die dreistufige Variable *Reporting* die den Vergleich der Selbstauskunft der Patienten zu den medizinischen Aufzeichnungen widergibt, zeigt über alle Patienten- und Erkrankungsgruppen hinweg folgende Verteilung (Abbildung 14): Von allen Patienten hat mehr als jeder Vierte keinerlei Abweichungen zu dem Register gezeigt. Fast jeder Fünfte hat mindestens eine Diagnose zu wenig angegeben, mehr als jeder Dritte mindestens eine zu viel und mehr als jeder Fünfte mindestens einmal eine Erkrankung zu viel und gleichzeitig mindestens eine Erkrankung zu wenig angegeben.

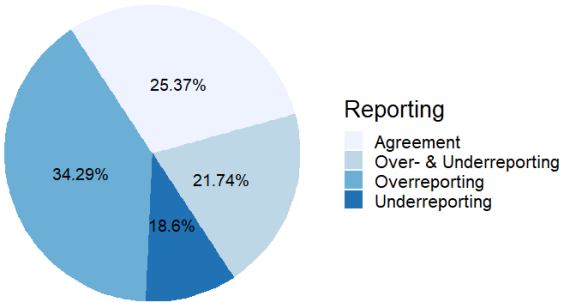

Abbildung 14: Häufigkeitsverteilung des Reportings

Konkret zeigt sich dies in Abbildung 15: In der Summe liegen nach dem elektronischen Register Gelenks-, arthritische und onkologische Erkrankungen am häufigsten vor, sowie Leber- und Nierenerkrankungen sowie Epilepsie am seltensten. Arthrose- und Rückenerkrankungen werden am häufigsten von Patient:innen genannt, obwohl sie nicht vorliegen. Vorliegende Gelenks- und Darmerkrankungen werden am häufigsten nicht genannt.

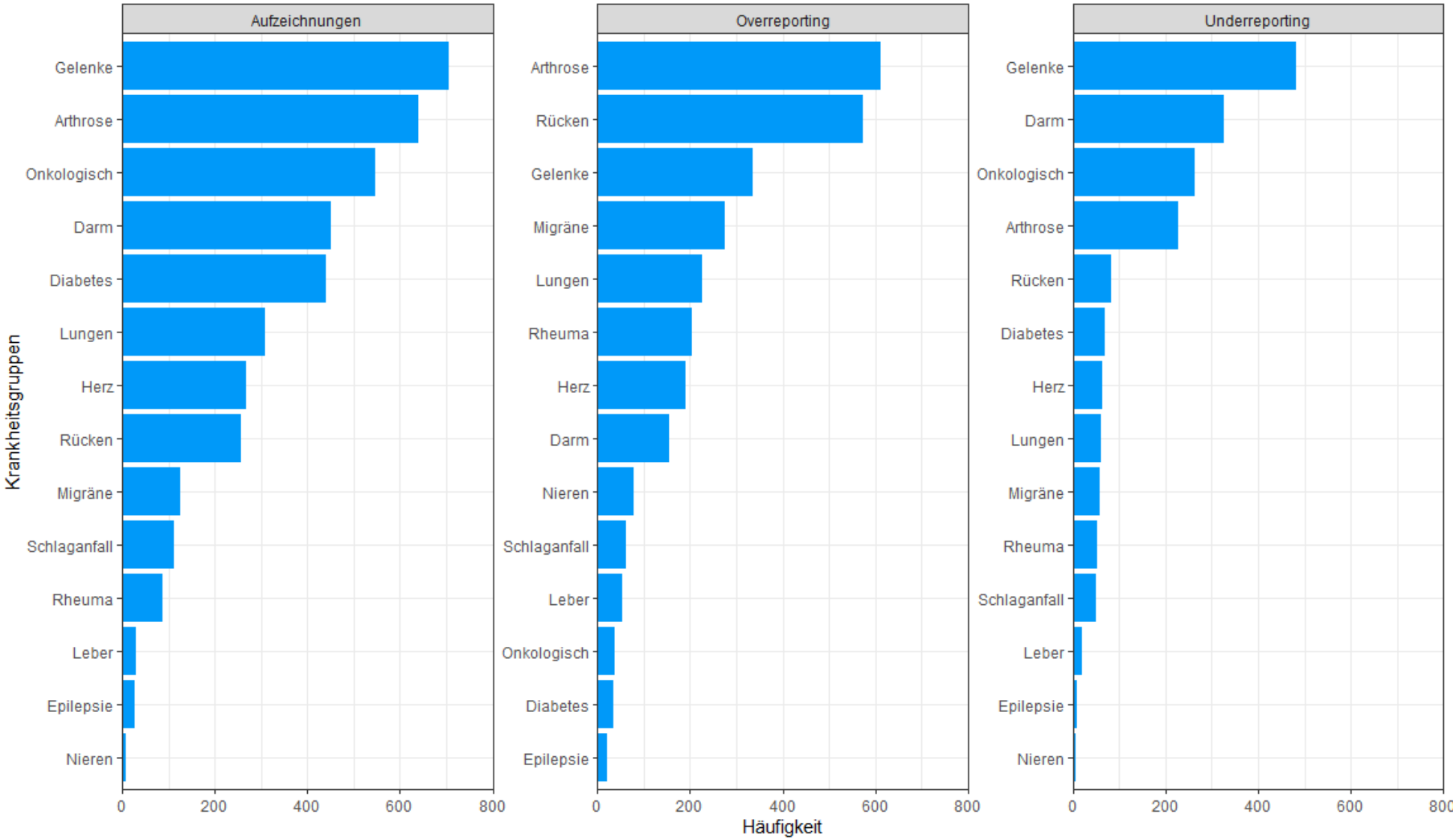

Abbildung 15: Häufigkeitsverteilung registrierter Diagnosen sowie des Over- und Underreportings

Zu den anderen Variablen zeigt sich folgendes Bild: Patient:innen ab einem Alter von 55 wurden in diese Studie eingeschlossen. Die entsprechende Verteilung wird in Abbildung 16 dargestellt.

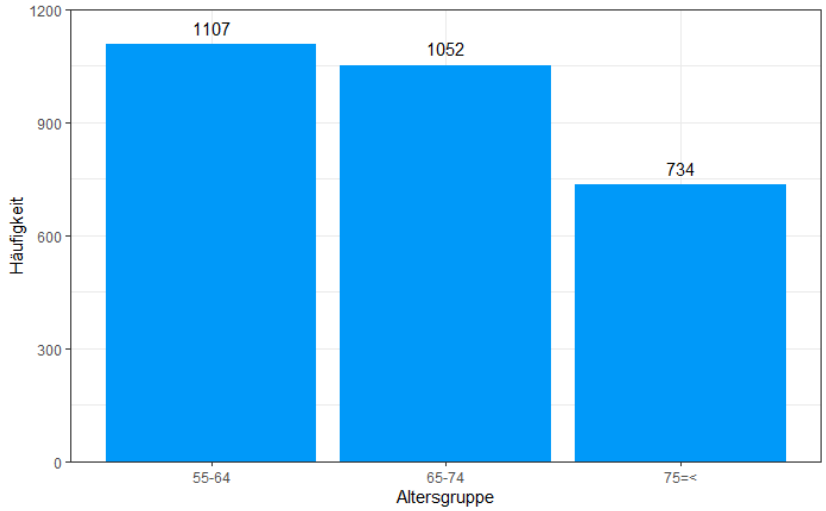

Abbildung 16: Altersverteilung der in die Studie eingeschlossenen Patient:innen

Der Bildungsstand wurde in der Studie in drei Gruppen unterteilt, wobei ein mittlerer Bildungsstand für einen Realschulabschluss mit Berufsausbildung steht und ein hoher für einen Universitätsabschluss. Allerdings zeigte sich hierbei (Abbildung 17), dass es eine große Menge fehlender Werte gibt. Somit wurde diese Variable aus den kommenden Analysen ausgeschlossen.

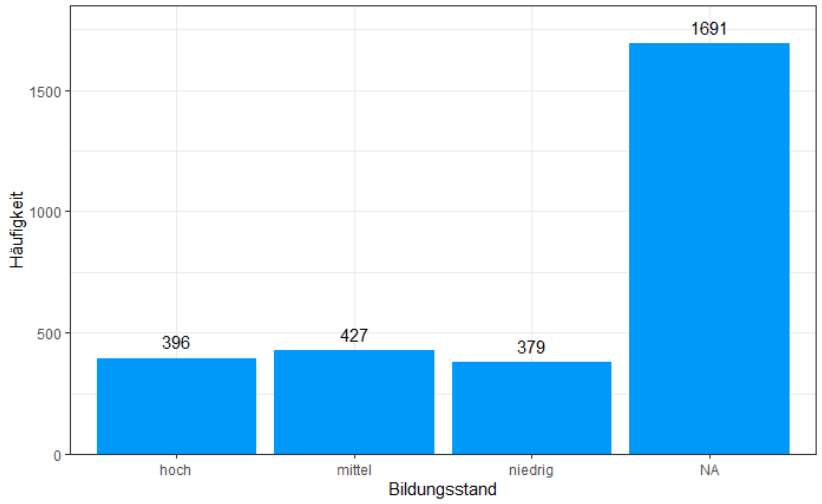

Abbildung 17: Verteilung des Bildungsstands der eingeschlossenen Patient:innen

Die Geschlechterverteilung ist mit einer leichten Mehrheit von Frauen (50.7%) ausgeglichen. Die Erfassung weiterer Geschlechtsidentitäten war nicht vorgesehen, wobei sich allerdings auch keine fehlenden Werte zeigten.

Die meisten Patient:innen hatten keine weitere chronische Erkrankung (33.8%) und eine einzelne weitere chronische Erkrankung lag bei 27.3% vor.

Physische Lebensqualität (in Form körperlicher Funktionsfähigkeit) und mentale Lebensqualität (in Form psychischem Wohlbefindens) zeigten im Mittel Werte von 46.42 und 52.32 auf einer möglichen Skala von *0* bis *100*, wobei ein höherer Wert eine höhere Angabe von Lebensqualität widergibt.

In den bisher dargestellten Ausschnitten der Daten liegt der Datensatz in einem sogenannten *Wide-Format* vor: In diesem Fall bedeutet dies, dass jede Zeile des Datensatzes eine/n Patient:in repräsentiert. Dieser besitzt hinsichtlich seiner Selbstauskunft bezüglich an ihn vergebener Diagnosen jeweils eine Spalte pro Erkrankungsgruppe. Allerdings können somit nicht in einer Analyse die Einflüsse der einzelnen Krankheitsgruppen untersucht werden, von denen laut vorangegangener Studien einzelne ein erhöhtes Risiko von *Under-* und *Overreporting* ausweisen. Um diesem Problem zu begegnen, wurde der Datensatz in ein *Long-Format* umgewandelt (siehe Abbildung 18): Hierbei hat nun Jede Krankheitsgruppe pro Patient:in eine Zeile bekommen. Da Selbstauskünfte bei 14 Krankheitsgruppen erhoben wurden, hat sich die Anzahl der Datensätze bei 2.893 Patienten auf 40.502 erhöht.

Patient	*Darmerkrankung*	*Nierenerkrankung*	*Herzerkrankung*
101	Overreporting	Underreporting	Agreement
102	Overreporting	Overreporting	Overreporting
103	Agreement	Agreement	Agreement

Patient	*Erkrankung*	*Reporting*
101	Darm	Overreporting
101	Niere	Underreporting
101	Herz	Agreement
102	Darm	Overreporting
102	Niere	Overreporting
102	Herz	Overreporting
103	Darm	Agreement
103	Niere	Agreement
103	Herz	Agreement

Abbildung 18: Schematische Umwandlung des Datensatzes von einem Wide- in ein Long-Format

Wenn man nun die Häufigkeiten auf der Ebene der Erkrankungsgruppen betrachtet (14 Erkrankungen pro Patient), ergibt sich ein anderes Bild (Abbildung 19): Bei weniger als fünf Prozent der Erkrankungen wurde eine Diagnose nicht erwähnt und bei über sieben Prozent zu viele aufgezählt. Diese Transformation führt zu einer starken Ungleichverteilung der Häufigkeiten in der *Reporting*-Variable.

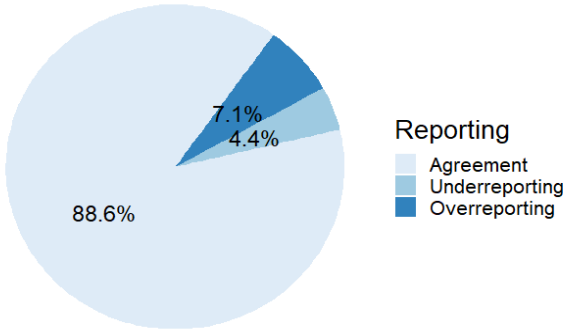

Abbildung 19: Verteilung von Over- und Under-Reporting sowie Agreement bei allen Patienten über alle Krankheitsgruppen hinweg

Die hier aufgeführten und in dem vorliegenden Datensatz verfügbaren Variablen werden in den folgenden Abschnitten zur

Beantwortung einer Fragestellung anhand einer inferenzstatistischen Methode sowie einer Fragestellung anhand einer Methode Maschinellen Lernens genutzt. Entsprechend den jeweils unterschiedlichen Begrifflichkeiten der beiden Forschungsmethoden wird in der Inferenzstatistik die Variable des *Reportings* als *abhängige Variable* und die auf einen Zusammenhang mit ihr getesteten Variablen als *unabhängige* bezeichnet. Beim Maschinellen Lernen wird *Reporting* als *Target-Variable* und die anderen als *Attribute* beziehungsweise *Features* angegeben.

5.3 Ergebnisse der Anwendung inferenzstatistischer Verfahren

In diesem Abschnitt werden die Analysen zur Identifikation von mit *Over-* und *Underreporting* assoziierten Faktoren anhand eines inferenzstatistischen Modells beschrieben.

Die für diese Analyse aufgestellte Fragestellung lautet: *Mit welchen Variablen ist Over- und Under-Reporting sowie Agreement von Patienten assoziiert?*

Hierfür wurden Variablen wie unter Abschnitt 5.1 beschrieben theoriegetrieben ausgewählt und anhand einer Zusammenhangshypothese getestet.

Das hinsichtlich des Testens von Zusammenhängen und der nominalskalierten Outcome-Variable angemessene Modell ist eine *Multinominale logistische Multilevel-Regressionsanalyse*. Die Wahl des Modells wird wie folgt begründet:

- *Multinominal-logistisch:* Die abhängige Outcome-Variable *Reporting* ist nominalskaliert und somit sind keine linearen, sondern logistische Regressionsmodelle angemessen[177]. Weiterhin liegt die nominalskalierte Outcome-Variable in drei Stufen vor, so dass eine Erweiterung der binär-logistischen Regressionsanalyse in Form einer multinominal-logistischen anzuwenden ist[178].
- *Multiple:* Es wird mehr als eine unabhängige Variable getestet und für diese Modellparameter in Form von Regressionsgewichten geschätzt.
- *Multilevel:* Die Patienten sind wie unter Abschnitt 5.1 dargestellt in Arztpraxen geclustert. Und da es für jeden Patienten einen Datenpunkt zu jeder der abgefragten Krankheiten gibt, sind die Krankheiten in die Patienten geclustert. Falls diese Datenstruktur in inferenzstatistischen Verfahren nicht in dieser Weise berücksichtigt würde, bestünde die Gefahr eines sogenannten *ökologischen Fehlschlusses*[179]: Dieser liegt vor, wenn man einen gefundenen Zusammenhang beziehungsweise einen statistischen Effekt fälschlicherweise auf der Ebene von Individuen (Level-1-Einheiten, in diesem Fall die Erkrankungsgruppen) interpretiert, obwohl dieser auf der Ebene von Gruppen vorliegt (Level-2-Einheiten, in diesem Fall Patienten).

Der konkrete Ablauf logistischer Regressionsanalysen jeglicher Art wird nach Backhaus et al.[180] in fünf Schritte unterteilt: 1. Modellspezifikation, 2. Schätzung der logistischen Regressionsfunktion, 3. Interpretation der Regressionskoeffizienten, 4. Prüfung des Gesamtmodells und 5. Prüfung der Merkmalsvariablen. Dieser Ablauf wird im Folgenden beschrieben.

[177] (Vgl. Backhaus et al. 2018, 268).
[178] (Vgl. Kwak und Clayton-Matthews 2002, 404).
[179] (Vgl. Eid, Gollwitzer und Schmitt 2017, 729–30).
[180] (Vgl. Backhaus et al. 2018, 272–73).

5.3.1 Modellspezifikation

Zunächst wurden die aufgrund der Annahmen zu testenden Variablen in Form eines Modells spezifiziert: Die *Reporting*-Variable als abhängige Variable, die weiteren unter Abschnitt 5.2 beschriebenen Variablen als unabhängige Variablen. Somit wird anhand des Modells getestet, ob und in welchem Ausmaß sich bei Veränderung einer unabhängigen Variable (beispielsweise zunehmendes *Alter*) die Wahrscheinlichkeit eines *Over*- oder *Underreportings* erhöht.

Zur Modellspezifikation sind für diese Arbeit drei Aspekte relevant: Der Umgang mit der Multilevel-Struktur, der Korrelation unabhängiger Variablen untereinander (*Multikollinearität*) und dem Umgang mit unabhängigen kategorialen Variablen.

Aufgrund der beschriebenen Multilevel-Struktur der Daten wird in der Literatur ein bestimmtes Vorgehen zur Testung der Variablen in Form von drei aufeinander aufbauenden Modellen empfohlen [181]: Anhand eines ersten Schritts werden anhand sogenannter *Intercept-only-Modelle* nur die höheren Hierarchie-Ebenen (hier Level 3 mit den in der Studie eingeschlossenen Praxen als Variable) getestet, ob die Outcome-Variable systematisch zwischen den Praxen variiert (hier würde sich zeigen, ob sich beispielsweise das *Overreporting* in bestimmten Praxen häuft). In einem zweiten Schritt wird die Level 2-Variable der *Patient:innen* hinzugefügt, um zu betrachten ob sich die Patienten systematisch unterscheiden (ob manche Patienten über die Krankheitsgruppen hinweg systematisch mit *Over*- oder *Underreporting* assoziiert werden können, beispielsweise als Person eine Neigung haben, Diagnosen unabhängig vom Inhalt zu verschweigen). Diese potenziellen systematischen Variationen werden anhand sogenannter *Intraklassen-Koeffizienten* (*Intraclass-Coefficients*: *ICC*) dargestellt, die einmal für die Praxen als auch für die Patient:innen je *Over*- und *Underreporting* berechnet werden: Dieser ICC gibt das Verhältnis der durch die unabhängigen Variablen erklärten Varianz der Outcome-Variablen zur Gesamtvarianz an. Aufgrund des nominalen Skalenniveaus der Outcome-Variable wird ein *ICC* für binäre Daten errechnet (siehe Formel 1)[182]: Je einen *ICC* für *Over*- und einen *ICC* für *Underreporting*.

Formel 1: Berechnung eines Intraklassenkoeffizienten für binäre Outcome-Variablen:

$$\frac{\tau^2}{\tau^2 + \frac{\pi^2}{3}}$$

Anhand eines *ICCs* wird auch erkannt, ob ein Multilevel-Modell notwendig zur Schätzung der Modellparameter ist: Wenn es keinen durch höhere Level erklärten bedeutsamen Anteil der Gesamtvarianz gibt, dann kann eine reduzierte Modellkomplexität in Form nicht-hierarchischer Modelle ausreichen, die eine geringere Rechenkapazität benötigen. In einem dritten Schritt wird ein sogenanntes *Random-Intercept-Modell mit Level 2- Prädiktoren* berechnet. Hierbei werden unter Berücksichtigung,

[181] (Vgl. Raudenbush und Bryk 2010, 23).
[182] (Vgl. Goldstein 2010).

dass sich die Outcome-Variable zwischen den Arztpraxen als auch zwischen den Patient:innen systematisch unterscheiden können, sowohl um die Krankheitsgruppen-Ebene (Level 1, also ob ein bestimmtes Reporting bei bestimmten Krankheitsgruppen systematisch häufiger auftritt) als auch unabhängige Variablen auf der Patienten-Ebene ergänzt: Inwiefern der *Bildungsgrad*, das *Alter*, die *Geschlechtszugehörigkeit* und so weiter wie in Abschnitt 5.2 dargelegt, einen Einfluss auf das *Reporting* haben.

Zusätzlich zu der Festlegung der Struktur der verschiedenen aufeinander aufbauenden Modelle müssen auch die einzelnen Variablen spezifiziert werden: Neben der theoriegetriebenen inhaltlichen Bedeutsamkeit der zu testenden Variablen müssen auch mathematische Vorbedingungen erfüllt sein. Unabhängige Variablen eines Modells dürfen nicht miteinander hoch korrelieren beziehungsweise dürfen keine linearen Abhängigkeiten zwischen ihnen bestehen (sogenannte *Multikollinearität*)[183]. Falls doch, würde die Schätzung der Regressionsparameter unzuverlässig und die in den Variablen vorhandene Information ließe sich nicht mehr eindeutig unabhängigen Variablen zuordnen, selbst wenn es einen Effekt einer unabhängigen Variable auf die abhängige gäbe[184]. Zur statistischen Absicherung des Einflusses unabhängiger Variablen wird daher die Testung auf *Multikollinearität* empfohlen: Eine Korrelationsanalyse der für diese Arbeit vorliegenden Daten zeigte, dass Korrelationen von circa -.50 zwischen den Variablen *mentale Lebensqualität*, *Vorliegen einer Angststörung* und *Vorliegen einer depressiven Störung* gegeben sind. Aufgrund der Höhe der Korrelationen und der inhaltlichen Nähe aller drei Variablen wurden für die inferenzstatistischen Analysen die beiden Variablen hinsichtlich des Vorliegens einer Angst- und depressiven Störung ausgenommen (Korrelationsmatrix in Anhang C).

Eine weitere Maßnahme im Rahmen der Modellspezifikation beinhaltet den Umgang mit kategorialen Variablen: Die Variable *Krankheitsgruppe* ist eine nominalskalierte Variable mit 14 Ausprägungen. Um nominalskalierte Variablen inferenzstatistisch anhand von Regressionsanalysen zu testen, muss eine dieser Ausprägungen als *Referenzkategorie* festgelegt werden. Dies wurde in dieser Arbeit anhand einer *Dummy-Kodierung* umgesetzt[185]: Hierfür wird die Variable in diesem Fall in 13 Variablen aufgeteilt, in der jede *Krankheitsgruppe* jeweils eine Variable darstellt. Die Referenzkategorie wird nicht durch eine eigene Variable dargestellt. Als Referenzkategorie wurde *Epilepsie* gewählt, da bei dieser geringe Häufigkeiten von *Over-* und *Underreporting* vorliegen (siehe Abbildung 15). Somit wird das Auftreten von *Over-* und *Underreporting* in allen 13 Krankheitsgruppen gegenüber einer Krankheitsgruppe getestet, die in Form von Epilepsie für eine annähernd korrekte Selbstauskunft von Patient:innen steht.

[183] (Vgl. Döring und Bortz 2016, 848).
[184] (Vgl. Backhaus et al. 2018, 98–99).
[185] (Vgl. Bortz und Schuster 2010, 363–64).

5.3.2 Schätzung der logistischen Regressionsfunktion

Die Schätzung der Parameter in Form von Regressionsgewichten wurde auf Basis des inferenzstatistischen Paradigmas der *Bayes-Statistik* (siehe 2.2.2) anhand des *R*-Paktes *brms*[186] durchgeführt[187]. Es wurde sich unter anderem für dieses Paradigma entschieden, da derzeit keine R-Pakete des *klassischen Fehler-Paradigmas* für eine *multinominale Regressionsanalyse* mit Multilevel-Struktur zur Verfügung steht.

Der Zusammenhang zwischen der abhängigen Variablen und den unabhängigen Variablen wird in Form von Regressionsgewichten angegeben. Für jede Variable wird in der Bayes-Statistik neben dem Regressionsgewicht ein *Glaubwürdigkeitsintervall (95%)* geschätzt. Wenn dieses *Glaubwürdigkeitsintervall* keinen Wert von *0* miteinschließt, wird die entsprechende Variable mit ihrem Regressionsgewicht als statistisch signifikant mit der abhängigen Variablen assoziiert eingestuft. Das Ausmaß des Zusammenhangs zwischen unabhängigen und abhängiger Variable wird anhand von Effektstärken in Form von *Odd Ratios* und ihnen zugehöriger *Konfidenzintervalle* (*95%*) angegeben. Sollte sich beispielsweise bei der unabhängigen Variable *Alter* ein Odds Ratio von 2.5 ergeben, wird dieses derart interpretiert, dass sich mit jeder Zunahme von einem Altersjahr das Risiko für *Overreporting* erhöht.

Die Parameterschätzungen für die Regressionsgewichte wurde in Form einer *Markov Chain Monte Carlo-Methode* (*MCMC*) vorgenommen: Hierbei wird eine Wahrscheinlichkeitsverteilung der Regressionsgewichte in der Population angenommen. Aus dieser werden anhand von Simulationen Stichproben gezogen (in der Analyse in Form der Anzahl der *Iterationen* ausgedrückt) und das Regressionsgewicht aus dem Mittelwert aller Stichproben geschätzt.

Ob die Parameterschätzungen als zuverlässig anzusehen sind, wird mit folgendem Vorgehen überprüft: Es können mehrere sogenannte *Ketten* mit jeweiligen Simulationen berechnet werden (mit jeweils unterschiedlichen Ausgangsbedingungen). Wenn diese *Ketten* zu gleichen Ergebnissen führen, spricht man von *Konvergenz* der *Ketten* und kann von zuverlässigen Ergebnissen ausgehen. Diese *Konvergenz* wird anhand verschiedener Koeffizienten angezeigt[188]:

- $\hat{R}$: Dieser gibt eine Übereinstimmung der Parameterschätzungen zwischen den *Ketten* als auch innerhalb der *Ketten* an und sollte nahe 1.00 liegen.
- *Tail-ESS* und *Bulk-ESS* (*ESS: Effective Sample Size*): Diese Koeffizienten dienen der Einschätzung, ob eine effektive Stichprobe innerhalb der Ketten zur Parameterschätzung gegeben ist. *Tail-* und *Bulk-ESS* sollten jeweils 100mal höher als die Anzahl der *Ketten* sein.

[186] (Bürkner 2017).
[187] Syntaxen der einzelnen Modelle in Anhang D
[188] (Vgl. Vehtari et al. 2021).

Für diese Arbeit wurden zwei *Ketten* geschätzt. Die Anzahl der Iterationen wurde mit zunehmender Modellkomplexität erhöht. Wie in Tabelle 4 ersichtlich, liegen alle *Konvergenz*-Koeffizienten in einem Bereich, in dem von zuverlässigen Schätzungen ausgegangen werden kann. In jedem Modell werden $\hat{R}$ und *Bulk*- sowie *Tail-ESS* für *Over*- als auch *Underreporting* ausgegeben wobei jede einzelne getestete unabhängige Variable einen *Bulk*- und *Tail-ESS*-Wert erhält.

Tabelle 4: Kennzahlen zur Konvergenz der MCMC-Analysen

	$\hat{R}$	*Bulk-ESS*	*Tail-ESS*	*Anzahl Ketten* *Anzahl Iterationen*
Interzept only-Modell 1				
Underreporting	1.00	480 - 1652	428 - 1019	2 Ketten
Overreporting	1.00	486 - 1365	755 - 1059	1.250 Iterationen
Interzept only-Modell 2				
Underreporting	1.00	390 - 1210	514 – 1614	2 Ketten
Overreporting	1.00	364 - 1503	502 - 1541	2.000 Iterationen
Random-Intercept-Modell				
Underreporting	1.00	611- 8085	796 - 2872	2 Ketten
Overreporting	1.00 – 1.01	558 - 5345	969 - 3086	4.000 Iterationen

5.3.3 Interpretation der Regressionskoeffizienten

Für diesen Schritt werden die ersten konkreten Ergebnisse der Analysen in Form der beschriebenen Effektstärken betrachtet. Nach Chen et al. deute ein *Odd Ratio* von *1* auf keinen Effekt hin, von etwa *2* auf einen kleinen Effekt, *3* auf einen mittleren und ein Effekt von mindestens *7* auf einen starken Effekt[189].

Extrem starke Effekte zeigen sich bei einzelnen Krankheitsgruppen in Hinblick auf *Underreporting* (siehe Abbildung 20): Die Wahrscheinlichkeit Gelenks-, Darm- sowie Arthroseerkrankungen nicht anzugeben ist im Vergleich zu der Referenzvariable Epilepsie (als eine Erkrankungsgruppe mit geringen *Over*- und *Underreporting*-Häufigkeiten) zwischen 40mal und mehr als 100mal so hoch. Bis auf Nierenerkrankungen (die zusammen mit *mentaler Lebensqualität* den Wert von *1* umschließt und somit keinen Effekt beinhalten), haben alle weiteren Erkrankungsgruppen eine höhere Wahrscheinlichkeit von Patient:innen nicht genannt zu werden. Auch die Anzahl von Erkrankungen geht mit einer mehr als zweifachen Wahrscheinlichkeit je zusätzlicher Erkrankung mit *Underreporting* einher (*Odds Ratio* von 2.14), die Zugehörigkeit zum männlichen Geschlecht geht mit einer leicht erhöhten Wahrscheinlichkeit einher (1.18) wie die physische Lebensqualität und das Alter.

[189] (Vgl. Henian Chen, P. Cohen und S. Chen 2010).

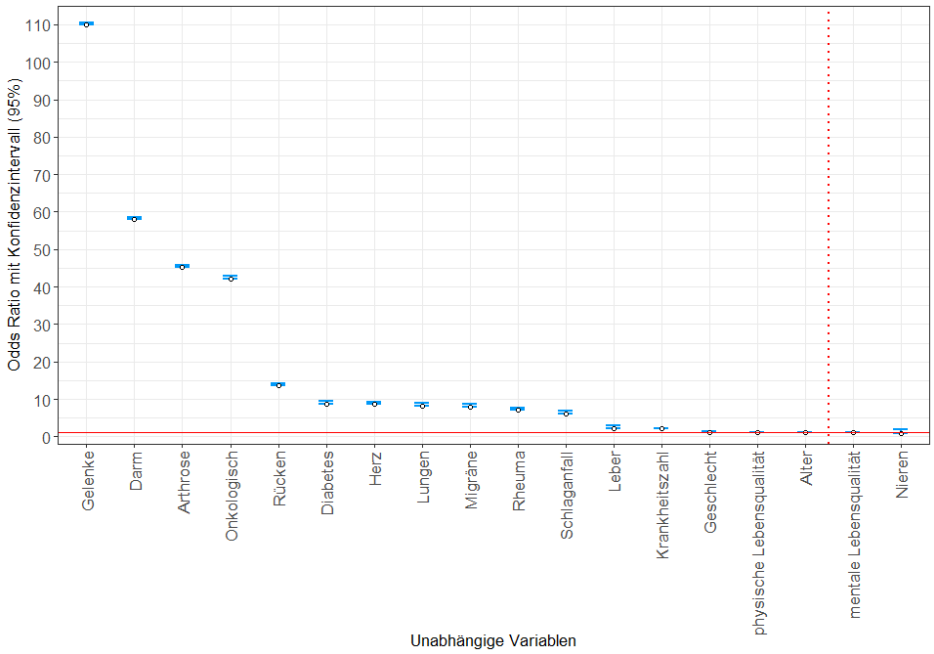

Abbildung 20: Die Verteilung der Odds-Ratios der unabhängigen Variablen im Hinblick auf Underreporting

Ebenfalls stark erhöht gegenüber Epilepsie ist die Wahrscheinlichkeit, Arthrose-, Rücken-, Gelenks- als auch weitere Erkrankungen anzugeben, auch wenn sie nicht offiziell vergeben wurden (*Overreporting*, siehe Abbildung 21). Auch männliches Geschlecht ist mit einer grundsätzlich erhöhten Wahrscheinlichkeit verbunden. Leichte negative Effekte zeigen *mentale* sowie *physische Lebensqualität* als auch die Anzahl der Erkrankungen. Dies bedeutet, dass als je besser die jeweilige *Lebensqualität* angegeben wird und je mehr weitere Erkrankungen vorliegen, desto geringer ist die Wahrscheinlichkeit für die Angabe nicht vergebener Diagnosen. Dagegen zeigt *Alter* bei *Overreporting* keinen Effekt.

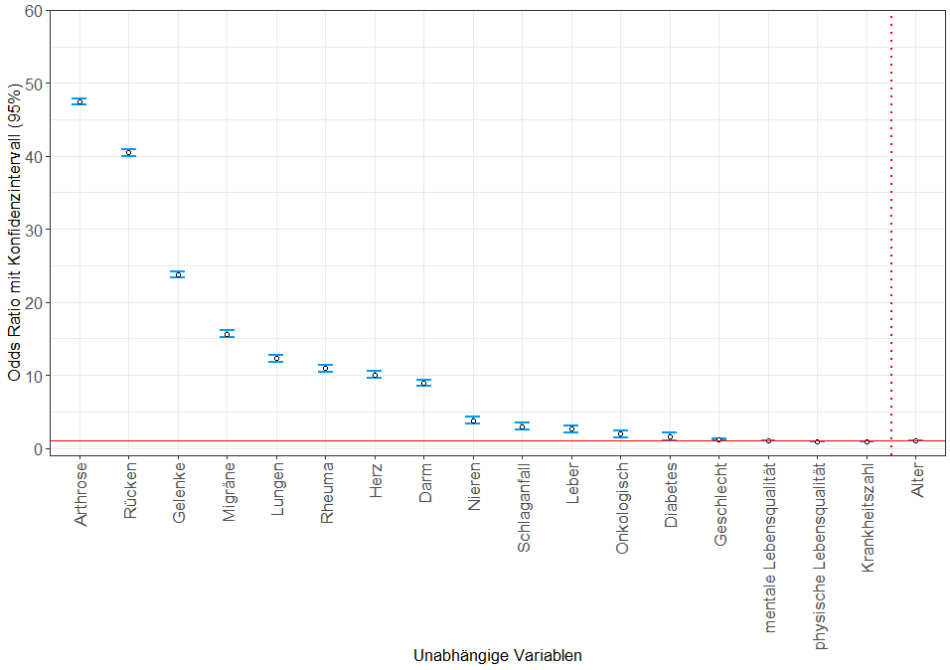

Abbildung 21: Die Verteilung der Odds-Ratios der unabhängigen Variablen im Hinblick auf Overreporting

Zusätzlich zu den Effekten der unabhängigen Variablen geben die *ICCs* der Arztpraxen und der Patient:innen weitere Auskunft über das Zustandekommen des *Reportings*: Drei bis fünf Prozent der Varianz kann im *Intercept-Only-Modell 1* auf die unterschiedlichen Arztpraxen zurückgeführt werden (Angaben von *Over*- und *Underreporting* sowie *Agreement*, die systematisch zwischen den Arztpraxen variieren). Diese reduziert sich, nachdem im *Modell 2* die Patient:innen als systematische Varianzquelle hinzugezogen werden. Hierbei zeigt sich, dass knapp 20 Prozent der Varianz durch diese erklärt werden kann (zum Beispiel durch die Patient:innen, deren Angaben grundsätzlich mit dem Register übereinstimmen, aber auch durch diejenigen, bei denen hinsichtlich mehrerer Krankheitsgruppen *Over*- und *Underreporting* vorliegt). Der *ICC* bei *Underreporting* reduziert sich deutlich, sobald die Level 2-Prädiktoren als auch die Krankheitsgruppen mit in das Modell aufgenommen werden. Bei *Overreporting* bleibt der *ICC* nahezu gleich. Detaillierte Ergebnisse des *Random Intercept-Modells* sind in Anhang E aufgeführt.

Tabelle 5: Die Intraklassenkoeffizienten der drei Modelle

		Intercept-Only-Modell 1	*Intercept-only-Modell 2*	*Random-Intercept-Modell mit Level 2- Prädiktoren*
Praxis	*Overreporting*	0.038	0.021	0.018
	Underreporting	0.049	0.026	0.012
Patient:innen	*Overreporting*	-	0.182	0.18
	Underreporting	-	0.194	0.021

5.3.4 Prüfung des Gesamtmodells

In diesem Schritt wird die Schätzung der Regressionsfunktion als Ganzes hinsichtlich ihrer Güte überprüft. Hierbei wird global ausgedrückt, wie gut sie als ein Modell der Realität geeignet ist[190]. Hierfür wird zumeist unter anderem ein Bestimmtheitsmaß R^2 genutzt. Für Modelle auf Basis der *Bayes-Statistik* wurde das *Bayesian* R^2 entwickelt[191]. Dieses repräsentiert das Maß der vorhergesagten Werte der Outcome-Variable relativ zu der Summe der Varianz der vorhergesagten Werte und der erwarteten Fehlervarianz. Somit wird eine Aussage getroffen, wie gut das Modell die Outcome Variable vorhersagt. Allerdings gibt es keine entsprechende Anwendung in dem verwendeten Paket für eine multinominal-skalierte Outcome-Variable.

[190] (Vgl. Backhaus et al. 2018, 74).
[191] (Vgl. Gelman et al. 2019).

5.3.5 Prüfung der Merkmalsvariablen

In diesem letzten Schritt wird die Güte der Regressionsfunktion lokal geprüft. Hierbei wird beurteilt, ob die einzelnen unabhängigen Variablen mit der Outcome-Variable statistisch signifikant in der Population zusammenhängen[192].

Für die Testung der statistischen Signifikanz werden die in der Regressionsfunktion geschätzten Regressionsgewichte und die den Regressionsgewichten zugehörigen *Glaubwürdigkeitsintervalle* der Variablen herangezogen. Die *Glaubwürdigkeitsintervalle* geben an, ob davon auszugehen ist, ob das geschätzte Regressionsgewicht mit einer bestimmten Wahrscheinlichkeit (in diesem Fall 95%) in diesem Intervall auch in der Population liegt.

Wenn nun ein *Glaubwürdigkeitsintervall* den Wert *0* beinhaltet, wird das Regressionsgewicht als nicht statistisch angenommen. Allerdings ist anzumerken, dass bei logistischen Regressionsanalysen die Höhe der Regressionsgewichte keine Indikatoren der Höhe des Zusammenhangs zwischen abhängigen und unabhängigen Variablen darstellen. Dies ist die Aussage der *Odd Ratios* in Schritt 3.

Aufgrund des multinominalen Niveaus der Outcome-Variable wird die statistische Signifikanz der unabhängigen Variablen einmal auf das Eintreten von *Over*- und einmal von *Underreporting* bezogen: Beim *Underreporting* zeigten *mentale Lebensqualität* und Nierenerkrankungen keine statistische Signifikanz (Abbildung 22).

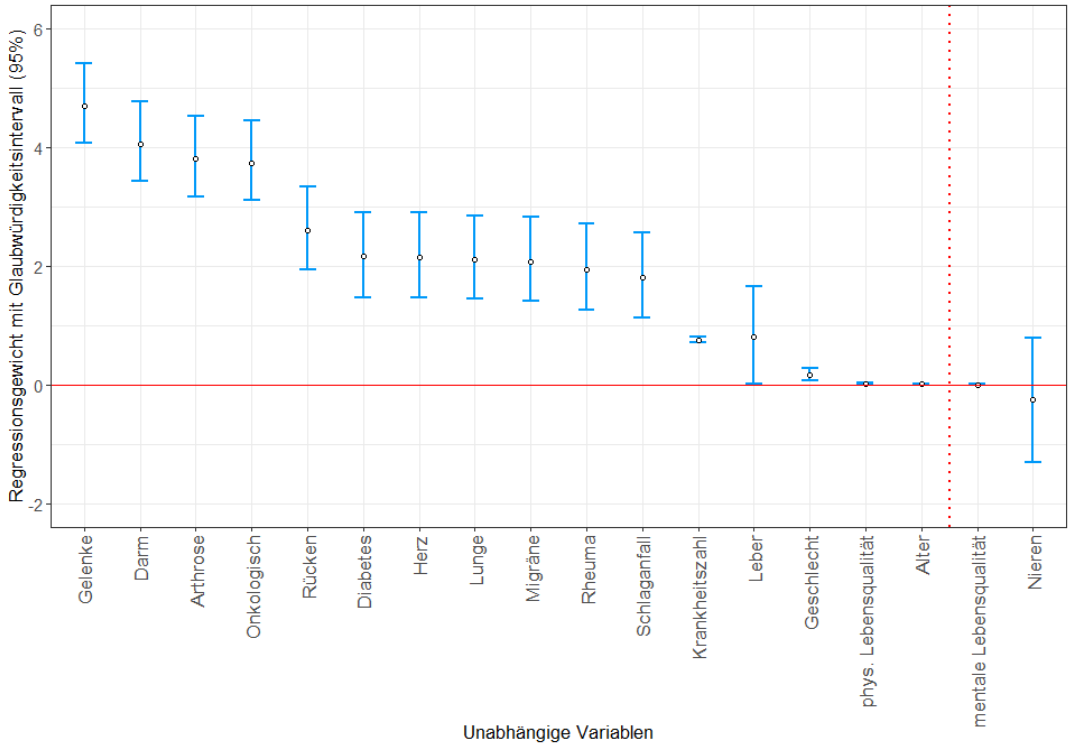

Abbildung 22: Darstellung der Regressionsgewichte mit jeweiligem Glaubwürdigkeitsintervall für Overreporting

[192] Dies ist der eigentliche inferenzstatistische Part.

Beim *Overreporting* zeigt die Variable *Alter* als einzige keinen statistisch signifikanten Zusammenhang (Abbildung 23).

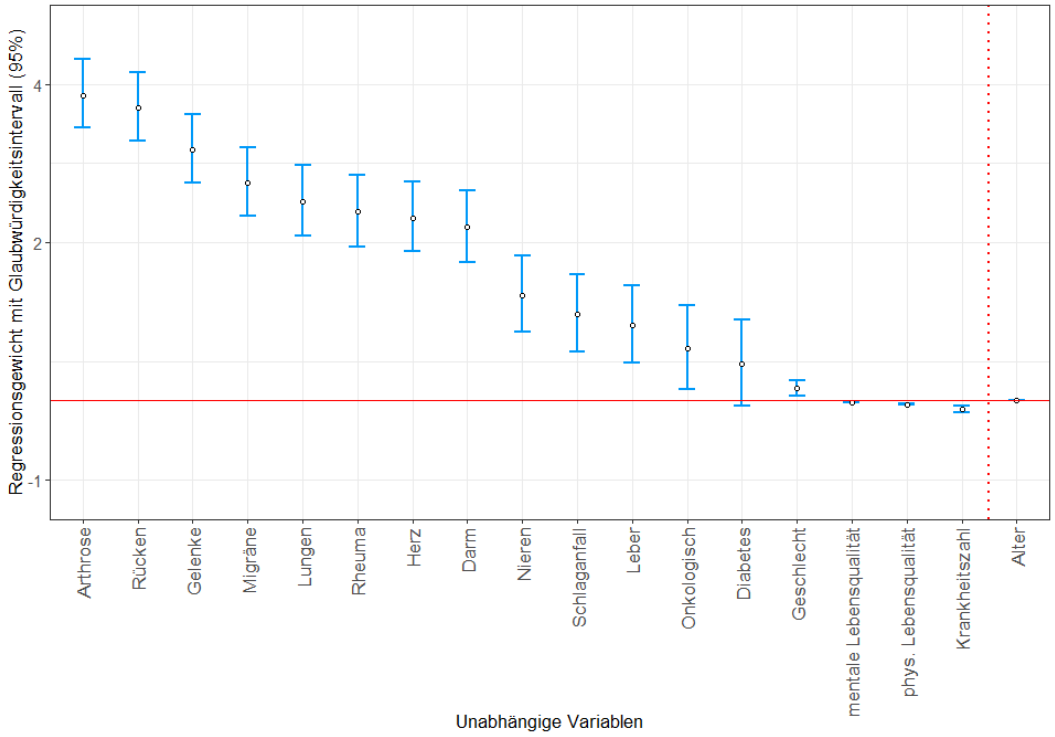

Abbildung 23: Darstellung der Regressionsgewichte mit jeweiligem Glaubwürdigkeitsintervall für Underreporting

5.3.6 Zusammenfassung

Ein Großteil der in dieser Analyse betrachteten unabhängigen Variablen ist statistisch signifikant mit dem Auftreten von *Under-* und *Overreporting* assoziiert. Bei diesen Methoden ist zu betrachten, dass die Erkrankungsgruppen mit statistischer Signifikanz Erkrankungen gegenüber der Wahrscheinlichkeit des Auftretens von *Over-* oder *Underreportings* bei Epilepsie getestet wurden (als eine Krankheitsgruppe mit geringen *Over-*, *Underreporting*-Raten). In Teilen scheinen manche unabhängige Variablen unterschiedlich bedeutsam für *Over-* und *Underreporting* zu sein (wie die Anzahl von Erkrankungen von Patient:innen), während andere ähnliche Effekte aufweisen (wie ein erhöhtes Risiko sowohl bei *Over-* als auch *Underreporting* bei männlichem gegenüber weiblichem Geschlecht).

Diese Ergebnisse der Parameterschätzungen können als verlässlich angesehen werden, da beide Ketten nach $\hat{R}$ und *ESS* gut konvergierten.

Zusätzlich sind Interaktionseffekte wahrscheinlich (beispielsweise zwischen *Alter*, *Erkrankungsanzahl* und *männlichem Geschlecht*). Aufgrund der mit Interaktionseffekten einhergehenden Modellkomplexität würde die Berechnungszeit viel Zeit in Anspruch nehmen und daher wurde im Rahmen dieser Arbeit darauf verzichtet.

5.4 Anwendung eines Verfahrens Maschinellen Lernens

Wie in Abschnitt 2.2.3 dargelegt, fokussiert sich die Anwendung Maschinellen Lernens auf die Vorhersage einer Outcome-Variable. In dem Rahmen dieser Masterarbeit wird anhand dieser Forschungsmethoden versucht zu überprüfen, ob die in dem inferenzstatistischen Verfahren als relevant identifizierten Variablen für solch eine Vorhersage geeignet sind.

Dementsprechend lautet die Forschungsfrage für diesen Abschnitt: *Kann aufgrund der vorab als relevant identifizierten Variablen das Reporting von Patient:innen korrekt vorhergesagt werden?*

Zur Beantwortung dieser Fragestellung wird auf Basis der Begrifflichkeiten aus Abschnitt 2.2.1.5 ein *Batch Model-Based Supervised Learning System* angewandt:

- *Batch:* Es liegt für das Training ein fester Datensatz vor, der nicht mit zusätzlichen Datensätzen angereichert wird.
- *Model-based:* Vorhersagen werden auf Basis von geschätzten Parametern getroffen.
- *Supervised:* Es liegen gelabelte Daten vor, an denen der Datensatz trainiert wird.

Es gibt verschiedene Bezugsmodelle, an denen sich die Entwicklung eines Systems Maschinellen Lernens orientieren kann: Zum Beispiel *CRISP-DM*[193], *KDD*[194] und *SEMMA*[195 & 196]. Allerdings besitzen *CRISP-DM* und *KDD* einen Business-Fokus[197], während die hier zu testende Anwendung einen reinen Erkenntniszweck beinhaltet. Dementsprechend wird das *SEMMA*-Bezugsmodell für diese Arbeit verwandt.

SEMMA beschreibt einen Ablauf von Data Mining-Prozessen zur Identifikation von Mustern in Daten und gliedert sich in folgende aufeinander aufbauende fünf Schritte: *Sample*, *Explore*, *Modify*, *Model* und *Assess*, deren Umsetzung in den folgenden Abschnitten beschrieben wird (siehe Abbildung 24).

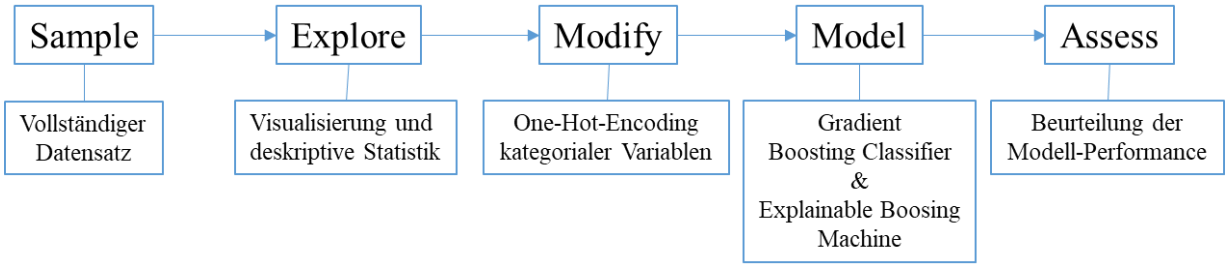

Abbildung 24: Ablauf des SEMMA-Modells und die Umsetzung der einzelnen Schritte in dieser Arbeit

[193] (Vgl. Schröer, Kruse und Gómez 2021).
[194] (Vgl. Fayyad und Stolorz 1997).
[195] (Vgl. SAS Institute Inc. 2017a).
[196] (Vgl. SAS Institute Inc. 2017b, 321–24).
[197] (Vgl. Dåderman und Rosander 2018, 28).

5.4.1 Sampling

Zunächst wird der Datensatz erstellt oder aufgeteilt. Dieser Schritt ist für diese Arbeit nicht notwendig, da der Datensatz bereits erstellt wurde und alle Datenpunkte für diesen Anwendungsfall genutzt werden.

5.4.2 Explore

Ein Datensatz wird auf erwartete Beziehungen, unerwartete Trends und Anomalien untersucht, um ein Verständnis der Daten zu erhalten und Ideen zu generieren.

In diesem Anwendungsfall liegt ein theoriegetriebenes Vorgehen vor und somit sind konkrete Vorstellungen vorab formuliert worden. Ein exploratives Vorgehen, um daraus ein entsprechendes Modell abzuleiten, ist in dieser Arbeit demnach nicht nötig, und es wurde sich auf die Visualisierung und Beschreibung der eingeschlossenen Variablen beschränkt

5.4.3 Modify

In diesem Schritt werden Variablen im Hinblick auf die Wahl des später durchzuführenden Modells ausgewählt, transformiert oder erstellt.

Eine Auswahl der Variablen fand wie erwähnt theoriegetrieben statt. Die Berücksichtigung von *Multikollinearität* hat in dieser Analyse keine Erfordernis, da *Multikollinearität* keinen Einfluss auf Vorhersage-Werte besitzt[198].

Hinsichtlich der Transformation von Variablen wird bei Modellen Maschinellen Lernens wie bei inferenzstatischen Modellen besondere Beachtung auf kategoriale Variablen gelegt. In dieser Arbeit liegen drei kategoriale Variablen mit mehr als zwei Stufen vor: Zum einen besitzt das Attribut der Krankheitsgruppen 14 Stufen. Für Modelle Maschinellen Lernens ist in diesem Fall keine *Dummy-Kodierung* wie für inferenzstatistische Modelle erforderlich. Allerdings muss die Variable auf andere Weise kodiert werden, um angemessen in das Modell aufgenommen zu werden: Für viele Modelle Maschinellen Lernen müssen die *Features* der *Attribute* anhand einer Zahl dargestellt werden, um von dem Algorithmus nicht als *ordinal* oder *metrisch* skaliert angesehen zu werden. Hierfür kann das Verfahren des *One-Hot-Encodings* genutzt werden: Die *Features* eines *Attributs* werden in jeweils eine eigene Variable transformiert. Im Gegensatz zur *Dummy-Kodierung* wird kein *Feature* als eine Referenzkategorie ausgewählt. Zum anderen liegen die Variablen *Arztpraxis* und *Patient:innen* mit mehr als zwei Stufen vor. Im Gegensatz zu dem inferenzstatischen Modell werden diese beiden Variablen in diesem Fall nicht mitaufgenommen: Es sind 2.893 Patient:innen und 145 Arztpraxen und dies würde bei dem *One-Hot-Encoding* zu ebenso vielen zusätzlichen Spalten führen, die die Modellkomplexität stark erhöhen würden.

[198] (Vgl. Bortz und Schuster 2010, 355–56).

5.4.4 Model

Hier wird das Modell spezifiziert, anhand dessen die Muster zur Outcome-Vorhersage geschätzt werden.

Als Methode Maschinellen Lernens wird ein *Gradient Boosting-Classifier* gewählt. Die Wahl dieser Methode wird wie folgt begründet:

- *Classifier:* Da es sich um eine dreistufig nominalskalierte Target-Variable handelt wird ein Algorithmus gewählt, der die Daten einem *Klassen-Label* zuordnen kann.
- *Boosting: Boosting* zählt zu den *Ensemble-Methoden*, die auf der Grundannahme beruhen, dass eine Gruppe von Modellen eine bessere *Performance* bietet als das beste Modell einzeln. Hierbei werden aus dem Datensatz Teil-Datensätze gebildet und für jeden eine Vorhersage berechnet. Hierbei werden die Prädiktoren sequenziell trainiert. Dies bedeutet, dass jeder neu geschätzte Prädiktor den vorhergehenden korrigiert.
- *Gradient:* Diese Korrektur kann je nach *Boosting*-Methode auf verschiedene Weisen vorgenommen werden. Bei der *Gradient Boosting*-Methode wird versucht, den Prädiktor auf Basis der Residuen („residual errors“) des Vorgängers anzupassen.

Wie aus der Visualisierung der Daten in Abbildung 19 ersichtlich, liegt bei der Outcome-Variable im Long Format eine Ungleichverteilung der Häufigkeiten der einzelnen Ausprägungen vor. Solch eine Ungleichverteilung kann zu einer Fehlklassifikation der *Klassen* führen, in denen die geringeren Häufigkeiten vorliegen. Auch wenn *Gradient Boosting*-Methoden bei einer Ungleichverteilung von Outcome-Variablen eine gute Performance erreichen sollen, kann bei sehr kleinen Datenmengen oder stark ungleich balancierten Datensätzen bei einem *Random Split* in Trainings- und Testdaten eine *Klasse* komplett von einem *Split* ausgeschlossen werden. Diesem wird in dieser Arbeit versucht durch folgende Maßnahmen entgegen zu wirken:

- *Stratify-Parameter* beim *Split* in Test- und Trainingsdaten: Hierbei wird der *Split* derart durchgeführt, so dass das Verhältnis der Klassen in beiden Sätzen gleich ist.
- *Sample Weights*: Hierbei werden Datensätze je nach ihrer Zugehörigkeit zu der Klasse mit der höheren oder niedrigeren Häufigkeit unterschiedlich gewichtet. Diejenigen mit den geringeren Häufigkeiten werden höher gewichtet.

Neben diesen Aspekten gilt es grundsätzlich, bei der Modellspezifikation die bestmögliche *Hyperparameter-Konfiguration* zu ermitteln: *Hyperparameter* sind Parameter, die vor dem Training festgelegt werden und sich an den Eigenschaften der Daten sowie der Kapazität des Algorithmus für den Lernprozess orientieren und somit einen Einfluss auf die *Performance* besitzen. Um die bestmögliche Konfiguration von *Hyperparametern* zu finden, gibt es verschiedene Strategien: Eine

davon ist die *Grid-Search*. Hierbei wird jede mögliche Kombination von Hyperparametern anhand des Trainingsdatensatzes auf Verbesserungen der *Performance* hin getestet.[199]

Dementsprechend wurden für diese Arbeit sechs Modelle berechnet und gegeneinander getestet:

- Ein *Baseline-Modell*, in dem die Hyperparameter-Grundeinstellungen der genutzten Software beibehalten wurden.
- Vier Modelle, in denen verschiedene Konstellationen folgender *Hyperparameter* getestet wurden: Anzahl der Boosting-Stufen (*n_estimators*), der *Learning Rates*, der maximalen Anzahl von *Features* um einen bestmöglichen *Split* zu vollziehen (*max_features*) und der maximalen Anzahl von Knoten (*max_depth*)[200].
- Ein finales Modell mit der bestmöglichen Konstellation an *Hyperparametern*[201].

Für die Analyse und der *Hyperparameteroptimierung* aller Modelle wurde als Software die *Python*-Bibliothek *sklearn*[202] genutzt.

Die Ergebnisse der Analysen werden in Tabelle 6 dargestellt. Grundsätzlich kann festgehalten werden, dass die Optimierung der *Hyperparameter* die *Modell-Performance* verbesserte. Als Aussage über die Güte der Performance werden die *Accuracy* (wie gut trifft das Modell Vorhersagen über alle Klassen), die *Precision* (alle korrekt als positiv klassifizierten Datenpunkte im Verhältnis zu allen entweder korrekt oder fälschlich als positiv klassifizierten Datenpunkten), der *Recall* (das Verhältnis aller korrekt positiv klassifizierten Datenpunkte zu allen tatsächlich als positiv zu klassifizierenden Datenpunkten) und der *F1-Score* (eine Art gewichteter Mittelwert von *Präzision* und *Recall*) herangezogen. Alle diese Kennzahlen können Werte zwischen 0 und 1 annehmen, wobei ein höherer Wert eine höhere Güte darstellt.

Im Folgenden werden die Ergebnisse des finalen Modells gegenüber dem *Baseline-Modell* beschreiben und in Tabelle 6 dargestellt:

- Modell insgesamt: Allgemein betrachtet würde man bei einer von 0.60 auf 0.88 gesteigerten *Accuracy* von einem tendenziell guten, aber verbesserungswürdigen Modell sprechen.
- *Agreement*: Hierbei zeigt sich ein guter *F1-Score*. Besonders der *Recall* ist sehr hoch, während die *Precision* anzeigt, dass es auch noch einige fälschlich positiv klassifizierten Datenpunkte gibt.
- *Overreporting*: Für diese Klasse ergaben sich durch die Veränderung der *Hyperparameter* zwei gegenläufige Entwicklungen, nämlich eine moderate Steigerung der *Precision* aber ein deutliches Absinken des *Recalls*.

[199] (Vgl. Agrawal 2021, Kap. 1 & 2).
[200] Syntaxen des Hyperparameter-Tunings in Anhang F
[201] Syntax des finalen Modells in Anhang G
[202] (Pedregosa et al. 2011).

- *Underreporting*: Hier zeigte sich ein ähnliches Bild wie bei dem *Overreporting*.

Tabelle 6: Die Perfomance-Kennzahlen links vor der Hyperparameteroptimierung und rechts danach

	Baseline-Modell			*Finales Modell*		
	Precision	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*
Agreement	0.96	0.59	0.73	0.89	0.98	0.93
Overreporting	0.17	0.62	0.27	0.28	0.07	0.11
Underreporting	0.17	0.78	0.28	0.39	0.15	0.21
Accuracy	0.60			0.88		

Neben einer angemessenen *Accuracy* und verwandter Kennzahlen gilt es wie unter Abschnitt 2.2.1.3 (S. 12) beschrieben, insbesondere in der Medizin interpretierbare Modelle aufzustellen. Diese *Interpretierbarkeit* kann auf zwei Ebenen betrachtet werden: Auf der Ebene *globaler Erklärbarkeit* (*Global Explanation*) und der *lokalen* (*Local Explanation*).

Globale Erklärbarkeit beinhaltet das Verständnis des gesamten Modells in Form der Summe der *Input-Features*. Dieser Einfluss der *Features* wird je *Feature* gemittelt als ein einzelnes Gewicht in Form einer *Feature Importance* dargestellt. Diese gibt in Prozent an, inwieweit sich der Vorhersage-Fehler durch das Entfernen des jeweiligen Features erhöht. Die Bibliothek *sklearn* gibt für den *Gradient Boosting-Classifier* eine Ausgabe dieser *Feature Importance* an. Die entsprechenden Ergebnisse für die hier durchgeführten Analysen, werden in Abbildung 25 dargestellt. Hierbei zeigte sich, dass insbesondere die *Lebensqualität* in *physischer* als auch *mentaler* Form einen deutlichen Beitrag zur Vorhersage leisteten, aber auch die *Krankheitszahl* und das *Alter*.

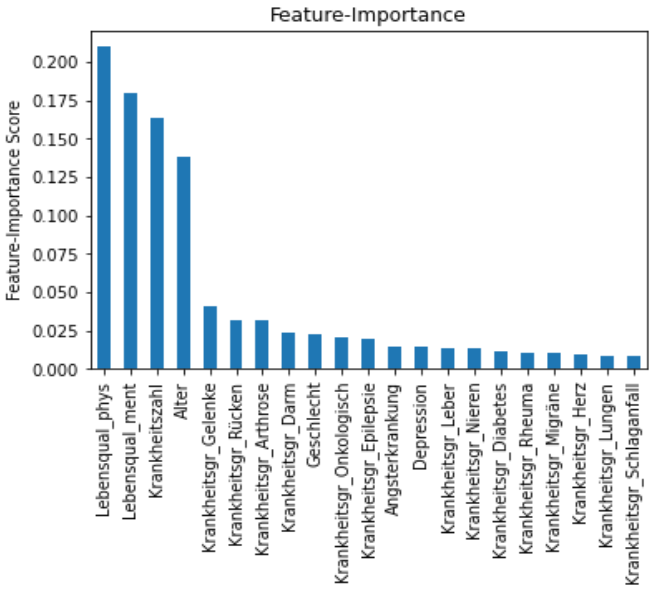

Abbildung 25: Feature-Importance nach der Hyperparameteroptimierung

Die Bibliothek *sklearn* bietet *globale Erklärungen* in dieser allgemeinen Form an, allerdings nicht in differenzierterer Form. Eine hierfür entsprechende *Python*-Bibliothek stellt *InterpetML* dar, die explizit ein besseres Verständnis sowohl für *globale* als auch *lokale Erklärbarkeit* bieten möchte[203]. Im Rahmen von *InterpretML* wird auch eine *Explainable Boosting Machine (EBM)* angeboten, die hier als

[203] (Nori et al. 2019).

entsprechendes Modell mit besserer *Interpretierbarkeit* zum vorab verwandten *Gradient Boosting-Classifier* angewandt wurde[204].

Die Ergebnisse der *Explainable Boosting Machine* zeigen folgende Ergebnisse: In Abbildung 26 kann betrachtet werden, wie sich zum Beispiel die *Feature Importance* der *physischen Lebensqualität* auf die Abstufungen der Outcome-Variable auswirken. Je besser die *Lebensqualität*, desto wahrscheinlicher ist eine Übereinstimmung zwischen Selbstauskunft und elektronisch gespeicherten Diagnosen und eines *Underreportings*. Je schlechter die *Lebensqualität*, desto wahrscheinlicher ist ein *Overreporting*, das mit Steigerung der Lebensqualität deutlich sinkt.

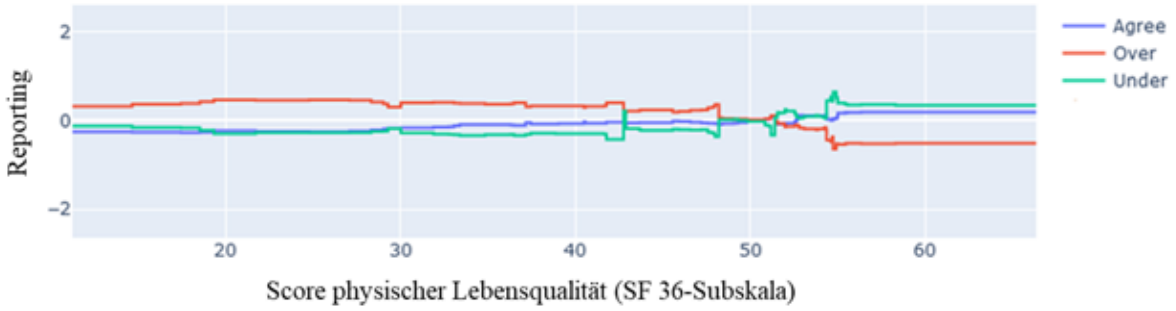

Abbildung 26: Wahrscheinlichkeit der Art des Reporting aufgrund der Ausprägungen der Patient:innen-Angaben auf der Skala physischer Lebensqualität (Grafik entnommen und angepasst aus den Analysen mit InterpretML)

Lokale Erklärbarkeit behandelt den Einfluss der Input-Features auf die Vorhersage eines Outcomes auf individueller Ebene (hier pro Krankheitsgruppe eines Patienten). Berechnet wird eine individuelle Vorhersage bei *InterpretML* durch sog. *Term Contributions* (gewichtete *Features*)[205]. Ein Ergebnis lokaler Interpretierbarkeit wird in Abbildung 27 veranschaulicht.

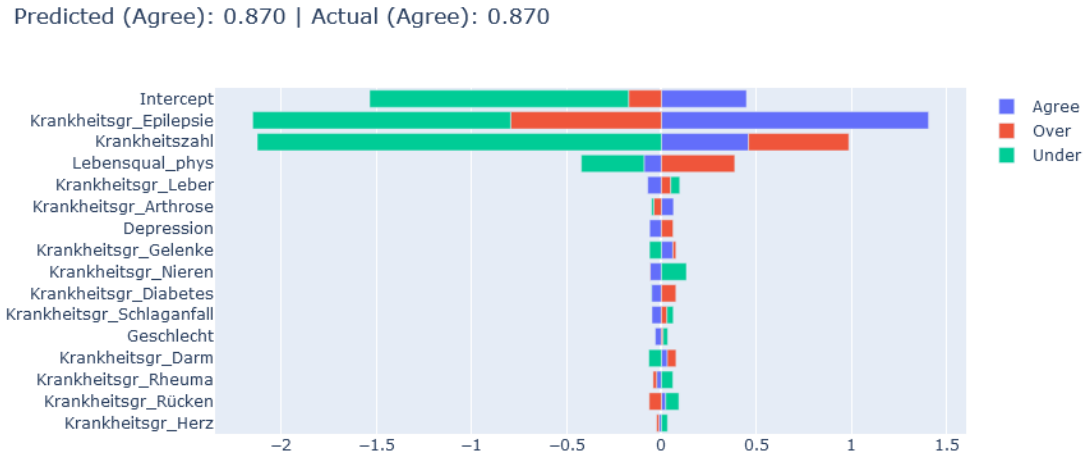

Abbildung 27: Beispiel lokaler Erklärbarkeit an einem Datenpunkt mit der Vorhersage und gegebenem Agreement

[204] Syntax des EBM-Modells in Anhang H
[205] (Nori et al. 2019, 4).

Die *Perfomance*-Kennzahlen dieses EBM-Modells sind in Tabelle 7 dargestellt. Sie zeigen ein ähnliches Ergebnis wie das Baseline-Modell anhand von *sklearn*. Potenzielle Gründe für diesen Unterschied, werden im kommenden Abschnitt diskutiert.

Tabelle 7: Performance-Kennzahlen des EBM-Modells

	Precision	*Recall*	*F1-Score*
Agree	0.96	0.59	0.73
Overreporting	0.16	0.62	0.26
Underreporting	0.17	0.74	0.28
Accuracy	0.60		

5.4.5 Assess (technische Evaluation)

In diesem Abschnitt werden die Verlässlichkeit und der Nutzen der Modell-Ergebnisse evaluiert.

Hinsichtlich der Modellergebnisse kann festgehalten werden, dass allein auf die *Accuracy* bezogen das Modell verbesserungswürdig ist, aber dennoch eine eher gute *Performance* besitzt. Allerdings wird dieses Ergebnis großteilig von der größten Klasse der Outcome-Variable (*Agreement*) bestimmt. Bei den anderen beiden Klassen (*Over-* und *Underreporting*) ergibt sich ein schlechtes Klassifikationsergebnis: Bei dem *Baseline-Modell* zeigte sich zunächst eine sehr niedrige *Precision* und nach der *Hyperparameteroptimierung* ein sehr niedriger *Recall*.

Dieses kann gegebenenfalls anhand von drei Problemen erklärt werden: Zum einen anhand des Missverhältnisses der Häufigkeiten zwischen den Klassen, die zu einer Fehlspezifikation der Klassen mit den geringeren Häufigkeiten führen kann. Zum anderen wurden vier potenziell relevante *Attribute* nicht in das Modell aufgenommen: Der *Bildungsstand* aufgrund zu vieler fehlender Werte und ein Maß zur Erfassung der *Health Literacy* der Patienten wurde in dem Studiendesign nicht vorgesehen. Ebenso wurden die Arztpraxen als auch die Patient:innen von der Analyse ausgeschlossen, da ein One-Hot-Encoding zu viele Variablen erzeugt hätte. Weiterhin wurde der Prozess zur *Hyperparameteroptimierung* aufgrund der Dauer der Berechnungszeit ab einem bestimmten Punkt nicht mehr weitergeführt.

Die letzten beiden Punkte würden dem beim Maschinellen Lernen potenziell zu begegnenden Problem des *Underfittings* entsprechen[206]: Hierbei wäre das Modell zu einfach, um die Struktur in den Daten zu lernen und es bräuchte stärkere Modelle mit mehr Parametern, die Aufnahme besserer Attribute und/oder eine verstärkte Freisetzung von Hyperparametern.

Ein weiteres mit Modellen Maschinellen Lernens verbundenes Problem liegt in dem des *Overfittings* (wenn das Modell zu stark an die Trainingsdaten angepasst ist und nicht auf andere Datensätze

[206] (Vgl. Géron 2019, Kap. 1).

generalisieren kann): Eine Beurteilung hierüber steht aus, da hierfür die in dieser Arbeit entwickelten Modelle in Replikationsstudien genutzt werden müssten.

Somit kann nicht abschließend beurteilt werden, ob *Over-* und *Underreporting* ausreichend gut vorhergesagt werden könnten. Eine Analyse mit einer größeren Stichprobe der beiden Klassen als auch die Hinzunahme weiterer Variablen wäre nötig.

Wie ersichtlich ist die Performance des *EBM-Modells* trotz Umsetzung der Empfehlungen zur *Hyperparameteroptimierung*[207] geringer als die des *Gradient Boosting-Classifiers*. Als Erklärung kann angeführt werden, dass für *EBM-Modelle* mit einer Outcome-Variable mit mehreren Klassen keine Interaktionen im Gegensatz zu dem Modell von *sklearn* zwischen den Attributen zugelassen werden.

6 Ableitung weiterer Anwendungsgebiete

Aus der Umsetzung der drei vorangegangenen Teilziele kann ein weiteres Anwendungsgebiet abgleitet werden: Die Prüfung von Interventionen in der allgemeinmedizinischen Versorgung. Hierunter kann ein Studiendesign verstanden werden, in dem beispielsweise ein bestimmtes Praxismanagementsystem oder Behandlungskonzept gegenüber einer Routinebehandlung getestet wird. Als Outcome-Variable wird hierbei häufig die potenzielle Veränderung der Krankheitslast von Patient:innen spezifiziert (wie die Verringerung depressiver Symptome[208] oder die Verringerung der Anzahl thromboembolischer Ereignisse einer Therapienebenwirkung[209]).

Dieses Anwendungsgebiet ist bei der Literaturrecherche zur Identifikation von Publikationen der allgemeinmedizinischen Versorgungsforschung nicht vorgekommen und wurde auch nicht in den Interviews genannt. In dem Screening allgemeinmedizinischer Fachzeitschriften kamen solche Studien gehäuft vor, allerdings ohne Anwendungsbezug zu Maschinellem Lernen. Aufgrund der prototypischen Durchführung und Evaluation des Anwendungsfalls wurde allerdings ersichtlich, dass auch für solche Interventionsstudien Maschinelles Lernen potenziell relevant sein kann: Durch ein Maß wie der *Feature Importance* könnte auch der Kontrast einer Interventionsmaßnahme gegenüber einer Routinebehandlung quantifiziert werden.

[207] (Vgl. Microsoft Research 2021).
[208] (Vgl. Gensichen et al. 2009).
[209] (Vgl. Mertens et al. 2019).

7 Evaluation der Ergebnisse

Das Hauptziel dieser Arbeit besteht in der Identifikation des Anwendungspotenzials Maschinellen Lernens in der allgemeinmedizinischen Versorgungsforschung. Dieses Ziel wurde versucht anhand von vier Teilzielen zu erreichen:

Das Teilziel 1 umfasste eine systematische Literaturrecherche, ein Literaturscreening als auch Interviews mit Expert:innen. Die Literaturrecherche war umfassend und die Suchstrategie mitsamt den aus ihr resultierenden Ergebnissen kann als angemessen eingestuft werden, da qualitativ hochwertige Datenbanken mit einer umfassenden Abfrage durchsucht wurden. Das Literaturscreening umfasste zwei Jahrgänge einer deutschen und zweier internationaler Fachzeitschriften. Hierbei konnte ein umfassender Überblick über die aktuell verwendeten Forschungsmethoden erhalten werden, unter die Maschinelles Lernen allerdings nicht fiel. Die Interviews wurden bei einer Stichprobengröße von sechs Personen durchgeführt. Als Limitation kann angesehen werden, dass ein allgemeiner umfassender Überblick Maschinellen Lernens in der Allgemeinmedizin nur durch Interviews mit einem größeren Anteil nationaler oder internationaler Expert:innen hätte erreicht werden können. Dieses war aber nicht möglich. Diese Limitation kann relativiert werden, da die eingeschlossenen Interviewpartner großteilig gut vernetzt sind und somit auch Einblicke in die Arbeiten und Entwicklungen anderer Institute und Arbeitsgruppen besitzen. Als eine weitere Limitation bei der Durchführung der Interviews ist anzusehen, dass kein *Pretest* eingeplant wurde. Auf Basis dieses Pretests hätten differenziertere Fragen entwickelt werden können, insbesondere bezüglich potenzieller Anwendungsgebiete Maschinellen Lernens. Diese Limitation kann sich auch einschränkend auf die Teilziele 2 und 4 bezüglich der Ableitung von Anwendungsgebieten ausgewirkt haben.

Die Entwicklung und Durchführung eines Anwendungsfalls anhand Maschinellen Lernens basierte auf einem aus dem Teilziel 1 abgeleiteten Anwendungsgebiet. Die *Performance* des Modells ist als mäßig anzusehen, was zum einen an der nicht balancierten Häufigkeitsverteilung in den Klassen der Outcome-Variable liegen könnte. Diesem wurde versucht vorab mit einer *Stratifizierung* der Daten bei dem *Split* in Trainings- und Testdaten zu begegnen als auch einer entsprechenden Gewichtung der Klassen bei der Schätzung des Modells. Zum anderen kann die mäßige *Modell-Performance* durch das Fehlen potenziell relevanter *Attribute* in dem Modell liegen. Als eine weitere Möglichkeit kann gelten, dass keine gute Vorhersage möglich ist, da es keine systematisch erklärbare Fundierung von *Over-* und *Underreporting* gibt und dieses Phänomen eher aus zufälligen Variationen entsteht. Grundsätzlich kann allerdings festgehalten werden, dass in Bezug auf die Stichprobengröße und auch der Variablenauswahl ein für die allgemeinmedizinische Versorgungsforschung repräsentativer als auch relevanter Datensatz vorlag: Patient:innen und Arztpraxen werden für eine bestimmte Studie rekrutiert und viele Variablen spezifisch für diese Studie gesammelt. Datensätze, die *Big Data*-Eigenschaften besitzen, kommen in der allgemeinmedizinischen Versorgungsforschung so gut wie nicht vor.

Als Kennzahlen der Relevanz von Variablen können in der Inferenzstatistik *Odds Ratios* und bei Maschinellem Lernen die *Feature Importance* angesehen werden. Hier soll nun noch betrachtet werden, ob die Ergebnisse beider Methoden in Bezug auf die Relevanz von Variablen eine Übereinstimmung zeigen:

- Grundsätzlich muss festgehalten werden, dass das inferenzstatistische Modell Effekte getrennt für *Over-* und *Underreporting* anzeigt, während diese Trennung bei dem Modell Maschinellem Lernens nicht vorgenommen wurde
- Zunächst scheinen sich am Beispiel der *mentalen Lebensqualität* unterschiedliche Ergebnisse zwischen beiden Methoden zu zeigen: Im inferenzstatistischen Modell ist sie bei *Underreporting* statistisch nicht signifikant und bei *Overreporting* zeigt sich trotz statistischer Signifikanz kein starker Effekt. Dagegen wird sie für die Vorhersage anhand Maschinellen Lernens als relativ relevant eingestuft. Jedoch müssen sich diese Ergebnisse nicht gegenseitig ausschließen, da inferenzstatistisch *Over-* und *Underreporting* jeweils getrennt gegen *Agreement* getestet wurde, aber beide Abstufungen nicht gegeneinander. Somit kann mentale Lebensqualität bei einer Vorhersage aller drei Klassen relevant sein. Das gleiche kann auch für die Variable *Alter* gelten.
- Gelenks- und arthritische Erkrankungen zeigen inferenzstatistisch starke Effekte, da sie gegenüber dem Effekt einer Referenzkategorie getestet wurden. Dies ist bei Maschinellem Lernen nicht der Fall, aber beide Erkrankungsgruppen sind auch bei diesem als relativ relevant angezeigt.

Daher müssen sich die inferenzstatischen und Maschinelles Lernen-Ergebnisse nicht widersprechen.

Die in dieser Arbeit durchgeführte Vorgehensweise und die gezogenen Schlussfolgerungen in Bezug auf die Domäne wurden zum Abschluss von einer Expertin der allgemeinmedizinischen Versorgungsforschung hinsichtlich ihrer Angemessenheit reflektiert[210]: Die systematische Literaturrecherche in Form der Wahl der Meta-Datenbank und der Suchstrategie wurden von ihr als angemessen eingestuft, die Ableitung der Anwendungsgebiete aus den gefunden Anwendungsfällen als plausibel und der für den prototypischen Anwendungsfall gewählte Datensatz als für die allgemeinmedizinische Versorgungsforschung repräsentativ. Die im prototypischen Anwendungsfall gesetzten Fragestellungen seien nach ihrer Aussage typisch für diese Domäne. Die grundsätzliche Vorgehensweise, anhand inferenzstatistischer Methoden relevante Variablen zu identifizieren und diese mit Methoden Maschinellen Lernens auf ihre Fähigkeit hin zu testen, gute Vorhersagen zu treffen, sei laut ihrer Aussage ein potenziell interessanter Ansatz.

[210] (Vgl. van den Akker 2022).

8 Zusammenfassung und Ausblick

Im Hinblick auf das Hauptziel dieser Arbeit kann geschlussfolgert werden, dass Maschinelles Lernen als ergänzende Forschungsmethode für die Domäne der allgemeinmedizinischen Versorgungsforschung potenziell relevant ist: Für wissenschaftliche Domänen erscheint ein theoriegetriebener Einsatz angemessen, wenn die verwendeten Daten spezifisch für Studien erhoben wurden. Dies geht eher mit einer begrenzten Anzahl von erhobenen Variablen sowie eingeschlossenen Patient:innen und Arztpraxen einher. In diesem Kontext kann Maschinelles Lernen ergänzend zu anderen Forschungsmethoden einen Beitrag zur Absicherung empirischer Erkenntnisse liefern:

- Inferenzstatistische Methoden können überprüfen, ob und in welchem Ausmaß vorab theoretisch abgeleitete Variablen mit einem Phänomen in Form einer Outcome-Variable assoziiert sind oder sie kausal erklären können.
- Methoden Maschinellen Lernens können erfassen, ob und in welchem Ausmaß die theoretisch abgeleiteten Variablen das entsprechende Phänomen vorhersagen können.
- Qualitative Auswertungsverfahren können eine Vorarbeit zur Generierung von Theorien leisten oder Erkenntnisse von Methoden der Inferenzstatistik und Maschinellen Lernens tiefer elaborieren.

Es hat sich im Rahmen dieser Arbeit gezeigt, dass Maschinelles Lernen für eine umfassende Menge von Anwendungsgebieten in der Versorgungsforschung relevant sein kann (von der *Bedarfsermittlung* über *Patientensicherheit* zu *Interventionsstudien*) und für numerische als auch nicht-numerische Daten geeignet ist.

Der für diese Arbeit genutzte Anwendungsfall hat potenzielle Einsatzmöglichkeiten aufgezeigt, sowie gleichzeitig auch potenzielle Schwächen: Das in dieser Studie entwickelte Modell zeigte keine gute *Performance*. Insofern steht eine erfolgreiche Anwendung Maschinellen Lernens in der allgemeinmedizinischen Versorgungsforschung noch aus. Hierfür sind neben einer guten theoretischen Fundierung eine ausreichende Stichprobengröße als auch eine angemessene Anzahl von Attributen bedeutsam.

Perspektivisch kann bei Modellen mit einer angemessenen und generalisierbaren *Performance* angedacht werden, deren Algorithmen automatisiert in die klinischen Praxis zur Identifikation von Risikopatienten zu überführen: Auf den Anwendungsfall bezogen könnten diese Patient:innen identifizieren, die potenziell keinen Überblick über ihren Gesundheitszustand besitzen.

Eine andere Perspektive hinsichtlich des Einsatzes Maschinellen Lernens liefern Maas et al.[211], die ein Rahmenmodell zur Verbindung daten- und theoriegetriebener Forschung auf großen Datensätzen basierend entwickelt haben: Laut ihrer Aussage könne eine reine Betonung von Datenanalysen ohne

[211] (Vgl. Maass et al. 2018).

eine domänenbezogene Theorie nur situationsspezifische Fragen beantworten, aber nicht zu dauerhaftem Wissen im wissenschaftlichen Sinne beisteuern. Dagegen würde ein ausschließlicher Schwerpunkt auf Theorien mit potenziell kleinen Datensätzen unter Umständen verhindern, zu wichtigen Erkenntnissen zu gelangen. Hierbei stelle die Verfügbarkeit von *Big Data* eine Voraussetzung dar. Wie in Abbildung 28 ersichtlich, würden hierfür anhand datengetriebener Forschung *Big Data*-Datensätze nach Mustern untersucht und daraus Erklärungen und Interpretationen abgeleitet, die in eine Theorie in Hinblick auf ihre Entwicklung, Verbesserung oder Bestätigung einfließen würden. Umgedreht würde theoriegetriebene Forschung Konstrukte und sie umfassende Hypothesen aufstellen, die wiederum Daten zu ihrer Überprüfung bräuchten. Allerdings wären beide Ansätze vom Wissen und Kompetenzen her nicht für einzelne Forschende zu bewältigen. Somit bräuchte es zusätzlich Forschende in der Domäne der Informationstechnologie, die eine Brücke zwischen beiden Ansätzen bilden könnten[212].

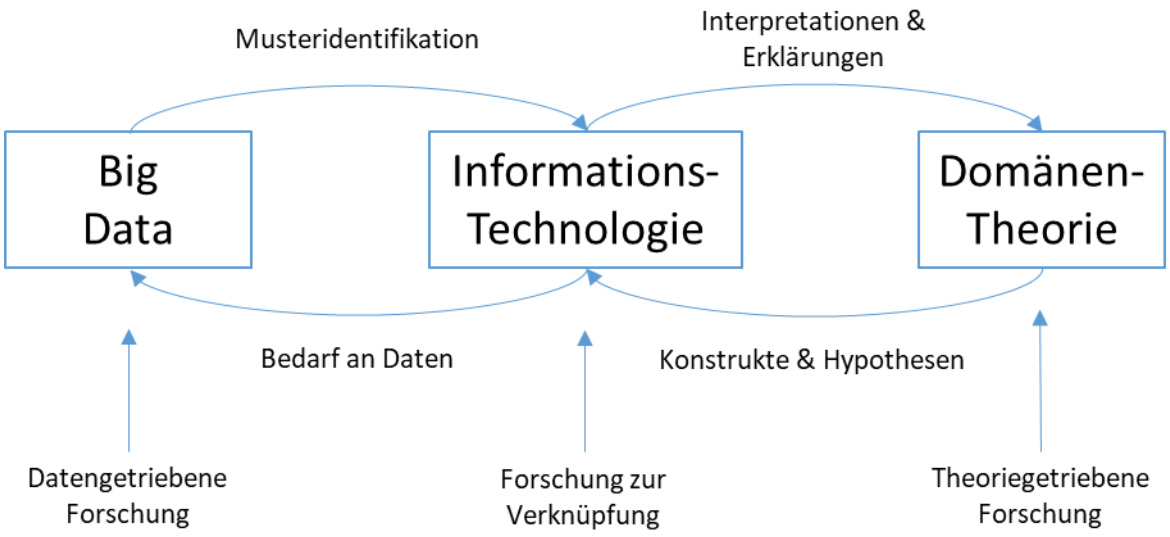

Abbildung 28: Verbindung datengetriebener Forschung auf Basis von Big Data und theoriegetriebener Forschung anhand von Informationstechnologie (eigene Abbildung in Anlehnung an Maas et al.[213])

Schlussendlich kann festgehalten werden, dass Maschinelles Lernen in der allgemeinmedizinischen Versorgungsforschung das Potenzial besitzt, herkömmliche theoriegetriebene Forschung zu ergänzen und deren Ergebnisse zu überprüfen, konkrete Anwendungen in der klinischen Praxis zu entwickeln, als auch in einem für die Allgemeinmedizin in der Zukunft liegenden interdisziplinären Kontext von Big Data in Kombination mit anderen Forschungsmethoden eine Rolle zu finden.

[212] (Vgl. Maass et al. 2018, 1259).
[213] (Vgl. Maass et al. 2018, 1265).

9 Literaturverzeichnis

Adadi, Amina, Safae Adadi und Mohammed Berrada. 2019. „Gastroenterology Meets Machine Learning: Status Quo and Quo Vadis.“ *Advances in Bioinformatics* 2019:1870975. https://doi.org/10.1155/2019/1870975.

Adombi, Adoubi Vincent De Paul, Romain Chesnaux und Marie-Amélie Boucher. 2021. „Review: Theory-Guided Machine Learning Applied to Hydrogeology—State of the Art, Opportunities and Future Challenges.“ *Hydrogeol J* 29 (8): 2671–83. https://doi.org/10.1007/s10040-021-02403-2.

Agrawal, Tanay. 2021. *Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient.* Berkeley, CA: Apress L. P.

Anand, Gopesh, Eric C. Larson und Joseph T. Mahoney. 2020. „Thomas Kuhn on Paradigms.“ *Prod Oper Manag* 29 (7): 1650–57. https://doi.org/10.1111/poms.13188.

Anderson, Jeffrey P., Jignesh R. Parikh, Daniel K. Shenfeld, Vladimir Ivanov, Casey Marks, Bruce W. Church, Jason M. Laramie et al. 2015. „Reverse Engineering and Evaluation of Prediction Models for Progression to Type 2 Diabetes: An Application of Machine Learning Using Electronic Health Records.“ *Journal of Diabetes Science and Technology* 10 (1): 6–18. https://doi.org/10.1177/1932296815620200.

Angelov, Plamen P., Eduardo A. Soares, Richard Jiang, Nicholas I. Arnold und Peter M. Atkinson. 2021. „Explainable artificial intelligence: an analytical review.“ *WIREs Data Mining Knowl Discov* 11 (5). https://doi.org/10.1002/widm.1424.

Armstrong, Grayson W. und Alice C. Lorch. 2020. „A(Eye): A Review of Current Applications of Artificial Intelligence and Machine Learning in Ophthalmology.“ *International Ophthalmology Clinics* 60 (1): 57–71. https://doi.org/10.1097/IIO.0000000000000298.

Backhaus, Klaus, Bernd Erichson, Wulff Plinke und Rolf Weiber. 2018. *Multivariate Analysemethoden.* Berlin, Heidelberg: Springer Berlin Heidelberg.

Bandyopadhyay, Prasanta S. und Malcolm R. Forster. 2011. *Philosophy of statistics.* Handbook of philosophy of science v. 7. Oxford, Oxford: North-Holland.

Barman, Debaditya und Nirmalya Chowdhury. 2020. „A novel semi supervised approach for text classification.“ *Int. j. inf. tecnol.* 12 (4): 1147–57. https://doi.org/10.1007/s41870-018-0137-9.

Bhowmick, Alexy und Shyamanta M. Hazarika. 2018. „E-Mail Spam Filtering: A Review of Techniques and Trends.“ In *Advances in Electronics, Communication and Computing*, 583–90: Springer, Singapore. https://link.springer.com/chapter/10.1007/978-981-10-4765-7_61.

Bogner, Alexander, Beate Littig und Wolfgang Menz. 2014. *Interviews mit Experten.* Wiesbaden: Springer Fachmedien Wiesbaden.

Bortz, Jürgen. 2005. *Statistik für Human- und Sozialwissenschaftler: Mit … 242 Tabellen.* 6., vollst. überarb. und aktualisierte Aufl. Springer-Lehrbuch. Heidelberg: Springer. https://ebookcentral.proquest.com/lib/gbv/detail.action?docID=1156268.

Bortz, Jürgen und Christof Schuster. 2010. *Statistik für Human- und Sozialwissenschaftler: Mit … 163 Tabellen.* 7., vollst. überarb. und erw. Aufl. Springer-Lehrbuch. Berlin, Heidelberg, New York: Springer. http://www.blickinsbuch.de/item/66fbce70c339aee077cc240a3b2bed04.

Brunton, Steven L. und J. Nathan Kutz. 2019. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control.* Cambridge, New York, NY, Port Melbourne, New Delhi, Singapore: Cambridge University Press.

Bürkner, Paul-Christian. 2017. „Brms : An R Package for Bayesian Multilevel Models Using Stan.“ *J. Stat. Soft.* 80 (1): 1–28. https://doi.org/10.18637/jss.v080.i01.

Bzdok, Danilo, Naomi Altman und Martin Krzywinski. 2018. „Statistics Versus Machine Learning.“ *Nature methods* 15 (4): 233–34. https://doi.org/10.1038/nmeth.4642.

Carrón, Javier, Yolanda Campos-Roca, Mario Madruga und Carlos J. Pérez. 2021. „A Mobile-Assisted Voice Condition Analysis System for Parkinson's Disease: Assessment of Usability Conditions.“ *Biomedical engineering online* 20 (1): 114. https://doi.org/10.1186/s12938-021-00951-y.

Caruana, Rich und Alexandru Niculescu-Mizil. 2006. „An Empirical Comparison of Supervised Learning Algorithms.“ In *Proceedings of the 23rd International Conference on Machine Learning*, hrsg. von William Cohen, 161–68. ACM Other conferences. New York, NY: ACM.

Chahar, Ravita und Deepinder Kaur. 2020. „A systematic review of the machine learning algorithms for the computational analysis in different domains.“ *International Journal of Advanced Technology and Engineering Explorat* 7 (71): 147–64. https://www.researchgate.net/profile/ravita-chahar/publication/346499690_a_systematic_review_of_the_machine_learning_algorithms_for_the_computational_analysis_in_different_domains/links/5fe2c32b92851c13feb198f3/a-systematic-review-of-the-machine-learning-algorithms-for-the-computational-analysis-in-different-domains.pdf.

Chen, Henian, Patricia Cohen und Sophie Chen. 2010. „How Big Is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies.“ *Communications in Statistics - Simulation and Computation* 39 (4): 860–64. https://doi.org/10.1080/03610911003650383.

Cheng, Heng-Tze Cheng. 2016. „Google AI Blog: Wide & Deep Learning: Better Together with TensorFlow.“ Zugriff am 15. November 2021. https://ai.googleblog.com/2016/06/wide-deep-learning-better-together-with.html.

Chin, Y. P. H., Z. Y. Hou, M. Y. Lee, H. M. Chu, H. H. Wang, Y. T. Lin, A. Gittin et al. 2020. „A Patient-Oriented, General-Practitioner-Level, Deep-Learning-Based Cutaneous Pigmented Lesion Risk Classifier on a Smartphone.“ *The British journal of dermatology* 182 (6): 1498–1500. https://doi.org/10.1111/bjd.18859.

Chmiel, F. P., D. K. Burns, M. Azor, F. Borca, M. J. Boniface, Z. D. Zlatev, N. M. White, T. W. V. Daniels und M. Kiuber. 2021. „Using Explainable Machine Learning to Identify Patients at Risk of Reattendance at Discharge from Emergency Departments.“ *Scientific reports* 11 (1): 21513. https://doi.org/10.1038/s41598-021-00937-9.

Christina, V., S. Karpagavalli und G. Suganya. 2010. „Email Spam Filtering using Supervised Machine Learning Techniques.“ *International Journal on Computer Science and Engineering* 2 (9): 3126–29.

Cox, Christopher R., Emma H. Moscardini, Alex S. Cohen und Raymond P. Tucker. 2020. „Machine Learning for Suicidology: A Practical Review of Exploratory and Hypothesis-Driven Approaches.“ *Clinical psychology review* 82:101940. https://doi.org/10.1016/j.cpr.2020.101940.

Dåderman, A. und S. Rosander. 2018. „Evaluating Frameworks for Implementing Machine Learning in Signal Processing : A Comparative Study of CRISP-DM, SEMMA and KDD.“ Bachelor Thesis, KTH, School of Electrical Engineering and Computer Science (EECS). Zugriff am 14. März 2022. http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-235408.

Deutsche Gesellschaft für Allgemeinmedizin und Familienmedizin e. V. 2002. „Fachdefinition: Beschluss der Jahreshauptversammlung.“ Zugriff am 4. Oktober 2021. https://www.degam.de/fachdefinition.html.

Deutsche Gesellschaft für Allgemeinmedizin und Familienmedizin e. V. 2021. „Zeitschrift für Allgemeinmedizin.“ Zugriff am 10. März 2022. https://www.online-zfa.de/.

Deutsche Gesellschaft für Allgemeinmedizin und Familienmedizin e.V. 2012. „Allgemeinmedizin - spezialisiert auf den ganzen Menschen: Positionen zur Zukunft der Allgemeinmedizin und der hausärztlichen Praxis.“ Zugriff am 19. April 2022. https://www.degam.de/files/Inhalte/Degam-Inhalte/Ueber_uns/Positionspapiere/DEGAM_Zukunftspositionen.pdf.

De-yu, Chao. 2021. „What is the difference between Traditional Programming and Machine Learning?“. *MLearning.ai*, 13. Mai 2021. https://medium.com/mlearning-ai/what-is-the-difference-between-traditional-programming-and-machine-learning-f6128ed4f595.

Diaz-Bone, Rainer. 2019. *Statistik für Soziologen.* 5., überarbeitete Auflage. utb-studi-e-book 2782. München: UVK Verlag. https:elibraryutbdedoibook10361989783838552101.

DocCheck Medical Services GmbH. 2022a. „Grundlagenforschung - DocCheck Flexikon.“ Zugriff am 1. März 2022. https://flexikon.doccheck.com/de/Grundlagenforschung.

DocCheck Medical Services GmbH. 2022b. „Medizin - DocCheck Flexikon.“ Zugriff am 6. Januar 2022. https://flexikon.doccheck.com/de/Humanmedizin.

Donepudi, Praveen Kumar. 2020. „Impact of Machine Learning in Neurosurgery: A Systematic Review of Related Literature.“ *1* 8 (1): 13–20. https://doi.org/10.18034/mjmbr.v8i1.520.

Döring, Nicola und Jürgen Bortz. 2016. *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Unter Mitarbeit von Sandra Pöschl-Günther. 5. vollständig überarbeitete, aktualisierte und erweiterte Auflage 2016. SpringerLink Bücher. Berlin, Heidelberg: Springer Berlin Heidelberg. http://link.springer.com/book/10.1007/978-3-642-41089-5.

Doupe, Patrick, James Faghmous und Sanjay Basu. 2019. „Machine Learning for Health Services Researchers.“ *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 22 (7): 808–15. https://doi.org/10.1016/j.jval.2019.02.012.

Droste, S. 2008. „Informations- und Wissensmanagement.“ In *Health technology assessment: Konzepte, Methoden, Praxis für Wissenschaft und Entscheidungsfindung*, hrsg. von Perleth M. Busse R. Gerhardus A. Gibis B. Lühmann D. 1. Aufl., 99–134. Berliner Schriftenreihe Gesundheitswissenschaften. Berlin: Medizinisch Wiss. Verl.-Ges.

Droste, S. und C.-M. Dintsios. 2011. „Informationsgewinnung für gesundheitsökonomische Evaluationen im Rahmen von HTA-Berichten.“ *Gesundh ökon Qual manag* 16 (01): 35–57. https://doi.org/10.1055/s-0029-1245460.

Dudchenko, Aleksei, Matthias Ganzinger und Georgy Kopanitsa. 2020. „Machine Learning Algorithms in Cardiology Domain: A Systematic Review.“ *TOBIOIJ* 13 (1): 25–40. https://doi.org/10.2174/1875036202013010025.

Eid, Michael, Mario Gollwitzer und Manfred Schmitt. 2017. *Statistik und Forschungsmethoden: Lehrbuch. Mit Online-Material.* Originalausgabe, 5., korrigierte Aufl. Weinheim: Beltz. http://nbn-resolving.org/urn:nbn:de:bsz:31-epflicht-1119447.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau und Sebastian Thrun. 2017. „Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks.“ *Nature* 542 (7639): 115–18. https://doi.org/10.1038/nature21056.

Fayyad, Usama und Paul Stolorz. 1997. „Data mining and KDD: Promise and challenges.“ *Future Generation Computer Systems* 13 (2-3): 99–115. https://doi.org/10.1016/S0167-739X(97)00015-0.

fh Gesundheitsberufe OÖ. „Primärversorgung (Primary Health Care): Ein Überblick.“ Wissen.schafft.Gesundheit. Unveröffentlichtes Manuskript.

Ford, Elizabeth, Philip Rooney, Seb Oliver, Richard Hoile, Peter Hurley, Sube Banerjee, Harm van Marwijk und Jackie Cassell. 2019. „Identifying Undetected Dementia in UK Primary Care Patients:

A Retrospective Case-Control Study Comparing Machine-Learning and Standard Epidemiological Approaches.“ *BMC Med Inform Decis Mak* 19 (1): 248. https://doi.org/10.1186/s12911-019-0991-9.

Franke, Alexa. 2012. *Modelle von Gesundheit und Krankheit.* 3., überarbeitete Auflage. Programmbereich Gesundheit / Verlag Hans Huber. Bern: Verlag Hans Huber.

Gelman, Andrew, Ben Goodrich, Jonah Gabry und Aki Vehtari. 2019. „R-squared for Bayesian Regression Models.“ *The American Statistician* 73 (3): 307–9. https://doi.org/10.1080/00031305.2018.1549100.

Gensichen, Jochen, Michael von Korff, Monika Peitz, Christiane Muth, Martin Beyer, Corina Güthlin, Marion Torge et al. 2009. „Case Management for Depression by Health Care Assistants in Small Primary Care Practices: A Cluster Randomized Trial.“ *Annals of Internal Medicine* 151 (6): 369–78. https://doi.org/10.7326/0003-4819-151-6-200909150-00001.

Géron, Aurélien. 2019. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* Second edition. Beijing, Boston, Farnham, Sebastopol, Tokyo: O'Reilly. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=5892320.

Gigerenzer, Gerd. 2004. „Mindless statistics.“ *The Journal of Socio-Economics* 33 (5): 587–606. https://doi.org/10.1016/j.socec.2004.09.033.

Glassner, Andrew. 2021. *Deep Learning: A Visual Approach.* Erscheinungsort nicht ermittelbar: No Starch Press. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=6179842.

Goldman, N. 2003. „Evaluating the quality of self-reports of hypertension and diabetes.“ *Journal of clinical epidemiology* 56 (2): 148–54. https://doi.org/10.1016/S0895-4356(02)00580-2.

Goldstein, Harvey. 2010. *Multilevel statistical models.* 4. Aufl. Ltd, Chichester, UK: John Wiley & Sons.

Goodfellow, Ian, Yoshua Bengio und Aaron Courville. 2016. *Deep Learning:* MIT Press.

Gour, Rinu. 2019. „Wide and Deep learning With TensorFlow in 10 Min.“ https://medium.com/@rinu.gour123/wide-and-deep-learning-with-tensorflow-in-10-min-4eb897dbcaf6.

Gupta, Nitin, Shashank Mujumdar, Hima Patel, Satoshi Masuda, Naveen Panwar, Sambaran Bandyopadhyay, Sameep Mehta et al. 2021. „Data Quality for Machine Learning Tasks.“ In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, hrsg. von Feida Zhu, 4040–41. ACM Digital Library. New York,NY,United States: Association for Computing Machinery.

Harcourt, Houghton Mifflin. 2008. *The American Heritage Medical Dictionary.* 4. Aufl. Houghton Mifflin Company.

Hassan, Rondik J. und Adnan Mohsin Abdulazeez. 2021. „Deep Learning Convolutional Neural Network for Face Recognition: A Review.“ *International Journal of Science and Business* 5 (2): 114–27.

Irving Rootman, Deborah Gordon-El-Bihbety, Jim Frankish, Heather Hemming und Barbara Ronson. *National Literacy and National Literacy and Health Resear Health Research Pr ch Program ogram Needs assessment and Needs assessment and Envir Environmental scan onmental scan*. https://www.researchgate.net/profile/margot-kaszap/publication/240629559_national_literacy_and_national_literacy_and_health_resear_health_research_pr_ch_program_ogram_needs_assessment_and_needs_assessment_and_envir_environmental_scan_onmental_scan.

Johnsson, T. 1992. „A Procedure for Stepwise Regression Analysis.“ *Statistical Papers* 33 (1): 21–29. https://doi.org/10.1007/BF02925308.

Karpatne, Anuj, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova und Vipin Kumar. 2017. „Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data.“ *IEEE Trans. Knowl. Data Eng.* 29 (10): 2318–31. https://doi.org/10.1109/TKDE.2017.2720168.

Keleher, Helen und Virginia Hagger. 2007. „Health Literacy in Primary Health Care.“ *Aust. J. Prim. Health* 13 (2): 24. https://doi.org/10.1071/py07020.

Kelle, Udo. 2014. „Mixed Methods.“ In *Handbuch Methoden der empirischen Sozialforschung*, hrsg. von Nina Baur und Jörg Blasius, 153–66. Wiesbaden: Springer Fachmedien Wiesbaden.

Klaus, Georg, Hrsg. 1976. *Philosophisches Wörterbuch.* 12., neubearb. und durchges. Aufl. Westberlin: Das Europ. Buch.

Koleva, Nancy. 2020. „When and When Not to Use Deep Learning.“ Zugriff am 9. Dezember 2021. https://blog.dataiku.com/when-and-when-not-to-use-deep-learning.

Kostev, Karel, Tong Wu, Yue Wang, Kal Chaudhuri, Russel Reeve und Christian Tanislav. 2021. „Predicting the Risk of Ischemic Stroke in Patients Treated with Novel Oral Anticoagulants: A Machine Learning Approach.“ *Neuroepidemiology* 55 (5): 387–92. https://doi.org/10.1159/000517512.

Kriegsman, Didi M.W., Brenda W.J.H. Penninx, Jacques Th.M. van Eijk, A.Joan P. Boeke und Dorly J.H. Deeg. 1996. „Self-reports and general practitioner information on the presence of chronic diseases in community dwelling elderly.“ *Journal of clinical epidemiology* 49 (12): 1407–17. https://doi.org/10.1016/S0895-4356(96)00274-0.

Kuckartz, Udo und Stefan Rädiker. 2022. *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung: Grundlagentexte Methoden.* 5. Auflage. Weinheim, Basel: Beltz Juventa. http://www.content-select.com/index.php?id=bib_view&ean=9783779955337.

Kuhlmey, A. 2011. „Versorgungsforschung zur angemessenen Gesundheitsversorgung im Alter.“ [Health care research to improve the quality of health care provision for older people]. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz* 54 (8): 915–21. https://doi.org/10.1007/s00103-011-1314-1.

Kwak, Chanyeong und Alan Clayton-Matthews. 2002. „Multinomial Logistic Regression.“ *Nursing Research* 51 (6): 404–10. https://journals.lww.com/nursingresearchonline/fulltext/2002/11000/multinomial_logistic_regression.9.aspx.

Lehmann, E. L. 2011. *Fisher, Neyman, and the Creation of Classical Statistics.* New York, NY: Springer Science+Business Media LLC. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=770143.

Lemley, Kevin V. 2019. „Machine Learning Comes to Nephrology.“ *JASN* 30 (10): 1780–81. https://doi.org/10.1681/ASN.2019070664.

López Seguí, Francesc, Ricardo Ander Egg Aguilar, Gabriel de Maeztu, Anna García-Altés, Francesc García Cuyàs, Sandra Walsh, Marta Sagarra Castro und Josep Vidal-Alaball. 2020. „Teleconsultations Between Patients and Healthcare Professionals in Primary Care in Catalonia: The Evaluation of Text Classification Algorithms Using Supervised Machine Learning.“ *International journal of environmental research and public health* 17 (3). https://doi.org/10.3390/ijerph17031093.

Loyola-Gonzalez, Octavio. 2019. „Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View.“ *IEEE Access* 7:154096–113. https://doi.org/10.1109/access.2019.2949286.

Lundberg, Scott M., Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal und Su-In Lee. 2020. „From Local Explanations to Global Understanding with Explainable AI for Trees.“ *Nat Mach Intell* 2 (1): 56–67. https://doi.org/10.1038/s42256-019-0138-9.

Maass, Wolfgang, Jeffrey Parsons, Sandeep Purao, Veda C. Storey und Carson Woo. 2018. „Data-Driven Meets Theory-Driven Research in the Era of Big Data: Opportunities and Challenges for Information Systems Research.“ *JAIS*, 1253–73. https://doi.org/10.17705/1jais.00526.

Mackey, Tim K., Vidya Purushothaman, Michael Haupt, Matthew C. Nali und Jiawei Li. 2021. „Application of unsupervised machine learning to identify and characterise hydroxychloroquine misinformation on Twitter.“ *The Lancet Digital Health* 3 (2): e72-e75. https://doi.org/10.1016/S2589-7500(20)30318-6.

Malik, Adnan S., Grigorios Giamouzis, Vasiliki V. Georgiopoulou, Lucy V. Fike, Andreas P. Kalogeropoulos, Catherine R. Norton, Dan Sorescu et al. 2011. „Patient Perception Versus Medical Record Entry of Health-Related Conditions Among Patients with Heart Failure.“ *The American journal of cardiology* 107 (4): 569–72. https://doi.org/10.1016/j.amjcard.2010.10.017.

Manhart, Klaus. 1996. „Artificial Intelligence Modelling: Data Driven and Theory Driven Approaches.“ In *Social Science Microsimulation*, hrsg. von Klaus G. Troitzsch, Ulrich Mueller, Nigel Gilbert und Jim E. Doran. 1st ed. 1996, 416–31. Springer eBook Collection. Berlin, Heidelberg: Springer Berlin Heidelberg; Imprint Springer.

Masis, Serg. 2021. *Interpretable Machine Learning with Python.* 1st edition. Erscheinungsort nicht ermittelbar, Boston, MA: Packt Publishing; Safari. https://learning.oreilly.com/library/view/-/9781800203907/?ar.

Merkin, Sharon Stein, Kerri Cavanaugh, J. Craig Longenecker, Nancy E. Fink, Andrew S. Levey und Neil R. Powe. 2007. „Agreement of Self-Reported Comorbid Conditions with Medical and Physician Reports Varied by Disease Among End-Stage Renal Disease Patients.“ *Journal of clinical epidemiology* 60 (6): 634–42. https://doi.org/10.1016/j.jclinepi.2006.09.003.

Mertens, Cornelia, Andrea Siebenhofer, Andrea Berghold, Gudrun Pregartner, Lisa-Rebekka Ulrich, Karola Mergenthal, Birgit Kemperdick et al. 2019. „Differences in the Quality of Oral Anticoagulation Therapy with Vitamin K Antagonists in German GP Practices - Results of the Cluster-Randomized PICANT Trial (Primary Care Management for Optimized Antithrombotic Treatment).“ *BMC health services research* 19 (1): 539. https://doi.org/10.1186/s12913-019-4372-y.

Microsoft Research. 2021. „Interpret: Should I be parameter tuning EBMs (and if so, what parameters should I tune)?“. Zugriff am 30. April 2022. https://interpret.ml/docs/faq.html.

Mitchell, Tom M. 1997. *Machine Learning.* McGraw-Hill series in computer science. New York, NY: McGraw-Hill.

Montáns, Francisco J., Francisco Chinesta, Rafael Gómez-Bombarelli und J. Nathan Kutz. 2019. „Data-driven modeling and learning in science and engineering.“ *Comptes Rendus Mécanique* 347 (11): 845–55. https://doi.org/10.1016/j.crme.2019.11.009.

Murdoch, W. James, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl und Bin Yu. 2019. „Definitions, Methods, and Applications in Interpretable Machine Learning.“ *Proceedings of the National Academy of Sciences of the United States of America* 116 (44): 22071–80. https://doi.org/10.1073/pnas.1900654116.

National Center for Biotechnology Information. 1988. „PubMed.“ Zugriff am 7. Oktober 2021. https://www.ncbi.nlm.nih.gov/.

Nguyen, Chi Nhan und Oliver Zeigermann. 2021. *Machine Learning – kurz & gut: Eine Einführung mit Python, Pandas und Scikit-Learn:* O'Reilly.

Nguyen, Hoang, Le-Minh Kieu, Tao Wen und Chen Cai. 2018. „Deep Learning Methods in Transportation Domain: A Review.“ *IET Intelligent Transport Systems* 12 (9): 998–1004. https://doi.org/10.1049/iet-its.2018.0064.

Nikolaou, Vasilis, Sebastiano Massaro, Wolfgang Garn, Masoud Fakhimi, Lampros Stergioulas und David B. Price. 2021. „Fast Decliner Phenotype of Chronic Obstructive Pulmonary Disease (COPD): Applying Machine Learning for Predicting Lung Function Loss.“ *BMJ open respiratory research* 8 (1). https://doi.org/10.1136/bmjresp-2021-000980.

Nori, Harsha, Samuel Jenkins, Paul Koch und Rich Caruana. 2019. „InterpretML: A Unified Framework for Machine Learning Interpretability.“ Unveröffentlichtes Manuskript. https://arxiv.org/pdf/1909.09223.

Okura, Yuji, Lynn H. Urban, Douglas W. Mahoney, Steven J. Jacobsen und Richard J. Rodeheffer. 2004. „Agreement Between Self-Report Questionnaires and Medical Record Data Was Substantial for Diabetes, Hypertension, Myocardial Infarction and Stroke but Not for Heart Failure.“ *Journal of clinical epidemiology* 57 (10): 1096–1103. https://doi.org/10.1016/j.jclinepi.2004.04.005.

Olson, Randal S., William La Cava, Patryk Orzechowski, Ryan J. Urbanowicz und Jason H. Moore. 2017. „PMLB: A Large Benchmark Suite for Machine Learning Evaluation and Comparison.“ *BioData mining* 10 (36). https://doi.org/10.1186/s13040-017-0154-4.

Ossom Williamson, Peace und Christian I. J. Minter. 2019. „Exploring PubMed as a Reliable Resource for Scholarly Communications Services.“ *Journal of the Medical Library Association : JMLA* 107 (1): 16–29. https://doi.org/10.5195/jmla.2019.433.

Patterson, Josh und Adam Gibson. 2017. *Deep Learning: A Practitioner's Approach.* First edition. Beijing, Boston, Farnham, Sebastopol, Tokyo: O'Reilly. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1564784.

Peckham, Hannah, Nina M. de Gruijter, Charles Raine, Anna Radziszewska, Coziana Ciurtin, Lucy R. Wedderburn, Elizabeth C. Rosser, Kate Webb und Claire T. Deakin. 2020. „Male Sex Identified by Global COVID-19 Meta-Analysis as a Risk Factor for Death and ITU Admission.“ *Nat Commun* 11 (1): 6317. https://doi.org/10.1038/s41467-020-19741-6.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, R. Weiss et al. 2011. „Scikit-learn: Machine learning in Python.“ *the Journal of machine Learning research* 12: 2825–30. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https://githubhelp.com.

Pernet, Cyril. 2015. „Null hypothesis significance testing: a guide to commonly misunderstood concepts and recommendations for good practice.“ *F1000Res* 4:621. https://doi.org/10.12688/f1000research.6963.5.

Plaue, Matthias, Hrsg. 2021. *Data Science.* Berlin, Heidelberg: Springer Berlin Heidelberg.

Porta, Miquel S., Hrsg. 2014. *A Dictionary of Epidemiology.* Sixth edition. Oxford quick reference. Oxford: Oxford University Press. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=765961.

Raudenbush, Stephen W. und Anthony S. Bryk. 2010. *Hierarchical Linear Models: Applications and Data Analysis Methods.* 2. ed., [Nachdr.]. Advanced quantitative techniques in the social sciences 1. Thousand Oaks, Calif. Sage Publ.

Ray, Susmita. 2019. „A Quick Review of Machine Learning Algorithms.“ In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*: IEEE.

Rinne, Horst. 2008. *Taschenbuch der Statistik.* 4., vollst. überarb. und erw. Aufl. Frankfurt am Main: Deutsch.

Röhrig, Bernd, Jean-Baptist Du Prel, Daniel Wachtlin und Maria Blettner. 2009. „Types of Study in Medical Research: Part 3 of a Series on Evaluation of Scientific Publications.“ *Deutsches Arzteblatt international* 106 (15): 262–68. https://doi.org/10.3238/arztebl.2009.0262.

Royal College of GEneral Practitioners. 2022. „British Journal of General Practice.“ Zugriff am 10. März 2022. https://bjgp.org/.

Rudin, Cynthia. 2019. „Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.“ *Nat Mach Intell* 1 (5): 206–15. https://doi.org/10.1038/s42256-019-0048-x.

Salem, Hesham, Daniele Soria, Jonathan N. Lund und Amir Awwad. 2021. „A Systematic Review of the Applications of Expert Systems (ES) and Machine Learning (ML) in Clinical Urology.“ *BMC Med Inform Decis Mak* 21 (1): 223. https://doi.org/10.1186/s12911-021-01585-9.

Samuel, A. L. 1959. „Some Studies in Machine Learning Using the Game of Checkers.“ *IBM J. Res. & Dev.* 3 (3): 210–29. https://doi.org/10.1147/rd.33.0210.

SAS Institute Inc. 2017a. „Introduction to SEMMA.“ Zugriff am 29. April 2022. https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm.

SAS Institute Inc. 2017b. „SAS® Enterprise Miner™ 14.3: Reference Help.“ Unveröffentlichtes Manuskript, zuletzt geprüft am 29. April 2022. https://documentation.sas.com/api/docsets/emref/14.3/content/emref.pdf.

Schäfer, Thomas. 2010. *Statistik I: Deskriptive und Explorative Datenanalyse.* Springer eBook Collection Humanities, Social Science. Wiesbaden: VS Verlag für Sozialwissenschaften.

Schiff, Gordon D., Lynn A. Volk, Mayya Volodarskaya, Deborah H. Williams, Lake Walsh, Sara G. Myers, David W. Bates und Ronen Rozenblum. 2017. „Screening for Medication Errors Using an Outlier Detection System.“ *Journal of the American Medical Informatics Association : JAMIA* 24 (2): 281–87. https://doi.org/10.1093/jamia/ocw171.

Schrappe, M., G. Glaeske, M. Gottwik, R. Kilian, K. Papadimitriou, C. Scheidt-Nave, K. D. Schulz, Ziegenhagen, D. und H. Pfaff. 2005. „Memorandum zur Versorgungsforschung in Deutschland: Konzeptionelle, methodische und strukturelle Voraussetzungen der Versorgungsforschung.“ Unveröffentlichtes Manuskript, zuletzt geprüft am 4. Oktober 2021. https://dnvf.de/files/theme_files/pdf/PDF-Publikationen/2.%20Memorandum%202005.pdf.

Schröer, Christoph, Felix Kruse und Jorge Marx Gómez. 2021. „A Systematic Literature Review on Applying CRISP-DM Process Model.“ *Procedia Computer Science* 181:526–34. https://doi.org/10.1016/j.procs.2021.01.199.

Sokolova, Marina und Guy Lapalme. 2009. „A systematic analysis of performance measures for classification tasks.“ *Information Processing & Management* 45 (4): 427–37. https://doi.org/10.1016/j.ipm.2009.03.002.

Srinivas, Sharan. 2020. „A Machine Learning-Based Approach for Predicting Patient Punctuality in Ambulatory Care Centers.“ *International journal of environmental research and public health* 17 (10). https://doi.org/10.3390/ijerph17103703.

Stang, Andreas und Bernd Kowall. 2020. „Fisher’s Significance Test: A Gentle Introduction.“ GMS Medizinische Informatik, Biometrie und Epidemiologie; 16(1):Doc03. https://doi.org/10.3205/MIBE000206.

Topol, Eric J. und Abraham Verghese. 2019. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. First edition. New York, NY: Basic Books.

Uexküll, Thure von und Wolfgang Wesiack. 1988. *Theorie der Humanmedizin: Grundlagen ärztlichen Denkens und Handelns.* München, Wien, Baltimore: Urban u. Schwarzenberg.

Valliani, Aly Al-Amyn, Daniel Ranti und Eric Karl Oermann. 2019. „Deep Learning and Neurology: A Systematic Review.“ *Neurol Ther* 8 (2): 351–65. https://doi.org/10.1007/s40120-019-00153-8.

van den Akker, Marjan. 2022. Interviewt durch M. Paulitsch. 28. April 2022. Frankfurt am Main.

van den Akker, Marjan, Mark G. Spigt, Lore de Raeve, Ben van Steenkiste, Job F. M. Metsemakers, Ernst J. van Voorst und Hein de Vries. 2008. „The SMILE Study: A Study of Medical Information and Lifestyles in Eindhoven, the Rationale and Contents of a Large Prospective Dynamic Cohort Study.“ *BMC public health* 8:19. https://doi.org/10.1186/1471-2458-8-19.

van den Akker, Marjan, Ben van Steenkiste, Eefke Krutwagen und Job F. M. Metsemakers. 2015. „Disease or No Disease? Disagreement on Diagnoses Between Self-Reports and Medical Records of Adult Patients.“ *The European journal of general practice* 21 (1): 45–51. https://doi.org/10.3109/13814788.2014.907266.

van Engelen, Jesper E. und Holger H. Hoos. 2020. „A survey on semi-supervised learning.“ *Mach Learn* 109 (2): 373–440. https://doi.org/10.1007/s10994-019-05855-6.

Vehtari, Aki, Andrew Gelman, Daniel Simpson, Bob Carpenter und Paul-Christian Bürkner. 2021. „Rank-Normalization, Folding, and Localization: An Improved Rˆ for Assessing Convergence of MCMC (with Discussion).“ *Bayesian Anal.* 16 (2). https://doi.org/10.1214/20-BA1221.

Walsh, Colin G., Jessica D. Ribeiro und Joseph C. Franklin. 2017. „Predicting Risk of Suicide Attempts over Time Through Machine Learning.“ *Clinical Psychological Science* 5 (3): 457–69. https://doi.org/10.1177/2167702617691560.

Wassermann, Sandra. 2015. „Das qualitative Experteninterview.“ In *Methoden der Experten- und Stakeholdereinbindung in der sozialwissenschaftlichen Forschung*, hrsg. von Marlen Niederberger und Sandra Wassermann, 51–67. Wiesbaden: Springer Fachmedien Wiesbaden.

WONCA Europe. 2022. „European Journal of General Practice.“ Zugriff am 10. März 2022.

Wong, Edwin S., Linnaea Schuttner und Ashok Reddy. 2020. „Does Machine Learning Improve Prediction of VA Primary Care Reliance?“. *The American journal of managed care* 26 (1): 40–44. https://doi.org/10.37765/ajmc.2020.42144.

Zhang, Aijun. 2022. „Dr. Aijun Zhang.“ Zugriff am 19. April 2022. http://statsoft.org/.

Anhänge

Anhang A: Interviewleitfaden

Interviewleitfaden Expert:innen in der Allgemeinmedizin

Forschungsfrage: Wie verbreitet und fundiert ist theoretisches Grundlagenwissen hinsichtlich Maschinellen Lernens (ML) im wissenschaftlichen Zweig der Allgemeinmedizin?

Einleitungs- bzw. Vorstellungsphase:

- Vorstellung der Interviewer:in/institutioneller Kontext: Im Rahmen einer Masterarbeit eines Studiums (Master of Data Science & Business Analytics) an der Hochschule der Medien in Stuttgart.
- Erläuterung des Themas der eigenen Untersuchung: Das Potenzial Maschinellen Lernens in der allgemeinmedizinischen Versorgungsforschung.
- Zeitlicher Interviewrahmen: Mindestens 15, höchstens 25 Minuten.
- Erläuterung des Interviewablaufs bzw. „erwünschter Antwortformen: Sechs offene Fragen bzgl. Ihrer Einschätzung und Erfahrung. Keine Testfragen (kein richtig oder falsch).
- Erlaubnis zur Tonbandaufzeichnung, ggfs. Anonymitätszusicherung.

Interview:

Leitfrage	Themenkomplex	Interviewfragen	Unterfrage
Inwieweit besitzen Expert:innen im Fach der Allgemeinmedizin theoretisches ML-Grundlagenwissen?	Theoretisches Grundlagenwissen bzgl. Maschinellem Lernen	• Was verstehen Sie unter Maschinellem Lernen? • Kennen Sie eine Anwendung basierend auf Maschinellem Lernen in Ihrem Alltag? • Kennen Sie eine Anwendung im allgemeinmedizinischen Kontext, die nicht zwingend in Ihrer Forschungsdomäne angesiedelt sein muss?	• Wenn ja, welche wären das?
Inwieweit haben Expert:innen der Allgemeinmedizin praktische Erfahrungen mit ML gemacht (und wenn ja, welche)?	Praktische Erfahrungen mit Maschinellem Lernen	• Haben Sie bereits praktische Erfahrungen mit Maschinellem Lernen in ihrem beruflichen Kontext gemacht?	• Haben Sie in der Vergangenheit für eine wissenschaftliche Arbeit eine ML-Anwendung erstellt und/oder eine Analyse durchgeführt? ○ Wenn ja, können Sie dies näher erläutern? • Sind Sie bei einer wissenschaftlichen Arbeit Autor/Koautor

			gewesen, die solch eine Anwendung/Analyse verwandt hat? o Wenn ja, können Sie dies näher erläutern? • Haben Sie spezifische Literatur in der eigenen Arbeit zitiert, die sich mit Maschinellem Lernen beschäftigt? o Wenn ja, können Sie dies näher erläutern? • Haben Sie Ideen um Einsatz von ML/KI im Kontext der eigenen Arbeit?
	Weitere Gedanken des Befragten	• Möchten Sie noch etwas hinzufügen?	

Ende des Interviews

Anhang B: Kerninhalte der Antworten in den Interviews

Interview-Partner:in	Theoretisches Grundlagenwissen bzgl. Maschinellem Lernen			Praktische Erfahrung mit Maschinellem Lernen	
	Was ist Maschinelles Lernen?	*Anwendungen aus dem Alltag*	*Anwendung im allgemeinemedizinischen Kontext*	*Im beruflichen Kontext in Vergangenheit (Durchführung, Koautor, spezifische Literatur zitiert)*	*Ideen im Kotext der eigenen Arbeit*
1.	Maschinelles Lernen ist eine Art von künstliche Intelligenz… dass man eine Mascine mit mehr Informationen füttert…Vielleicht sollten Maschinen durch die zunehmende Zahl von weiteren Informationen und Daten besser werden?	Nicht wissentlich	Nein.	Nein.	Ich bin offen für Neues
2.	Simulation des menschlichen Lernprozesses also auf automatisierter Ebene, die also der Einsatz von Algorithmen wie Computer passiert, die den menschlichen Lernprozess sozusagen simulieren oder analog nachvollziehen. Indem sie eben unter anderem auf Mustererkennung rekurrieren	Nein.	…wir hatten vor über 20 Jahren mit einer bestimmten Software, die bestimmte Hirnstrukturen als Volumen bestimmter Hirnstrukturen ermittelt…	…würde mir spontan nichts einfallen…	Wir sind ja immer sehr nah am am Alltag, also gerade mit diesen psychosoziale Interventionen beispielsweise, aber vielleicht auch ein bisschen Gesundheitssystemforschung geht um Versorgungsforschung in dem soeben im klassischen Sinne also große Mengen von Routinedaten auf eine bestimmte Lage hin zu untersuchen oder zu explorieren. Und das könnte ich mir vorstellen, dass es sinnvoll eingesetzt werden kann.
3.	Da fällt mir spontan ein, wie Programmierung von Robotern oder wie also, dass die programmiert werden zu bestimmten Dingen, die sie ausführen und vielleicht sogar sozusagen die Kompetenz kriegen, was dazuzulernen.	Also mir würde jetzt spontan zum Beispiel so etwas wie die ADA-App einfallen, die irgendwo, also, wo verschiedene Patienten sozusagen Daten eingeben. Und die App selber hat bestimmte Algorithmus Netz AG, die dann erkennen, welche Krankheit ist. Und dadurch, dass ganz viele Leute Sachen einspeisen wird quasi dieses Programm Apple immer komplexer und findet auch immer mehr Krankheitsbilder.	Also nö, tatsächlich, wenn dann nur sozusagen durch solche Privatanwender genommen,	Da fällt mir jetzt nichts ein.	Automatisierte Terminfindung und oder so Verwaltung. Wenn ich jetzt eigentlich so viel mehr von der Weiterbildung, dann könnte es ja sein. Nachdem wenn bestimmte Anwender sich den Podcast zum Beispiel anhören oder bestimmte Folge vom bestimmte Apps Beiträge auf Instagram, das dann automatisch bestimmte andere Seiten verlinkt werdenDas gegückt haben. kriegen Sie jetzt einen Link.
4.	Im Rahmen von Algorithmen und künstlicher Intelligenz	Zum Energiesparen	Nein	Nicht im wissenschaftlichen Bereich	zum Beispiel bei seltene Diagnosen. Bei diffusen Symptomen schwere Verläufe vorherzusagen und zur Unterstützung jüngerer Kollegen in der Ausbildung. Zur Unterstützung korrekte Diagnosen vorherzusagen.
5.	Ich würde sagen, Computeranwendung die sich automatisch verbessern.	Nein, ich kenne keine direkte Anwendung.	Im eHealth-Studium in welchen Bereichen es verwendet werden kann. Risiken und Chancen.	Nein.	Eher praktische als medizinische. Grenze in der Allgemeinmedizin. Zwei Sachen: Vorhersage der Krankmeldungsdauer, zur Anwendung zur Entlastung des Patienten in administrativer Hinsicht entlasten, dass er nicht alle 4 Wochen vorstellig werden muss. Beispielsweise bei Krebspatienten Vorhersagen, ob Krebspatienten zuerst zum Hausarzt oder zuerst zu anderen Spezialisten.
6.	Bei Maschinellem Lernen denke ich vor allem an Big Data, komplizierte Fragen, automatiesierte Auskünfte.	Internetsuche. wahrscheinlich auch Routenplaner.	Nein	Eine Publikation bei Data Mining.	Es müssten komplizierte Fragen sein, und ich könnte mir vorstellen für die Überinstimmung von Patienten welche Erkrankungen sie haben und welche auch in elektronical medical Records aufgezeichnet sind, welche es gibt. Wir wissen, dass das häufig nicht übereinstimmt.Die Gründe dafür sind baer noch unbekannt. Und ob es da vielleicht auch so Muster gibt, dwelche Erkrankungen von Patienten eher erwähnt werden.

Anhang C: Korrelationsmatrix unabhängiger Variablen

	Geschlecht	*Alter*	*Angst-erkrankung*	*Depression*	*physische Lebens-qualität*	*mentale Lebens-qualität*	*Krankheits-zahl*	*Lungen*	*Herz*	*Darm*	*Leber*
Geschlecht	1,000	-0,046	0,045	-0,016	-0,056	-0,042	0,057	-0,022	-0,061	0,010	0,011
Alter	-0,046	1,000	0,003	0,104	-0,287	-0,038	0,281	0,001	0,165	0,106	-0,019
Angsterkrankung	0,045	0,003	1,000	0,517	-0,213	-0,515	0,088	0,028	0,025	0,081	0,035
Depression	-0,016	0,104	0,517	1,000	-0,308	-0,498	0,137	0,054	0,093	0,064	0,041
physische Lebensqualität	-0,056	-0,287	-0,213	-0,308	1,000	0,100	-0,386	-0,099	-0,165	-0,126	-0,043
mentale Lebensqualität	-0,042	-0,038	-0,515	-0,498	0,100	1,000	-0,076	-0,035	-0,063	-0,092	-0,041
Krankheitszahl	0,057	0,281	0,088	0,137	-0,386	-0,076	1,000	0,046	0,106	0,242	0,026
Lungen	-0,022	0,001	0,028	0,054	-0,099	-0,035	0,046	1,000	0,034	0,031	0,028
Herz	-0,061	0,165	0,025	0,093	-0,165	-0,063	0,106	0,034	1,000	0,046	0,028
Darm	0,010	0,106	0,081	0,064	-0,126	-0,092	0,242	0,031	0,046	1,000	0,012
Leber	0,011	-0,019	0,035	0,041	-0,043	-0,041	0,026	0,028	0,028	0,012	1,000
Nieren	-0,027	0,050	0,022	0,039	-0,090	-0,033	0,080	0,018	0,062	0,038	0,075
Diabetes	-0,029	0,058	0,050	0,053	-0,051	-0,016	0,118	0,011	-0,005	0,004	0,041
Onkologisch	0,000	0,174	-0,015	0,030	-0,095	-0,020	0,254	0,030	0,055	0,027	0,034
Epilepsie	0,002	-0,005	0,003	0,036	-0,031	-0,050	0,032	0,033	0,064	-0,009	0,048
Migräne	0,150	-0,024	0,041	0,040	-0,037	-0,047	0,039	-0,013	0,000	0,041	0,033
Schlaganfall	-0,037	0,079	0,046	0,082	-0,095	-0,072	0,124	0,030	0,046	0,056	0,014
Gelenke	0,095	0,072	0,067	0,051	-0,156	-0,045	0,272	0,051	0,009	0,066	0,057
Rheuma	0,088	0,076	0,119	0,130	-0,258	-0,064	0,216	0,021	0,054	0,054	0,027
Arthrose	0,067	0,054	0,066	0,060	-0,173	-0,049	0,071	0,059	0,022	0,057	0,015
Rücken	0,043	0,013	0,080	0,084	-0,205	-0,027	0,065	0,061	0,024	0,044	0,055

	Nieren	*Diabetes*	*Onkologisch*	*Epilepsie*	*Migräne*	*Schlaganfall*	*Gelenke*	*Rheuma*	*Arthrose*	*Rücken*
Geschlecht	-0,027	-0,029	0,000	0,002	0,150	-0,037	0,095	0,088	0,067	0,043
Alter	0,050	0,058	0,174	-0,005	-0,024	0,079	0,072	0,076	0,054	0,013
Angsterkrankung	0,022	0,050	-0,015	0,003	0,041	0,046	0,067	0,119	0,066	0,080
Depression	0,039	0,053	0,030	0,036	0,040	0,082	0,051	0,130	0,060	0,084
physische Lebensqualität	-0,090	-0,051	-0,095	-0,031	-0,037	-0,095	-0,156	-0,258	-0,173	-0,205
mentale Lebensqualität	-0,033	-0,016	-0,020	-0,050	-0,047	-0,072	-0,045	-0,064	-0,049	-0,027
Krankheitszahl	0,080	0,118	0,254	0,032	0,039	0,124	0,272	0,216	0,071	0,065
Lungen	0,018	0,011	0,030	0,033	-0,013	0,030	0,051	0,021	0,059	0,061
Herz	0,062	-0,005	0,055	0,064	0,000	0,046	0,009	0,054	0,022	0,024
Darm	0,038	0,004	0,027	-0,009	0,041	0,056	0,066	0,054	0,057	0,044
Leber	0,075	0,041	0,034	0,048	0,033	0,014	0,057	0,027	0,015	0,055
Nieren	1,000	0,110	0,041	0,060	-0,011	0,019	0,014	0,060	0,001	0,019
Diabetes	0,110	1,000	0,056	-0,001	-0,004	0,010	0,025	0,027	0,009	0,018
Onkologisch	0,041	0,056	1,000	0,067	0,000	0,036	0,042	0,020	0,024	-0,019
Epilepsie	0,060	-0,001	0,067	1,000	0,004	0,032	0,000	0,004	0,005	-0,017
Migräne	-0,011	-0,004	0,000	0,004	1,000	0,018	0,046	0,046	0,036	0,096
Schlaganfall	0,019	0,010	0,036	0,032	0,018	1,000	0,018	0,078	0,055	0,011
Gelenke	0,014	0,025	0,042	0,000	0,046	0,018	1,000	0,153	0,247	0,084
Rheuma	0,060	0,027	0,020	0,004	0,046	0,078	0,153	1,000	0,118	0,089
Arthrose	0,001	0,009	0,024	0,005	0,036	0,055	0,247	0,118	1,000	0,087
Rücken	0,019	0,018	-0,019	-0,017	0,096	0,011	0,084	0,089	0,087	1,000

Anhang D: Syntaxen der verwendeten inferenzstatistischen Modelle

Intercept-only Modell 1:

```
modell_1 <- brm(over_under_agree ~ 1 + (1|Praxis), data = inf_mod_long, family = categorical(link = "logit", refcat = "0"), warmup = 600, iter = 1250, chains = 2, seed=123, cores = 2)
```

Intercept-only Modell 2:

```
modell_2 <- brm(over_under_agree ~ 1 + (1|Praxis/Teilnehmer), data = inf_mod_long, family = categorical(link = "logit", refcat = "0"), warmup = 800, iter = 2000, chains = 2, seed=123, cores = 2)
```

Random-Intercept-Modell mit Level 2-Prädiktoren:

```
modell_4 <- brm(over_under_agree ~ 1 + (1|Praxis/Teilnehmer) + Krankheitsgr + Geschlecht + Alter + Lebensqual_phys + Lebensqual_ment + Krankheitszahl, data = inf_mod_long, family = categorical(link = "logit", refcat = "Agree"), warmup = 1000, iter = 4000, chains = 2, seed=123, cores = 3)
```

Anhang E: Ergebnisse des Random-Intercept-Modells mir Level 2-Prädiktoren

Underreporting:

	Regressions-gewicht	Standard-fehler	unteres Glabwürdig-keitsintervall (95%)	oberes Glaubwürdig-keitsintervall (95%)	$\hat{R}$	Bulk-ESS	Tail-ESS
Intercept	-9,56	0,48	-10,54	-8,65	1	1087	1868
Arthrose	3,81	0,35	3,17	4,53	1,01	602	920
Darm	4,06	0,35	3,43	4,78	1,01	602	909
Diabetes	2,16	0,36	1,46	2,9	1	650	1036
Gelenke	4,7	0,34	4,07	5,41	1,01	598	913
Herz	2,15	0,36	1,47	2,91	1,01	664	1000
Leber	0,81	0,41	0,02	1,65	1	814	1517
Lungen	2,11	0,36	1,45	2,85	1	651	1147
Migräne	2,08	0,36	1,41	2,83	1	665	1034
Nieren	-0,24	0,53	-1,3	0,79	1	1217	1850
Onkologisch	3,74	0,35	3,11	4,46	1,01	609	1023
Rheuma	1,95	0,37	1,26	2,71	1,01	644	1036
Rücken	2,61	0,36	1,95	3,34	1,01	641	951
Schlaganfall	1,81	0,37	1,13	2,57	1	688	987
Geschlecht	0,17	0,06	0,07	0,28	1	7573	4241
Alter	0,01	0	0,01	0,02	1	11167	4490
physische Lebensqualität	0,02	0	0,02	0,03	1	8690	4364
mentale Lebensqualität	0	0	0	0,01	1	13894	4628
Krankheitszahl	0,76	0,02	0,72	0,8	1	5476	4260

Overreporting

	Regressions-gewicht	Standard-fehler	unteres Glabwürdig-keitsintervall (95%)	oberes Glaubwürdig-keitsintervall (95%)	$\hat{R}$	Bulk-ESS	Tail-ESS
Intercept	-2,37	0,37	-3,09	-1,66	1	1335	3048
Arthrose	3,86	0,22	3,45	4,31	1	587	1319
Darm	2,19	0,23	1,75	2,66	1	625	1527
Diabetes	0,47	0,28	-0,07	1,03	1	906	1851
Gelenke	3,17	0,22	2,75	3,62	1	595	1181
Herz	2,31	0,22	1,89	2,77	1	612	1380
Leber	0,96	0,25	0,48	1,46	1	761	1829
Lungen	2,51	0,23	2,08	2,97	1	614	1264
Migräne	2,75	0,23	2,33	3,2	1	604	1324
Nieren	1,33	0,24	0,87	1,83	1	699	1539
Onkologisch	0,66	0,27	0,14	1,21	1	850	1706
Rheuma	2,39	0,23	1,95	2,85	1	626	1391
Rücken	3,7	0,22	3,28	4,14	1	600	1333
Schlaganfall	1,09	0,25	0,62	1,59	1	752	1636
Geschlecht	0,16	0,05	0,06	0,26	1	4776	3696
Alter	0	0	0	0,01	1	5388	4915
physische Lebensqualität	-0,05	0	-0,05	-0,04	1	4884	4193
mentale Lebensqualität	-0,02	0	-0,02	-0,01	1	5870	4344
Krankheitszahl	-0,11	0,02	-0,15	-0,07	1	4845	4476

Anhang F: Hyperparameter-Tuning des Gradient Boosting-Modells anhand einer Grid-Search mit *sklearn*

1. Tuning n_estimators & Learning rate

```
test_1 = {'learning_rate':[0.15,0.1,0.05,0.01,0.005,0.001],
'n_estimators':[100,250,500,750,1000,1250,1500,1750,2000,2250,2500]}

tuning = GridSearchCV(estimator = GradientBoostingClassifier(max_depth = 3, max_features = 'sqrt',
min_samples_split = 2, min_samples_leaf = 1, subsample = 1, random_state = 10),  param_grid =
test_1, scoring = 'accuracy', n_jobs = 4, cv = 5)

tuning.fit(X_train,y_train, sample_weight = sample_weights)

tuning.best_params_, tuning.best_score_
```

2. Tuning max_depth

```
test_2 = {'max_depth':[2,3,4,5,6,7,8,10,11,12,13,14]}

tuning = GridSearchCV(estimator = GradientBoostingClassifier(learning_rate = 0.15, n_estimators =
2500, max_features = 'sqrt', min_samples_split = 2, min_samples_leaf = 1, subsample = 1,
random_state = 10), param_grid = test_2, scoring = 'accuracy', n_jobs = 4, cv = 5)

tuning.fit(X_train,y_train, sample_weight = sample_weights)

tuning.best_params_, tuning.best_score_
```

3. Tuning max_features

```
test_3 = {'max_features':[2,3,4,5,6,7,8,9,10,11,12]}

tuning = GridSearchCV(estimator = GradientBoostingClassifier(learning_rate = 0.15, n_estimators =
2500, max_depth = 14, min_samples_split = 2, min_samples_leaf = 1, subsample = 1, random_state =
10), param_grid = test_3, scoring = 'accuracy', n_jobs = 4, cv = 5)

tuning.fit(X_train,y_train, sample_weight = sample_weights)

tuning.best_params_, tuning.best_score_
```

4. Tuning subsample

```
test_4 = {'subsample':[0.7,0.75,0.8,0.85,0.9,0.95,1]}

tuning = GridSearchCV(estimator = GradientBoostingClassifier(learning_rate = 0.15, n_estimators =
2500, max_depth = 14, max_features = 5, min_samples_split = 2, min_samples_leaf = 1,
random_state = 10), param_grid = test_4, scoring = 'accuracy', n_jobs = 4, cv = 5)

tuning.fit(X_train,y_train, sample_weight = sample_weights)

tuning.best_params_, tuning.best_score_
```

5. Tuning min_samples_split und min_samples_leaf

```
test_5 = {'min_samples_split':[2,4,6,8,10,20,40,60,100], 'min_samples_leaf':[1,3,5,7,9]}


tuning = GridSearchCV(estimator = GradientBoostingClassifier(learning_rate = 0.15, n_estimators = 2500, max_depth = 14, max_features = 5, subsample = 1, random_state = 10),  param_grid = test_5, scoring = 'accuracy', n_jobs = 4, cv = 5)

tuning.fit(X_train,y_train, sample_weight = sample_weights)

tuning.best_params_, tuning.best_score_
```

Anhang G: Syntax des finalen Gradient Boosting-Modells mit *sklearn*

```
final = GradientBoostingClassifier(learning_rate = 0.15, n_estimators = 2500, max_depth = 14,
max_features = 5, subsample = 1, min_samples_split = 2, min_samples_leaf = 1, random_state = 10)

final.fit(X_train, y_train, sample_weight = sample_weights)

pred = final.predict(X_test)

print(classification_report(y_test, pred))
```

Anhang H: Syntax des Explainable Boosting Machine-Modells mit *InterpretML*

```
ebm = ExplainableBoostingClassifier(learning_rate = 0.15, max_rounds = 500, min_samples_leaf = 5, random_state=10)

ebm.fit(X_train, y_train, sample_weight = sample_weights)

pred=ebm.predict(X_test)

print(classification_report(y_test, pred))
```