Methoden und Werkzeuge zur privatheitsbewahrenden Synthese medizinischer Forschungsdaten

Masterarbeit

im Studiengang Medizinisches Informationsmanagement

an der

Hochschule Hannover

vorgelegt am

27. Oktober 2021

von

Christin Schober

aus Münster

Erstprüfer: Prof. Dr.-Ing. Oliver J. Bott

Zweitprüfer: Matthias Katzensteiner, Master of Arts

Ehrenwörtliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Hannover, 27. Oktober 2021 Unterschrift

Zusammenfassung

In der medizinischen Forschung nimmt die Bedeutung, langfristigen Zugriff auf hochqualitative medizinische Daten zu erhalten, stetig zu. Aus wissenschaftlicher, ethischer und besonders auch aus rechtlicher Sicht darf die Privatheit betroffener Individuen dabei nicht verletzt werden.

In dieser Masterarbeit wurde ein synthetischer Datensatz erzeugt, der sowohl auf die Kriterien einer guten Datenqualität als auch das Offenlegungsrisiko geprüft wurde. Im Rahmen einer Literaturrecherche wurden zunächst Methoden zur Erzeugung synthetischer Daten, Evaluierungstechniken zur Prüfung der Datenqualität synthetischer Daten und Anonymisierungstechniken ermittelt und zusammengetragen. Mit einem Teildatensatz des *MIMIC-III*-Datensatzes wurde anschließend mit dem Tool *DataSynthesizer* ein neuer Datensatz synthetisiert.

Die beiden Datensätze wurden mittels *Kolmogorov-Smirnow-Test*, *Kullback-Leibler-Divergenz* und der *Paarweisen-Korrelations-Differenz* verglichen. Für die kategorischen Attribute konnte eine deutliche Übereinstimmung in der Werteverteilung nachgewiesen werden. Für die numerischen Attribute waren in den Verteilungen Unterschiede, welche mit Fehlwerten im ursprünglichen Datensatz assoziiert wurden.

Für die Prüfung der Privatheit der Daten wurde für unterschiedliche Szenarien für den ursprunggebenden Datensatz eine höhere *k-Anonymität* und für den synthetischen Datensatz eine höhere *ℓ-Diversity* ermittelt.

Zudem wurden in beiden Datensätzen übereinstimmende Objekte ermittelt. Für eine vorab aus dem realen Datensatz erstellte Kontrollgruppe wurde ein mehr als zwei Mal höheres *Relatives Risiko* und eine 2,9-fach höhere Chance (*Odds-Ratio)* ermittelt, ein identisches Objekt zu identifizieren, als für den synthetischen Datensatz.

Summary

In medical research, the importance of obtaining long-term access to high-quality medical data is constantly increasing. From a scientific, ethical, and especially from a legal point of view, the privacy of the individuals concerned must not be violated.

In this master thesis, a synthetic data set was generated, which was tested for both, the criteria of good data quality and the disclosure risk. A literature review was conducted to first identify and compile methods for generating synthetic data, evaluation techniques for testing the data quality of synthetic data, and anonymization techniques. A subset of the MIMIC III dataset was then used to synthesize a new dataset using the *DataSynthesizer* tool.

The two datasets were compared using *Kolmogorov-Smirnov test*, *Kullback-Leibler divergence*, and *pairwise correlation difference*. For the categorical attributes, significant agreement in the distribution of values was demonstrated. For the numeric attributes, differences in the distributions were found, which were associated with spurious values in the original data set.

For testing the privacy of the data, a higher *k-anonymity* was found for the original dataset and a higher *ℓ-diversity* for the synthetic dataset for different scenarios.

In addition, matching objects were determined in both datasets. In comparison to the synthesized dataset, a more than two times higher *relative risk* and a 2.9 times higher chance (*odds ratio*) of identifying an identical object were determined for a previously separated control group from the real dataset.

Inhaltsverzeichnis

Abkürzungsverzeichnis

1:1 ..... *Eins-zu-Eins-Übereinstimmungen*
25%Q ..... *unteres Quartil*
75%Q ..... *oberes Quartil*
Anf ..... *Anforderungen*
BN ..... Bayes'sches Netzwerk
CrCl ..... *Cross-Classification*
CSV ..... *Comma-separated values*
DEK ..... *Datenethikkommission*
DP ..... *Differential Privacy*
DS ..... *Datensatz*
DSGVO ..... *EU-Datenschutz-Grundverordnung*
FA ..... *Folgeaufenthalt auf der Intensivstation*
GAN ..... *Generative Adversarial Network*
HIPAA ..... *Health Insurance Portability and Accountability Act*
ICD ..... *International Statistical Classification of Diseases and Related Health Problems*
IDs ..... *Identifikationsnummern*
KDWH ..... *Klinisches Data Warehouse*
KL-Divergenz ..... *Kullback-Leibler-Divergenz*
Kon ..... *konfrontiere Objekte*
KS-Test ..... *Kolmogorow-Smirnow-Test*
lit ..... *littera*
Max ..... *Maximum*
medBGAN ..... *Bounary-Seeking-GAN*
medGAN ..... Medical GAN
medWGAN ..... *Wasserstein-GAN mit Gradientenstrafe*
MeSH ..... *Medical Subject Headings*
MIMIC-III ..... *Medical Information Mart for Intensive Care III*
Min ..... *Minimum*
ML ..... *Machine-Learning*
MS ..... *Microsoft*
MW ..... *arithmetischer Mittelwert*
N ..... *Anzahl der Häufigkeiten*
PKD ..... Paarweise-Korrelation Differenz
RS ..... *Real-Synthetisch*
s. ..... *siehe*
SQL ..... *Structured Query Language*
SR ..... *Synthetisch-Real*
Std ..... *Standardabweichung*
UI ..... *web User Interface*
WSL ..... *Windows-Subsystem für Linux*
ZDIN ..... *Zentrums für digitale Innovationen Niedersachen*
ZLG ..... *Zukunftslabor Gesundheit*

Tabellenverzeichnis

Abbildungsverzeichnis

1 Einleitung

Die medizinische Forschung wird durch die Verarbeitung und Analyse von sensiblen medizinischen Daten vorangetrieben. Eine Vielzahl der Forschungsvorhaben, die zum Zwecke des medizinischen Fortschritts erfolgen, hängen mit der Erhebung von Daten betroffener Personen zusammen. Häufig können retrospektiv Erkenntnisse aus Daten gewonnen werden, die ursprünglich für einen anderen Zweck erhoben wurden. Der Austausch von Informationen zwischen verschiedenen Disziplinen der medizinischen Forschung, wie z. B. verschiedenen medizinischen Fachrichtungen, Bereichen der Biologie und Chemie oder medizinisch-technischen Bereichen wie der Medizinischen Informatik, kann die Forschung zusätzlich antreiben. Für diese Vorhaben sind große Datenmengen von guter Qualität erforderlich. Deshalb ist es wichtig, medizinische Daten langfristig und institutionsübergreifend zur Verfügung stellen zu können (1–3).

In der Forschung werden neben den medizinischen Daten von behandelten bzw. untersuchten Personen weitere sensible und personenbezogene Daten für die Ermittlung von neuen Erkenntnissen verarbeitet. Bei personenbezogenen Daten handelt es sich um sämtliche Informationen, die eine Identifizierung einer Person zulassen (4), wie z. B. der Name, der Geburtstag, der Wohnort, das Geschlecht, bestimmte Diagnosen oder der sozial-ökonomische Status. Die Wiederherstellung des Personenbezugs über die vorhandenen Daten zur betroffenen Personen wird als Re-Identifizierung bezeichnet (5).

1.1 Datenschutz

Die Einhaltung der Datenschutzregelungen, wie sie in der aktuell gültigen EU-Datenschutz-Grundverordnung (DSGVO) (6) festgelegt sind, ist im Umgang mit sensiblen personenbezogenen Daten in der medizinischen Forschung unabdingbar. Im Folgenden werden einige Aspekte der DSGVO erklärt, um die Thematik dieser Arbeit einzuordnen.

1.1.1 **Einwilligung und Zweckbindung**

Personenbezogene Daten dürfen nach Artikel 9 Absatz 2 lit. DSGVO nicht ohne die ausdrückliche Einwilligung der betroffenen Person verarbeitet werden. Die Angabe des Verwendungszwecks für die Verarbeitung der Daten muss in der Einwilligung festgelegt werden. Die erhobenen Daten unterliegen einer Zweckbindung und dürfen nur für die Zwecke verwendet werden, welche in der Einwilligung mit der betroffenen Person vereinbart worden sind. Sollen die Daten für ein anderes Bestreben verarbeitet und analysiert werden, ist ein weiteres Einverständnis von der betroffenen Person einzuholen.

Für die langfristige Bereitstellung qualitativ hochwertiger Daten sind die Verwendungszwecke zum Zeitpunkt der Einwilligung jedoch nicht bekannt. Daher ist eine personengebundene Verarbeitung zur langfristigen Bereitstellung nach der DSGVO nicht möglich (3).

Für die medizinische Forschung ist in den Landesdatenschutz- bzw. Landeskrankenhausgesetzen geregelt, dass eine Einwilligung für die zukünftige Nutzung von personenbezogenen Gesundheitsdaten eingeholt werden kann. Über eine *Breite Einwilligung (Broad Consent)* kann das Einverständnis untersuchter Personen aus der Klinik und klinischen Studien eingeholt werden, die eine Weiterverarbeitung der Daten in pseudonymisierter Form erlaubt. Diese Einwilligung betrifft, neben dem aktuellen Verwendungszweck, auch die Verwendung für zukünftige, zum Zeitpunkt der Einwilligung nicht bekannte, Forschungsvorhaben. Die Medizininformatikinitiative hat für den *Broad Consent* einen bundesweit einheitlichen Text für eine breit formulierte Einwilligung verfasst (7).

1.1.2 **Datenschutzrechtliche Aufklärung**

In einer datenschutzrechtlichen Aufklärung muss erklärt werden, welche Daten zu welchem Zweck von welchen Verantwortlichen verarbeitet werden und die Dauer der Speicherung genannt werden. Außerdem müssen die betroffenen Personen über ihre geltenden Rechte, wie z. B. das Recht auf *Korrektheit der Daten* und das *Recht auf Vergessenwerden* im Zusammenhang mit der Einwilligung aufgeklärt werden (3).

1.1.3 **Das Recht auf Vergessenwerden**

Dieses Recht besteht für betroffene Personen nach Kapitel 3 Artikel 17 DSGVO als Recht auf Löschung personenbezogener Daten. Gründe für das Recht auf Vergessenwerden sind z. B., dass die erhobenen Daten für den vereinbarten Zweck nicht mehr benötigt werden, die betroffene Person ihre Einwilligung zurückzieht oder sie einen Widerspruch gegen die Verarbeitung der Daten einlegt.

Im Sinne des Urteils in der Rechtssache *C -131/12* des Europäischen Gerichtshofs vom 13. Mai 2014 meint der Begriff „Löschen" aber nicht zwangsläufig das tatsächliche Löschen der betroffenen Daten, sondern erlaubt auch die Durchführung einer irreversiblen Anonymisierung (3,8).

Medizinische Daten, deren in anderen Gesetzen geregelten Aufbewahrungspflicht abgelaufen ist, wie z. B. die in § 630f Absatz 3 BGB geregelte Aufbewahrungsfrist von 10 Jahren, müssen entsprechend der DSGVO gelöscht werden. Dies hat unter der Annahme zu geschehen, dass zu diesem Zeitpunkt keine der vereinbarten Verwendungszwecke für die Erhebung der Daten mehr vorhanden sind (3,9).

1.1.4 **Aufheben des Personenbezugs**

Wird beabsichtigt Daten z. B. nach Ablauf der Aufbewahrungsfrist für andere Zwecke zu verwenden, besteht die Möglichkeit, den Personenbezug der Daten nach Artikel 6 Absatz 4 DSGVO aufzuheben. Der grundsätzliche Vorgang zur Aufhebung des Personenbezugs wird auch als De-Identifizierung bezeichnet. Eine vorübergehende Aufhebung des Personenbezugs lässt sich durch eine Pseudonymisierung erreichen. Eine unumkehrbare Aufhebung ist durch eine Anonymisierung der personenbezogenen Daten möglich (5).

Pseudonymisierung

Artikel 4 Absatz 5 DSGVO definiert die Pseudonymisierung als Verarbeitungsweise personenbezogener Daten, bei der Daten ohne die Hinzuziehung zusätzlicher Informationen einer spezifischen betroffenen Person nicht mehr zugeordnet werden können. Die zusätzlichen Informationen werden gesondert aufbewahrt und unterliegen technischen und organisatorischen Maßnahmen, die gewährleisten, dass die personenbezogenen

Daten nicht ohne eine zusätzliche Verarbeitung einer Person zugeordnet werden können (5). Das bedeutet, personenbezogene Daten, wie z. B. der Name und Wohnort, können etwa durch einen Code aus Buchstaben oder Zahlen ersetzt werden, um eine Re-Identifizierung an einer dritten datenverarbeitenden Stelle zu verhindern. Hierzu wird beispielsweise eine Codierungstabelle erstellt, die für jedes Datum sowohl die personenbezogenen Daten als auch den zugeordneten Code als Schlüssel enthält und dementsprechend die Re-Identifizierung möglich macht.

Die Option einer Re-Identifizierung pseudonymisierter Daten durch den berechtigen Inhaber ist in einigen Fällen sogar beabsichtigt. Eine Re-Identifizierung ist z. B. vorgesehen, wenn die pseudonymisierten Daten durch die datenerhebende Stelle regelmäßig aktualisiert werden und diese in pseudonymisierter Form an eine dritte Stelle weitergegeben werden. In diesem Fall ist der in der Einwilligung vereinbarte Verwendungszweck noch nicht abgeschlossen und der Datenschutz gilt gegenüber der dritten Stelle.

Die Daten unterliegen auch weiterhin der DSGVO, da der direkte Bezug zur betroffenen Person bei pseudonymisierten Daten nur vorübergehend eingeschränkt ist. Da diese Einschränkung unter der Hinzuziehung zusätzlicher Informationen wiederhergestellt werden kann, wie es z. B. über eine Identifizierungsliste erfolgt, gilt die Pseudonymisierung auch nicht als Technik zur Aufhebung des Personenbezugs (5).

Anonymisierung

Für die Anonymisierung ist in der DSGVO keine eindeutige Definition enthalten, wird jedoch im Erwägungsgrund 26 zur DSGVO erwähnt:

> *„Die Grundsätze des Datenschutzes sollten für alle Informationen gelten, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen. (…) Um festzustellen, ob eine natürliche Person identifizierbar ist, sollten alle Mittel berücksichtigt werden, die von dem Verantwortlichen oder einer anderen Person nach allgemeinem Ermessen wahrscheinlich genutzt werden, um die natürliche Person direkt oder indirekt zu identifizieren, wie beispielsweise das Aussondern. Bei der Feststellung, ob Mittel nach allgemeinem Ermessen wahrscheinlich zur Identifizierung der natürlichen Person genutzt werden, sollten alle*

> *objektiven Faktoren, wie die Kosten der Identifizierung und der dafür erforderliche Zeitaufwand, herangezogen werden, wobei die zum Zeitpunkt der Verarbeitung verfügbare Technologie und technologische Entwicklungen zu berücksichtigen sind. Die Grundsätze des Datenschutzes sollten daher nicht für anonyme Informationen gelten, d.h. für Informationen, die sich nicht auf eine identifizierte oder identifizierbare natürliche Person beziehen, oder personenbezogene Daten, die in einer Weise anonymisiert worden sind, dass die betroffene Person nicht oder nicht mehr identifiziert werden kann. Diese Verordnung betrifft somit nicht die Verarbeitung solcher anonymer Daten, auch für statistische oder für Forschungszwecke."* (10)

Die Stelle, an der die Anonymisierung der Daten vorgenommen wird, muss prüfen, welches Ausmaß an finanziellen und zeitlichen Mitteln nötig wäre, um Rückschlüsse zu den betroffenen Personen zu erlangen. Hierbei sollen Technologien und Methoden erprobt und berücksichtigt werden, die zum Zeitpunkt der Verarbeitung dem Stand der Technik entsprechen. Ist eine ausreichende Anonymisierung erfolgt, indem alle personenbezogenen Daten entfernt wurden und ein Personenbezug nicht wiederhergestellt werden kann, unterliegen die Daten nicht weiter dem Anwendungsbereich der DSGVO (5).

Einige Untersuchungen ließen darauf schließen, dass sich ein Personenbezug durch Anwendung eines Anonymisierungsverfahrens nicht in jedem Fall aufheben lässt, speziell bei großen Datenmengen aus verschiedenen Datenquellen (3). Dies gilt teilweise sogar bei der Kombination mehrerer Verfahren (5). Für die Re-Identifizierung wäre allerdings ein erheblicher Aufwand erforderlich, der eine Datenbeschaffung aus verschiedenen Quellen und deren korrektes Management bedeuten würde.

Bei einer Zusammenführung von Daten aus einer US-Volkszählung von 1990 ergab eine Studie, dass sich in diesem Szenario eine Re-Identifizierung in sehr vielen Fällen allein mit dem Geschlecht, der Postleitzahl und dem Geburtsdatum vornehmen ließ. Dabei konnten 87 % der amerikanischen Bevölkerung anhand dieser Daten identifiziert werden (11). Hierbei sollte betrachtet werden, dass die Bevölkerungsdichte in den USA mit 33,4 Einwohnern/km^2 (12) erheblich geringer ist als z. B. in Deutschland mit 232,5 Ein-

wohner/km^2 (12). Ist die Bevölkerungsdichte niedriger, steigt die Aussagekraft bestimmter Merkmale, wie z. B. der Postleitzahl, da weniger Personen aus einer Region existieren, denen die personenbezogenen Daten zugeordnet werden können. Daher muss bei einer Anonymisierung immer auch eine Kontextbetrachtung durchgeführt werden, um das erforderliche Maß an kritischen Parametern festzustellen.

Generell gilt, je mehr Daten von einer Person erhoben werden, desto größer ist die Wahrscheinlichkeit der Re-Identifizierung (3). Z. B. ist ein Rückschluss auf eine betroffene Person anhand einer über mehrere Jahre dokumentierten individuellen Patientengeschichte samt Diagnosen, Operationen und Verordnungen wahrscheinlicher, selbst wenn personenbezogene Daten, wie der Name, Geburtstag und Wohnort unkenntlich gemacht wurden. Mithilfe der Anzahl und des Umfangs der sonstigen erhobenen vorhandenen Daten kann eine Re-Identifikation zur Person möglich sein.

Auf andere Weise ist es auch möglich anhand einer seltenen Krankengeschichte, z. B. ursächlich eines Gendefektes, Personen anhand anonymisierter Daten zu identifizieren (13). Das lässt sich z. B. durch die geringe Anzahl der auftretenden Fälle und der damit zusammenhängenden individuellen Historie erklären.

Eine weitere Folge der Anonymisierung durch die Reduzierung oder Vertauschung der Daten, sowie das Hinzufügen von Zufallswerten, kann eine Minderung der Datenqualität sein. So können bei der Verarbeitung der Daten Verzerrungen und Bias entstehen (14). Gegebenenfalls können durch die Reduzierung oder Verrauschung zunächst irrelevant erscheinender Daten kausale Zusammenhänge übersehen werden, welche für wissenschaftliche Erkenntnisse relevant sein könnten.

1.2 **Big Data**

Die Speicherung, Verwaltung und Analyse großer komplexer Datenmengen wird unter dem Begriff *Big Data* zusammengefasst. Die Verarbeitung solcher großen Datenmengen hat in den letzten Jahren in der medizinischen Forschung an Bedeutung gewonnen (3).

Vor allem in der *Big Data*-Forschung sollen unbekannte Zusammenhänge entdeckt werden, die nicht mit dem ursprünglichen Verwendungszweck in Verbindung stehen. Da die Qualität der zu verarbeitenden Daten hoch sein muss, stehen die beiden Aspekte, die Privatheit und die Qualität der Daten, im direkten Konflikt zueinander. Dementsprechend ist nach rechtlichen Vorgaben für die Datenverarbeitung von *Big Data* eine gültige Einwilligung notwendig (3).

1.3 Synthetische Daten

Um für den langfristigen Einsatz in der Forschung Daten von guter Qualität zu erhalten, lassen sich künstliche Daten mit Hilfe von Algorithmen erzeugen. Das Resultat ist ein künstlicher Datensatz (DS)[1], welcher keine Bezüge zu realen Tatsachenbeständen und Personen aus einem Originaldatensatz haben soll. Die erzeugten Daten können dann problemlos unter Einhaltung sämtlicher Datenschutzanforderungen institutionsübergreifend zur Verarbeitung und Analyse zur Verfügung gestellt werden (15). Die erzeugten Daten werden als *synthetische Daten* bezeichnet.

1.3.1 Eigenschaften und Einsatz

Synthetische Daten können prinzipiell anhand jeglicher Daten, z. B. auch aus Bild- und Textdateien generiert werden. Auch die zu generierende Anzahl der künstlichen Objekte[2] lässt sich beliebig planen, unabhängig davon welchen Umfang der Quelldatensatz hat (14). Ein weiterer Vorteil gegenüber Realdaten ist das Ausbleiben von Fehlwerten in quantitativ künstlich generierten Objekten. Das kann z. B. bei Verhaltensprüfungen neu entwickelter technischer Systeme gegenüber potentiell fehlerhaften Realdaten Vorteile haben (16).

Demnach finden *synthetische Daten* als Testdaten, z. B. zur Simulation, in der Software- und Systemprüfung und in der medizinischen Ausbildung ihren Einsatz, ohne Bedenken,

[1] Die Gesamtheit einer Datensammlung wird in dieser Arbeit als *Datensatz* definiert.

[2] Einzelne Elemente eines Datensatzes (Tupel) werden in dieser Arbeit als *Objekte* definiert.

den Datenschutz oder Persönlichkeitsrechte zu verletzen (17). In einer Studie, die bewerten sollte, ob *synthetische Daten* und Originaldaten bei einer Analyse dieselben oder ähnliche Ergebnisse erreichen, konnten Azizi et al. dies bestätigen. Sie schließen daraus, dass *synthetische Daten* stellvertretend für Realdaten in Studien verwendet werden können (18).

Bestenfalls weisen die synthetisch erzeugten Daten die gleiche Datenqualität auf, wie der Originaldatensatz und führen bei Analysen zu den gleichen Aussagen. Kriterien, die Angaben über die Güte des verwendeten Modells ermöglichen, lassen sich auf die gleiche Weise und parallel beim realen, ursprunggebenden DS und beim künstlichen DS untersuchen. Auch Garantien zur Bewahrung der Privatheit können mathematisch ermittelt werden (14).

Das Ersetzen nur einer Teilmenge der Daten eines Originaldatensatzes ist ebenfalls möglich. Hierbei können Daten mit höherer Datenqualität erzeugt werden, jedoch besteht eine größere Wahrscheinlichkeit für eine Re-Identifizierung betroffener Personen aus dem Originaldatensatz. Das Risiko einer Re-Identifizierung sei nach Aussage von Ali und Dyakov jedoch geringer als bei Anonymisierungsverfahren (14).

Erste Ideen für die Erzeugung synthetischer Daten wurden Anfang der 1990er Jahre entwickelt (19). In den von D.B. Rubin entwickelten Verfahren werden Werte von den Originaldaten so ersetzt, dass sich die statistische Verteilung relevanter Merkmale von der Verteilung des ursprünglichen Datensatzes möglichst nicht unterscheidet und keine Verzerrungen entstehen (14).

Durch die zunehmende Verarbeitung von *Big Data* und den Fortschritt in der Forschung, wie z. B. im Bereich der *künstlichen Intelligenz,* hat die Synthese von Daten an Bedeutung gewonnen. Dies hängt mit der erforderlichen Einhaltung des Datenschutzes und der Verhinderung von Diskriminierung zusammen.

Auch für *synthetische Daten* ist, vergleichbar mit anonymisierten Daten, die DSGVO nicht mehr anwendbar. Für die Verwendung synthetischer Daten wird auch vorausgesetzt, dass eine Umkehrbarkeit zu den Realdaten und den betroffenen Personen ausgeschlossen werden kann (14).

Die *Datenethikkommission* (DEK)[3] sieht für die Forschung auf dem Feld synthetischer Daten Potential. Der Einsatz von *synthetischen Daten* ist in Bezug auf die Datenqualität, Vermeidung von Verzerrungen und für den Erhalt der Privatheit betroffener Personen ein Ansatz, den es lohnt, weiter zu prüfen:

> *„Die Datenethikkommission empfiehlt der Bundesregierung, die Forschung im Bereich synthetischer Daten zu fördern. Dabei besteht u.a. Forschungsbedarf zu der Frage, inwieweit und in welchen Kontexten synthetische Daten die Verarbeitung echter Daten ersetzen können und wie eng die synthetischen Daten an die Eigenschaften von echten Daten angelehnt sein sollen. Die DEK empfiehlt, die Erzeugung und den Einsatz synthetischer Daten weiter zu untersuchen, beispielsweise im Hinblick auf ihre Datenqualität und auf die Vermeidung von Verzerrungen (Bias) und Diskriminierung.“ (16)*

1.3.2 **Methoden und Tools**

Für die Generierung von synthetischen Daten gibt es unterschiedliche Methoden. Eine Möglichkeit synthetische Daten zu generieren, liegt in der Erstellung einer künstlichen Repräsentation eines realen Datensatzes. Für die Erzeugung werden die Eigenschaften der Daten, wie z. B. Dateninhalte, die statistische Verteilung und Deskription, analysiert und erklärt. Algorithmisch wird auf Basis dieser Informationen der synthetische DS mit

[3] Die DEK (18.07.2018 gegründet) entwickelt im Auftrag der Bundesregierung auf wissenschaftlicher und technischer Basis ethische Maßstäbe und Leitlinien, sowie konkrete Handlungsempfehlungen zum Schutz jeder einzelnen Person, zur Wahrung des gesellschaftlichen Zusammenlebens und die Sicherung und Förderung des Wohlstands im Informatikzeitalter (16,20).

den gleichen bzw. möglichst ähnlichen Eigenschaften erstellt, ohne Bezüge zu den realen Daten aufrecht zu erhalten (14). Dieses Ziel kann z. B. mit Methoden erreicht werden, die mit *Generative Adversarial Networks* (GANs) arbeiten (21). Ein GAN ist ein *Machine-Learning*-Modell, das aus *Künstlichen Neuronalen Netzen* Daten generiert (22).

Eine andere Methode erstellt die synthetischen Daten anhand von öffentlich verfügbaren Statistiken, wie z. B. epidemiologischen Informationen, medizinischen Leitlinien und Zensusdaten. Auch die statistische Verteilung und der Umfang der Daten können den Bedürfnissen bzw. dem Vorhaben in der Forschung und der geplanten Verarbeitung angepasst werden (17). Diese Methodik generiert realistische Daten ohne Bezug oder Nutzung von personenbezogenen Realdaten. Ein Beispiel für eine Methode zur Generierung synthetischer Gesundheitsdaten ohne einen Quelldatensatz ist das *Publicly Available Data Approach to the Realistic Synthetic EHR* (PADARSER)-Framework (17,23), welches auch als Ursprung für die Entwicklung weiterer Methoden diente. Die *Content Modeling for Synthetic E-Health Records* (CoMSER)-Methode (24) ist eine dieser Weiterentwicklungen.

Für die Generierung eines synthetischen Datensatzes sind verschiedene Werkzeuge und Programmbibliotheken verfügbar. Die Generierung synthetischer Daten ist sowohl mit Hilfe von *Open Source* Programmen als auch über kommerzielle Software möglich. Beispiele für Open Source Tools zur Generierung medizinischer synthetischer Daten sind *Synthea*[4] und *MedGAN*[5]. Die Erzeugung der synthetischen Daten erfolgt mit diesen Tools auf unterschiedliche Weise. Mit *Synthea* werden vollständige Patientengeschichten mit

[4] *Synthea* (25) erzeugt synthetische, realistische Patientendaten in verschiedenen Formaten, wie z.B. *HL7 FHIR* (26) oder *Comma-separated values* (*CSV)*.

[5] *MedGAN* ist ein Framework das synthetische Patentendaten mit Hilfe mit aggregierten Merkmalen erzeugt (21).

Hilfe von *Agentenbasierter Modellierung*[6] auf Basis von PADARSER-Frameworks generiert. Das Tool kann Daten für ausgewählte Krankheitsbilder erzeugen. Das zweite Tool *MedGAN* generiert die Daten auf Basis von GAN und verwendet für die Datenherstellung einen Originaldatensatz.
Die Firma *MDClone* bietet als Dienstleitung eine Plattform namens *ADAMS* an, um privatheitsbewahrende, realistische Daten generieren zu lassen (28). Die Daten werden mithilfe von vorhandenen Kohorten erzeugt und als eine künstliche Population bereitgestellt (23).

Verglichen mit den Tools zur Erzeugung synthetischer Daten gibt es auch Software, die eine Anonymisierung von personenbezogenen Daten in unterschiedlicher Methodik vornimmt. Beispiele für solche Anwendungen sind *PySyft*[7] oder *Anonimatron*[8]. Während *PySyft* durch eine verschlüsselte Verarbeitung bei *Deep-Learning*-Methoden den Datenschutz gewährleistet (29), ersetzt *Anonimatron* Werte mit Personenbezug durch kryptische Synonyme (31).

1.3.3 **Vergleichsschema für Pseudonymisierung, Anonymisierung und Synthese**

In Abbildung 1 sind zur Veranschaulichung die drei erwähnten Verfahren zur Sicherstellung des Datenschutzes schematisch abgebildet. Die Pseudonymisierung ist im Vergleich zu den anderen Verfahren umkehrbar. Die Farbgebung der Figuren im Originaldatensatz steht für die personenbezogenen Merkmale der einzelnen Personen. Diese Merkmale sind sowohl bei der Pseudonymisierung als auch bei der Anonymisierung nicht mehr vorhanden. Die Anzahl der vorhandenen Figuren (Objekten) ist bei diesen beiden Ver-

[6] *Agentenbasierte Modellierung* meint die computerbasierte Erstellung von Modellen und Simulationen komplexer Sachverhalte (27).

[7] *PySyft* (29) ist eine Bibliothek der Programmiersprache *Python* (30) zur Erstellung von anonymisierten Daten.

[8] *Anonimatron* anonymisiert Daten und verfälscht personenbezogene Daten, wie z.B. Namen und Email-Adressen (31).

fahren gleich wie beim Originaldatensatz, da sich die Datenmenge durch die Verwendung der Verfahren nicht ändert. Es könnten höchstens (zufällig) ausgewählte Objekte entfernt werden. Von diesen beiden Verfahren abgegrenzt, ist auf der rechten Seite die Synthese dargestellt. Die unterschiedlichen Farben der Figuren bilden die vorhandenen personenbezogenen Merkmale ab, die sich von den originalen Daten unterscheiden. Die eckige Form der Figuren symbolisiert, dass es sich nicht um Realdaten, sondern um künstlich erstellte Daten handelt. Die Anzahl der Figuren ist höher, da sich mittels *Synthese* eine beliebige Anzahl von Objekten in den Datensätzen generieren lässt.

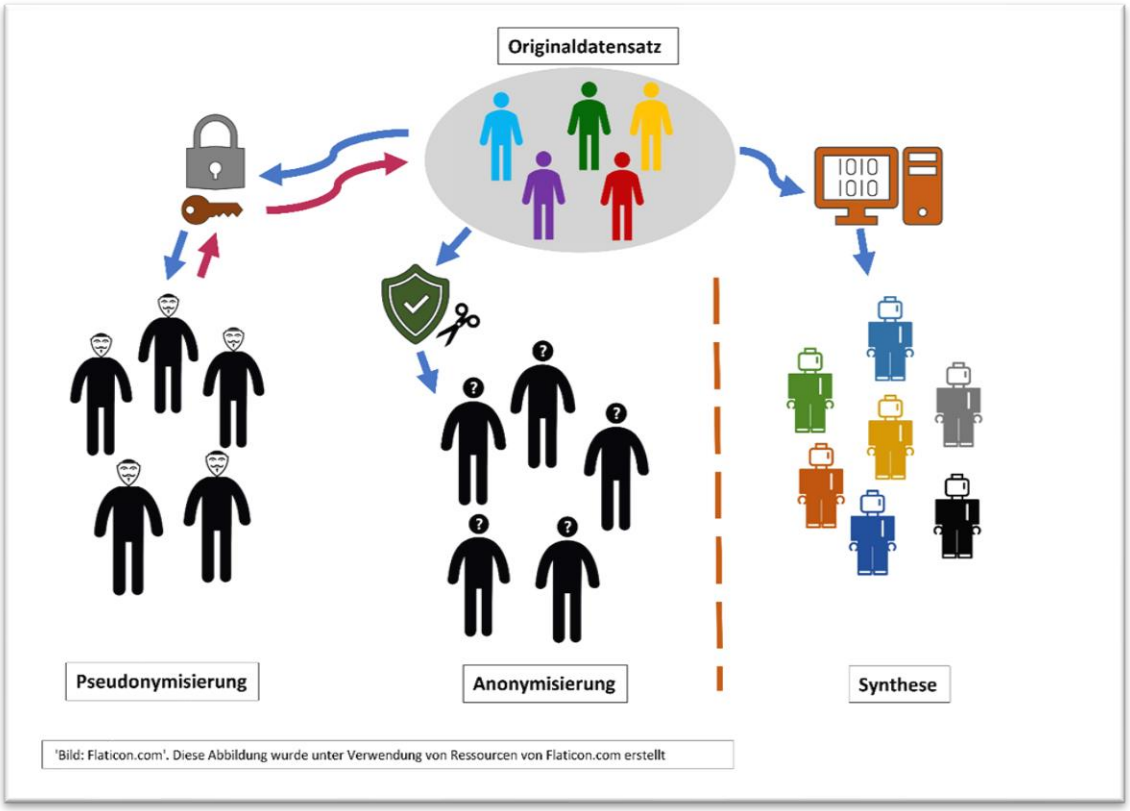


Abbildung 1:Vergleichsschema Pseudonymisierung, Anonymisierung und Synthese (eigene Darstellung)

1.4 Motivation der Arbeit

Die Motivation dieser Arbeit ist die Identifikation von Methoden, die der medizinischen Forschung unter Wahrung der Datenschutzvorschriften langfristig qualitativ hochwertige synthetische Daten zur Verfügung stellen können.

Im Rahmen dieser Arbeit werden vorhandene Werkzeuge zur Generierung synthetischer Daten verglichen. Aus diesem Vergleich wird ein Tool ausgewählt, welches die vorab ermittelten Anforderungen, speziell in Hinblick auf die Einhaltung des Datenschutzes und die Qualität der Daten, am vollständigsten erfüllen kann. Mit diesem Tool wird dann auf Basis eines frei verfügbaren Datensatzes (*MIMIC-III*-Datensatz) ein synthetischer DS erstellt. Anschließend werden die mit dem Modell erzeugten Daten sowohl auf die Datenqualität als auch auf die Einhaltung des Datenschutzes, bzw. den Grad eines potenziellen Verstoßes, geprüft.

1.4.1 **ScreenReject-Projekt**

Als Anwendungsbeispiel für diese Arbeit dient das *ScreenReject*-Projekt. Das *ScreenReject*-Projekt ist ein Verbundprojekt der *Hochschule Hannover*, dem Institut der Transfusionsmedizin der *Medizinischen Hochschule Hannover* und dem Institut für Technische Chemie der *Leibniz-Universität Hannover.* Das übergeordnete Ziel dieses Projektes ist der Schutz vor Organabstoßung und dem damit verbundenen Organverlust nach einer Nierentransplantation (32,33).

Im Rahmen dieses Projektes ist ein *Klinisches Data Warehouse* (*KDWH*) aufgebaut worden. Das *KDWH* ist in einem Teilprojekt an der Hochschule Hannover entstanden und soll Daten für die Entstehung eines Expertensystems für die Abstoßungsdiagnostik langfristig bereitstellen und die Entwicklung unterstützen. In das *KDWH* werden alle relevanten klinischen Daten aus den verschiedenen klinischen Primärsystemen überführt und in einer elektronischen Patientenakte, auf Basis des international bekannten *openEHR*[9]-Standards, bereitgestellt (32,33).

1.4.2 ***MIMIC-III*-Datensatz**

Für die Untersuchung und die Generierung der synthetischen Daten werden für diese Arbeit Daten aus einer frei verfügbaren Datenbank verwendet. Die Datenbank *Medical*

[9] *openEHR* ist ein internationaler offener Interoperabilitätsstandard für die Erstellung, Speicherung, Verwaltung und den Abruf von Gesundheitsdaten in elektronischen Patientenakten (34).

Information Mart for Intensive Care III (*MIMIC-III*) erfasst ca. 40.000 Personen, die im Zeitraum von 2001 bis 2012 auf den Intensivstationen *des Beth Israel Deaconess Medical Center* (Boston, USA) behandelt wurden. Der Datenumfang erfasst z. B. stündliche Vitaldatenmessungen, Laborergebnisse, Bildgebungsberichte, Medikationen und die Mortalität der Patienten (35).

1.5 **Problemstellung**

Aus den beschriebenen Sachverhalten sind folgende Problemstellungen abgeleitet:

Problem 1: Eine offene Frage ist, welche Anforderungen an die Datenqualität und die Vermeidung der Re-Identifizierung zu stellen sind und wie diese zu prüfen sind.

Problem 2: Unbekannt ist, welche verfügbaren Werkzeuge diese Anforderungen in welchem Umfang erfüllen können.

Problem 3: Nicht bekannt ist, bis zu welchem Grad synthetisch erzeugte Daten gegenüber den Ursprungs- bzw. Originaldaten im statistischen Vergleich, bei Auswertungen und Interpretationen zu gleichen bzw. annähernden Resultaten führen.

Problem 4: Unklar ist, ob und wie umfangreich Rückschlüsse der synthetisch generierten Daten auf die personenbezogenen Daten des Originaldatensatzes sind und ob eine Re-Identifizierung ausgeschlossen werden kann.

1.6 **Zielsetzung**

Im Einzelnen wird das Erreichen folgender Ziele angestrebt:

Zu Problem 1 :

Ziel 1: Zusammenstellung qualitativer und quantitativer Anforderungen an Art, Umfang und Qualität der zu generierenden synthetischen Forschungsdaten.

Zu Problem 2:

Ziel 2: Identifikation und vergleichende Darstellung geeigneter Tools für die Generierung synthetischer Daten und Auswahl eines Favoriten nach den in Ziel 1 festgelegten Anforderungen.

Ziel 3: Synthese eines neuen Datensatzes auf Basis der bereitgestellten Eingangsdaten aus der *MIMIC-III*-Datenbank.

Zu Problem 3:

Ziel 4: Analyse, Beschreibung und Vergleich des ursprünglichen und des synthetischen Datensatzes in Hinblick auf die im Rahmen von Ziel 1 erarbeiteten Anforderungen an die Qualität der Daten.

Zu Problem 4:

Ziel 5: Untersuchung der Güte des erstellten Modells in Bezug auf die Einhaltung des Datenschutzes (Ausschlusses eines Personenbezugs) der in Ziel 3 generierten synthetischen Daten.

1.7 **Fragestellung**

Folgende Fragen lassen sich von der Zielsetzung ableiten und sollen in dieser Arbeit beantwortet werden:

Zu Ziel 1:

Frage 1.1: Welche Anforderungen in Bezug auf Art, Umfang und Qualität der generierten Daten lassen sich aus deren potenzieller Verwendung für datenbasierte Fragestellungen medizinischer Forschung ableiten?

Zu Ziel 2:

Frage 2.1: Welche Methoden und Werkzeuge für die privatheitsbewahrende Synthese von medizinischen Daten sind aktuell in der Literatur bekannt bzw. als Software verfügbar?

Frage 2.2: Welche der verfügbaren Tools können sowohl die ermittelten Anforderungen aus Frage 1.1 an die Daten als auch die ermittelten technischen Anforderungen aus Frage 2.1 am ehesten erfüllen?

Zu Ziel 3:

Frage 3.1: Wie werden Daten mit dem ausgewählten Tool synthetisiert?

Zu Ziel 4:

Frage 4.1: Wie kann die Qualität der synthetisierten und der Originaldaten formalisiert beschrieben werden, damit ein Vergleich möglich ist und welche Methoden sind geeignet, diesen Vergleich durchzuführen?

Frage 4.2: Wie umfangreich bzw. relevant sind die möglichen auftretenden Abweichungen bei den in Frage 4.1 definierten Vergleichsmethoden?

Zu Ziel 5:

Frage 5.1: Ist sichergestellt, dass die synthetischen Daten keine Rückschlüsse auf die personenbezogenen Daten des Originaldatensatzes ermöglichen und somit keine Re-Identifizierung zulassen?

1.8 Gliederung der Arbeit

In der Einleitung werden die wichtigsten Zusammenhänge zum Thema Datenschutz und synthetische Daten erklärt. Auch die Motivation für diese Arbeit wird in diesem Kapitel erläutert und zugehörige Informationen zum *ScreenReject*-Project und zum verwendeten Datensatz gegeben. Anschließend folgt die Analyse der vorhandenen Problemstellen und die daraus hervorgehende Ableitung der Ziele für die Masterarbeit.

Im zweiten Kapitel folgt eine Aufbereitung der grundlegenden Informationen, um das Verständnis zum Thema dieser Arbeit zu verbessern. Z. B. gibt dieses Kapitel Informationen zur Datenqualität und ihrer Evaluierung. Auch verschiedene Methoden zur Anonymisierung und die damit verbundenen Termini werden in diesem Kapitel erklärt. Im Anschluss werden weitere Erläuterungen zu Verfahren zur Erzeugung synthetischer Daten gegeben. Abschließend werden die ermittelten Anforderungen, sowohl an die zu erstellenden Daten als auch an das zu verwendende Tool erörtert.

Die Vorgehensweise bei der Synthese und der Identifizierung des für die Synthese verwendeten Tools werden im dritten Kapitel präsentiert. Im Methodenteil werden auch die Eigenschaften der verwendeten Datensatze und die Inbetriebnahme des für die Synthese verwendeten Tools und die Bereitstellung der Arbeitsumgebung beschrieben. Zuletzt wird dargelegt, wie der erzeugte Datensatz auf Datenqualität und Einhaltung des Datenschutzes geprüft werden.

Das vierte Kapitel werden zunächst die für die Synthese verwendeten Daten beschrieben. Anschließend werden die aus der Synthese resultierenden Ergebnisse präsentiert. Zunächst wird der Originaldatensatz beschrieben und anschließend mit dem synthetischen Datensatz verglichen. Darauf folgt die Prüfung des verwendeten Modells bezüglich der Einhaltung der Datenschutzanforderungen.

Das letzte Kapitel schließt die Arbeit mit der Prüfung und Diskussion der Ergebnisse ab, nachdem die Eignung der verwendeten Methoden erörtert worden ist.

2 Grundlagen

In diesem Kapitel werden die Ergebnisse aus der Literaturrecherche und die daraus entnommenen Informationen für das Verständnis dieser Masterarbeit erläutert und die in der Recherche ermittelten verfügbaren Tools zur Generierung synthetischer Daten gegenübergestellt. Außerdem werden die erarbeiteten Anforderungen an den zu verwendenden und zu erzeugenden Datensatz, sowie das für die Synthese verwendete Tool beschrieben.

2.1 Literaturrecherche

In den folgenden Abschnitten wird das Vorgehen bei der Literaturrecherche erläutert. Zunächst werden die für die Recherche verwendeten Begriffe dargelegt und anschließend wird das PRISMA-Diagramm zur Literaturrecherche präsentiert.

2.1.1 Suchbegriffe für die Literaturrecherche

Für die Literaturrecherche sind die Schlüsselbegriffe und die dazugehörigen *Medical Subject Headings* (MeSH)[10]-Terme aus dem Titel dieser Arbeit „***Methoden*** und ***Werkzeuge*** zur ***privatheitsbewahrenden Synthese medizinischer Forschungsdaten***“ verwendet worden. Die ausgewählten Suchbegriffe sind mit den englischen Begriffen und den MeSH-Termen in Tabelle 1 aufgelistet. Das Ziel, eine gute Datenqualität zur erreichen, ist dem Ziel einer Wahrung der Privatsphäre gleichgestellt. Deshalb ist die Liste der Suchbegriffe um diesen Begriff ergänzt worden. Synonyme für die Suchbegriffe sind in der Suchsyntax nicht berücksichtigt worden, weil bereits ausreichend viele Treffer erzielt werden konnten.

[10] Medical Subject Heading (MeSH) sind ein von der Medizin und Biowissenschaften genutztes Schlagwortregister, dass der Erschließung von Büchern und Zeitschriftenartikeln dient (36).

Tabelle 1: Suchbegriffe für die Literaturrecherche

Begriff	**Englischer Begriff**	**MeSH-Term**
Methoden	Methods	Methods
Werkzeuge	Tools	Software
Privatheitsbewahrend	Privacy preserving	Confidentiality
Synthese/synthetische Daten	Synthetic data	-
Medizinische Forschungsdaten	Medical research data	Big Data
Datenqualität	Data Quality, Data Accuracy	Data Accuracy

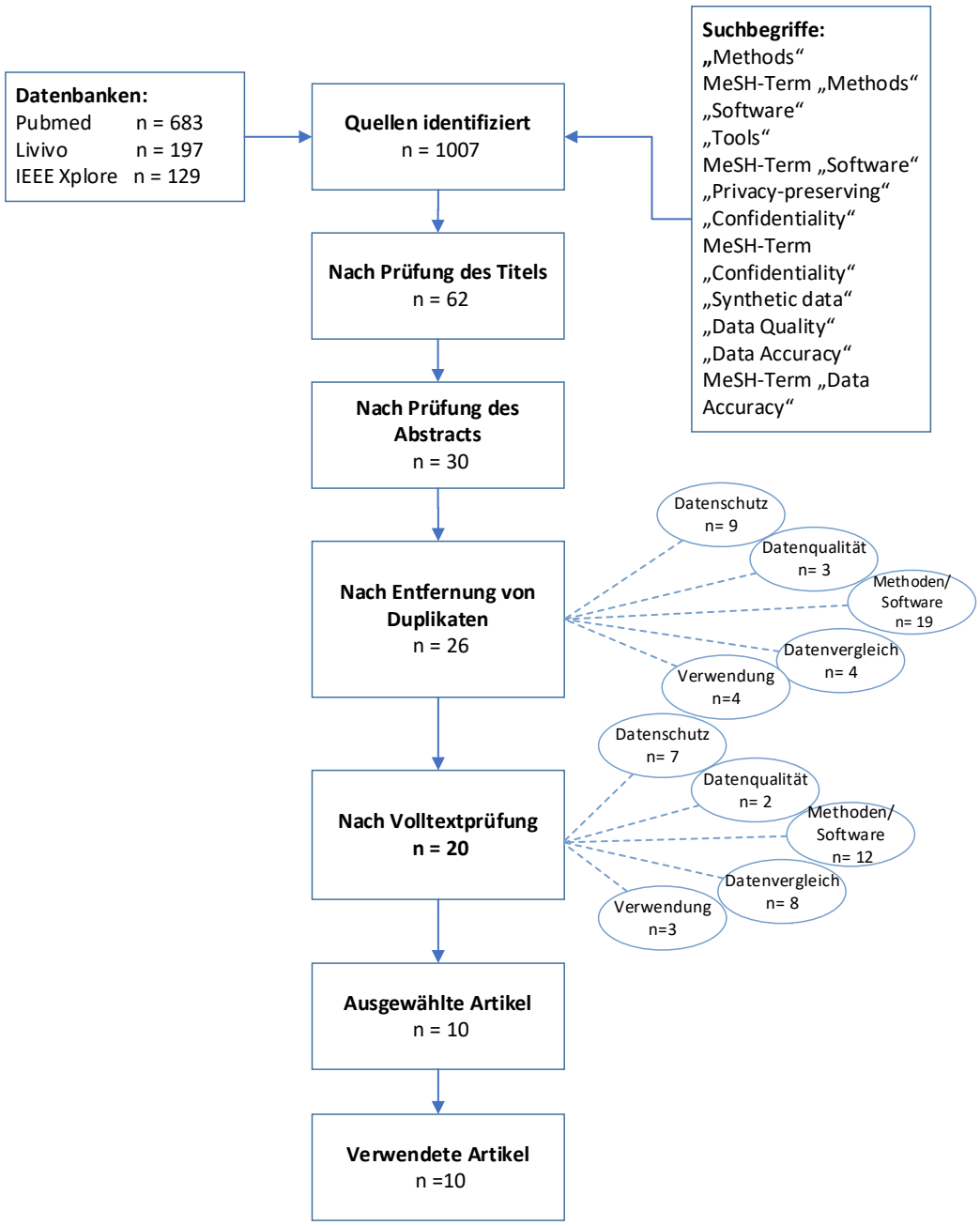
2.1.2 **Darstellung der Literatursuche**

Für die am 14.05.2021 durchgeführte Literaturrecherche ist in den drei Forschungsdatenbanken *Pubmed*, *Livivo* und *IEEE Xplore* gesucht worden. Um das Vorgehen nachzuvollziehen und das Ergebnis reproduzieren zu können, ist die Recherche konzeptuell in einem PRISMA-Diagramm[11] visualisiert worden. In Abbildung 2 ist das Schema dargestellt, das die Vorgehensweise und die Selektion der gefundenen und letztendlich verwendeten Literatur beschreibt.

Während der Durchsicht der Abstracts sind die Inhalte und Themen der gefundenen Referenzen den möglichen Bearbeitungsgebieten *Datenschutz*, *Datenqualität*, *Methoden und Software*, *Datenvergleich* und *Verwendung* zugeordnet worden, um eine schnellere Verfügbarkeit der Quellen zu erhalten. Dem Gebiet *Datenschutz* sind Veröffentlichungen über Methoden zur Prüfung des Datenschutzes zugeordnet worden. Artikel, die Evaluierungsmethoden der Datenqualität thematisiert haben, sind dem Gebiet *Datenqualität* zugeordnet worden. Veröffentlichungen, in denen sich die Forschenden mit Methoden zur Erzeugung synthetischer Daten gewidmet haben, sind dem Gebiet *Methoden und Software* zugeordnet worden. Publikationen, die dem Gebiet *Datenvergleich* zugeordnet wurden, beziehen sich inhaltlich auf den Vergleich eines originalen mit einem synthetischen DS.

[11] Ein PRISMA-Diagramm stellt den Informationsfluss einer systematischen Recherche schematisch in unterschiedlichen Phasen der Literaturüberprüfung dar (37).

Dem letzten Themengebiet sind wissenschaftliche Artikel zugeordnet, in denen Beispiele für die Verwendung von synthetischen Daten beschrieben worden sind. Die Ergebnisse der Literarturrecherche werden sowohl für die Identifizierung möglicher Messverfahren zur Einhaltung des Schutzes personenbezogener Daten und zum Vergleich der generierten Daten mit den Originaldaten als auch für den Vergleich der Methoden und Tools zur Generierung synthetischer Daten verwendet.



PubMed Search: ((("Synthetic data"[All Fields] AND ("Software"[All Fields] OR "Software"[MeSH Terms] OR "Methods"[All Fields] OR "Methods"[MeSH Terms])) OR ("Synthetic data"[All Fields] AND ("Privacy preserving"[All Fields] OR "Confidentiality"[MeSH Terms] OR "Confidentiality"[All Fields])) OR (("Data Accuracy"[All Fields] OR "Data Accuracy"[MeSH Terms] OR "Data Quality"[All Fields]) AND "Synthetic data"[All Fields])) AND "2016/05/14 00:00":"3000/01/01 05:00"[Date - Publication]) AND (y_5[Filter])

Livivo: ((("synthetic data") OR KW=("synthetic data")) AND "Software") AND PY=2011:2021

Xplore:("All Metadata":„Synthetic data“) AND ("All Metadata":„Methods“) OR ("Mesh_Terms":„Methods“) AND ("All Metadata":„Software“) OR ("Mesh_Terms":„Software“) AND ("All Metadata":„Privacy-preserving“) OR ("All Metadata":„Confidentiality“) OR ("Mesh_Terms":„Confidentiality“) AND ("All Metadata":„Data Quality“) OR ("Mesh_Terms":„Data Accuracy“) OR ("All Metadata":„Data Accuracy“)
Filters Applied: data privacydata mining
Journals
IEEE Journal of Biomedical and Health InformaticsIEEE Transactions on Industrial InformaticsIEEE Transactions on Cloud ComputingIEEE/ ACM Transactions on Computational Biology and BioinformaticsIEEE Transactions on Pattern Analysis and Machine Intelligence

Abbildung 2: Schematische Darstellung und Suchsyntax der Literatursuche

2.2 **Datenqualität**

Gegenüber der Erhaltung der Privatheit ist eine hohe Qualität der generierten synthetischen Daten für diese Arbeit ein gleichgesetztes Ziel. Im folgenden Abschnitt werden der Begriff Datenqualität für diese Arbeit definiert und die Kriterien für die Messbarkeit der Datenqualität erarbeitet.

Eine schlechte Datenlage kann in Analysen zu Verzerrungen führen, die dann Fehlentscheidungen in der medizinischen Versorgung zur Folge haben können. Daraus können wirtschaftliche oder in der Medizin auch gesundheitliche Schäden für betroffene Personen entstehen (38), sodass die Notwendigkeit besteht die ermittelten Daten in einem kontinuierlichen Prozess auf ihre Qualität zu prüfen.

Allgemein kann festgestellt werden, dass eine ausreichende Datenqualität erreicht ist, wenn die Daten dem Bedürfnis der Verwendung entsprechen, was auf diesem Gebiet als *„fit for use“* bezeichnet wird. Wang und Strong verwenden für diese Verwendungsweise die Formulierung *„for the task at hand“* (40, 41). Um diese Bedürfnisse zu ermitteln, können die Personen, die ein Interesse an den Daten haben, über die Anforderungen befragt werden (41).

In der Literatur fehlen einheitliche Definitionen für die Datenqualität und die damit zusammenhängenden Kriterien (38). Trotzdem soll die Qualität von Daten anhand messbarer Kriterien ermittelt werden können.

2.2.1 **Definition von Datenqualität**

Für den Begriff Datenqualität gibt es mehrere Definitionen, unterschiedlicher Ursprünge. Im Nachfolgenden sind zwei Definitionen von Datenqualität aufgeführt, deren Aussagen ähnlich sind.

Nach VG Würthele definiert sich Datenqualität folgendermaßen (38):

> *„Mehrdimensionales Mass* [sic] *für die Eignung von Daten, den an ihre Erfassung/Generierung gebundenen Zweck zu erfüllen. Diese Eignung kann sich über die Zeit ändern, wenn sich die Bedürfnisse ändern.“*

Eine weitere Definition ist durch V Wandt erstellt worden (42):

> *„Datenqualität ist die Bewertung von Datenbeständen hinsichtlich ihrer Eignung, einen bestimmten Zweck zu erfüllen („fitness for use“). Als Kriterien gelten dabei die Korrektheit, die Relevanz und die Verlässlichkeit der Daten, sowie ihre Konsistenz und Verfügbarkeit auf verschiedenen Systemen.“*

Beide Definitionen meinen, dass mehrere Kriterien für die Qualität der Daten eine Rolle spielen können. Diese sind abhängig davon, für welche Verwendung die Daten erhoben bzw. zusammengestellt werden sollen. Diese Anforderungen können jedoch im Laufe der Zeit variieren. Demnach werden die Kriterien der Datenqualität für jedes Vorhaben individuell festgelegt und können sich während der Durchführung noch verändern. Wandt geht bereits auf Qualitätskriterien ein, die für die Messung der Datenqualität entscheidend sein können.

2.2.2 **Kriterien der Datenqualität**

In der Literatur werden mehrere Kriterien zur Messbarkeit der Datenqualität eingeführt. In einer Umfrage mit 30 Studierenden, welche bereits Erfahrungen im Datenmanagement gesammelt haben, ermittelten Wang und Strong für die Datenqualität relevante Kriterien (39). Das Ergebnis der ermittelten Qualitätskriterien wird in Tabelle 2 präsentiert. Die aus der Umfrage ermittelten Kriterien werden zur Messung der Datenqualität in der Literatur häufig erwähnt und für die Analyse angewandt.

Tabelle 2: Datenqualitätskriterien (39,41)

Kategorie	Qualitätskriterium	Beschreibung
Intrinsische Datenqualität	Glaubwürdigkeit (believability)	Akzeptanz, Wahrheitsgehalt, Realität und Glaubwürdigkeit der Daten
	Genauigkeit (believability)	Korrektheit, Zuverlässigkeit und Fehlerfreiheit der Daten
	Objektivität (objectivity)	Vorurteilsfreiheit und Diplomatie der Daten
	Vertrauenswürdigkeit	Seriosität der der Datenquellen
Kontextuelle Datenqualität	Mehrwert (value-added)	Wertschöpfung durch die Verarbeitung der Daten
	Relevanz (relevancy)	Bedeutsamkeit und Gewichtigkeit der Daten
	Aktualität (timeliness)	Ausreichendes Alter der Daten für den Verwendungszweck
	Vollständigkeit (completeness)	Ausreichende Breite, Tiefe und Umfang der Daten
	Angemessene Datenmenge (appropriate amount of data)	Für den Verwendungszweck ausreichende, aber nicht unnötig große Menge erhobener Daten
Repräsentative Datenqualität	Interpretierbarkeit (interpretability)	Eindeutige Definition über die Angabe der Daten (Sprache, Messeinheit)
	Verständlichkeit (understandability)	Eindeutigkeit und Klarheit der Daten
	Konsistente Darstellung (consistent representation)	Einheitliche Formatangabe und Kompatibilität der Daten
	Kompakte Darstellung (concise representation)	Angabe erforderlicher Daten ohne Überladung
Zugriffsqualität	Zugänglichkeit (accessibility)	Verfügbarkeit und Abrufbarkeit der Daten
	Zugriffsicherheit	Eingeschränkter Zugang zu schützenswerten Daten

In Publikationen werden die Kategorien *Vollständigkeit*, *Fehlerfreiheit*, *Konsistenz* und *Aktualität* besonders intensiv diskutiert und sind demnach bei der Prüfung der Datenqualität besonders zu beachten (41).

Z. B. wurden diese Kriterien in einer Umfrage aus dem Jahre 2002, die im Rahmen der Dissertation von Helfert in fünf größeren Unternehmen in Deutschland, Österreich und der Schweiz durchgeführt wurde, am häufigsten genannt (41,43).

2.2.3 **Metriken zur Evaluierung der Datensätze**

In den folgenden Abschnitten werden die in der Literaturrecherche ermittelten Veröffentlichungen über Verfahren zur Prüfung der Datenqualität von synthetischen Daten präsentiert. Aus der Literaturrecherche resultierte, dass sich eine gute Datenqualität eines synthetischen Datensatzes durch möglichst gleichgesetzte Eigenschaften gegenüber dem realen DS auszeichnet. Auch nach den in Kapitel 2.2.2 beschriebenen Definitionen müssen für jeden Anwendungsfall die Qualitätskriterien individuell festgelegt werden. Deshalb werden die Metriken für die Prüfung der Datenqualität und des Datenvergleichs zwischen dem realen und synthetischen Datensätzen zusammengefasst erläutert.

In Tabelle 3 werden die Metriken zur Evaluierung der Datensätze, die in der Literaturrecherche identifiziert wurden, zusammengefasst dargestellt. Ausgewählte Metriken, wie z. B. die nach Goncalves et al., werden anschließend erläutert.

Tabelle 3: Übersicht der Metriken zur Evaluierung

Publikation	**Autoren**	**Beschriebene Metriken**
Generating high-fidelty synthetic patient data for assessing machine learning healthcare software, *2020* (13)	Tucker et al.	Kolmogorov-Smirnov-Test Kullback-Leibler-Divergenz Chi-Quadrat-Test (Signifikanztest)
Synthesizing electronic health records using improved generative networks, 2018 (21)	Baowaly et al.	Kolmogorov-Smirnov-Test Dimension-wise probability Dimension-wise average
Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies, *2020* (44)	Benaim et al.	Vergleich von Verhältnissen, Odds-Ratios, Hazard Ratio und Überlebenszeitanalysen
Generation and evaluation of synthetic patient data, *2020* (45)	Goncalves et al.	Kullback-Leibler-Divergenz, Paarweise-Korrelations-Differenz, Log-Cluster Metrik,

Publikation	Autoren	Beschriebene Metriken
		Support-Coverage-Metrik, Cross-Classification
A Method for Generating Synthetic Electronic Medical Record Text, *2021* (46)	Guan et al.	Self-BLEU
Anonymization Through Data Synthesis Using Adversarial Networks (ADS-GAN), 2020 (47)	Yoon, Jingsun et al.	Jensen-Shannon Divergenz, Wasserstein Distanz
Privacy-Preserving Generative Deep Neural Networks Support Clinical Sharing, (48)*, 2019*	Beaulieu-Jones et al.	Paarweise-Pearson Korrelationen

Evaluationsmetriken nach Goncalves et al.

Für die Prüfung der Qualität und den Vergleich zweier Datensätze haben Goncalves et al. eine Reihe an Metriken aufgestellt. Die Sammlung eignet sich für die Gegenüberstellung der realen Daten und den daraus generierten synthetischen Daten.

Kullback-Leibler (KL)-Divergenz

Die KL-Divergenz ist ein Maß, das die Unterschiede der statistischen Verteilungen von zwei Datensätzen feststellt (13, 50). Dafür wird die Verteilung über ein Paar von Datensätzen mit den gleichen Eigenschaften, z. B. von einem realen und einem synthetischen DS, für eine bestimmte Variable berechnet und anschließend wird die Ähnlichkeit der Verteilung gemessen.

Sind beide Verteilungen identisch, ist die KL-Divergenz gleich 0. Je höher der Wert der KL-Divergenz ist, desto größer ist die Diskrepanz zwischen den beiden Verteilungen. Die KL-Divergenz muss für jede Variable einzeln berechnet werden und misst keine Abhängigkeiten zwischen den Variablen (45).

Paarweise-Korrelations-Differenz (PKD)

Bei der PKD wird ermittelt, wie stark sich die Korrelationen zwischen den Attributen von zwei Datensätzen, z. B. eines realen und eines synthetischen Datensatzes, unterscheiden. Die PKD wird als Frobeniusnorm[12] des *Pearson*-Korrelationskoeffizienten gemessen. Je näher die PKD an 0 ist, desto größer sind die Übereinstimmungen der linearen Korrelationen der beiden Datensätze (45).

Hierbei kann die PKD sowohl einen Wert im positiven als auch im negativen Bereich annehmen, der im Wertebereich zwischen -2 und 2 liegt. Liegen z. B. zwei zu vergleichende Pearson-Korrelationskoeffizienten mit stark abweichenden Werten 0,98 und -0,99 vor, so beträgt die PKD 1,97. Die PKD kann in einer *Heatmap*[13] visualisiert werden.

Log-Cluster Metrik

Die *Log-Cluster Metrik* ermittelt die Ähnlichkeit der zugrundeliegenden *Cluster*[14]-Struktur zweier Datensätze, wie z. B. eines realen und synthetischen Datensatzes. Für die Berechnung werden beide Datensätze zusammengeführt und anschließend eine *Clusteranalyse* mit einer definierten Anzahl von *Clustern* unter der Verwendung des *k-Means-Algorithmus*[15] durchgeführt. Goncalves et al. legen die Anzahl der Cluster auf 20 fest.

Je größer die Übereinstimmungen in den Clustern beider Datensätze sind, desto weniger Unterschiede liegen in den Verteilungen der beiden verglichenen Datensätze vor (45). Anhand der mathematischen Formulierung resultiert aus zwei vollständig identischen

[12] Die Frobeniusnorm ist eine Matrixnorm, die sich aus der Wurzel der Summe der Betragsquadrate aller Matrixelemente definiert (50).

[13] Eine *Heatmap* ist eine zweidimensionale Visualisierung, die unter den Einsatz von Farben unterschiedliche Werte darstellen (51), wie z.B. die Korrelationen von Merkmalen im paarweisen Vergleich eines Datensatzes.

[14] Ein Cluster ist eine homogene Gruppe, die bezüglich eines Merkmals gleichartig sind (52).

[15] Der k-Means-Algorithmus ist ein Rechenverfahren des unüberwachten Maschinellen Lernens, das für die Clusteranalyse, einer Gruppierung von Objekten, eingesetzt wird (53).

Datensätzen ein U_C-Wert von $-\infty$. Aus der Formel kann ebenso abgeleitet werden, dass bei zwei Datensätzen mit der gleichen Größe, in denen keinerlei Überschneidungen in den Clustern vorliegen (jedes Cluster enthält ausschließlich synthetische oder ausschließlich reale Daten), der maximal erreichbare U_C-Wert bei $\log_{10}(0{,}25) \approx -0{,}6$ läge (45).

Support-Coverage-Metrik

Die *Support-Coverage-Metrik* erfasst, wie viele der im realen DS vorhandenen Variablen auch im synthetischen DS vorhanden sind. Dabei wird das Verhältnis und die Anzahl der Ausprägungen mit berücksichtigt (45).

Cross-Classification (CrCl)

Eine weitere Metrik zum Messen, wie ähnlich die statistischen Abhängigkeiten zweier Datensätze sind, ist die *Cross-Classification*. Anders als bei der PKD, die eine statistische Abhängigkeit anhand des Pearson-Korrelationskoeffizienten misst, werden bei der *CrCl* die Abhängigkeiten über die Vorhersagen ermittelt, die über andere Variablen für eine Variable mit einem Klassifikator generiert werden. Das bedeutet, ausgehend von den anderen Variablen, wird eine Variable klassifiziert und das Klassifikationsergebnis mit dem ursprünglichen Objekt verglichen. Goncalves et al. nutzen für das Klassifikationsmodell einen Entscheidungsbaum[16] als Klassifikator (45).

Goncalves et al. stellen zwei Metriken der Kreuzklassifikation vor:

Die *CrCl-Real-Synthetisch (RS)* nutzt für das Training der Kreuzklassifikation den realen DS und prüft das erstellte Modell für die Vorhersage sowohl mit den realen Daten als auch mit den synthetischen Daten. Diese Metrik kann verwendet werden, um die statistischen Eigenschaften beider Datensätze zu vergleichen.

[16] Ein Entscheidungsbaum (engl. Decision Tree) ist eine im überwachten Maschinellen Lernen verbreitete Methode der Klassifikation/Regression von Datensätzen (54).

Im Ablauf werden die verfügbaren Daten in Trainings- und Testdatensätze aufgeteilt. Auf den Trainingsdatensatz, der aus einem Teil der realen Daten besteht, wird ein Klassifikator trainiert. Um das Modell zu testen wird der Testdatensatz verwendet, der sowohl aus dem zurückgehaltenen Teil des realen Datensatzes als auch dem synthetisch generierten DS besteht. Die Leistung, im Sinne der Güte der Klassifikation, wird für beide Datensätze festgelegt. Der Wert *CrCl-RS* definiert sich aus dem Verhältnis der Modellperformance bei den synthetischen Testdaten und zur Modellperformance bei den zurückgehaltenen realen Testdaten.

Um die Klassifizierung durchzuführen, wird eine der Variablen als Zielvariable verwendet, während die übrigen als Prädiktoren dienen. Dieses Vorgehen wird für jede Variable als Zielvariable wiederholt. Als Richtwert dient der Durchschnittswert der ermittelten Leistung. Der ideale Wert für eine Kreuzklassifikation liegt bei einem Wert von 1 (45).

Bei der *CrCl-Synthetisch-Real (SR)* wird, im Gegensatz zur *CrCl-RS,* als Trainingsdatensatz ein Teil der synthetischen Daten genutzt. Die Prüfung des Klassifikationsmodells erfolgt demnach mit den zurückgehaltenen synthetischen Daten und dem Real-Datensatz. Diese Metrik kann verwendet werden, um festzustellen, ob Schlüsse, die aus erstellten statistischen Modellen bzw. dem *Machine-Learning* (ML)-Modellen gezogen wurden, plausibel sind. Dies bezieht sich z. B. auf ML-Modelle, die mit synthetischen Datensätzen trainiert wurden. Die *CrCl-SR* kann prüfen, ob diese Modelle ebenfalls sicher auf reale Datensätze angewendet werden können (45).

Das Vorgehen und die Berechnung der *CrCl-SR* erfolgt auf gleiche Weise, wie bei der *CrCl-RS.*

Kolmogorov-Smirnow-Test

Um die statistischen Wahrscheinlichkeitsverteilungen des realen und des synthetischen Datensatzes zu vergleichen, beschreiben Baowaly et al. und Tucker et al. den *Kolmogorov-Smirnow-Test* (KS-Test).

Der KS-Test ist ein nichtparametrischer Anpassungsgütetest. Er findet Verwendung in der Feststellung, ob sich zwei Verteilungen zweier Datensätze oder eine Verteilung von einer angenommen Verteilung unterscheiden (55).

Baowaly et al. berücksichtigen für den KS-Test zum einen die dimensionsspezifische Wahrscheinlichkeit, wobei die Berechnungen numerischer Größen exemplarisch wie folgt vorgenommen wird:

$$Dimensionsspez.Wahrscheinlichkeit = \frac{Anzahl\ der\ Patienten\ mit\ Erkrankung/Behandlung}{Gesamtanzahl}$$

Zum anderen wird der dimensionsspezifische Durchschnitt berechnet. Dieser bezieht sich auf den Mittelwert eines Merkmals von einer Dimension (Patienten mit einer bestimmten Ausprägung (z. B.: Diagnose/Behandlung):

$$Dimensionsspez.Durchschnitt = \frac{Summe\ eines\ Merkmals\ (Spalte)}{Anzahl\ der\ Datensätze}$$

Der KS-Test wird für eine Stichprobe beider Datensätze durchgeführt, um festzustellen, ob für beide Datensätze eine ähnliche Verteilung vorliegt. Die Nullhypothese wird verworfen, wenn der *p-Wert* ≤ 0,05 ist (21). Bei einem signifikanten Ergebnis liegt ein Unterschied in der Verteilung beider Datensätze vor.

2.3 Bewahrung des Datenschutzes

Auch für die Methoden zur Prüfung des Datenschutzes ist, vergleichbar mit Tabelle 3, eine Übersicht der aus der Literaturrecherche hervorgehenden Methoden erstellt worden. In Tabelle 4 ist eine Übersicht von Publikationen gelistet, die Methoden zur Privatheitswahrung bzw. Messung des Offenlegungsrisikos beschreiben.

Tabelle 4: Methoden zur Überprüfung des Datenschutzes

Publikation	Autoren	Beschriebene Methoden
Anonymization Through Data Synthesis Using Adversarial Networks (ADS-GAN), 2020 (47)	Yoon et al.	Differential Privacy (47)
Big Data Privacy in Biomedical Research*, 2020* (56)	Wang et al.	ε-Differential Privacy k-Anonymity (56)
Privacy-Preserving Generative Deep Neural Networks Support Clinical Sharing, *2019* (48)	Beaulieu-Jones et al.	ε, δ-Differential Privacy (48)
Evaluating Identify Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation, *2020* (57)	Emam et al.	Methodik zur Bewertung des Identitäts-Offenlegungsrisikos von vollsynthetischen Daten
Generating high-fidelty synthetic patient data for assessing machine learning healthcare software*, 2020* (13)	Tucker et al.	Re-Identifikationsrisiken durch Ausreißeranalysen mit Distanz-Metriken erkennen

In den folgenden Abschnitten wird zunächst die wichtigste Terminologie im Zusammenhang mit der Einhaltung des Datenschutzes bzw. der Aufhebung des Personenbezuges erklärt und anschließend werden ausgewählte Methoden zur Prüfung der Privatheit erläutert.

2.3.1 **Direkte Identifikatoren**

Direkte Identifikatoren, auch eindeutige Identifikatoren genannt, sind Attribute bzw. Merkmale, die eine eindeutige Re-Identifikation einer Person zulassen (58).

Neben dem Patientennamen sind die Patientenidentifikationsnummer oder auch die Krankenversicherungsnummer Beispiele für direkte Identifikatoren aus dem medizinischen Bereich.

2.3.2 **Quasi-Identifikatoren**

Quasi-Identifikatoren, auch als Indirekte Identifikatoren bekannt, sind mehrere Attribute bzw. Merkmale, die in ihrer Kombination und im Zusammenhang mit externen Informationen Rückschlüsse auf eine bestimmte Person ermöglichen oder absichern (58,59).

Diese Merkmale lassen bei anonymisierten Daten scheinbar keine Rückschlüsse zu, können aber durch andere, z. B. öffentlich zugängliche Informationen oder weiteren Datenquellen, eine Re-Identifizierung möglich machen. Bei diesen Merkmalen handelt es sich z. B. um Informationen zur ethnischen Herkunft, Geburtsdatum, Geschlecht oder Postleitzahl der betroffenen Personen (59).

2.3.3 **Sensible Attribute**

Sensible Attribute sind Informationen, die durch die De-Identifizierung zu schützen sind. Die Offenlegung sensibler Attribute kann für die betroffene Person zu Nachteilen mit finanziellen, persönlichen oder gesellschaftlichen Auswirkungen führen, wenn sie mit ihr in Zusammenhang gebracht werden (57,58).

Sensible Attribute in der medizinischen Forschung stellen z. B. Diagnosen, Operationen bzw. Behandlungsprozeduren oder verordnete Medikamente dar (57,58). Je nach Sachverhalt kann ein *Quasi-Identifikator* auch ein sensibles Attribut sein. So kann die ethnische Herkunft oder die Konfessionszugehörigkeit sowohl als *Quasi-Identifikator* als auch als sensibles Attribut definiert werden.

2.3.4 *k*-Anonymität

Um eine Aussage treffen zu können, wie hoch die Wahrscheinlichkeit ist, dass sich zwei Datensätze verknüpfen lassen, kann die *k-Anonymität* angewendet werden. Mit diesem Datenschutzmodell kann die Wahrscheinlichkeit für eine Re-Identifizierung aus den verwendeten Datensätzen berechnet werden (5).

Die Eigenschaft der *k-Anonymität* ist erreicht, wenn die Informationen für jedes im DS enthaltene Objekt von mindestens *k-1* anderen enthaltenen Objekten nicht unterschieden werden können (47). Der zugehörige Parameter ist der *k*-Wert und drückt aus, wie viele Objekte des Original-Datensatzes innerhalb einer Äquivalenzklasse[17] quasi-identifizierender Attribute mindestens vorkommen müssen. Um dies zu erreichen werden die identifizierenden Attribute gelöscht und die quasi-identifizierenden Attribute verallgemeinert. Z. B. wird die Postleitzahl 31234 in 31* verallgemeinert oder das Alter wird in Gruppen zusammengefasst. Der *k*-Wert ist eine natürliche Zahl und repräsentiert bei einem höheren Wert eine stärkere Anonymität (5).

Da in den Äquivalenzklassen bei den sensiblen Attributen identische Attributwerte vorhanden sein können, weist sich dieses Vorkommen als Schwäche der *k-Anonymität* aus. Es kann z. B. nicht ausgeschlossen werden, dass bei allen *k* verschiedenen Objekten einer bestimmten Äquivalenzklasse von Patientendatensätzen das Attribut „Diagnose“ den Wert „HIV“ hat und dadurch für einen Patienten die Diagnose bekannt ist, wenn er dieser Äquivalenzklasse zugeordnet werden kann.

Infolgedessen sind aus der *k-Anonymität* die *ℓ-Diversity* und *t-Closeness* als Weiterentwicklungen entstanden (5).

[17] Eine Äquivalenzklasse ist eine Menge aller Elemente, die im Vergleich zur Grundmenge in einem oder mehreren Merkmalen gleichwertig sind (60).

2.3.5 ℓ -Diversity

Bei der ℓ-*Diversity* wird das Datenschutz-Modell in Bezug auf die Homogenität in den Äquivalenzklassen verbessert, indem die Verbindung eines sensiblen Attributs mit einer anderen Person des Datensatzes geschützt wird. Das wird durch das Verstecken des sensiblen Attributs erreicht, indem sichergestellt wird, dass in jeder Äquivalenzklasse mindestens ℓ verschiedene Ausprägungen zu einem Attribut vorhanden sind (z. B. ℓ verschiedene Diagnosen). So werden mehrere bzw. weitere Informationen benötigt um eine betroffene Person identifizieren zu können (5).

2.3.6 *t*-Closeness

Aufbauend auf der ℓ- *Diversity* verfeinert die *t-Closeness* den Ansatz weiter. Bei diesem Prinzip werden die Äquivalenzklassen so gebildet, dass die Merkmalswerte der ursprunggebenden Daten ähnlich verteilt sind. Das kann erreicht werden, indem eine Bedingung eingeführt wird, die ℓ verschiedene Werte in einer Äquivalenzklasse voraussetzt und vorschreibt, wie oft jeder Wert in jeder Äquivalenzklasse vertreten sein soll, um eine Verteilung zu erhalten, die der Verteilung der ursprünglichen Daten entspricht (5). Weitere Verfeinerungen der *t-Closeness* beziehen zudem die Semantik der Werte sensibler Attribute ein, um z. B. zu verhindern, dass zwar eine Äquivalenzklasse entsteht, die ℓ verschiedene Attributwerte enthält, diese Attributwerte aber alle einem gemeinsamen Oberbegriff (z. B. Unfruchtbarkeit) zuzuordnen sind (61).

2.3.7 Anonymisierungstechniken

Um einen anonymisierten DS zu erstellen, gibt es verschiedene Techniken. Neben der *formalen Anonymisierung*, bei der ausschließlich die direkten Identifikatoren entfernt werden, und die deswegen als unsicher eingestuft wird, sind weitere Techniken entwickelt worden, um eine Re-Identifikation ausschließen zu können. Die Techniken können kombiniert verwendet werden, um einen anonymisierten DS zu erhalten (62). Im Folgenden werden mögliche Techniken beschrieben, mit denen die Anonymisierung eines DS erreicht werden kann.

Verringerung der repräsentativen Personen

Die Anzahl der im DS erfassten Personen lässt sich verkleinern, indem einzelne Objekte mit Ausreißern oder seltenen Merkmalswerten oder -kombinationen entfernt werden. Auch durch das Auslassen eines festgelegten Prozentsatzes oder das Einbeziehen nur einer Stichprobe aus dem Gesamtdatensatz kann der Personenbezug verringert werden (62).

Veränderung der Merkmalsausprägungen

Für die Veränderung der Merkmalsausprägungen einzelner Objekte gibt es verschiedene Optionen (62). In Tabelle 5 werden die verschiedenen Techniken kurz erläutert.

Tabelle 5: Techniken zur Veränderung der Merkmalsprägungen (62)

Technik	**Beschreibung**
Verrauschen	Hinzufügen künstlicher Messfehler, ohne Veränderung der statistischen Aussage des Gesamtdatensatzes, z. B. durch das Verfälschen eines Geburtsdatums.
Vergröbern	Zusammenfassen von Wertebereichen, z. B. durch die Erstellung von Altersgruppen.
Mikro-Aggregation	Ersetzen der Einträge von Objekten durch die Mittelwerte kleiner Gruppen von Objekten.
Zufälliges Vertauschen	Vertauschen zufälliger Einträge eines Merkmals ohne Veränderung der statistischen Verteilung.

2.3.8 Angriffe auf anonymisierte Daten nach vorhergehenden Methoden

In der Literatur sind verschiedene Angriffe auf die *k-Anonymität* bekannt und definiert worden. Bei dem *Angriff mit Hintergrundwissen* handelt es sich um einen dieser Angriffe. Mit dieser Art von Angriffen ist gemeint, dass eine angreifende Person durch das ihr bekannte Wissen und Fachkenntnisse Personen aus einem anonymisierten DS identifizieren kann. Beispielsweise lässt die Diagnose Prostatakarzinom Rückschlüsse auf das Geschlecht und das etwaige Alter einer betroffenen Person zu, wenn weitere Merkmale offenliegen (63).

Ein weiterer Angriffstyp ist der *Homogenitätsangriff*. Hier können Rückschlüsse auf Personen einer Äquivalenzklasse hergestellt werden, bei denen in allen Attributen identische Merkmale vorliegen. Diese Art Angriff ist besonders bei Klassen mit kleinem *k-Wert* erfolgreich (63). Eine Person lässt sich z. B. dadurch identifizieren, dass zwei männliche Personen mit derselben Postleitzahl, demselben Geburtsjahr und einer identischen Diagnose einer Äquivalenzklasse zugeordnet sind (siehe (s.) Beispiel zu Schwächen der *k-Anonymität* in Kapitel 2.3.4).

Weitere mögliche Angriffe sind als *Unsorted-Matching-Angriff* und *Komplementärveröffentlichung* definiert worden. Ein *Unsorted-Matching-Angriff* bezieht sich auf die in den Tabellen vorliegenden und erhaltenen Reihenfolgen der enthaltenen Tupel und den verknüpften Tabellen in relationalen Datenbanken. So können anhand der Position der Zeile in der Tabelle zusammengehörige Tupel über mehrere Tabellen verknüpft werden (63).

Bei der *Komplementärveröffentlichung* werden zwei aus derselben Datenquelle gewonnene, unterschiedlich *k*-anonymisierte Tabellen miteinander kombiniert. Die daraus entstehende Folge ist, dass die zu schützenden Merkmale offengelegt werden können (63).

2.3.9 **Differential Privacy**

Die *Differential Privacy (DP)* ist ein übergeordnetes Maß für die Ermittlung der Wahrscheinlichkeit, dass eine Person als Teil in einer Datenbank erfasst ist (63). Die von Informatikern entwickelte Technik hat sich als eine Standardmethode für die Wahrung der Privatsphäre etabliert, wobei die gesamte Abdeckung möglicher offenzulegender Informationen möglich ist (56). Im Falle eines Angriffs sollen durch DP die Identität oder sensible Informationen von Personen nicht preisgegeben werden.

Das grundsätzliche Konzept von DP ist, dass sich das Ergebnis einer Datenbankabfrage auf einen zu schützenden DS durch das Hinzufügen, Löschen oder Ändern einzelner Objekte nicht wesentlich ändert. Um sicherzustellen, dass die statistischen Eigenschaften des Datensatzes im Wesentlichen unverändert bleiben, werden die Abfrageergebnisse typischerweise durch Verrauschung der Daten verändert (56).

Im Folgenden werden die DP- Varianten *ε-Differential Privacy* und *(ε, δ)-Differential Privacy* vorgestellt.

ε-Differential Privacy

Als mathematische Funktion soll ε-DP den Grad der De-Identifizierbarkeit messbar machen, indem eine Aussage darüber getroffen wird, wie hoch die Wahrscheinlichkeit einer Re-Identifizierung einzelner Personen ist. Das Risiko der Re-Identifizierung wird durch den Parameter Epsilon (ε) ausgedrückt (5). Der Parameter ε misst die Veränderung der Daten, die durch das Hinzufügen, Entfernen oder Verändern vorgenommen wird.

Bei der ε-Differential Privacy wird die Wahrscheinlichkeit gemessen, dass in zwei Datenbanken mit der gleichen Datenabfrage das gleiche Ergebnis erreicht werden kann. Je kleiner ε ist, desto größer ist der Schutz vor einer Re-Identifizierung. Der Wert, den ε für eine Anonymisierung erhalten soll, muss für jeden Fall individuell entschieden werden (5,56).

Trotzdem sind in der Umsetzung Einschränkungen vorhanden, da DP nicht direkt für eine endliche Anzahl von Objekten berechnet werden kann. Das ist nach Yoon et al. dadurch begründet, dass das zugrundeliegende Wahrscheinlichkeitsmaß für einen beliebigen Teildatensatz aus einer unendlichen Anzahl an Objekten definiert ist. DP könne durch die Generierung synthetischer Daten nur theoretisch gewährleistet werden, da die DP bei einem endlichen synthetischen DS nicht geprüft werden kann (47).

Nach Yoon et al. liegt bei den unter Einhaltung der DP erzeugten, synthetischen Daten eine schlechte Datenqualität vor. Bezüglich einer vergleichbaren Verteilung gegenüber dem ursprunggebenden realen Daten konnten bei den generierten Daten bedeutende Unterschiede nachgewiesen werden, sodass sie für ihre vorgesehenen Verwendungszwecke je nach Anforderungen nicht in jedem Fall nutzbar seien. Yoon et al. begründen dies mit der strukturellen Architektur von DP, die keinen endlichen DS vorsieht (47).

(ε,*δ*)-Differential Privacy

Eine Erweiterung der ε-Differential Privacy ist die (ε, *δ*)-DP. Hier ist die Funktion um den Parameter Delta (*δ)* erweitert worden.

Auch hier misst der Parameter ε die Veränderung der Daten. Der zweite Parameter *δ* begrenzt den ε-Parameter, um ein Offenlegungsrisiko möglichst gering zu halten. Mit anderen Worten wird durch ε die Wahrscheinlichkeit ausgedrückt, Informationen zu verlieren und *δ* misst gleichzeitig, wie hoch die Wahrscheinlichkeit ist, die Privatsphäre zu verletzen.

Auch bei dieser sehr strengen Variante müssen die Parameter für ein zufriedenstellendes Ergebnis fallspezifisch festgelegt werden (48).

2.3.10 Methodik zur Bewertung des Identitäts-Offenlegungsrisikos von vollsynthetischen Daten nach Emam et al.

Emam et al. haben eine Methode entwickelt, um das Offenlegungsrisiko von Identitäten von vollsynthetischen Daten zu bewerten (57). In dem Modell wird angenommen, dass die synthetischen Daten identifiziert werden können, wenn sie mit Personen aus dem Originaldatensatz abgeglichen werden. Für den Abgleich werden die *Quasi-Identifikatoren* vorgesehen. Bei dem Modell werden mehrere Komponenten berücksichtigt, wie z. B. Übereinstimmungen zwischen den realen und synthetischen Daten zu finden und die Richtung des Abgleichs. Auch die Chance eines Angreifers neue Informationen zu erhalten, wird in dem von Emam et al. entwickelten Modell berücksichtigt.

Die Autoren konnten mithilfe der von ihnen entwickelten Methode an zwei Datensätzen unterschiedlichen Umfangs beweisen, dass die aus dem Realdatensatz generierten synthetischen Daten ein vier- bzw. fünfmal niedrigeres Offenlegungsrisiko haben (57).

2.4 Generierung synthetischer Daten

Im folgenden Abschnitt werden verschiedene Techniken für die Generierung synthetischer Daten vorgestellt. Anschließend werden die für den Vergleich ausgewählten Tools vorgestellt, welche auf Basis einer der betrachteten Techniken entwickelt wurden.

Die in der Einleitung erwähnte Methodik, synthetische Daten mithilfe von Zensusdaten zu generieren, wie es über die agentenbasierte Modellierung erfolgt, ist für die Zielsetzung dieser Arbeit nicht relevant. Dies ist damit zu begründen, dass der synthetische DS anhand eines bestehenden Datensatzes in die identische Datenstruktur generiert wird. Diese Methode und die dazugehörigen Werkzeuge, wie z. B. das bereits erwähnte Framework *Synthea* (17) oder das *R*-Paket *SimPop*[18] werden im weiteren Ablauf dieser Arbeit und den Methodenvergleich nicht weiter berücksichtigt.

2.4.1 Methoden zur Generierung synthetischer Daten

Der Literaturrecherche wurden verschiedene Methoden entnommen, mit denen die Generierung von synthetischen Daten möglich ist. In den folgenden Abschnitten werden einige dieser Methoden erläutert.

Bayes'sches Netzwerk (BN)

Ein BN ist ein probabilistisches, die Wahrscheinlichkeit berücksichtigendes, grafisches Modell, das in jedem seiner Knoten eine Zufallsvariable repräsentiert. Die Kanten zwischen den Knoten beschreiben bedingte Abhängigkeiten zwischen den entsprechenden Zufallsvariablen (45,65).

[18] SimPop ist eine Bibliothek für die Programmiersprache R, um auf Basis modellbasierter Methoden anhand von Hilfsdaten eine simulierte Population erstellt (64).

Der Aufbau eines BN erfolgt in einem Prozess, der aus zwei Schritten besteht. Im ersten Schritt erlernt das Netzwerk einen *azyklischen Graphen*[19] aus den Daten, der die paarweise bedingten Abhängigkeiten und Unabhängigkeiten zwischen den Variablen definiert. Im zweiten Schritt werden die *bedingten Wahrscheinlichkeiten* für jede Variable über die maximale Wahrscheinlichkeit geschätzt (45). In Tabelle 6 werden die Vor- und Nachteile eines BN gegenübergestellt.

Tabelle 6: Vor- und Nachteile eines Bayes'schen Netzwerks (45)

Vorteile eines BN	**Nachteile eines BN**
Gute Skalierung mit der Dimensionalität des Datensatzes	Vereinfachte Darstellung der Abhängigkeitsstruktur trotz Faktorisierung der vollständigen gemeinsamen Verteilung
Der erstellte azyklische Graph eignet sich für die Prüfung kausaler Zusammenhänge zwischen den Variablen	
Rechnerisch effizient	

Um synthetische Daten mit einem BN zu generieren werden die Graphenstrukturen und Wahrscheinlichkeitsverteilungen für die Generierung aus dem realen DS abgeleitet (45).

Generative Adversarial Network

Eine weitere Methode zur Erzeugung synthetischer Daten sind GAN, zu Deutsch *generierende gegnerische Netzwerke*. Mit diesem Ansatz können neben Mikrodatensätzen[20] auch synthetische Daten z. B. aus Bildern oder Texten generiert werden (22,45).

Ein GAN arbeitet mit zwei verschiedenen *Neuronalen Netzwerken*, die als *Generator* und *Diskriminator* bezeichnet werden. Das erste Netz, der *Generator*, erzeugt die synthetischen Daten anhand des realen Datensatzes. Das zweite Netz, der Diskriminator, klas-

[19] Als azyklischen Graph wird ein gerichteter Graph bezeichnet, der keinen verbindenten Zyklus enthält. Er lässt sich in eine Richtung durch die Ergänzung aller Kanten erweitern (66).

[20] Mikrodaten, synonym Einzelangaben, sind Angaben über einen Sachverhalt eines Merkmalsträgers, z.B. einer Person (Beruf, Gesundheitszustand) (67).

sifiziert die erzeugten Daten in reale oder synthetisch erzeugte Daten. Durch die Verknüpfung der beiden *Neuronalen Netze* und ein zahlreiches Wiederholen der Arbeitsschritte nimmt die Qualität der synthetischen Daten in Bezug auf die Eigenschaften der realen Daten weiter zu. Ziel dieses Prozesses ist es, dass der Diskriminator die realen Daten nicht mehr von den synthetischen Daten unterscheiden kann und somit realistische synthetische Daten erzeugt werden (21,22,45).

In der Literatur finden sich verschiedene GAN, um verschiedene Datentypen, wie z. B. Bild- und Textdateien, Zeitreihendaten oder *EHR*, zu generieren (68). In Tabelle 7 sind einige Vorteile von GAN zusammengetragen.

Tabelle 7: Vor- und Nachteile von GAN (45)

Vorteile eines GAN	**Nachteile eines GAN**
Fordert keine strengen probabilistischen Modellannahmen	Schwierig zu trainieren, da der Min-Max-Optimierungsprozess instabil sein kann
Leichte Erweiterung gemischter Datentypen ist möglich, z. B. stetiger und kategorieller Variablen	Kann Verteilung von diskreten Daten während eines Trainings nicht erlernen (21)
	Einschränkungen bei der Bearbeitung diskreter Daten

Parametrische bzw. nicht-parametrische Modelle

Eine weitere Möglichkeit für die Erzeugung synthetischer Daten erfolgt über die Erstellung statistischer, parametrischer und nicht-parametrischer Modelle in Form von Klassifizierungs- bzw. Regressionsmodellen.

Parametrische Modelle werden eingesetzt, wenn die verwendeten Daten eine Normalverteilung vorweisen. Nicht-parametrische Modelle hingegen beziehen sich auf schief verteilte Daten, wobei diese Modelle auch normalverteilte Daten erlauben.

Die Modellentwicklung erfolgt über Methoden des maschinellen Lernens. Für die Erstellung synthetischer Daten werden z. B. *Klassifikations- und Regressionsbäume (Entscheidungsbäume)* eingesetzt (69).

2.4.2 Tools zur Generierung synthetischer Daten

Durch die Literaturrecherche wurden verschiedene Tools, Bibliotheken und Frameworks für die Generierung synthetischer Daten identifiziert. Im folgenden Abschnitt werden die Werkzeuge vorgestellt, die für das Auswahlverfahren anhand der Anforderungsanalyse berücksichtigt werden. Die Auswahl der berücksichtigten Tools ist damit zu begründen, dass das erzeugte Ergebnis für diese Arbeit zielführend ist und eine Dokumentation der Tools, z. B. auf *GitHub* [21] oder über Publikationen vorliegt.

DataSynthesizer

Das Tool *DataSynthesizer* ist mit einem BN entwickelt worden und wurde in der Programmiersprache *Python 3*[22] implementiert. Die *Open Source* Anwendung *DataSynthesizer* verfügt über eine benutzerfreundliche Weboberfläche, um die Parameter für das Netzwerk zu adjustieren. So kann das Tool generisch und mit wenigen Eingaben des Benutzers genutzt werden. Der Datenimport erfolgt mit Daten in der *Ersten Normalform*[23] (73) im CSV-Format. Außerdem lässt sich der *DataSynthesizer* über ein klassisches *Python*-Skript verwenden, ohne die Anwendung im Browser aufzurufen (73).

Die Daten sollen für jedes Attribut in homogener Form vorliegen. Das Tool kann verschiedene Datentypen, wie z. B. *String*, *Integer*, *Float* und *Date Time* erfassen und in synthetische Daten umsetzen.

Zusätzlich besteht die Möglichkeit, bei der Generierung der synthetischen Daten eine Verrauschung mit ε-DP zu durchzuführen. So besteht die Option ggf. direkte Identifikatoren, wie z. B. den Namen, die Steuer-Identifizierungsnummer oder andere Zeichenketten, die nicht einer kategorischen Aufteilung unterliegen, durch kryptische Werte zu

[21] *GitHub* ist eine Verwaltungsplattform für IT-Entwicklungs- und Softwareprojekte (70).

[22] *Python 3* wurde im Dezember 2008 eingeführt und ist derzeit die aktuellste Version der Programmiersprache (30,71).

[23] Die *Erste Normalform* liegt vor, wenn alle Informationen in einer Tabelle atomar sind (72).

ersetzen. Als Resultat entsteht ein synthetischer DS mit beliebiger Größe, der sensible Informationen nicht offenlegt und dem realen DS statistisch ähnlich ist (74).

Für die Generierung können drei verschiedene Verfahren gewählt werden. Das erste Verfahren ist sehr rechenintensiv und untersucht, ob zwischen den verschiedenen Attributen des Datensatzes Korrelationen vorliegen *(Correlated attribute mode)*. Das zweite Verfahren generiert Daten ohne die Berücksichtigung der Korrelationen (*Independent atrribute mode)*. Dieses Verfahren eignet sich vor allem bei kleineren Datenmengen oder wenn eine Generierung im *Correlated attribute mode* bei größeren Datenmengen zu rechenintensiv ist. Als dritte Variante gibt es den Zufallsmodus *(Random Mode)*, der typkonsistente, Zufallswerte erzeugt, die dem Datenmodell der Ursprungsdaten entsprechen, jedoch von der Verteilung abweichen. Dieser Modus eignet sich, wenn die vorliegenden Daten sehr sensibel sind (74).

Zur Überprüfung der erzeugten Daten erstellt *DataSynthesizer* vergleichende Statistiken für jedes Attribut beider Datensätze als *Histogramm.* Im Korrelationsmodus wird zusätzlich eine *Heatmap* angefertigt, welche die paarweisen Korrelationen aller Attribute anzeigt. Diese wird anhand eines für den Attributtypen passenden Korrelations-Koeffizienten und der KL-Divergenz erstellt (74).

Eine Demoversion des Tools *DataSynthesizer* mit Probedatensätzen ist online verfügbar (75).

Medical GAN (medGAN)

Das Framework *medGAN* funktioniert auf Basis eines GAN und erzeugt realistische synthetische *EHR*-Daten anhand realer *EHR*-Datensätze. Das Tool *medGAN* zeichnet sich dadurch aus, dass es hochdimensionale Datensätze verarbeiten kann.

Zum einem werden diskrete Zähldaten, also z. B. wie oft ein Patient mit einem bestimmten *International Statistical Classification of Diseases and Related Health Problems* (ICD)-Code mit einer Erkrankung oder medizinischen Maßnahme assoziiert wird, verarbeitet. Zum anderen werden auch binäre Daten verarbeitet, das heißt, wie oft z. B. bestimmte

ICD-Codes bei einer Person anwesend bzw. abwesend sind (21). Das Framework ist in *Python* programmiert und über GitHub verfügbar (76).

Damit die Verarbeitung diskreter Daten mit *medGAN* möglich ist, haben die Entwickler Choi et al. einen *Autoencoder*[24] in das GAN implementiert, damit es die Verteilung hochdimensionaler diskreter Daten erlernen kann, indem hervorstechende Merkmale diskreter Variablen erfasst werden (78).

Zusätzlich ist in *medGAN* die von Choi et al. benannte Minibatch-Mitteilungsmethode integriert, um ein *Overfitting* beim Training mit geringen Datenmengen zu verhindern. Auch ein Modus-Kollaps soll durch die Minibatch-Mitteilungsmethoden verhindert werden. Dies tritt dann auf, wenn ein GAN dazu neigt, Daten zu generieren deren Werte sich ähneln (21,78).

Das Framework *medGAN* erzeugt aggregierte diskrete Informationen von virtuellen Patienten, die aus EHRs im Längsschnitt, einer Aneinanderreihung aller Informationen, abgeleitet werden. Die mit *medGAN* erzeugten Daten erreichen vergleichbare statistische Eigenschaften und eine vergleichbare Qualität wie der reale DS (45,78).

SynthEHR

Ein weiteres Framework zur Erzeugung synthetischer Daten, das auf der GAN-Methode basiert, ist *SynthEHR.* Die Entwickler Baowaly et al. wollten gegenüber *medGAN* ein verbessertes Netzwerk zur Erstellung synthetischer EHR bereitstellen. Das Ziel der Autoren war, synthetische Gesundheitsdaten zu erzeugen, deren Eigenschaften noch näher an den realen Daten sind (21).

Dafür ist das Framework *medGAN* über ein *Wasserstein-GAN mit Gradientenstrafe* (medWGAN) und ein *Boundary-Seeking-GAN* (medBGAN) weiterentwickelt worden.

[24] Ein Autoencoder ist ein neuronales Netz, das versucht Eingangsinformationen zu komprimieren und mit reduzierten Informationen im Ausgang wieder korrekt nachzubilden (77).

Für das *medWGAN* ist im *medGAN*-Framework das Netzwerk durch das *WGAN-gradient penalty (GP)* ersetzt worden. Das Ziel ist es, ein stabileres Training zu ermöglichen und bessere Stichproben aus dem Modell zu erhalten. Als weitere Variante ist das vorhandene Framework mit dem *BGAN*-Algorithmus modifiziert worden. So profitiert *SynthEHR* neben einer besseren Performance davon, dass ein Training mit einem DS möglich ist, der auch stetige Daten enthält (21).

SynthPop

Das von Nowok et al. entwickelte *R*-Paket *SynthPop* kann synthetische Daten mit *parametrischen* und *nicht-parametrischen statistischen Methoden* generieren. Das Package ist ursprünglich für die Generierung synthetischer Populationsdaten entwickelt worden, eignet sich aber auch für Datensätze anderen Ursprungs (69), wie z. B. elektronische Gesundheitsdaten und andere Mikrodaten.

Ziel bei der Generierung synthetischer Daten mit *SynthPop* ist es, im synthetischen DS die Verteilung jeder Variable der Verteilung des realen Datensatzes anzunähern. Dies erfolgt mit Hilfe von Regressionsmodellen, die anhand der realen Daten erstellt werden. Hierfür werden die Modelle so konditioniert, dass die bei der Generierung eines Variablenwertes berücksichtigten Kovariablen in der Synthesereihenfolge weiter zunehmen und die letzte Variable auf alle anderen konditioniert ist (69,79). Das bedeutet, dass Attribute, die weiter vorne positioniert sind, eine niedrigere Anzahl an Kovariablen haben als Attribute, die in der Reihenfolge weiter hinten stehen. Das letzte Attribut ist demnach auf alle Variablen konditioniert. Auf diese Weise können parametrische und nichtparametrische Modelle erstellt werden, die für jede Variable logische Zusammenhänge und fehlende Daten berücksichtigen (79).

Für das *SynthPop*-Paket gibt es eine Neuimplementierung in *Python*, welche es sogar erlaubt, das Offenlegungsrisiko zu beeinflussen (80).

2.5 **Anforderungen**

In diesem Abschnitt werden die Anforderungen (Anf.) erläutert. Hier wird unterschieden, welche Anforderungen an den verwendeten DS, die zu erstellenden synthetischen Daten und an das benutzte Tool für die Synthese erfüllt werden sollen.

2.5.1 **Anforderungen an den verwendeten DS**

Die Generierung der synthetischen Daten soll anhand des *MIMIC-III*-Datensatzes vorgenommen werden, welcher in Kapitel 3.2.1 näher beschrieben wird. Während der Erstellung der Masterarbeit wurde ein Meeting des Projektes *Zukunftslabor Gesundheit* (ZLG) des *Zentrums für digitale Innovationen Niedersachsen* (ZDIN) besucht (81,82). Die Arbeitsgruppe, dessen Spezialisierung in der Sicherstellung des Datenschutzes liegt, haben im Rahmen des Projektes ein Review zur Anonymisierung von Gesundheitsdaten erstellt (58). Für die Prüfung verschiedener Anonymisierungsmethoden haben Olatunji et al. ebenfalls eine Extraktion aus dem *MIMIC-III*-Datensatz verwendet. Um gegebenenfalls zu einem späteren Zeitpunkt Vergleiche mit dieser und auch anderen Arbeiten ziehen zu können, wird für die Synthese die gleiche Datenextraktion verwendet. Die Extraktion der Daten wird in Kapitel 3.2.3 näher beschrieben.

2.5.2 **Anforderungen an die synthetischen Daten**

Die Anforderungen an die synthetischen Daten beziehen sich vor allem darauf, dass die generierten Daten möglichst ähnliche statistische Eigenschaften in Bezug auf Umfang und Verteilungen aufweisen sollen, wie im ursprünglichen DS. Dieses Ziel kann z. B. mit der paarweisen Korrelationsdifferenz zwischen den Datensätzen gemessen werden.

Andere Qualitätsmerkmale beziehen sich auf den Wunsch, dass eine Plausibilität zwischen den Datensätzen über die in Beziehung stehenden Tabellen besteht. Das bedeutet z. B., dass einer weiblichen Person bei der Synthese kein Prostatakarzinom als Diagnose zugewiesen werden kann, Zeitfolgen stimmen und Prozeduren passenden Diagnosen zugeordnet werden.

Als nächstes Qualitätskriterium soll das Offenlegungsrisiko anhand ermittelter Metriken gemessen werden. Hierfür soll mindestens eine möglichst hohe ℓ*-Diversity* im synthetischen Datensatz erreicht werden. Die ℓ*-Diversity* soll ermittelt werden, ohne die *Quasi-Identifikatoren*, etwa wie bei der Anonymisierung, zu gruppieren oder zu vergröbern. Das bedeutet, die Äquivalenzklassen sollen nicht, z. B. durch die Erstellung von Altersgruppen, zusammengefasst werden, um einen möglichst hohen Informationsgehalt zu bewahren.

2.5.3 **Anforderungen an das zu verwendende Tool**

Die Anforderungen an das Tool finden sich darin, dass es kostenfrei kommerziell verwendbar ist und als Open Source zur Verfügung steht. Außerdem sollte das Tool mindestens CSV-Dateien als Eingangsdaten akzeptieren und mindestens die gleiche Anzahl an Daten synthetisieren können, wie im realen DS vorhanden sind. Nach Möglichkeit soll das Tool über eine Windows-Anwendungsumgebung zur Verfügung stehen und anwendungsfallbezogen konfigurierbar sein. Als letzte Anforderung sollte das ausgewählte Tool mit aktuell verfügbaren Bibliotheken und Quellpaketen lauffähig sein und idealerweise kontinuierlich weiterentwickelt werden.

2.5.4 **Anforderungsliste**

Eine gekürzte Version der Anforderungsliste ist in Tabelle 8 zu sehen, die komplette Tabelle mit den zu erfüllenden Anforderungen ist in Anhang A - Anforderungsliste zu finden.

Die Anforderungen sind in Muss- und Wunschanforderungen unterschieden. So kann differenziert werden, welche Anforderungen zwingend erfüllt werden müssen und für welche Anforderungen unter Umständen optimalere Lösungen gefunden werden können.

Tabelle 8: Anforderungsliste (gekürzt)

Anf.-ID	Art der Anf.	Beschreibung	Werte/Daten/Erläuterungen M – Mussanforderung W - Wunschanforderung
1		**Allgemeiner Aufbau**	
1.1	M	MIMIC-III-DS verwenden	Extraktion aus dem MIMIC-III DS
(…)	(…)	(…)	(…)
1.9	M	Daten in Erster Normalform	Atomare Aufteilung der Informationen
2		**Datenextraktion aus MIMIC-III enthält:**	Vorhandene Attribute im verwendeten RD
2.1	W	age (DOB)	Berechnet aus date of birth
(…)	(…)	(…)	(…)
2.25	W	organ failure (sofa)	Maßzahl zur Beurteilung des Organversagens bei Sepsis
3		**Zielformat des Datensatzes**	
3.1	M	CSV	Ausgabeformat der synthetischen Datensatzes al CSV
(…)	(…)	(…)	(…)
4		**Statistischer Umfang**	
4.1	M	Gleiche Anzahl zu generierender Objekte wie im RD	Identischer Datenumfang zwischen RD und SD (1:1)
4.2	M	Nahe Übereinstimmung der Deskription mit dem Originaldatensatz	Gleiche Statistische Eigenschaften beider DS
(…)	(…)	(…)	(…)
4.2.5	M	Median	Ähnlicher Median beider DS
5		**Erfüllung des Datenschutzes**	
5.1	M	k-Anonymität	Erreichen eines möglichst hohen K-Wertes
(…)	(…)	(…)	(…)
6		**Datenqualität des SD**	
6.1	M	Teilweise SD ohne Offenlegungs-Risiko	Synthese der direkten Identifikatoren, Quasi-Identifikatoren und sensiblen Attributen
(…)	(…)	(…)	(…)
7		**Anforderungen an das Tool**	
7.1	W	Integrierbarkeit	Einsetzbar in anderen Systemen wie z. B. SQL Server Data Tool (SSDT)
(…)	(…)	(…)	(…)

3 Methodik

In diesem Kapitel werden die in dieser Arbeit verwendeten Methoden näher dargelegt. Zunächst wird das schrittweise Vorgehen bei der Datensynthese nach McLachlan beschrieben, die Bereitstellung der Daten dargelegt und die Durchführung der Nutzwertanalyse[25] erläutert. Auch die Einrichtung der benötigten Arbeitsumgebung für diese Arbeit wird in diesem Kapitel beschrieben. Zum Abschluss werden die verwendeten Verfahren zur Analyse der Datenqualität und des Offenlegungsrisikos erläutert.

3.1 Vorgehen bei der Synthese nach McLachlan

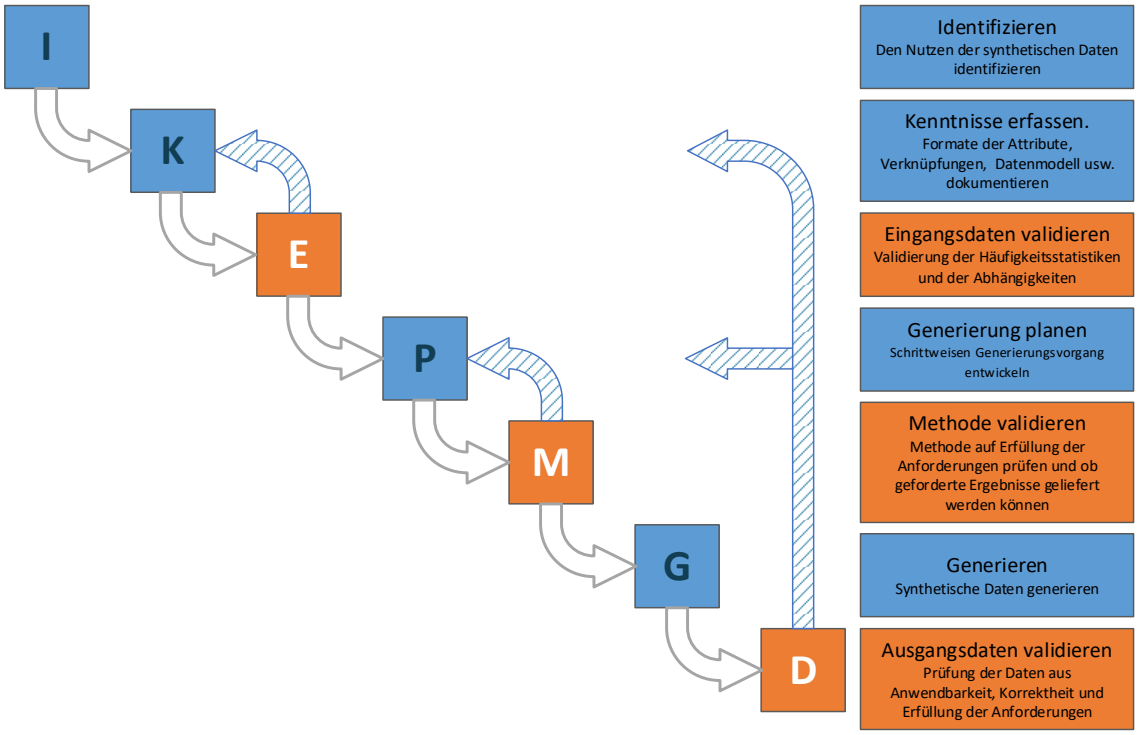


Abbildung 3: Schrittweises Vorgehen zur Erstellung synthetischer Daten (81)

Für die Generierung eines synthetischen Datensatzes hat McLachlan ein Vorgehensschema mit sieben Stufen beschrieben, das in Abbildung 3 mit den einzelnen Schritten

[25] Eine Nutzwertanalyse ist eine Entscheidungshilfe, die bei der Auswahl aus mehreren Alternativen hilft, bei denen subjektive Einflüsse ein Rolle spielen (83).

dargestellt ist (84). Bei diesem Vorgehen sind in drei Stufen Validierungen, im Schema orange gefärbt, vorgesehen. Die Validierungen sollen erfolgen, um einen synthetischen DS zu erhalten, der den Anforderungen entspricht. Treten an einer der Validierungsstufen Schwierigkeiten auf, sollen nach dem Stufenplan die vorherigen Schritte optimiert und wiederholt werden (84).

Der Ablauf zur Erzeugung des synthetischen Datensatzes dieser Arbeit orientiert sich an diesem Vorgehen. Im Folgenden werden die einzelnen Vorgehensschritte nach McLachlan (84) erörtert und die Planung der Umsetzung in dieser Arbeit dargelegt.

1. Identifizieren

Zu Beginn wird der Zweck der Verwendung des synthetischen Datensatzes festgelegt. McLachlan gibt an, dass die Problemstellung für den Verwendungszweck genau definiert werden muss. Der in der Literatur am häufigsten angegebene Grund für den Gebrauch synthetischer Daten ist, reale personenbezogene und somit sensible Daten zu ersetzen (84).

Der Verwendungszweck der Daten für diese Arbeit liegt in der Prüfung, ob die generierten Daten eine Re-Identifikation zulassen und in dem Vergleich der Datenqualität in Bezug auf den ursprunggebenden Datensatz. Die Anforderungen an die Daten für diese Arbeit sind in der Anforderungsanalyse, siehe Kapitel 2.5.2 beschrieben.

2. Kenntnisse erfassen

Der ursprunggebende DS sollte vor der Generierung synthetischer Daten umfassend analysiert und beschrieben werden. Hierfür sollen z. B. die Datentypen und Formate der Attribute, Verknüpfungen zu anderen Tabellen erfasst werden und deskriptive Statistiken der Eingangsdaten erstellt werden (84).

Für den verwendeten Teildatensatz wird ein Data Dictionary erstellt, s. Anhang B – Data Dictionary, Tabelle 26, indem die Informationen zu den einzelnen Attributen und den Verbindungen zwischen den Tabellen zusammengefasst sind.

3. Eingangsdaten validieren

In der nächsten Stufe werden die erstellten Informationen und Statistiken auf ihre Korrektheit, Plausibilität uns Aussagekraft geprüft. Das Stufenmodell sieht vor, dass die zweite Stufe bei vorkommenden Differenzen wiederholt wird (84).

Für die Validierung werden Histogramme und Boxplots über die kategorischen Attribute untersucht. Auch eine Deskription mit den statistischen Kenngrößen über die numerischen Daten in den *Pivot*-Tabellen[26] werden für die Übersicht bereitgestellt.

4. Generierung planen

Nach einer erfolgreichen Validierung des Datensatzes wird nach McLachlan ein geeignetes Verfahren für die Synthese gewählt bzw. erstellt. Er betont, dass die Daten an die Bedürfnisse der Methode für die Synthese angepasst werden müssen und diese unterschiedlich sein können. Dies stehe im direkten Zusammenhang mit der Qualität des Ergebnisses (84).

Die Auswahl des Tools für die Generierung der synthetischen Daten für diese Arbeit wird mit Hilfe einer Nutzwertanalyse getroffen. Die Wahl wird aus vier Anwendungen getroffen, welche in der Literaturrecherche ermittelt wurden. Die Planung der Nutzwertanalyse wird in Kapitel 3.3 beschrieben und das Ergebnis der Nutzwertanalyse wird in Kapitel 4.1 präsentiert.

[26] Eine Pivot-Tabelle ist eine spezielle Tabellenart, mit der sich in Tabellenform vorliegende Daten mit geringen Aufwand strukturieren, verdichten und analysieren lassen, ohne dass die Ausgangsdaten verändert werden (85).

5. Methode validieren

Ist für die Synthese eine eigene Methode entwickelt worden, sieht diese Stufe ihre Validierung vor. An dieser Stelle soll z. B. geprüft werden, ob das entwickelte Verfahren die gestellten Anforderungen erfüllen kann. Sollten an dieser Stelle Fehler oder Unvollkommenheiten auffallen, kann die Entwicklung aus der vorherigen Stufe überarbeitet werden (84).

Da das Tool für die Synthese in dieser Arbeit nicht selbst entwickelt worden ist und anhand der erstellten Anforderungen ermittelt wird, können an dieser Stelle nur möglicherweise bei der Synthese auftretende Probleme diskutiert und erörtert werden und gegebenenfalls die verwendeten Parameter für die Synthese nachadjustiert werden. Eine Anpassung der Methode ist nicht vorgesehen.

6. Synthetische Daten generieren

In der sechsten Stufe kann mit dem validierten Realdatensatz und der ausgewählten Methode die Synthese vorgenommen werden. An dieser Stelle werden die Realdaten, Bedingungen, Anforderungen und Einschränkungen an den Generierungsalgorithmus übergeben (84).

Die Synthese erfolgt mit der in Kapitel 3.4.5 beschriebenen Anwendung. Für die Synthese müssen die verschiedenen Parameter des Tools angepasst werden. Das bedeutet, diese Stufe fordert ein Vorgehen, bei denen Schritte optimiert und wiederholt werden müssen, um ein optimales Ergebnis zu erreichen.

7. Ausgangsdaten validieren

In der letzten Stufe werden die aus der Synthese resultierenden Daten validiert. McLachlan schreibt , dass hierfür sowohl die Struktur des Datensatzes als auch die statistischen Eigenschaften der Ausgangsdaten überprüft werden sollen (84).

In dieser Arbeit wird der synthetische DS neben der formalen Konsistenz, z. B. einer identischen Datenstruktur mit identischen Datentypen, und den statistischen Eigenschaften auch auf die Datenqualität und den Ausschluss einer Re-Identifikation bzw. die Privatheitswahrung hin untersucht. Die hierfür verwendeten Methoden werden in Kapitel 2.2.3 für die Prüfung der Datenqualität und in Kapitel 2.3 für die Einhaltung des Datenschutzes beschrieben.

3.2 **Daten für die Synthese**

Für die Generierung der synthetischen Daten ist ein öffentlich zugänglicher DS gewählt worden. In den folgenden Abschnitten werden nähere Informationen zum Umfang, zu der Bereitstellung und der Extraktion der verwendeten Daten gegeben.

3.2.1 **MIMIC-III-Datensatz**

Wie in Kapitel 1.4.2 erwähnt, wird für die Synthese der *Medical Information Mart for Intensive Care* (MIMIC)-III Datensatz verwendet. Die Datenbank beruht auf insgesamt 26 Tabellen, in denen neben den Informationen zu Laborparametern und Vitalparametern auch Informationen, wie z. B. Einweisungsgrund, Religion, Versicherung, ethnische Herkunft, erfasste Diagnosen und durchgeführte Prozeduren zu den dokumentierten Patienten enthalten sind (86).

Der Umfang des Datensatzes liegt bei 46.520 anonymisierten Personen, denen 58.976 Krankenhausaufenthalte und 14.567 *ICD-9*-Diagnosen zugeordnet sind. Zum Teil konnte für die Behandlungsfälle ein Datenpunkt pro Minute gemessen werden, sodass der DS mehrere Millionen (ca. 315 Millionen) gemessene Vital- und Laborparameter für die Nutzung wissenschaftlicher Analysen zur Verfügung stellt (35).

3.2.2 **Bereitstellung der MIMIC-III-Daten**

Der MIMIC-III-Datensatz ist über die *Physionet*[27]-Plattform abrufbar. Die Daten werden als CSV-Dateien mit den erforderlichen Skripten zum Import in ein Datenbanksystem bereitgestellt, wenn der Zugang zu den Daten erlangt werden kann. Interessierte Forschende können die Daten erhalten, indem sie online einen anerkannten Kurs zum Schutz menschlicher Forschungsteilnehmer nach Anforderungen der *Health Insurance Portability and Accountability Act* (HIPAA) absolvieren und einer Datennutzungsordnung zustimmen. Diese soll einen angemessenen Umgang mit den Daten bestätigen und untersagt eine Identifizierung einzelner im DS erfasster Personen (86). Unter Angabe des gewünschten Verwendungszwecks und einer Referenz der forschenden Einrichtung werden die Daten nach ein bis vier Wochen bereitgestellt.

3.2.3 **Datenextraktion**

Für die ausführlich erfassten Daten aus dem MIMIC-III-Datensatz finden sich in der Literatur Publikationen, in denen ein Teildatensatz für die vorgenommenen Analysen verwendet worden ist. Für die Analysen wurden die Teildatensätze dem Verwendungszweck angepasst und veröffentlicht (88–90).

Für die Erzeugung des synthetischen Datensatzes in dieser Arbeit ist die Datenextraktion von Tang et al. gewählt worden (90). Die Autoren haben den Teildatensatz für den Vergleich verschiedener Machine-Learning-Methoden erstellt (91). Für die Bereitstellung des Teildatensatzes ist auf der Plattform *GitHub* die Dokumentation mit Vorgehensbeschreibung und notwendigen Skripten veröffentlicht (90).

Tang et al. haben für die Auswahl des Teildatensatzes die 13 am häufigsten erfassten Laborwerte und die acht am häufigsten erfassten Vitalparameter berücksichtigt. Die teilweise minütlich erfassten Werte wurden zu stündlichen Werten zusammengefasst, sodass stündliche Zeitreihenvariablen vorhanden sind. Nicht ermittelte Parameter sind

[27] *Physionet* ist der Kurzname einer Forschungsressource, die darauf zielt, der Forschung und Lehre Zugriff auf biomedizinische Datensammlungen und Open Source-Projekten zu ermöglichen (87).

hinter den Zeitstempeln als fehlender Wert hinterlegt (91). In diesem Teildatensatz wurden aus den zusammengefassten Vital- und Laborparameter drei *Pivot*-Tabellen abgeleitet.

Für diese Arbeit ist die Extraktion des Datensatzes nach Tang et al. von der Arbeitsgruppe des ZLG ZDIN zur Verfügung gestellt worden. Diese Datenextraktion wurde ebenfalls für die Erarbeitung der Arbeitsergebnisse für das Review von Olatunji et al. verwendet (58). Einen ähnlichen DS zu verwenden, wurde in die Anforderungen aufgenommen, s. Tabelle 25: Anforderungstabelle.

Der Teildatensatz nach Tang et al. wird an die in der Anforderungsanalyse geforderten Attribute angepasst. Die Attribute in der angepassten Extraktion sind in einem Data Dictionary, s. Tabelle 26: Data Dictionary, beschrieben.

3.3 Auswahl des Tools

Um ein Tool aus den möglichen Vorschlägen auszuwählen, wird eine Nutzwertanalyse anhand der gesammelten Anforderungen an das Tool angefertigt. Eine Nutzwertanalyse ist ein Weg, objektiv zu bewerten, welche Lösungsmöglichkeit sich am besten eignet, um die gestellten Anforderungen zu erreichen (83).

Bei der Nutzwertanalyse wird wie folgt in sechs Schritten vorgegangen (83,92):

1. **Festlegung der Alternativen und Entscheidungsvarianten**

Für die Nutzwertanalyse werden die in Kapitel 2.4.2 beschriebenen Tools berücksichtigt. Die Vorauswahl der vier Tools für die Nutzwertanalyse ist im Anschluss an die Literaturrecherche aufgrund der Popularität und der vorhandenen Dokumentation auf verschiedenen Plattformen, wie z. B. GitHub, und in Publikationen getroffen worden. Die Auswahl der Tools ist in Tabelle 9 zur Übersicht aufgeführt.

Tabelle 9: Berücksichtigte Tools in der Nutzwertanalyse

Tool-ID	Name des Tools
1	DataSynthesizer
2	medGAN
3	SynthEHR
4	SynthPop

2. Definition der Bewertungskriterien

Für die Bewertungskriterien der Nutzwertanalyse werden die Anforderungen an das Tool (Punkt 6) aus der Anforderungsliste, s. Kapitel 2.5.4 und Anhang A - Anforderungsliste entnommen. Unter diesen Anforderungen sind drei Mussanforderungen (Anf. 6.2 (Open Source), Anf. 6.3 (kostenfreie kommerzielle Verwendbarkeit) und Anf. 6.5 (generische Nutzbarkeit/ fallbezogene Konfigurierbarkeit) definiert. Diese Anforderungen werden für die Nutzwertanalyse nicht berücksichtigt, da sie zwingend zu erfüllen sind. In Tabelle 10 sind die zu berücksichtigenden Kriterien für die Nutzwertanalyse aufgelistet.

Tabelle 10: Berücksichtigte Kriterien für die Nutzwertanalyse

Kriteriums-ID	Kriterium
I	Integrierbarkeit
A	Unbegrenzte Anzahl an Daten generieren
W	Betriebssystem: Windows
C	Eingangsdaten: CSV
B	Intuitive Bedienbarkeit
L	Mit verfügbaren Quellpakten lauffähig
K	Kontinuierliche Weiterentwicklung

3. Gewichtung der Bewertungskriterien

Für jedes Kriterium besteht eine spezifische Wichtigkeit für dessen Erfüllung. Deshalb erhält jedes Kriterium eine prozentuale Gewichtung, die aus der Bedeutsamkeit des Kriteriums abgeleitet ist. Insgesamt summieren sich alle Prozentsätze auf 100% (83).

Um die Gewichtung festzulegen, wird eine Gewichtungsmatrix, s. Tabelle 11, erstellt. Dies erfolgt, indem für alle Kriterien ein paarweiser Vergleich durchgeführt wird. Für jede Paarung wird ein Kriterium als das wichtigere definiert und im Kreuzungspunkt der Matrix eingetragen. Pro Zeile wird aufsummiert, wie oft das linksstehende Kriterium wichtiger als das obenstehende Kriterium ist. Anhand dieser Summe wird der Anteil der Gesamtsumme und damit der Prozentsatz für die Gewichtung festgelegt (93). Dabei werden auch Kreuzungspunkte eines Kriteriums mit sich selbst berücksichtigt, um eine 0 %-Gewichtung zu verhindern. Die Gewichtungsmatrix ist in Tabelle 11 abgebildet.

Tabelle 11: Gewichtungsmatrix für die Nutzwertanalyse

Kriterium wichtiger / als	**I**	**A**	**W**	**C**	**B**	**L**	**K**	Σ	Gewichtung (%)
I	I	I	I	C	B	L	I	4	14
A	I	A	W	A	B	L	A	3	11
W	I	W	W	W	B	L	K	3	11
C	C	A	W	C	B	L	K	2	7
B	B	B	B	B	B	L	B	6	21
L	L	L	L	L	L	L	L	7	25
K	I	A	K	K	B	L	K	3	11
Prüfsumme								**28**	**100**

4. Festlegung des Bewertungsmaßstabes

Für die Erfüllung der einzelnen Kriterien werden für jede Anwendung Punkte vergeben. Für eine ideale Erfüllung des Kriteriums werden vier Punkte vergeben und für ein unerfülltes Kriterium wird ein Punkt gegeben. In Tabelle 12 sind Erfüllungsgrade für jedes Kriterium definiert, um eine objektive und nachvollziehbare Bewertung zu erreichen.

Tabelle 12: Bewertungsmaßstab für die Nutzwertanalyse

Kriterium	4 Punkte	3 Punkte	2 Punkte	1 Punkt
I	Integrierbarkeit in alle Software-Systeme möglich	Integrierbarkeit in andere Software- Systeme möglich	Integrierbarkeit in andere Software-Systeme bedingt möglich	Keine Integration in andere Software-Systeme möglich
A	Generierung unbegrenzter Anzahl von Daten möglich	Generierung unbegrenzter Anzahl bedingt möglich	Generierung unbegrenzter Anzahl unter größerem Aufwand möglich	Anzahl der generierten Daten ist begrenzt
W	Als Portable-Version für alle Betriebssysteme verfügbar	Lauffähig unter MS Windows-Betriebssystem	Tool ist über eine virtuelle Maschine auf MS-Windows lauffähig	Anderes Betriebssystem als MS Windows notwendig
C	Weitere Eingangsformate neben CSV möglich	Eingangsformat CSV möglich	Eingangsformat CSV unter größerem Aufwand möglich	CSV-Dateien als Eingangsformat nicht möglich
B	Automatisierte Bedienung des Tools	Bedienung durch Eingabe gewünschter Parameter	Bedienung nach längerer Einarbeitung oder mit Hilfestellung	Keine intuitive Bedienbarkeit und keine Anleitung
L	Automatische Updates	Version mit aktuellen Quellpaketen verfügbar	Durch Aktualisierung des Quellcodes lauffähig	Nicht mit aktuell verfügbaren Quellpaketen lauffähig
K	Kontinuierliche Weiterentwicklung	Weiterentwicklung der Quellcodes erfolgt sporadisch	Eine gelegentliche Fehlerbehebung erfolgt	Eine Weiterentwicklung oder Fehlerbehebung erfolgt nicht

5. Bewertung der Alternativen

Um die am besten geeignete Lösung zu ermitteln, wird jede betrachtete Anwendung anhand der Erfüllungsgrade aus Tabelle 12 auf Erfüllung ausgewertet. Anschließend werden die vergebenen Punkte anhand der Gewichtungsmatrix, s. Tabelle 11, mit der jeweiligen Bedeutung des Kriteriums multipliziert.

6. Summierung und Auswahl

Durch das Summieren aller gewichteten Punkte je Lösung kann eine Rangliste erstellt werden, an deren Spitze diejenige Lösung steht, die mit den definierten Kriterien am besten übereinstimmt. Diese Auswahl wird dann für das weitere Vorgehen benutzt. Die höchste zu erreichende Punktzahl liegt bei 4,0 Punkten.

Das Ergebnis der Nutzwertanalyse wird Kapitel 4.1 präsentiert und erläutert.

3.4 Software für die Analyse und Synthese der Daten

Im folgenden Anschnitt werden die für die Arbeitsumgebung notwendigen Softwarepakete und deren Bereitstellung geschildert.

3.4.1 Ubuntu LTS for Windows

Die in Kapitel 2.4.2 vorgestellten Tools sind in der Programmiersprache *Python 3* implementiert. Für die Installation von *Python,* der notwendigen Bibliotheken und des verwendeten Tools für die Synthese werden das *Windows*-Subsystem für *Linux* (WSL) und das virtuelle Betriebssystem *Ubuntu 20.04 LTS on Windows* für die Kommandoeingabe verwendet (94). Dieses ist im Microsoft (MS)-Store kostenlos verfügbar.
Viele Bibliotheken können über ein einzeiliges Kommando installiert werden. Die *Python*-Bibliothek *Matplotlib* z. B. kann mithilfe der *PyPi*-Bibliothek mit folgenden Befehl installiert werden:

```
python -m pip install -U matplotlib
```

Auch die Befehlseingabe zum Start der verwendeten Programme erfolgt über die Kommandozeile.

Alternativ lässt sich *Python 3* für *Windows* über *Anaconda 3*[28] inklusive der benötigten Packages, s. unter anderen Tabelle 13, und *Jupyter Notebooks*, s. Kapitel 3.4.4 bereitstellen.

3.4.2 Python 3

Für die Datenbeschreibung, Datenanalyse und die Bereitstellung des für die Synthese verwendeten Datensatzes wird die Programmiersprache *Python 3* verwendet. Für die Durchführung werden verschiedene Bibliotheken benötigt, auf welche auch das ausgewählte Tool *DataSynthesizer* zugreift.

[28] *Anaconda 3* ist eine Sammlung verschiedener Tools und Bibliotheken, z. B. für die Entwicklung von *Python*-Projekten (95).

Python 3 ist die aktuelle Version der Programmiersprache, welche erstmals im Dezember 2008 veröffentlicht wurde (71). Für die Erstellung dieser Arbeit wird die Version *Python 3.9.5* verwendet.

3.4.3 **Python Bibliotheken**

Python-Bibliotheken, die für die Bearbeitung, Beschreibung und Analyse der Daten notwendig sind, müssen zusätzlich installiert werden. In Tabelle 13 sind die verwendeten Bibliotheken mit ihren wichtigsten Funktionen kurz beschrieben. Die Bibliotheken sind teilweise voneinander abhängig, sodass sie gemeinsam installiert werden, um lauffähig zu sein.

Tabelle 13: Verwendete Python Bibliotheken

Bibliothek	**Beschreibung**	**Beispiele für Funktionen**
***Pip* (96)**	Paketverwaltungsprogramm	Installation von Packages und Bibliotheken
***Pandas* (97)**	Modul zur Datenanalyse	Einlesen, bearbeiten, manipulieren, transformieren, aggregieren, bereinigen und verbinden tabellarischer Daten
***NumPy* (98)**	Open Source Bibliothek, die ermöglicht mehrdimensionale Arrays, Matrizen und Vektoren zu bearbeiten	Arrays erstellen und bearbeiten
***Seaborn* (99)**	Open Source Bibliothek zur Visualisierung von Daten in kurzen Kommandos basierend auf Matplotlib (99)	Bivariater Verteilungsdiagramme Regressionsdiagramme Balkendiagramme
***Matplotlib* (100)**	Diagrammbibliothek zur mathematischen Darstellung von den Daten	Histogramme, Streudiagramme, Balkendiagramme,
SciKit-Learn (101)	Softwarebibliothek für Maschinelles Lernen	Clustering-, Regressions- und Klassifizierungsalgorithmen, Cross Validation, Zusammenfassung und Darstellung von Daten

3.4.4 **Jupyter Notebook**

Für die Erstellung des *Python*-Quellcodes der Datenanalyse und Statistiken wird die Software *Jupyter Notebook* verwendet.

Das *Jupyter Notebook* wird als Open Source für die Erstellung von Datenanalysen und wissenschaftlichen Berechnungen von dem gemeinnützigen Projekt *Jupyter* zur Verfügung gestellt und über *GitHub* weiterentwickelt. Das *Jupyter*-Notebook kann neben Python auch für die Programmiersprachen *R* und *Julia* verwendet werden. Für die Datenanalysen und die Erstellung der Statistiken wird die Software *Jupyter Notebook* verwendet (102).

Jupyter Notebook lässt sich über die *Ubuntu 20.04 LTS on Windows*-Kommandozeile installieren und anschließend mit folgenden Befehl starten:

```
jupyter notebook
```

3.4.5 ***DataSynthesizer* – Bereitstellung und Konfiguration**

Anhand der Nutzwertanalyse, s. Kapitel 4.1, ist für die Synthese der Daten das Tool *DataSynthesizer* ausgewählt worden. Die Eigenschaften dieses Tools wurden bereits in Absatz 2.4.2 beschrieben.

Die Bereitstellung der Weboberfläche der Anwendung *DataSynthesizer* erfolgt über einen Befehl in die *Ubuntu*-Kommandozeile und wird über ein *web User Interface* (UI), einer Webschnittstelle, ausgeführt. Die Befehle für die Bereitstellung von *DataSynthesizer* sind auf *GitHub* dokumentiert (73).

Das Programm teilt sich in drei Bereiche, den *Data Describer*, den *Data Generator* und den *Model Inspector* auf. In Abbildung 4 ist die Systemarchitektur und der damit zusammenhängende Ablauf der Synthese schematisch dargestellt.

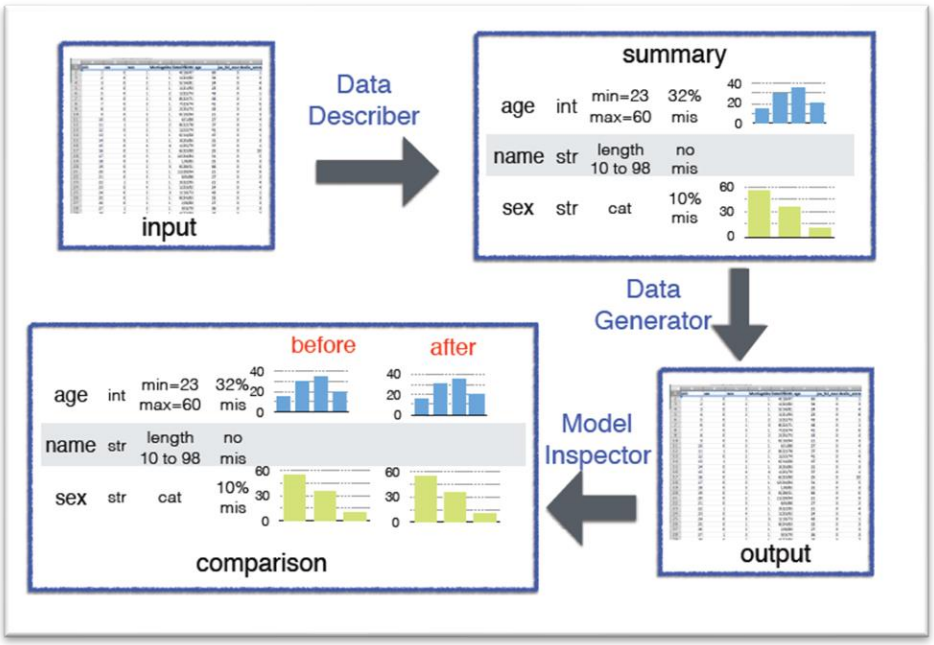


Abbildung 4: Systemarchitektur der Anwendung DataSynthesizer (74)

Zunächst werden die im CSV-Format importierten Realdaten von dem *Data Describer* verarbeitet und angezeigt. Für die Beschreibung der Daten analysiert der *Data Describer* für jedes Attribut den Datentyp und ob z. B., ein *Unique*, ein einmaliger Wert, ein numerischer Wert oder ein kategorisches Attribut vorliegt. Außerdem werden vom *Data Describer* die bei der Ausgabe angezeigten Statistiken, Histogramme zu jedem Attribut und eine *Heatmap* mit den paarweisen Korrelationen, erstellt. Dies erfolgt für einen Vergleich zunächst für die Realdaten und nach der Synthese für die synthetischen Daten (74).

In einem Formular, s. Abbildung 5, werden Voreinstellungen angezeigt, welche an den jeweiligen Bedarf angepasst werden können. In Tabelle 14 werden die Parameter zur Konfiguration der Synthese zusammengefasst und erläutert, welche Wertangaben empfohlen sind.

Für die Synthese wird der Modus verwendet, bei dem die Korrelation zwischen den Attributen berücksichtigt wird *(Correlated attribute mode)*, um die statistischen Zusammenhänge zwischen den Attributen im synthetischen DS zu erhalten.

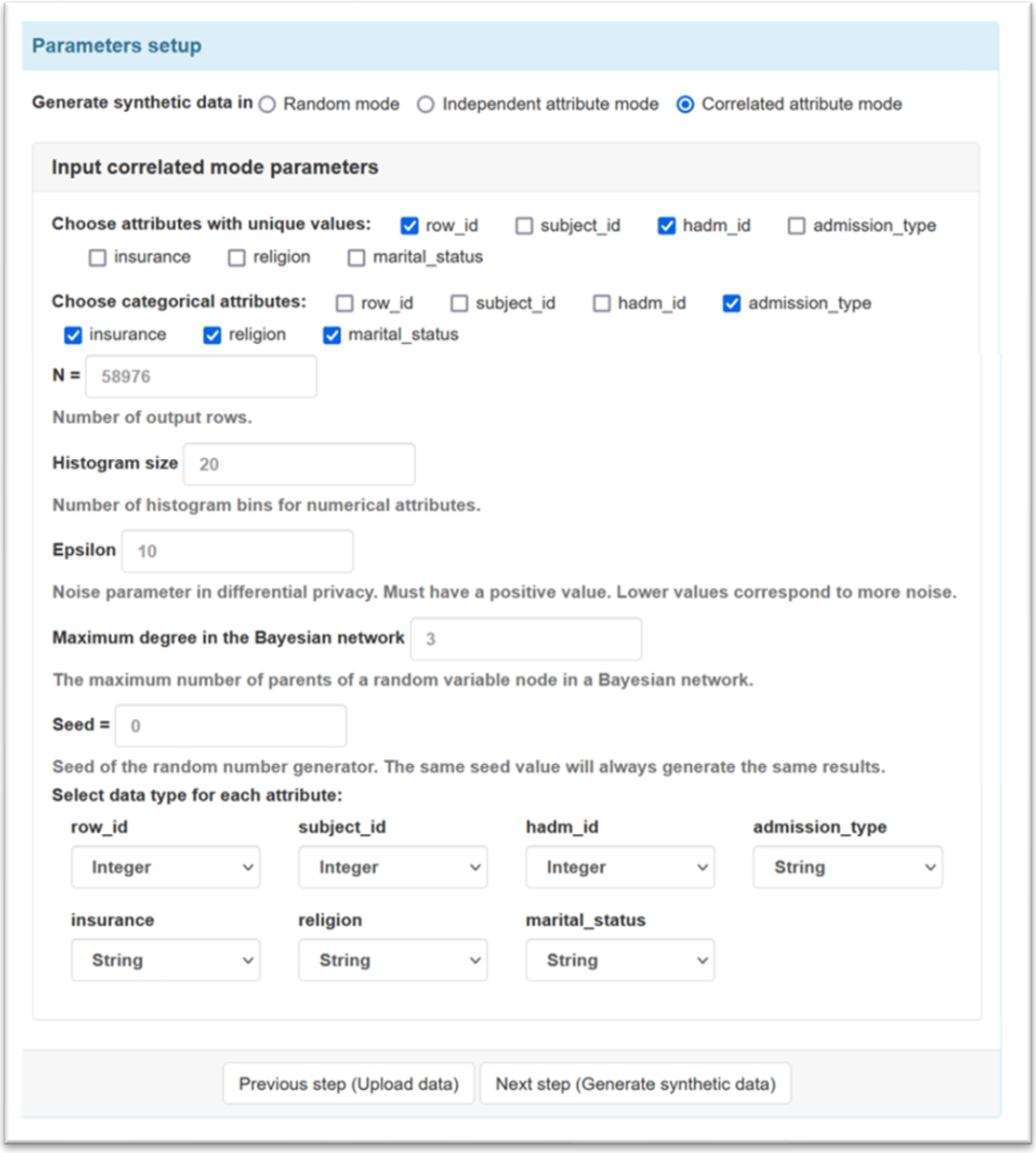


Abbildung 5: Parameterkonfiguration von DataSynthesizer

Der *Data Generator* erzeugt dann mit der durch den *Data Describer* erstellten Deskription der Realdaten ein Bayes'sches Netz und daraus wiederum den synthetischen DS. Dies erfolgt unter der Berücksichtigung der im Formular (Abbildung 5) angegebenen Parameter. In dem Formular können u.a. der Modus und die Parameter für die Synthese eingestellt werden.

Der *Model Inspector* bietet verschiedene Möglichkeiten, die erzeugten synthetischen Daten mit den Realdaten direkt zu vergleichen. Zunächst werden nach der Synthese

beide Datensätze mit den ersten fünf Objekten angezeigt. Außerdem generiert der Data Inspector für alle Attribute beider Datensätze Histogramme für den Vergleich der statistischen Verteilung. Als dritte Möglichkeit für den Vergleich wird eine *Heatmap* mit den paarweisen Korrelationen aller Attribute erzeugt (74).

Tabelle 14: Konfigurationsparameter von DataSynthesizer

Parameter	Beschreibung	Eigenschaften/ Einstellungsempfehlung
Generate synthetic data in	Auswahl, ob Attribute beliebig, unabhängig oder in Korrelation synthetisiert werden	Je nach Datenumfang und Anforderungen an die Daten anzupassen.
Choose attributes with unique values	Auswahl der Schlüssel-Attribute (Unique-Werte)	Werden von dem Programm ermittelt, sollten jedoch geprüft und ggf. angepasst werden.
Choose categorical attributes	Auswahl der kategorischen Attribute	Werden vom Programm ermittelt und können bei Bedarf angepasst werden.
N	Anzahl der synthetischen Objekte	Voreinstellung ist die Anzahl aus dem Originaldatensatz, kann aber beliebig gewählt werden.
Histogram size	Balkenanzahl des beschreibenden Histogramms	Anzahl ist nach Wunsch zu wählen.
Epsilon	Höhe der zu erreichenden ε-DP	Bei 0 generiert das Tool Daten ohne Verrauschen, sonst gilt: Je höher ε ist, desto geringer ist die Verrauschung (103).
Maximum degree in the Bayesian network	Maximale Anzahl der Elternknoten eines Knotens im BN	Im *Correlated attribute mode* ist der Default-Wert von 3 Elternknoten festgelegt. Mehr Knoten bedeuten eine längere Berechnungszeit. Dafür werden präzisere Korrelationen berechnet, da mehr Elternkoten berücksichtigt werden.

Parameter	Beschreibung	Eigenschaften/ Einstellungsempfehlung
Seed	Startwert eines Pseudo-Zufallsgenerators	Mit einem anderen Seed können mit dem gleichen BN reproduzierbar unterschiedliche Datensätze erzeugt werden.
Select data type for each attribute	Datentypen der Attribute	Vom Programm ermittelt und wenn erforderlich anpassbar.

Für die Erstellung des BN für die Erzeugung des synthetischen Datensatzes wird der *Correlated attribute mode* verwendet, um die statistischen Zusammenhänge zwischen den Attributen im synthetischen DS zu erhalten. Als Konfigurationsparameter wurden die empfohlenen Voreinstellungen verwendet. Das heißt es wurde ein synthetischer Datensatz mit der gleichen Anzahl an Objekten erzeugt, wobei kein Verrauschen der Daten vorgenommen wurde *(Ɛ = 0)* und die maximale Anzahl der Elternknoten beträgt dem voreingestellten Wert *(Maximum degree in the Bayesian network = 3)*. Als Startwert wurde auch der *Default*-Wert *(Seed = 0)* gewählt.

3.5 **Prüfung der Datenqualität**

Für die Beschreibung und den Vergleich der statistischen Eigenschaften und der Datenqualität werden für beide Datensätze tabellarische Beschreibungen über die statistischen Kenngrößen angefertigt und miteinander verglichen. Für die kategorialen Merkmale werden Häufigkeitstabellen erstellt. Für die Visualisierung der Dateneigenschaften werden *Heatmaps*, *Histogramme* und *Boxplots* von beiden Datensätzen angefertigt.

Für den Vergleich der statistischen Verteilung und die Ermittlung der Datenqualität werden ausgewählte Evaluationsmetriken, wie z. B. die PKD, der KS-Test , die KL-Divergenz und die *Log-Cluster-Metrik*, welche in Kapitel 2.2.3. beschrieben wurden, genutzt. Dazu werden entweder vorgefertigte Funktionen aus *Python*-Bibliotheken verwendet oder auf Basis der mathematischen Formulierungen der Metriken eigene Funktionen geschrieben.

3.5.1 **Differenz zwischen *Cramérs V***

Ähnlich, wie die in Kapitel 2.2.3 beschriebene PKD mit dem Pearson Korrelationskoeffizienten für die numerischen Merkmale, wird auch die Differenz der Zusammenhänge zwischen kategorischen Merkmalen ermittelt. Um die paarweisen statistischen Zusammenhänge zwischen den kategorischen Merkmalen der Datensätze zu messen, wird das Maß *Cramérs V* verwendet und mit einer *Heatmap* visualisiert.

Das Zusammenhangsmaß *Cramérs V* basiert auf dem Koeffizienten *Chi-Quadrat* (χ^2)[29] und gibt Auskunft über den Zusammenhang nominalskalierter Merkmale. Das Maß lässt sich für Kreuztabellen anwenden, deren Merkmale mehr als zwei Ausprägungen enthalten.

Der Ausgangswert des *Cramérs V* liegt im Bereich zwischen 0 und 1. Je höher der Wert ist, desto stärker ist der Zusammenhang zwischen den getesteten Merkmalen (105,106).

3.5.2 **Vorgehen bei der *Log-Cluster Metrik***

Für die in Kapitel 2.2.3 beschriebene *Log-Cluster Metrik* werden die beiden Datensätze vor dem Zusammenführen über eine *Boolean*-Variable als synthetisch *(1)* bzw. real *(0)* gekennzeichnet. Vor der Durchführung der Clusteranalyse wird die Variable in einem zusätzlichen *Panda-Dataframe* mit dem zusammengesetzten Schlüssel der IDs *subject_id* und *hadm_id* für die spätere Identifizierung separiert.

Nach der Clusteranalyse werden die beiden *Pandas-Dataframes* zusammengeführt, um die Zuordnung zum realen oder synthetischen Datensatz wiederherzustellen. Mit Hilfe eines Histogramms lassen sich die Zuordnungen in die verschiedenen Cluster vergleichen. Für die Berechnung des U_C-*Wertes* wurde eine Funktion erstellt, die anhand der von Goncalves et al. beschriebenen Formel erstellt wurde (45).

[29] *Chi-Quadrat (χ^2)* ist ein Maß, das Auskunft über den Zusammenhang von zwei nominal– oder ordinalskalierten Variablen gibt (104).

3.6 Prüfung der Privatheit

In den folgenden Abschnitten werden die Funktionen für die Prüfung der Privatheit erläutert.

3.6.1 Ermittlung der *k-Anonymität* und *ℓ-Diversity*

In der Anforderungsanalyse, s. Kapitel 2.5, ist das Erreichen einer möglichst hohen *k-Anonymität* bzw. *ℓ-Diversity* als Mussanforderung definiert. Für die Analyse wird sowohl der reale DS als auch der synthetische DS geprüft.

Für die Ermittlung der *ℓ-Diversity* wird eine Funktion erstellt, mit der geprüft werden kann, welche *Quasi-Identifikatoren* für welches *sensible Attribut* welche *ℓ-Diversity* erreichen. So können verschiedene Szenarien geprüft werden, indem ein *sensibles Attribut* mit unterschiedlichen *Quasi-Identifikatoren* kombiniert wird. Zusätzlich wird für dieselben Szenarien die erreichte *k-Anonymität* bestimmt.

Als ein Beispiel kann angenommen werden, dass die Religionszugehörigkeit das *sensible Attribut* ist. Dann kann geprüft werden, wie hoch die *ℓ-Diversity* bzw. *k-Anonymität* in Bezug auf dieses Merkmal ist, wenn z. B. erkennbare äußere Merkmale als *Quasi-Identifikatoren* angenommen werden. Darunter fallen das Geschlecht, der Einweisungsgrund und die ethnische Herkunft, die einem Angreifer jeweils durch Beobachtung bekannt sein könnten.

3.6.2 Übereinstimmungen finden

Für die Überprüfung übereinstimmender Objekte zwischen den beiden Datensätzen wird eine weitere Funktion implementiert. Die Funktion ist an die Idee von Eman et al. angelehnt (57), bei der geprüft wird, ob in den Stichproben Übereinstimmungen vorliegen.

Die erstellte Funktion prüft, wie viele Objekte im synthetischen DS mit Objekten aus dem realen DS identische Merkmale nachweisen. Dies erfolgt für eine unterschiedliche An-

zahl von Merkmalen. Um eine Kontrollgruppe zu erstellen, wird nach dem gleichen Vorgehen der reale DS gegeneinander auf identische Objekte geprüft. Das bedeutet in der Kontrollgruppe gibt es für jedes Objekt mindestens eine *Eins-zu-Eins*-Übereinstimmung. Eine schematische Übersicht über die Erstellung der beiden Vergleichsgruppen ist in Abbildung 6 zu sehen.

Aus diesen Resultaten werden die *Risiken* und *Wahrscheinlichkeiten (Odds)* berechnet, ob ein Objekt mit identischen Merkmalen in beiden Datensätzen vorliegt. Dann werden im Vergleich der beiden Gruppen das *Relative-Risiko*[30] und das *Chancen-Verhältnis (Odds-Ratio)*[31] für die Kontrollgruppe bestimmt, ob ein Objekt mit den identischen *Quasi-Identifikatoren* vorliegt. Ist das *Odds Ratio* größer als 1 (107), ist die Chance, dass eine Übereinstimmung in der Kontrollgruppe vorliegt größer, als für den Vergleich zwischen dem synthetischen und dem realen DS. Für das *Relative Risiko* gilt ebenfalls, dass bei einem Wert größer als 1 ein Risiko für die Kontrollgruppe vorliegt (107)

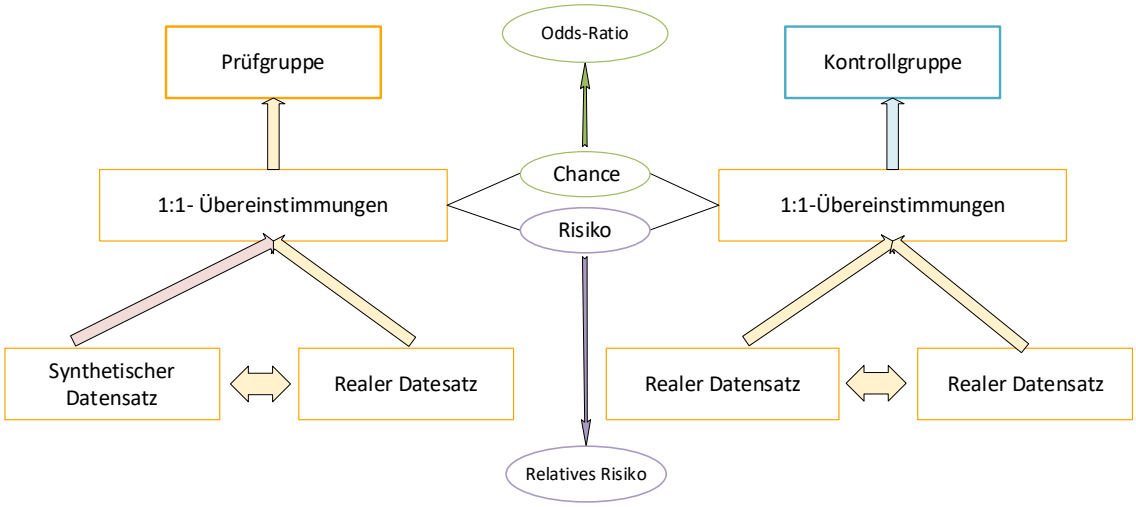


Abbildung 6: Erstellung der beiden Vergleichsgruppen (Gesamtdatensatz und Split).

[30] Das *Relative Risiko* wird als Verhältnis des Risikos bei den Exponierten zum Risiko bei den Nichtexponierten definiert (107).

[31] Das *Odds-Ratio* ist ein Vergleichsmaß, dass in der Epidemiologie eingesetzt wird und auf dem Odds basiert. Es lässt sich als Faktor interpretieren, wie hoch die Chance ist, dass ein Ereignis eintrifft. (107).

4 Ergebnisse

In diesem Kapitel werden die Ergebnisse dieser Arbeit präsentiert und erörtert. Zunächst wird die Auswahl des Tools anhand des Ergebnisses der Nutzwertanalyse begründet und dann das *Data Dictionary* vorgestellt. Außerdem werden die Besonderheiten beim Vorgehen der Synthese beschrieben und die Erstellung des letztendlich für die Synthese verwendeten Längsschnittes der Daten erörtert.

Dann werden die statistischen Beschreibungen der Daten des Originaldatensatzes und des synthetischen Datensatzes vorgestellt. Anschließend wird der Vergleich der Datensätze anhand der beschriebenen Methoden und die Analyse der Privatheitswahrung präsentiert.

4.1 Nutzwertanalyse für die Auswahl des Tools

In Tabelle 15 wird das Ergebnis der Nutzwertanalyse über das zu verwendende Tool für die Synthese vorgestellt. Aus der Nutzwertanalyse ging hervor, dass die Anwendung *DataSynthesizer* die in Tabelle 10 definierten Kriterien mit einer Punktzahl von 3,2 Punkten am ehesten erfüllt und deshalb für die Synthese verwendet wurde. Vor allem durch die intuitive Bedienung über die webbasierte Anwendungsoberfläche konnte *DataSynthesizer* in der Nutzwertanalyse überzeugen. In Tabelle 15 werden die bereits gewichteten Punkte, welche in Tabelle 11 gelistet wurden, angezeigt.

Der Versuch die beiden Tools *medGAN* und *SynthEHR* bereitzustellen scheiterte, weil diese mit dem verfügbaren Quellcode und den verwendeten Bibliotheken nicht mehr lauffähig sind. Die beiden vom Quellcode und Aufbau sehr ähnlichen Frameworks können ohne eine Modifizierung des Quellcodes und der Bibliotheken nicht bereitgestellt werden. Dies ist durch die Aktualisierungen der verwendeten Pakete begründet. Insbesondere ist

für die Implementierung der beiden Tools die Bibliothek *Tensorflow*[32] *1.x* verwendet worden, welche nicht mehr verfügbar ist. Auch der Versuch *Tensorflow 1.x* mit einer älteren Version als *Python 3.7* in einem *Virtual Enviroment*[33], zu installieren, ist nicht gelungen, weil die erste Version von *Tensorflow* nicht mehr verfügbar ist. Weil eine intuitive Bedienbarkeit nicht bewertet werden konnte, wurde die kleinstmögliche Punktzahl vergeben.

Die Bereitstellung der beiden anderen Tools *DataSynthesizer* und *SynthPop* hingegen ist mit Verwendung der aktuellen Versionen und Bibliotheken möglich gewesen.

Tabelle 15: Ergebnis der Nutzwertanalyse

	Data-Synthesizer	**medGAN**	**SynthEHR**	**SynthPop**
Kriterium		**Bewertung**		
Integrierbarkeit	0,4	0,4	0,4	0,4
Unbegrenzte Anzahl an Daten generieren	0,4	0,4	0,4	0,4
Betriebssystem: Windows	0,2	0,3	0,3	0,3
Eingangsdaten: CSV	0,3	0,3	0,3	0,3
Intuitive Bedienbarkeit	0,8	0,2	0,2	0,4
Mit verfügbaren Quellpakten lauffähig	0,8	0,3	0,3	0,8
Kontinuierliche Weiterentwicklung	0,2	0,1	0,1	0,2
Summe	**3,2**	**2,0**	**2,0**	**2,9**

[32] Tensorflow ist eine Open Source verfügbare Python-Kernbibliothek, welche die Entwicklung und das Training von Machine Learning-Modellen unterstützt (108).

[33] eine virtuelle Programmierumgebung für Python (109).

4.2 **Data Dictionary**

Um einen Überblick über die für die Synthese verwendete Extraktion aus dem *MIMIC-III*-DS zu erhalten, ist ein *Data Dictionary* erstellt worden. Ein Ausschnitt aus dem Data Dictionary ist in Tabelle 16 abgebildet, die gesamte Liste in Anhang B – Data Dictionary, zu finden.

Alle im extrahierten DS vorhandenen Attribute wurden mit ihren Eigenschaften, wie z. B. dem Datentyp, der Maßeinheit und den Ausprägungen der kategorischen Merkmale beschrieben. Insgesamt besteht der verwendete Teildatensatz aus fünf Tabellen. Zwei Tabellen, *admissions* und *icustay_detail*, enthalten kategorische Attribute, wie z. B. *Familienstand (marital_status)*, *Geschlecht (gender)* oder die *Krankenversicherung* (*insurance*).

Zusätzlich sind in drei *Pivot*-Tabellen 13 verschiedene Vital- und Laborparameter im Teildatensatz enthalten, die aus dem Gesamtdatensatz zusammengefasst worden sind, wie in Abschnitt 3.2.3. beschrieben wurde. Diese Parameter haben ihren Ursprung in der Tabelle *Lab-events* aus dem Gesamtdatensatz, in dessen Anordnung für jeden erhobenen Laborparameter ein Objekt mit Zeitstempel existiert. Zeitgleich erfasste Parameter wurden in den *Pivot*-Tabellen in einem Objekt zusammengefasst und für nicht erhobene Werte ein *Null*-Wert angegeben.

In der Tabelle *pivoted_bg* stehen diverse Laborparameter aus dem Blutbild, wie z. B. die *Sauerstoffsättigung (spo2)*, der *Kaliumwert (potassium)* oder der *Blutzuckerspiegel (glucose)*, die während der intensivmedizinischen Behandlung erfasst worden sind. Die Tabelle *pivoted_lab* erfasst Laborparameter, wie z. B. *Kreatininwerte (creatinine)*, *Albuminwerte (albumin)* oder *Blutharnstoffwerte (bun)*, die vor oder nach dem Aufenthalt auf der Intensivstation aufgenommen wurden. Vitalparameter aus den ersten 24 Stunden der Behandlung auf der Intensivstation, wie z. B. die *Herzfrequenz (heartrate)*, *Körpertemperatur (tempc)* und *systolischer/diastolischer Blutdruck (sysbp, diasp)* sind in der Tabelle *pivoted_vital* zusammengefasst (110). Bei den erfassten Attributen zwischen den drei *Pivot*-

Tabellen gibt es Doppelerfassungen, wie z. B. für die Attribute *Blutzucker* oder *Albuminwert*. Die Werte unterscheiden sich darin, dass die Messungen zu verschiedenen Zeitpunkten erfasst wurden bzw. unterschiedliche Zeitpunkte zusammenfassen.

Tabelle 16: Auszug aus dem Data Dictionary

Attribut	Datentyp	Ausprägungen/ Wertespezifikation	Beschreibung	PK	FK	NN	Index
Admissions							
row_id	Str		Identifikator für Zeile	x		x	x
subject_id	Str		Patienten-Identifikator		x		x
hadm_id	Int		Identifikator für Krankenhausaufenthalt		x	x	x
admission_type	Varchar(50)	"ELECTIVE" "EMERGENCY" "NEWBORN" "URGENT"	Art der Einweisung			x	
insurance	Varchar(50)	"Private" "Medicare" "Medicaid insurance" "Self Pay" "Government"	Versicherung			x	
(…)							
icustay_detail							
subject_id	Int		Patienten-identifikator		x	x	
(…)							
gender	varchar()	"M"/ "F"	Geschlecht			x	
Age	Int	Alter in Jahren	Patientenalter zum Erfassungszeitpunkt				
pivoted_bg							
hadm_id	Str		Verbindungsattribut (df_admissions)		x		
icustay_id	Str		Verbindungsattribut		x		
(…)							

Für die Generierung der synthetischen Daten sind die drei Tabellen *admissions*, *icustay_detail* und *pivoted_bg* verwendet worden. Diese sind für die Erstellung des benötigten Längsschnittes, s. Kapitel 4.4, des Datensatzes verwendet worden.

4.3 Besonderheiten bei der Synthese

Bei der Generierung der synthetischen Daten wurde nach dem Schema von McLachlan gearbeitet, das in Kapitel 3.1 beschrieben wurde. In diesem Schema sind Wiederholungen der Arbeitsstufen bei Unvollkommenheiten in den Validierungsstufen vorgesehen.

Um einen synthetischen Datensatz zu erzeugen, der über mehrere Tabellen in Relationen steht, sollten die Tabellen für die Synthese zunächst in eine Tabelle in erster Normalform, zusammengeführt werden.

Für den ersten Durchgang der Synthese wurden für die Eingangsdaten zunächst die drei Tabellen *admissions*, *icustay_detail* und *pivoted_bg* über die *subject_id,* die *hadm_id* bzw. die *icustay_id* über einen Join verbunden. Das Resultat war ein Datensatz in der ersten Normalform, bei dem aufgrund von Mehrfacherfassungen in der zusammengefassten Tabelle die gleichen Identifikationsnummern (IDs) für *subject_id*, *hadm_id* und *icustay_id* vorlagen.

In Tabelle 17 ist ein Ausschnitt der durch den Join erstellten Tabelle abgebildet. Die dargestellten Daten wurden jeweils demselben Subjekt desselben Falles desselben Krankenhausaufenthaltes zu einem jeweils anderen Zeitpunkt zugeordnet.

Tabelle 17: Ausschnitt aus der JOIN-Tabelle

subject_id	hadm_id	icustay_id	insurance	(…)	gender	(…)	spo2	(…)	glucose
533	100009	253656	Private	(…)	M	(…)	*NULL*	(…)	162
(…)	(…)	(…)	(…)	(…)	(…)	(…)	(…)	(…)	(…)
533	100009	253656	Private	(…)	M	(…)	100	(…)	148
677	100044	289655	Government	(…)	M	(…)	*NULL*	(…)	*NULL*
677	100044	289655	Government	(…)	M	(…)	*100*	(…)	*NULL*
1569	100045	260971	Medicare	(…)	F	(…)	97	(…)	*NULL*
7174	100062	215932	Private	(…)	F	(…)	*NULL*	(…)	*125*
7174	100062	215932	Private	(…)	F	(…)	*NULL*	(…)	*NULL*
7174	100062	215932	Private	(…)	F	(…)	*98*	(…)	*123*
3365	100103	200434	Medicare	(…)	F	(…)	*NULL*	(…)	129
(…)	(…)	(…)	(…)	(…)	(…)	(…)	(…)	(…)	(…)
3365	100103	200434	Medicare	(…)	F	(…)	*NULL*	(…)	170

Der *DataSynthesizer* hat bei der Synthese nicht erkannt, dass die Objekte mit identischen IDs dem gleichen Subjekt zugeordnet wurden. So hat die Anwendung bei der Synthese für

jedes im Datensatz vorhandene Objekt ein individuelles synthetisches Objekt mit einer neuen synthetischen ID als *Unique*-Wert erstellt. Würde der synthetische Datensatz im nächsten Schritt in das ursprüngliche Datenmodell zurückgeführt werden, lägen daher einzelne Objekte ohne Folgeaufenthalte vor.

4.4 Erstellung des Längsschnittes als Eingangsdaten

Eine den Anforderungen entsprechende Zusammenfassung der Tabellen über simple *Joins* in Structured Query Language (SQL) war nicht möglich, wie in Kapitel 4.3 näher erläutert wurde. Deswegen wurden die Daten der drei Tabellen *admissions, icustay_detail* und *bg_pivoted* für die Synthese im nächsten Versuch in einen Longitudinalschnitt, Synonym Längsschnitt[34], transformiert. Alle Informationen und Verlaufsdaten zu einem Patienten wurden als ein Objekt (in eine Zeile) in einer Tabelle zusammengefasst.

Der Längsschnitt wurde mit der *Python*-Bibliothek *Pandas* erstellt, indem zunächst die zusammengehörenden Objekte anhand der Identifikatoren *subject_id, hadm_id* und *icustay_id* mit der Funktion *merge()* zu einem *Inner Join* zusammengefasst. Erst wurden die Tabellen *icustay_detail* und *pivoted_bg* zusammengeführt und dann wurde die zusammengefasste Tabelle mit der Tabelle *admissions* verbunden. Objekte, denen keine Einträge in der Partnertabelle zugeordnet waren, entfielen an dieser Stelle durch den *Inner Join*.

Im nächsten Schritt wurden die in mehreren Objekten erfassten Krankenhausaufenthalte über die *subject_id* identifiziert, was in der Liste *multikeys* zusammengefasst wurde. Objekte bzw. Krankenhausaufenthalte, dessen Anzahl größer als vier *(multikey[obj] >= 4)* war, sind an dieser Stelle entfallen, um zu verhindern, dass der Längsschnitt zu viele Aufenthalte erfasst, in denen dann enorm viele Fehlwerte entstehen. Die Fehlwerte würden für jedes Objekt mit der steigenden Anzahl von Krankenhausaufenthalten zunehmen, da

[34] Solche Datensätze werden für die Durchführung von Längsschnittstudien (Longitundinalstudien) , z.B. in de Epidemiologie, verwendet und erfassen einen zeitlichen Verlauf mit wiederholten Messungen (111).

nicht für jedes Subjekt gleichviele Krankenhausaufenthalte vorhanden sind. Dies hätte zufolge, dass vom BN bei wenigen vorhandenen Werten keine Zusammenhänge erkannt werden und die Streuungen zwischen den Werten zunehmen .

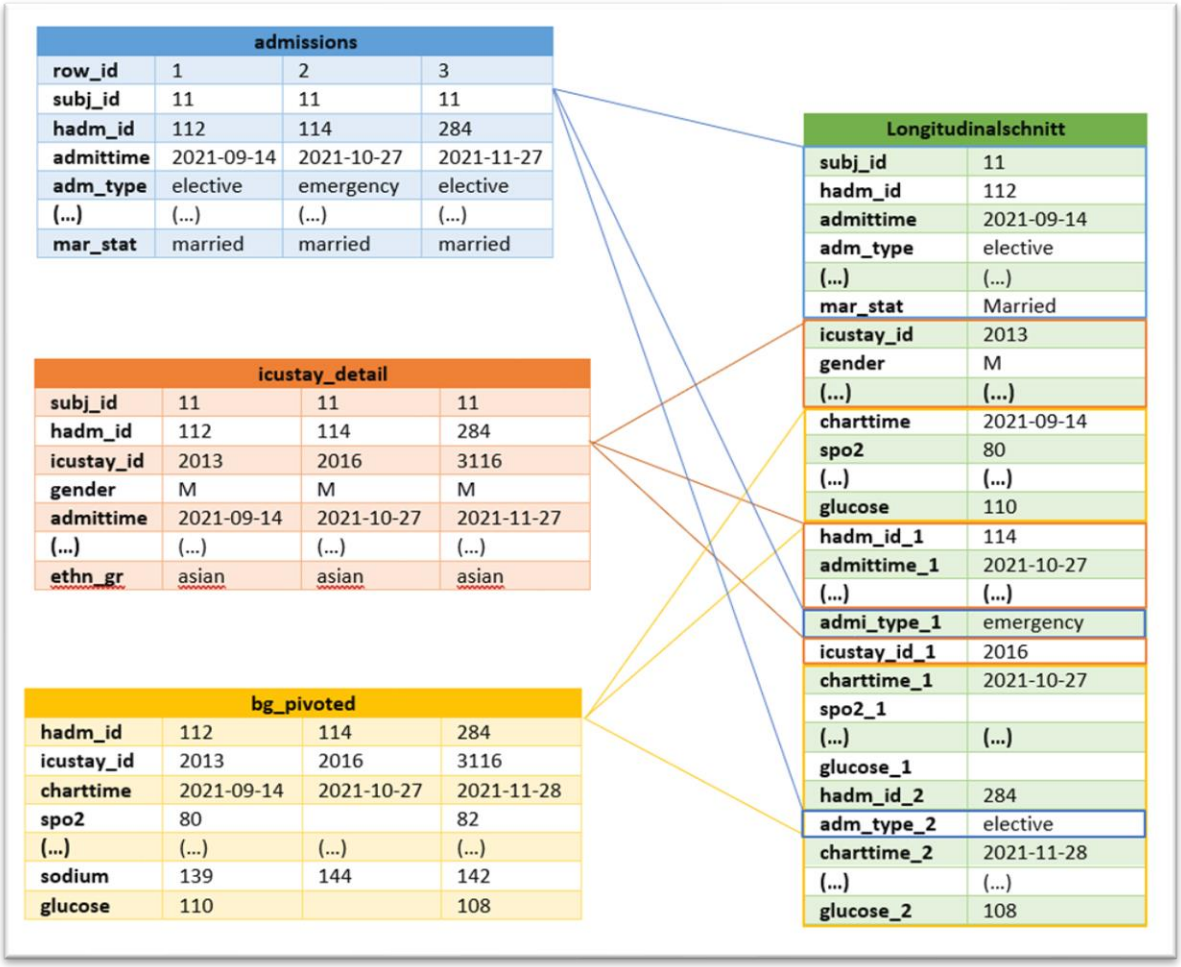


Abbildung 7: Erstellung des Longitudinalschnittes

Anschließend wurden die Folgewerte, z. B. zu verschiedenen Aufnahmen über die Schlüssel in chronologischer Reihenfolge aneinandergereiht. Doppelt erfasste Merkmale, welche sich nicht verändern, wie z. B die *Ethnische Zugehörigkeit (ethnicity_grouped)* wurden im Längsschnitt nur einmalig aufgeführt. Informationen, die über mehrere Tabellen doppelt erfasst wurden, wie z. B. das Erfassungsdatum *(admittime)* wurden im Längsschnitt auch nur einmalig aufgeführt. In Abbildung 7 wird anhand eines Objektes dargestellt, wie die Daten für die Synthese zusammengefasst wurden.

Das Ziel bei der Erstellung des Längsschnittes war es, jeden in der Tabelle *admissions* festgehaltenen Fall mit mindestens einem und maximal vier in der Tabelle *icustay_detail* erfassten Intensivstationsaufenthalten zu kombinieren. Dem Aufenthalt aus *icustay_detail* sollte eine Messreihe in der Tabelle *bg_pivoted* erfassten Blutlaborparameter zugeordnet werden können.

Da die Tabelle *icustay_detail* über die *icustay_id* auch auf die anderen beiden *Pivot*-Tabellen verweist, wurden durch den *Inner Join* nur Aufenthalte berücksichtigt, zu denen es Blutlaborparameter aus der Tabelle *pivoted_bg* gibt.

Um den Längsschnitt erzeugen zu können, wurden vorab für die Laborparameter in der Tabelle *pivoted_bg* je Krankenhausaufenthalt *(icustay)* die Mittelwerte aus allen zum Aufenthalt erfassten Laborparametern erzeugt (Mittelwerte der Zeitreihe). Dadurch sollte sowohl verhindert werden, dass der Längsschnitt durch mehrere Messreihen zu viele Attribute erhält und die Menge der Fehlwerte ansteigt. Dies ist vergleichbar mit der Reduktion der Krankenhausaufenthalte der Tabelle *icustay_id*. So ist zwischen den Tabellen *icustay_detail* und *pivoted_bg* aus einer *1:n*-Beziehung eine *1:1*-Beziehung entstanden. Insgesamt reduzierte sich der Datenumfang des Längsschnittes auf 26.063 Objekte, während die Tabelle *admissions*, die für den Längsschnitt verwendet wurde, 58.976 Objekte hatte.

4.5 **Vorstellung des realen Datensatzes**

In den folgenden Abschnitten werden zunächst die vorgenommene Validierung der Daten und anschließend die statistische Beschreibung des für die Synthese verwendeten Datensatzes erörtert. Für die Beschreibung der Verteilung der Daten wurden sowohl Tabellen als auch Grafiken angefertigt. Bei der Beschreibung der Daten wird zwischen numerischen Merkmalen, welche sich durch beliebig viele Zahlenausprägungen, und kategorischen Merkmalen, mit einer begrenzten Anzahl an Ausprägungen, unterschieden.

4.5.1 **Datenvalidation**

Zunächst wurden die Vitalparameter der Tabelle *pivoted_bg* anhand des arithmetischen Mittelwertes zusammengefasst, um für jeden Krankenhausaufenthalt einen singulären

Wert anstatt eine Messreihe zu erhalten. Dieser Schritt ist für die Erstellung des Längsschnittes notwendig gewesen und hat die Anzahl der Fehlwerte reduziert. Um die Streuung der Daten zu verringern und die Plausibilität der Daten zu erhöhen, wurden zunächst für die numerischen Attribute die extremen Ausreißer aus dem DS durch einen *Null*-Wert ersetzt.

Für das Merkmal *Alter in Jahren (age)* wurden die Werte auf ganze Lebensjahre gerundet und Werte, die höher als 299 Jahre waren, durch einen *Null*-Wert ersetzt. Für diese Werte wurde angenommen, dass das Alter bei der Datenerfassung unbekannt gewesen ist, weil das sonstige Maximum bei 89 Jahren lag. Für das Merkmal *Sauerstoffsättigung (spo2)* wurden Werte ersetzt, die kleiner als 20 % und größer als 101 % waren. Werte, bei denen das *PaO^2/FiO^2-Ratio (pao2fio2)* größer als 2.100 war, sind auch durch *Null*-Werte ersetzt worden. Für den *Natriumwert (sodium)* wurden Werte ersetzt, die nicht zwischen 90 und 300 mmol/l lagen, *Kaliumwerte (potassium)* größer als 30 mmol/l wurden auch als unplausible Ausreißer behandelt. Erfasste *Blutzuckerwerte (glucose)*, deren Wert größer als 1.500 mg/dl war, wurden ebenfalls durch *Null*-Werte ersetzt. Für das Merkmal *Bicarbonat (bicarbonate)* wurden keine Werte als Ausreißer berücksichtigt.

Auch die kategorisch erfassten Merkmale wurden validiert, da teils doppelt angelegte Kategorien und fehlende Werte vorlagen. Für das kategorische Merkmal Familienstatus wurden *Null*-Werte durch den Wert *unknown* ersetzt. Objekten, denen für die Religionszugehörigkeit *(religion)* kein Wert *(Null)*, der Wert *unobtainable* oder *not specified* zugordnet war, wurden ebenfalls unter *unknown* zusammengefasst. Für die anderen kategorischen Merkmale lagen keine fehlenden Werte oder Kategorien mit gleicher Bedeutung vor.

4.5.2 **Deskriptive Statistik**

Nachdem der Längsschnitt über die drei Tabellen *admissions*, *icustay_detail* und *pivoted_bg* erstellt wurden, ist die Deskription erstellt worden. Jeder Folgeaufenthalt auf der Intensivstation (FA) wurde in der Beschreibung einzeln erfasst.

In Tabelle 18 wird die statistische Verteilung der numerischen Merkmale mit Anzahl der Häufigkeiten (N), arithmetischem Mittelwert (MW), Standardabweichung (Std), Minimum

(Min), unterem Quartil (25%Q), Median, oberem Quartil (75%Q) und Maximum (Max) beschrieben.

Tabelle 18: Deskriptive Statistik der numerischen Merkmale-Realdatensatz

Merkmal	FA	N	MW	Std	Min	25% Q.	Median	75%Q.	Max
Alter (age)	0	25.082	59,52	21,32	0,00	50,00	63,00	75,00	89,00
	1	3.218	64,96	14,57	0,00	55,00	67,00	77,00	89,00
	2	775	65,32	14,04	0,00	56,00	67,00	76,00	89,00
	3	181	64,87	13,89	25,00	55,00	67,00	76,00	88,00
O2-Sättigung (spo2)	0	22.243	97,58	3,67	22,00	96,80	98,50	99,67	100,00
	1	3.003	97,11	3,78	54,00	96,00	98,00	99,40	100,00
	2	708	96,88	3,93	63,50	95,83	98,00	99,40	100,00
	3	170	96,80	4,32	71,88	95,05	98,42	100,00	100,00
Pa O2-fi O2-Ratio (pao2fio2ratio)	0	20.646	281,37	117,80	20,00	198,00	269,75	353,97	1.542,50
	1	2.468	259,00	114,58	35,00	177,50	245,31	327,77	1.216,25
	2	569	250,73	110,98	33,00	166,36	242,00	317,70	652,00
	3	139	254,11	119,32	40,00	161,67	248,57	334,50	537,14
Bicarbonat -wert (bicarbonate)	0	1.064	23,05	5,10	5,00	20,00	23,00	26,00	44,00
	1	120	23,95	5,97	7,00	21,00	24,00	28,00	41,00
	2	21	23,21	7,77	7,00	20,00	23,00	26,00	42,00
	3	4	21,83	3,28	18,00	20,50	21,67	23,00	26,00
Kaliumwert (potassium)	0	16.143	4,21	0,64	0,70	3,80	4,20	4,57	10,51
	1	1.851	4,22	0,71	2,00	3,77	4,12	4,60	10,10
	2	353	4,29	0,76	2,50	3,80	4,17	4,60	8,20
	3	90	4,40	0,72	2,90	3,90	4,30	4,90	6,30
Natriumwert (sodium)	0	13.574	136,85	4,15	92,27	135,00	137,00	139,00	177,00
	1	1.357	136,23	4,92	110,00	133,40	136,33	139,00	161,00
	2	236	136,27	6,18	118,00	133,00	136,00	139,00	169,00
	3	60	136,73	5,52	123,00	133,73	137,00	140,00	152,00
Blutzucker (glucose)	0	15.641	149,49	57,76	17,00	120,82	137,00	160,67	1.046,00
	1	1.695	147,17	59,07	12,00	115,00	134,00	163,09	899,50
	2	307	148,33	81,07	29,00	107,55	129,00	165,00	898,00
	3	83	145,66	59,97	31,00	109,00	131,21	178,36	381,00

Die Anzahl der Häufigkeiten nimmt für jedes Merkmal mit jedem FA ab. Das kann damit begründet werden, dass die Summe der erhobenen Krankenhausaufenthalte für jede behandelte Person unterschiedlich sein kann und nicht für jeden Aufenthalt alle Parameter erhoben wurden, sodass im DS Fehlwerte enthalten sind.

Bei der Altersverteilung *(age)* liegt in der Deskription für den ersten Aufenthalt *(FA 0)*, mit 21,32 eine höhere Standardabweichung vor als für die weiteren Folgeaufenthalte. Dies

kann damit begründet werden, dass für den ersten Aufenthalt Neugeborene (0 Jahre) erfasst sind und in der Verteilung als Ausreißer gelten. Dies ist z. B. in Abbildung 8 in den Boxplots zu *age_0, age_1,* und *age_3* zu erkennen. Hier ist dem Histogrammen zu entnehmen, dass die Erfassung Neugeborener (0-jähriger) zu späteren Zeitpunkten seltener auftreten und dann nicht mehr vorkommen. In allen Histogrammen ist eine leicht linksschiefe Verteilung zu erkennen. Die unterschiedliche Balkenbreite in den Histogrammen lässt sich damit begründen, dass die Streuungen in den FA unterschiedlich stark sind, aber die Diagramme auf einer gleichen Skala der *x*-Achse mit der gleichen Anzahl Balken (*n=10*) angeordnet sind.

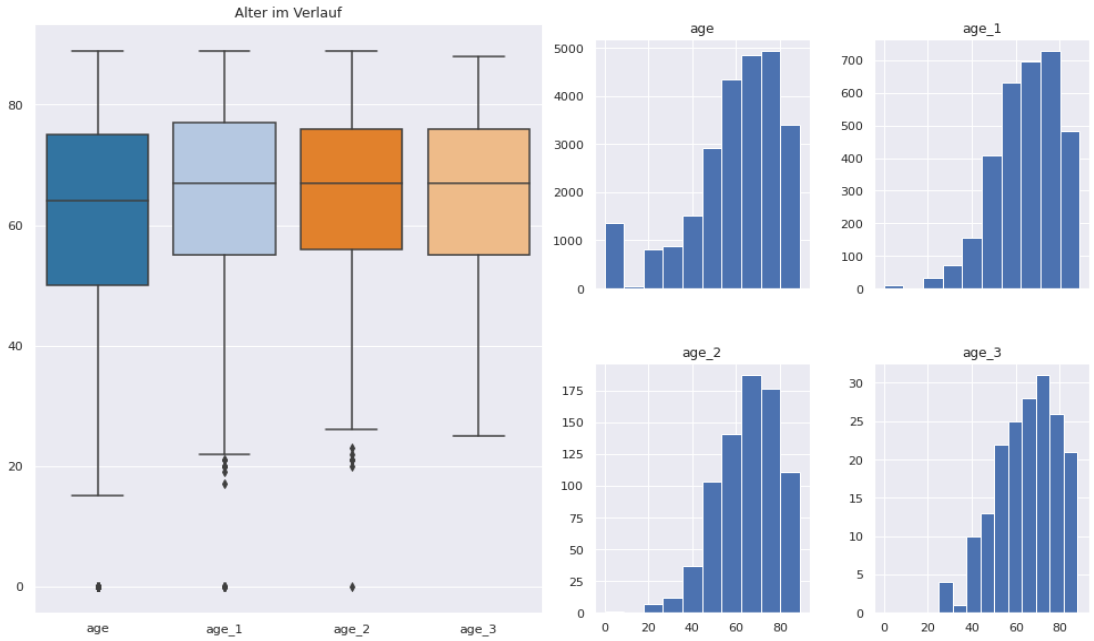


Abbildung 8: Boxplots und Histogramme zur Altersverteilung

Bei der Sauerstoffsättigung in Abbildung 9 liegen auch linksschiefe Verteilungen vor. Die Sauerstoffsättigung liegt für alle Aufenthalte in den überwiegenden Fällen bei 100 %. Demnach treten die Ausreißer bei den niedrigeren Werten auf.

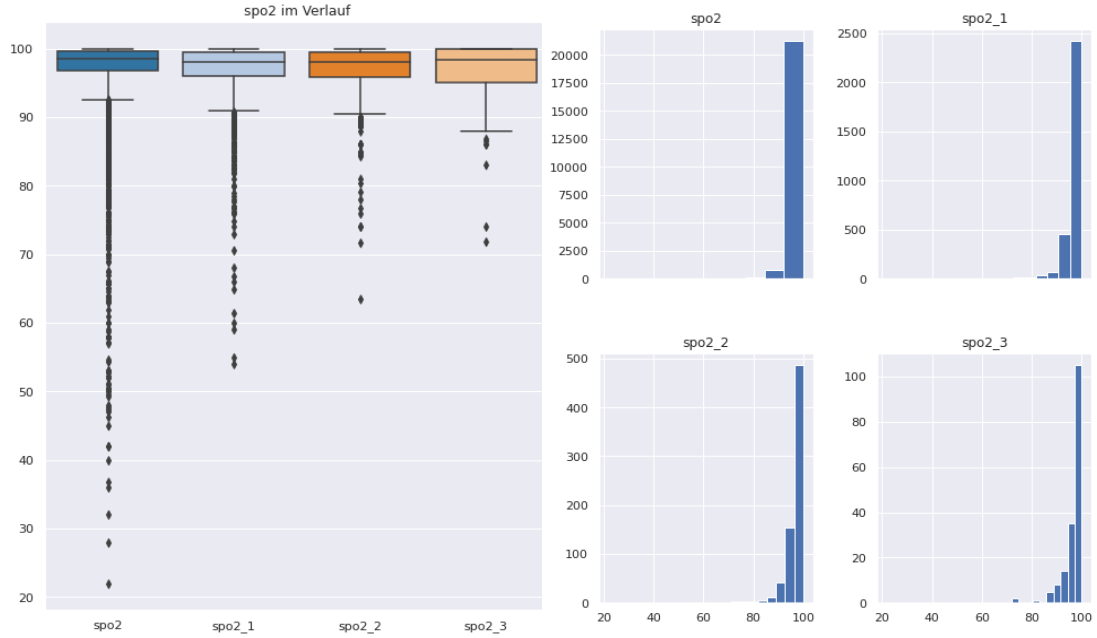


Abbildung 9: Boxplots und Histogramme zur Sauerstoffsättigung

Für die Datenverteilung zum *PaO^2/FiO^2-Ratio* in Abbildung 10 liegt eine rechtsschiefe Verteilung vor. Vor allem für den ersten Aufenthalt *(pao2fio2ratio_0)* gibt es viele Ausreißer.

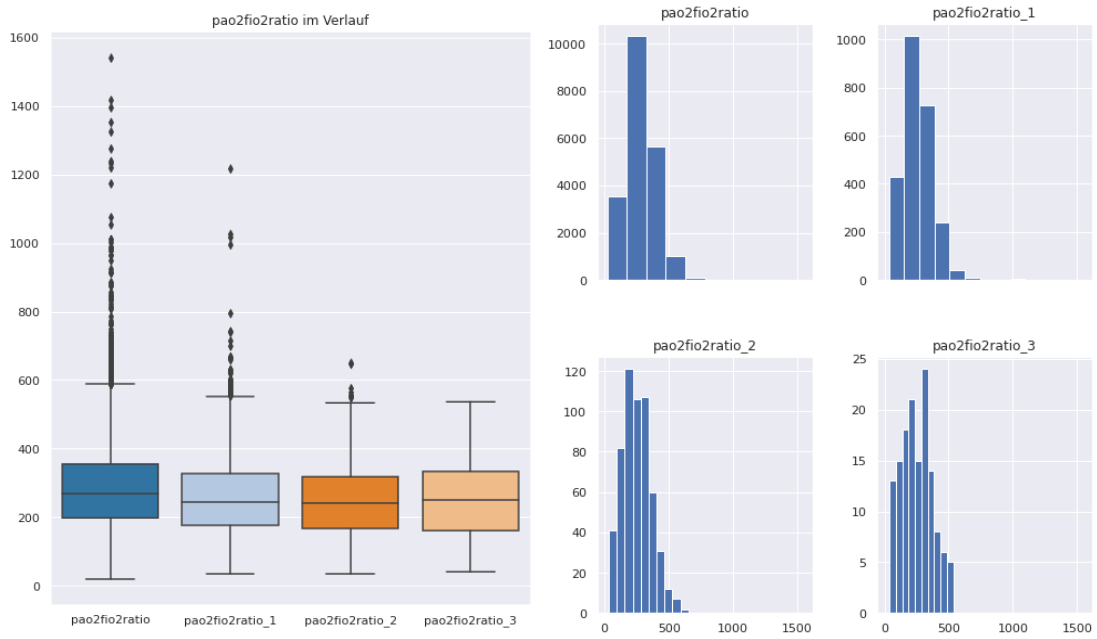


Abbildung 10: Boxplots und Histogramme zum PaO2/FiO2-Ratio

In Abbildung 11 ist die Datenverteilung zu den *Bicarbonatwerten* in allen Aufenthalten mit einer Normalverteilung zu sehen. Ausgenommen ist der Wert *bicarbonate_3 ,* in dessen Histogramm erkennbar ist, dass für diesen FA nur vier Werte vorliegen.

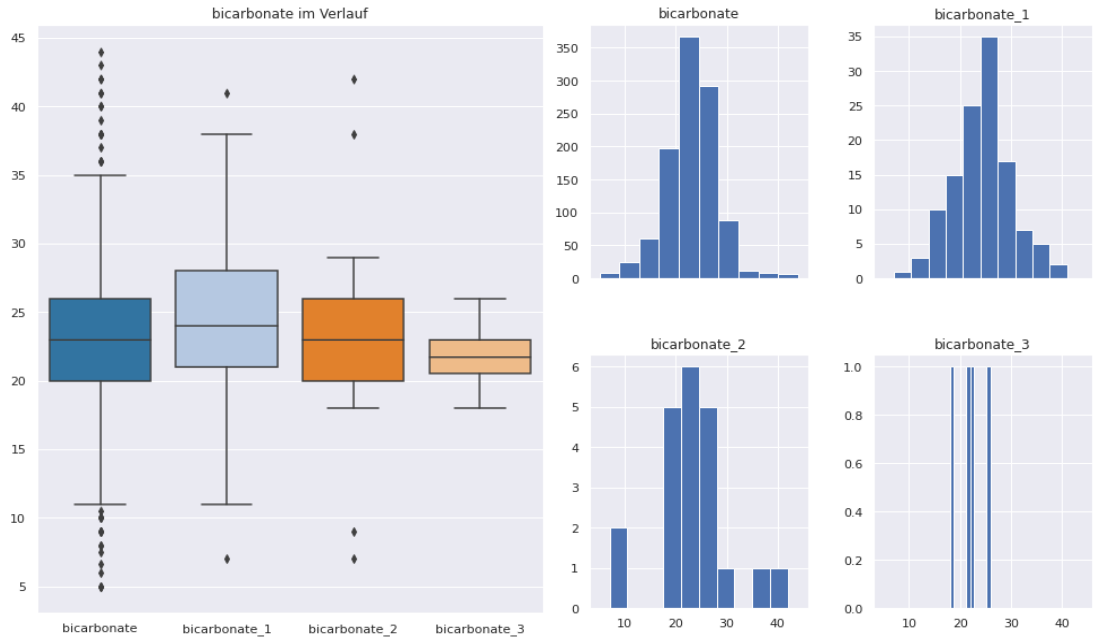


Abbildung 11: Boxplots und Histogramme zu Bicarbonat

Die statistische Verteilung des Merkmals *Kalium (potassium)* ist in Abbildung 12 dargestellt. Auch hier sin die Daten nahezu normalverteilt. Die Verteilung bleibt bei abnehmenden Häufigkeiten über die FA konstant.

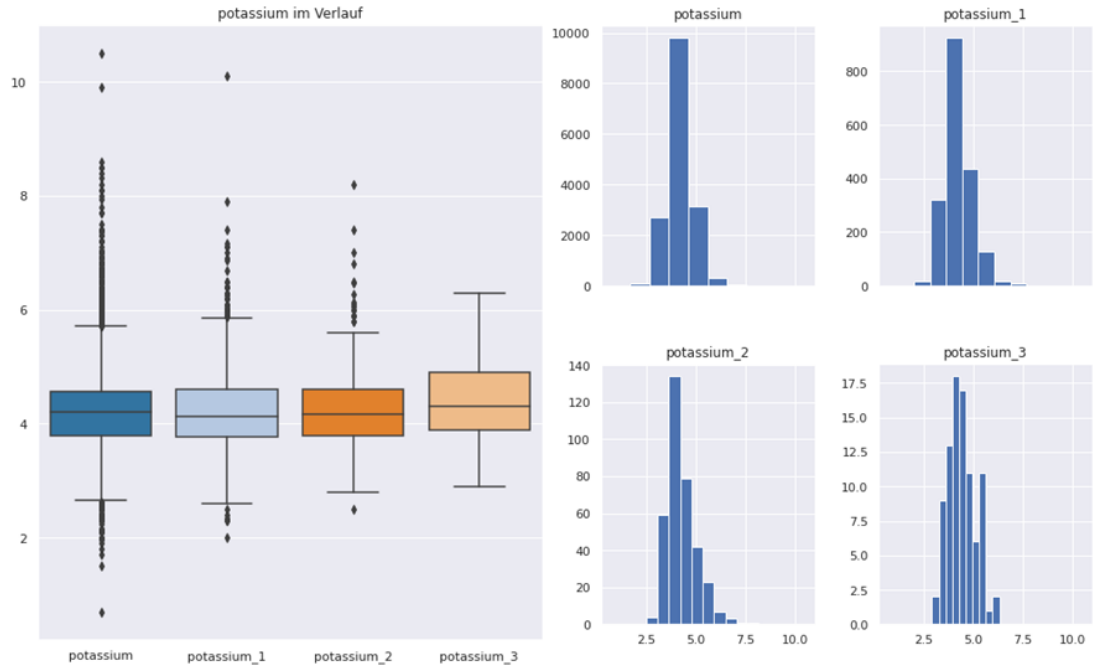


Abbildung 12: Boxplots und Histogramme zu Kalium (potassium)

Ein ähnliches Resultat liegt für die Verteilung des *Natriumwertes (sodium)* in Abbildung 13 vor. Auch hier sind die Daten normalverteilt und bleiben über die FA konstant.

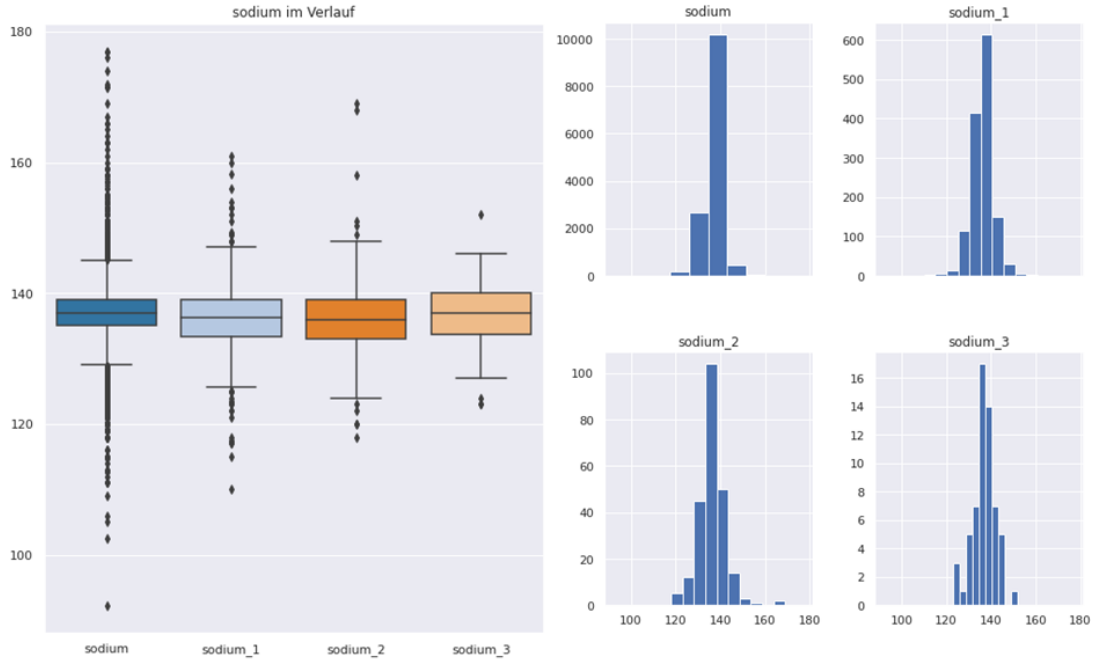


Abbildung 13: Boxplots und Histogramme zu Natrium (sodium)

Die Boxplots und Histogramme zur rechtsschiefen Verteilung der *Blutzuckerwerte (glucose)* sind in Abbildung 14 dargestellt. Auch hier bleiben die Werte über den Verlauf der FA konstant.

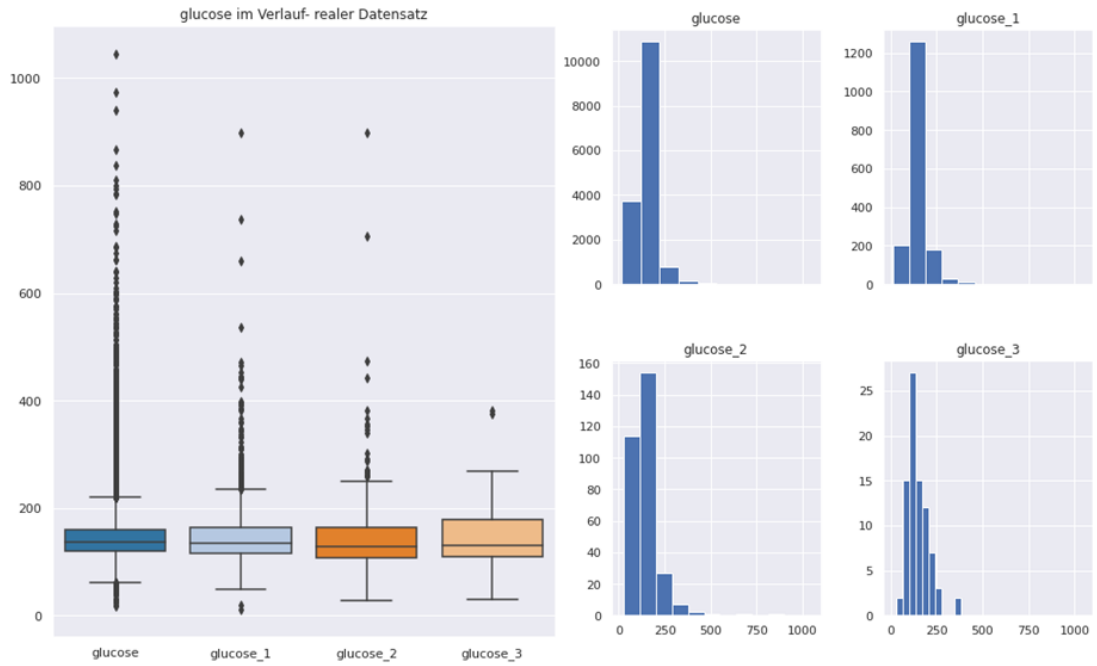


Abbildung 14: Boxplots und Histogramme zum Blutzuckerspiegel (glucose)

Insgesamt sind in der Datenverteilung viele Ausreißer ermittelt worden. Diese sind nicht entfernt worden, da nicht belegt werden konnte, dass die Werte unplausibel wären. Die Parameter sind während der intensivmedizinischen Versorgung erhoben worden. Bei einem schlechten medizinischen Allgemeinzustand konnte angenommen werden, dass die Parameter außerhalb des sonst üblichen Normalbereiches liegen können.

Für die Beschreibung der kategorischen Merkmale ist eine Häufigkeitstabelle erstellt worden, die für alle Ausprägungen jedes Merkmales die Anzahl der Häufigkeiten mit prozentualem Anteil beschreibt. Die Häufigkeiten beider Datensätze sind in Tabelle 20 gegenübergestellt.

4.5.3 **Zusammenhänge zwischen den Merkmalen**

Um die Korrelationen für den paarweisen Vergleich zwischen den Merkmalen zu ermitteln, wurden *Heatmaps* erstellt. Für die numerischen Merkmale ist der Pearson-Korrelationskoeffizient verwendet worden und für die Ermittlung der Zusammenhänge zwischen den kategorischen Merkmalen wurde das Zusammenhangsmaß *Cramérs V* berechnet.

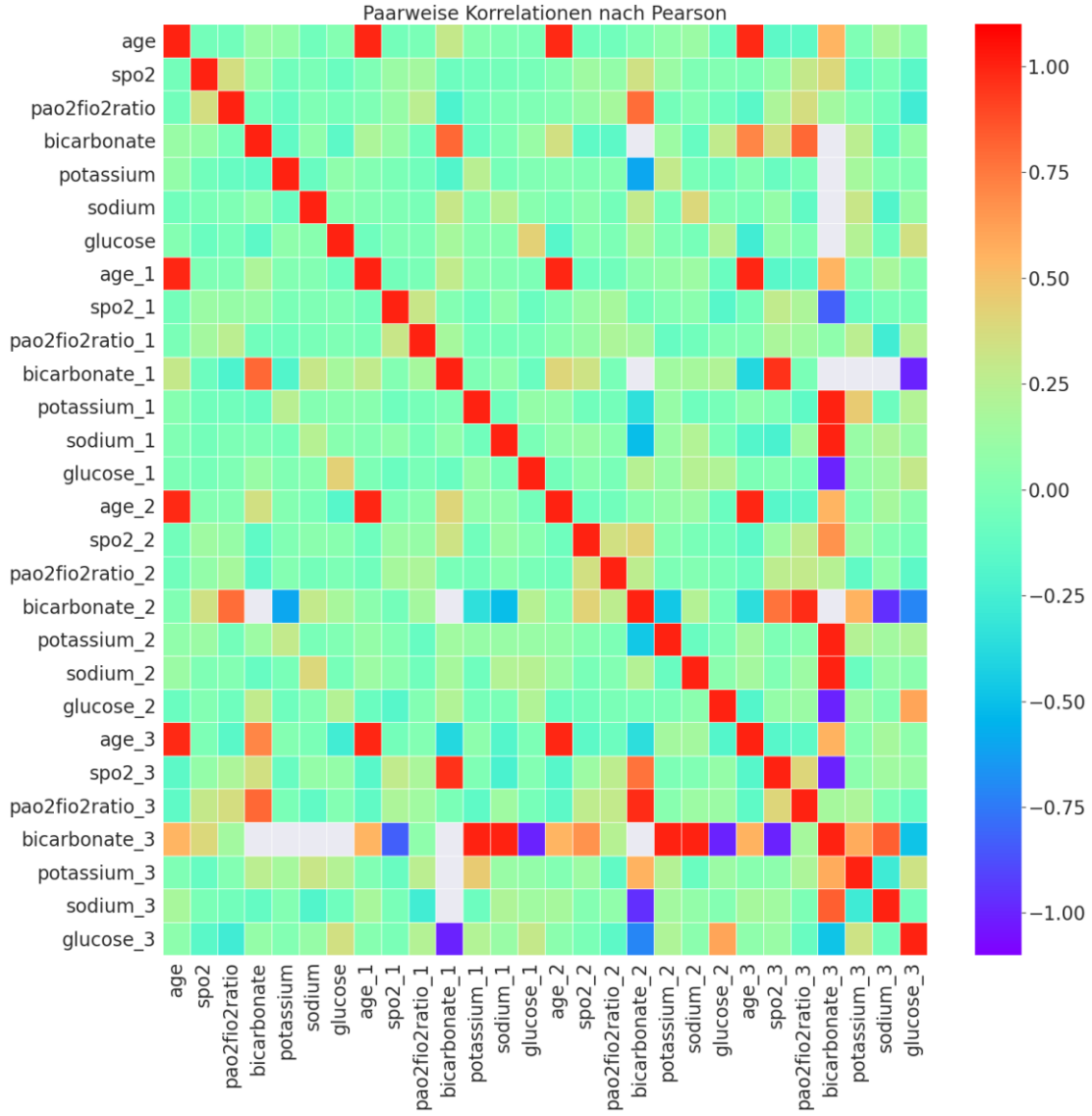


Abbildung 15: Paarweise Korrelationen nach Pearson (reale Daten)

In Abbildung 15 ist die *Heatmap* für die numerischen Merkmale aller Aufenthalte des Längsschnittes abgebildet. Hier sind z. B. positive Korrelationen in den paarweisen Ver-

gleichen zwischen dem verschiedenen Altersstufen *(age, age_1, age_2, age_3)* vorhanden. Auch negative Korrelationen, wie z. B zwischen den Attributen *glucose_3* und *bicarbonate_1* können in der *Heatmap* identifiziert werden.

Zwischen zehn der Paarungen, wie z. B. bicarbonate_2 – bicarbonate und bicarbonate_3 – potassium konnte in der Heatmap kein Korrelationskoeffizient ermittelt werden. Für einige dieser Werte, wie z. B. bicarbonate_3, liegt eine kleine Fallzahl (n=4), siehe Abbildung 18, vor. Hier ist auch zu erwähnen, dass z. B. für das Merkmal bicarbonate_3 zwar viele starke Korrelationen vorliegen, deren Aussagekraft aber aufgrund der kleinen Fallzahlen anzuzweifeln ist. Die berechneten Korrelationen könnten hier auch zufällig entstanden sein.

In Abbildung 16 ist eine Heatmap für die kategorischen Merkmale dargestellt. Für die paarweisen Vergleiche zwischen diesen Merkmalen wurden die statistischen Zusammenhänge mit dem Maß *Cramérs V* bestimmt.

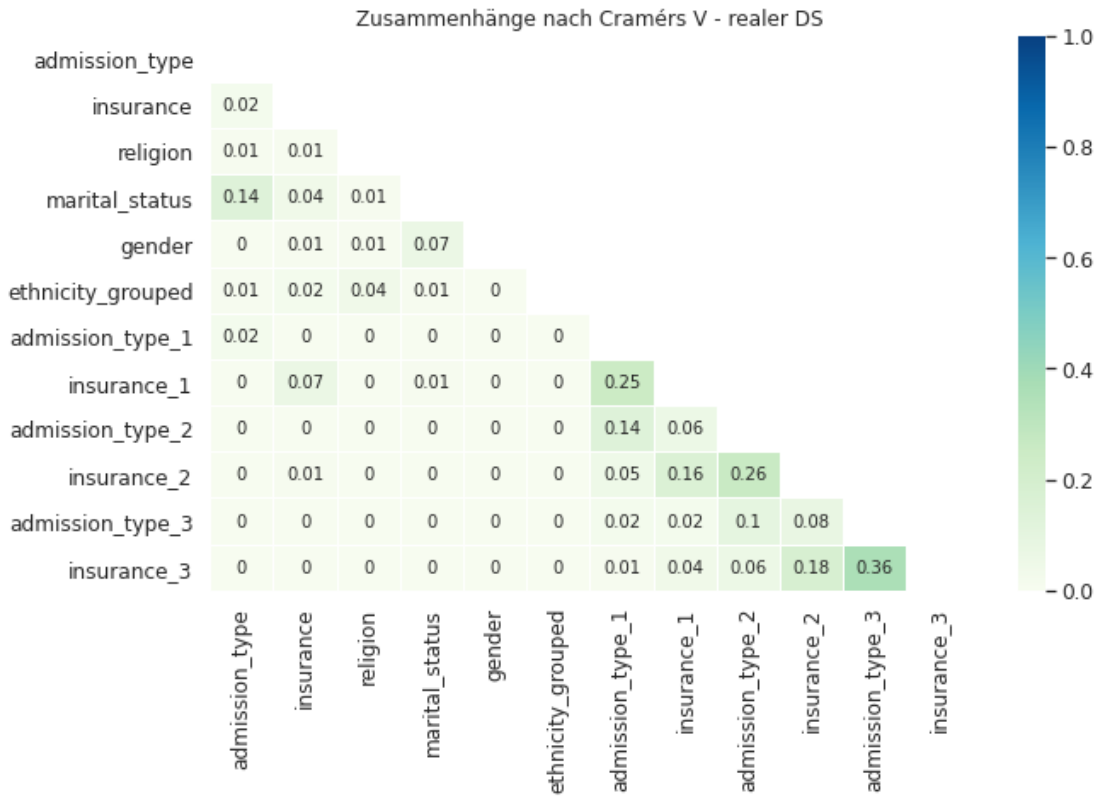


Abbildung 16: Zusammenhänge nach Cramérs V – realer DS

Insgesamt sind die Zusammenhänge in allen Vergleichen schwach. Für den Großteil der Vergleiche liegt das Maß unter 0,1. Zwischen den Paarungen *insurance_1* - *admission_type_1* und *insurance_2* - *admission_type_2* sind die Zusammenhänge, mit einem Maß von 0,25 und 0,26, etwas stärker. Der stärkste Zusammenhang konnte für den Vergleich zwischen *insurance_3* und *admission_type_3* mit einem Maß von 0,36 nachgewiesen werden.

4.6 **Vergleich der Datensätze**

In den folgenden Abschnitten wird der synthetisch generierte DS vorgestellt und gleichzeitig mit dem ursprunggebenden DS verglichen. Zunächst werden die statistischen Verteilungen verglichen und anschließend wird die Datenqualität anhand ausgewählter Evaluierungs-Metriken bewertet.

4.6.1 **Plausibilität der Zeitdaten im synthetischen Datensatz**

Um die Plausibilität des Datensatzes zu prüfen, wurden zunächst die Rohdaten betrachtet. Für die Zeit-Attribute *admittime* und *charttime* wurden keine plausiblen chronologischen Zeitfolgen durch die Synthese generiert. Es liegen z. B. Zeitpunkte der Folgeaufenthalte vor der Erstaufnahme oder die Zeitabstände betragen mehrere Jahrzehnte, die z. B. größer als 50 Jahre sein können. Aus diesem Grund wurden die Zeit-Attribute bei den Datenanalysen nicht berücksichtigt. Außerdem fiel bei der Prüfung der Rohdaten auf, dass für manche Einträge in den späteren FA *(_2, _3)* keine Zeitdaten zu entstandenen Laborparametern generiert wurden, sondern nur Laborparameter erfasst wurden, ohne dass eine Zeiterfassung vorliegt.

Auch zwischen den *Altersangaben in Jahren (age, age_1, age_2, age_3)* kamen bei der Synthese keine plausiblen aufsteigenden Werte zustande, sodass nachfolgende Werte kleiner wurden. Trotz der fehlenden Plausibilität wurden diese Merkmale bei der Datenanalyse mitberücksichtigt.

4.6.2 **Deskriptive Statistik der synthetischen Daten**

Nach der Synthese wurde für den synthetischen DS eine Beschreibung der Daten mit demselben Vorgehen wie für den realen DS durchgeführt. Die Anzahl der vorhandenen Objekte ist, wie in der Konfiguration für die Synthese angegeben wurde, in beiden Datensätzen mit 26.063 Objekten identisch.

In Tabelle 19 ist die statistische Beschreibung der numerischen Merkmale für den synthetischen DS dargestellt. Hier fällt auf, dass die Anzahl der Häufigkeiten in den späteren FA *(_2, _3)* größer sind als im realen DS. Auch die Anzahl der Häufigkeiten sind insgesamt für

alle Merkmale etwas größer als im realen DS. Für das *Alter (age_3)* z. B. liegen 399 Werte vor, während es im realen DS 181 Werte sind, für den *Bicarbonatwert* des letzten FA *(bicarbonate_3)* gibt es insgesamt 203 synthetische Werte, während im realen DS hier vier Werte erfasst sind.

Tabelle 19: Deskriptive Statistik der numerischen Merkmale-synthetischer DS

Merkmal	FA	N	MW	Std	Min	25%Q.	Median	75%Q.	Max
Alter (age)	0	25.088	59,52	20,94	0,00	50,00	63,00	75,00	89,00
	1	3.291	64,97	14,69	1,00	55,00	67,00	77,00	89,00
	2	884	62,20	17,25	0,00	53,00	65,00	75,00	89,00
	3	399	59,91	17,65	25,00	46,00	62,00	74,00	88,00
O2-Sättigung (spo2)	0	22.256	96,82	4,11	23,09	96,34	97,56	98,79	100,00
	1	3.087	96,85	3,75	54,29	95,95	97,95	98,95	100,00
	2	863	91,84	9,61	63,57	89,37	95,86	98,57	99,99
	3	403	89,55	8,63	72,11	82,19	91,73	97,50	99,99
Pa O2-Fi O2-Ratio (pao2fio2ratio)	0	20.682	283,97	126,06	20,06	196,91	273,82	360,52	1.542,22
	1	2.502	259,18	113,69	35,05	174,30	247,96	328,06	1.210,66
	2	731	282,67	140,25	33,44	177,72	268,31	368,80	647,36
	3	381	275,77	134,60	40,98	162,19	282,67	379,03	536,61
Bicarbonatwert (bicarbonate)	0	1.295	23,03	6,96	5,13	19,09	23,20	26,62	43,83
	1	329	24,37	8,75	7,45	18,31	24,46	29,99	40,74
	2	295	23,72	9,35	7,00	19,00	23,00	27,00	42,00
	3	203	21,42	2,80	18,00	18,00	21,33	22,00	26,00
Kaliumwert (potassium)	0	16.256	4,22	0,70	0,94	3,77	4,19	4,59	10,18
	1	1,925	4,20	0,72	2,02	3,74	4,13	4,60	10,03
	2	551	4,73	1,31	2,51	3,81	4,41	5,46	8,18
	3	325	4,51	0,92	2,91	3,79	4,42	5,30	6,28
Natriumwert (sodium)	0	13.611	136,92	4,68	96,11	134,92	136,95	139,10	176,62
	1	1453	136,10	4,97	112,65	133,46	136,31	138,78	160,23
	2	472	140,65	12,89	118,09	131,92	137,40	149,27	168,82
	3	296	137,10	8,17	123,06	130,25	137,44	143,42	151,97
Blutzucker (glucose)	0	15.663	150,96	62,90	17,27	121,23	143,61	165,99	1.006,38
	1	1.789	148,00	62,08	17,60	113,63	135,89	169,83	855,70
	2	491	273,04	223,79	32,51	115,54	170,92	365,27	896,24
	3	314	196,45	100,50	31,65	111,71	188,64	281,08	380,77

In beiden Deskriptionen sind in den statistischen Kenngrößen wenige Abweichungen zwischen den beiden Datensätzen zu identifizieren. Liegt ein Ausreißer vor, unterscheiden sich die Maximalwerte beider Datensätze voneinander. Das Maximum von *potassium_1* im realen DS mit 10,51 mmol/l unterscheidet sich z. B. zu dem Wert 10,03 mmol/l im synthetischen DS und der Wert von *glucose* im realen DS von 1.006,38 mg/dl zu 1.046,00

mg/dl im synthetischen DS. Die Standardabweichungen von *glucose_2* mit 273,06 und *glucose_3* mit 196, 45 sind im Vergleich zum realen DS größer und die Streuung der synthetischen Daten höher. Werden hierzu die Boxplots von *glucose_2* und *glucose_3* in Abbildung 17 verglichen, fällt eine höhere Streuung in den vergleichsweise hohen Boxplots auf.

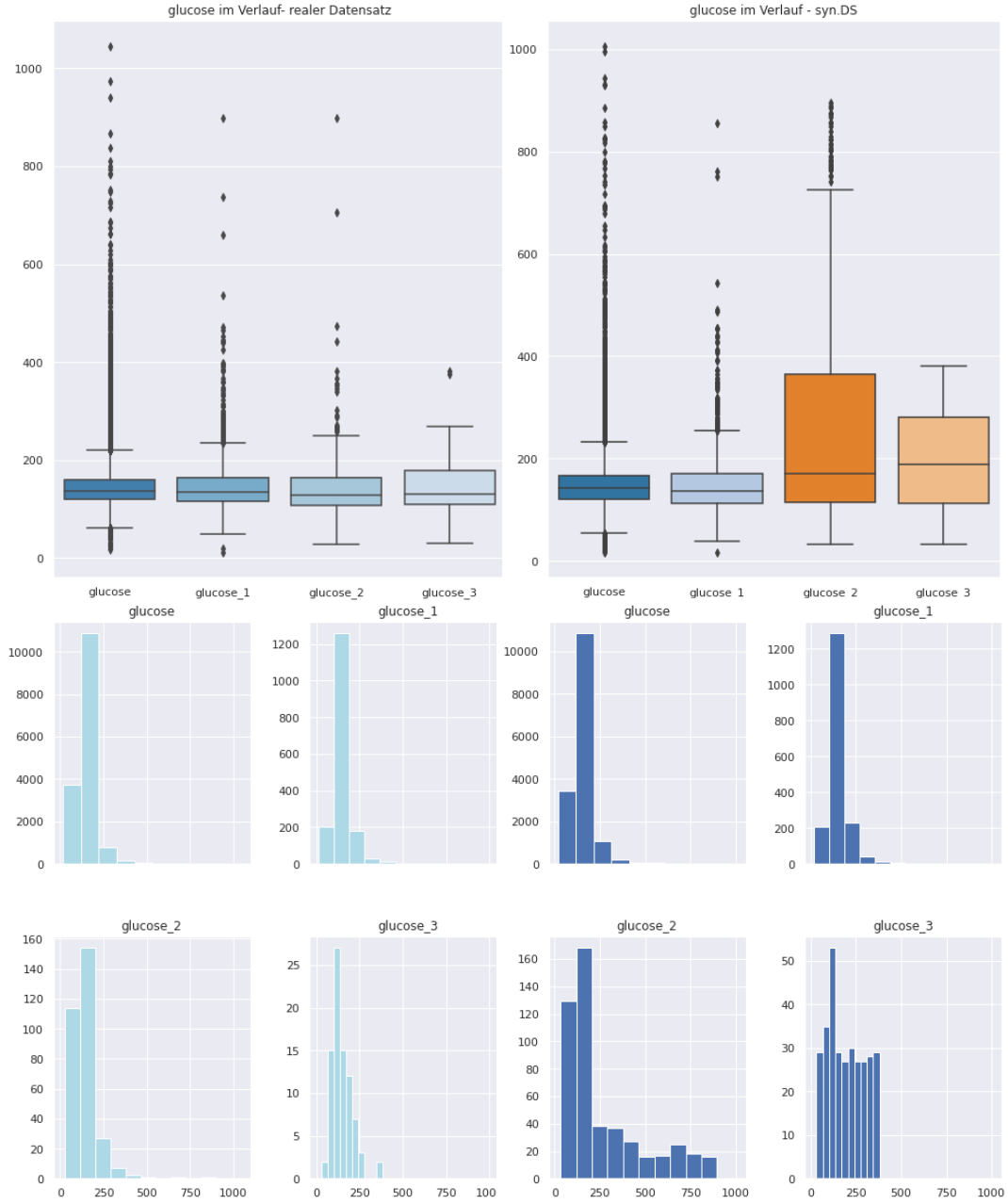


Abbildung 17: Glucose im Verlauf - Vergleich beider Datensätze

Wird die Altersverteilung des synthetischen Datensatzes mit der des realen Datensatzes in Abbildung 18 verglichen, streuen die Ausreißer in den ersten drei Aufenthalten stärker. Das bedeutet, während im realen DS alle Ausreißer bei null Jahren liegen, werden diese im synthetischen DS unterhalb des unteren Whiskers gestreut.

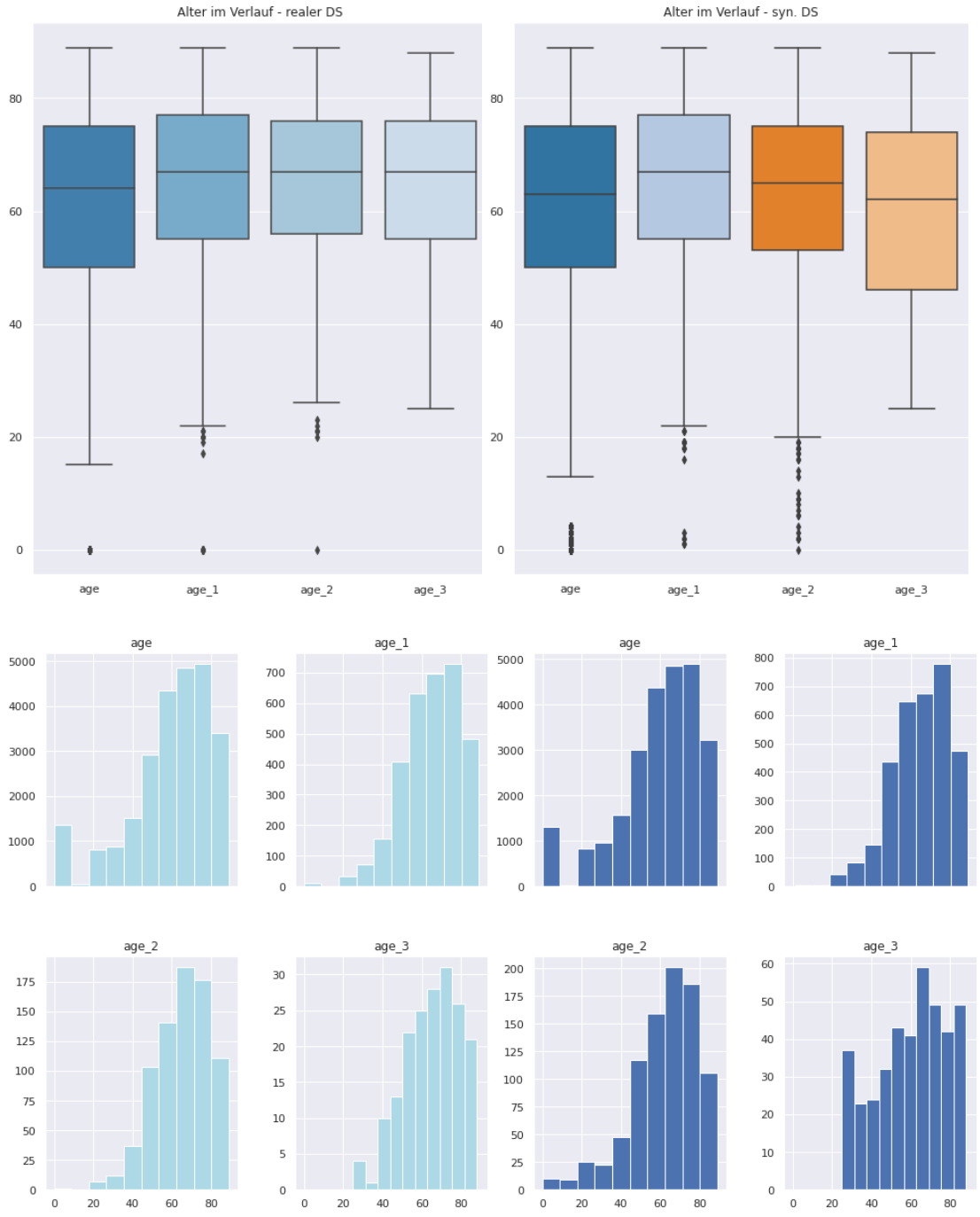


Abbildung 18: Alter im Verlauf - Vergleich beider Datensätze

Die Boxplots und Histogramme zu den weiteren Merkmalen des synthetischen Datensatzes sind in Anhang C – Datenbeschreibung hinterlegt.

4.6.3 **Häufigkeitsverteilung der kategorischen Merkmale beider Datensätze**

Um die Verteilung der kategorischen Merkmale zu vergleichen, sind in Tabelle 20 die Häufigkeitsverteilungen beider Datensätze mit prozentualem Anteil gegenübergestellt. In den meisten Ausprägungen der Kategorien liegt trotz leicht abweichender Häufigkeiten im synthetischen DS eine identische prozentuale Verteilung vor. Wird z. B. die Geschlechterverteilung betrachtet, liegen in beiden Datensätzen die Prozentanteile der männlichen Personen bei 59 % und der weiblichen Personen bei 41 %.

Ausgenommen hiervon sind einige Verteilungen der Merkmale *Einweisungsgrund (admission_type)* und *Versicherung (insurance)*. Hier liegen Differenzen zwischen den Häufigkeiten und den prozentualen Anteilen beider Datensätze vor. Für die Ausprägung *Versicherung (insurance_3)* z. B. ist die prozentuale Gewichtung für die Ausprägung *Medicare* im realen DS bei 73% *(n= 140)* und im synthetischen DS bei 41% *(n= 159)*.

Tabelle 20: Häufigkeitsverteilung der kategorischen Merkmale beider DS

		Realer Datensatz		**Synthetischer Datensatz**	
Merkmal	**Ausprägung**	**N**	**%**	**N**	**%**
Geschlecht (gender)	Männlich (M)	15.265	59	15.270	59
	Weiblich (F)	10.798	41	10.793	41
Familienstand (marital_ status)	Married	12.198	47	12.173	47
	divorced	1.507	5	1.567	6
	Single	5.765	22	5.747	22
	widowed	3.303	13	3.288	13
	life partner	6	*(0,002)* < 0,1	17	*(0,006)* < 0,1
	unknown	2.842	11	3.026	11
Religionszugehörigkeit (religion)	Catholic	9.394	36	9.312	36
	Protestant Quaker	3.088	12	3.120	12
	jewish	2.101	8	2.096	8
	Greek orthodox	186	0,7	173	0,6
	Episcopalian	338	1	343	1
	Christian scientist	172	0,6	175	0,6
	Methodist	6	*(0,002)* < 0,1	10	*(0,003)* < 0,1
	Jehova´s witness	64	*(0,02)* < 0,1	67	*(0,03)* < 0,1

			Realer Datensatz		Synthetischer Datensatz	
Merkmal	**Ausprägung**		**N**	**%**	**N**	**%**
	Buddhist		94	*(0,04)* < 0,1	97	(*0,04)* < 0,1
	Hindu		38	*(0,01)* < 0,1	46	*(0,01)* < 0,1
	Baptist		17	*(0,001)* < 0,1	20	*(0,01)* < 0,1
	7th day adventist		29	*(0,001)* < 0,1	35	*(0,001)* < 0,1
	Romanian east. Orth		42	*(0,01)* < 0,1	31	*(0,01)* < 0,1
	Unitarian- Universialist		57	*(0,02)* < 0,1	52	*(0,02)* < 0,1
	Muslim		70	*(0,02)* < 0,1	86	*(0,02)* < 0,1
	Hebrew		11	*(0,002)* < 0,1	9	*(0,005)* < 0,1
	Lutheran		1	*(0,0003)*< 0,1	6	*(0,001)* < 0,1
	other		1.223	5	1.170	5
	unknown		9.132	35	9.215	35
Ethnische Zugehörigkeit	white		18.330	70	18.256	70
	black		1.860	7	1.871	7
	asian		589	2	601	2
	hispanic		845	3	890	4
	native		17	*(0,006)*< 0,1	20	*(0,008)* <0,1
	other		775	3	782	3
	unknown		3.647	14	3.643	14
Einweisungsgrund (admission_type)	Elective	0	5.404	21	5.418	21
		1	2.801	15	508	15
		2	85	11	121	14
		3	10	5	76	20
	Emergency	0	18.590	71	18.595	71
		1	2.801	83	2.860	83
		2	696	87	669	79
		3	10	93	222	59
	urgent	0	742	3	590	3
		1	67	2	53	2
		2	16	2	16	4
		3	3	2	45	2
	Newborn	0	1.327	5	1.301	5
		1	4	0,1	3	0,1
		2	1	0,1	19	2
		3	0	0	0	0
Versicherung (insurance)	Medicare	0	13.304	51	13.274	51
		1	2.135	63	2.162	63
		2	555	70	516	60
		3	140	73	159	41
	Private	0	9.432	36	9.346	36
		1	886	26	901	26
		2	167	21	224	26
		3	30	16	101	26

			Realer Datensatz		Synthetischer Datensatz	
Merkmal	**Ausprägung**		**N**	**%**	**N**	**%**
	Medicaid insurance	0	2.257	9	2.328	9
		1	284	8	278	8
		2	64	8	90	10
		3	18	9	61	16
	Government	0	748	3	651	3
		1	59	2	54	2
		2	11	2	31	4
		3	4	2	35	17
	Self Pay	0	322	1	269	1
		1	10	0,2	10	0,5
		2	0	0	0	0
		3	0	0	0	0

4.6.4 **Zusammenhangsvergleich zwischen den Merkmalen beider Datensätze**

In der *Heatmap* der paarweisen Korrelationen zu den synthetischen Daten Abbildung 19 sind vergleichbar mit der *Heatmap* der realen Daten, s. Abbildung 15, starke Korrelationen zwischen dem *Alter* beim Aufenthalt *(age)* zu erkennen. Insgesamt liegen weniger starke Korrelationen zwischen den Merkmalen vor, bei denen im realen DS kleine Fallzahlen vorhanden sind, wie z. B *bicarbonate_3*. Das kann durch eine besseren Verteilung der Häufigkeiten zwischen den FA begründet werden. Demnach ist die Aussagekraft der paarweisen Korrelationen in diesen Fällen eher annehmbar als für den realen Datensatz. Für die synthetischen Daten liegen keine fehlenden Korrelationen in den paarweisen Vergleichen vor.

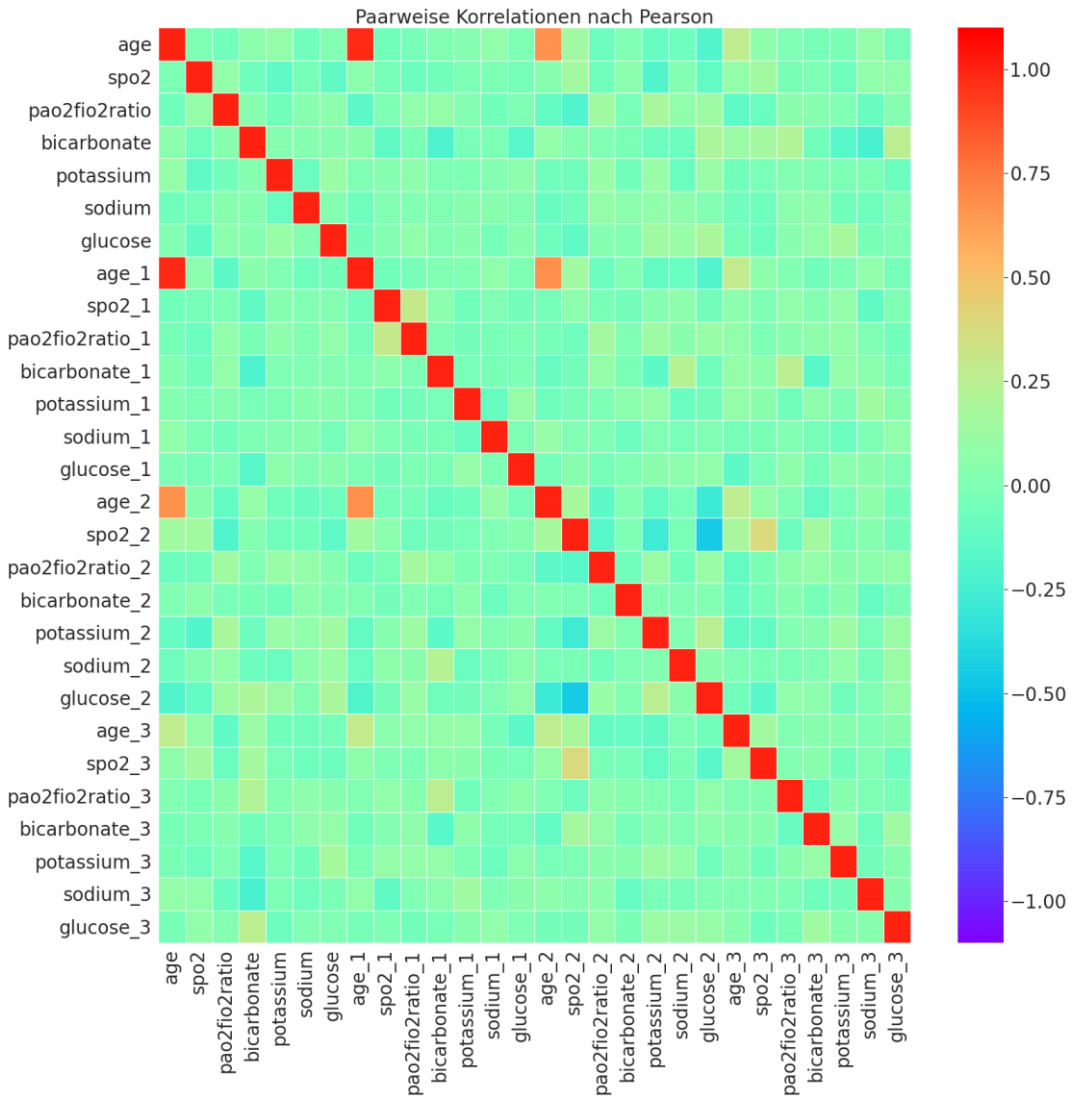


Abbildung 19: Paarweise Korrelationen der numerischen Merkmale - synthetischer DS

Um einen Vergleich zwischen den Korrelationen der beiden Datensätze zu erhalten ist in Abbildung 20 eine dritte *Heatmap* erstellt worden, in der die paarweisen Korrelationsdifferenzen nach Pearson vom realen zum synthetischen DS ermittelt wurden.

Vor allem für den ersten Aufenthalt, in dem eine größere Anzahl an Werten vorliegt, stimmen die Korrelationen miteinander stärker überein als in den FA, für die eine höhere Anzahl fehlender Werte vorliegt. Wird hier z. B. das Merkmal *bicarbonate_3* betrachtet, von dem bekannt ist, dass eine geringe Fallzahl (n=4) vorhanden ist, liegen für die Korrelationen starke Differenzen vor. Diese beziehen sich sowohl in den positiven als auch in den negativen Bereich. Das lässt darauf schließen, dass diese ermittelten Korrelationen dem

Zufall entsprechen können. Für die zehn Paarungen, für die im realen DS keine Korrelation ermittelt werden konnte, wurde auch die PKD nicht ermittelt.

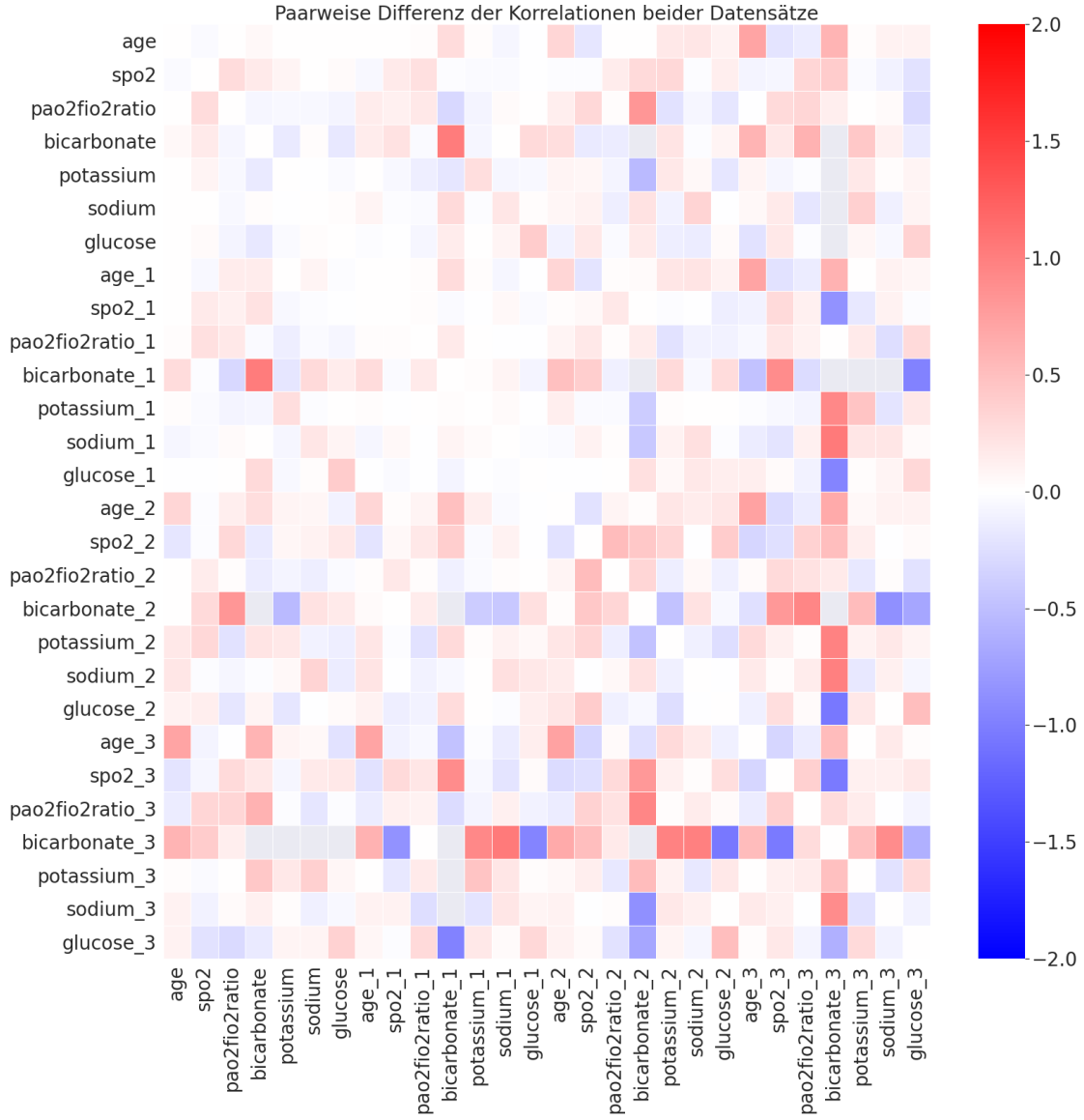


Abbildung 20: Paarweise Korrelationsdifferenz beider DS

Ähnlich wie für die numerischen Merkmale die PKD in den paarweisen Vergleichen zwischen den Datensätzen berechnet wurde, ist auch die Zusammenhangsdifferenz des *Cramérs V* für die kategorischen Merkmale berechnet worden. Die in Abbildung 21 erstellte *Heatmap* stellt die Differenzen des *Cramérs V* zwischen dem realen und synthetischen Datensatz nahe 0,00 dar. Das bedeutet, die Zusammenhänge zwischen kategorischen

Merkmalen sind in beiden Datensätzen nahezu gleich. Die höchste Differenz ist der Paarung *insurance_3 – admission_type_3* mit einem Maß von 0,12 zu entnehmen. Auch hier liegen im Vergleich zu den anderen Paarungen viele Fehlwerte vor.

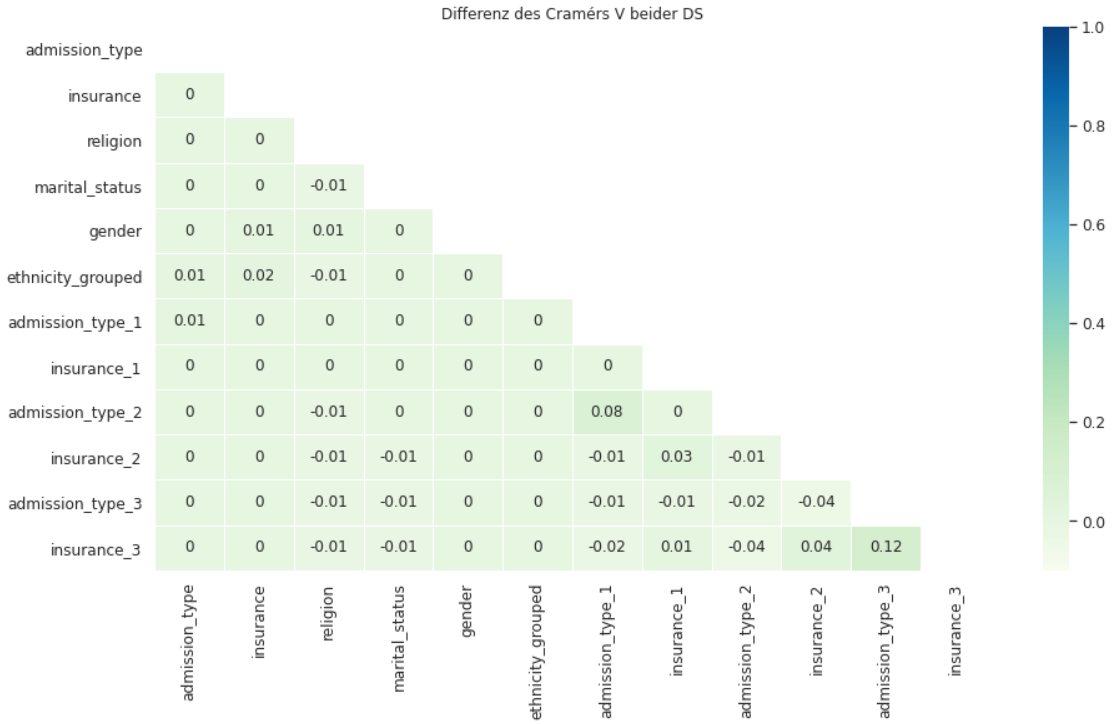


Abbildung 21: Differenz des Cramérs V zwischen beiden Datensätzen

4.6.5 **Kolmogorov Smirnov (KS)-Test und Kullback-Leibler (KL)-Divergenz**

Um die statistischen Verteilungen zwischen dem realen DS und dem synthetischen DS zu vergleichen, wurden die in Kapitel 2.2.3 beschriebenen Evaluationsmetriken KS-Test und die KL-Divergenz für jedes Merkmal berechnet. Die Resultate beider Tests zu den Merkmalen sind in Tabelle 21 aufgelistet.

Tabelle 21: Resultate aus dem KS-Test und der KL-Divergenz

		KS-Test		KL-Divergenz
Merkmal		**Statistik**	**p-Wert**	**Maß**
admission_type	0	0,0007	1,00	0,0003
	1	0,0014	1,00	<0,0001
	2	0,0476	0,33	0,0502
	3	0,1823	<0,01	0,4399

		KS-Test		KL-Divergenz
Merkmal		**Statistik**	**p-Wert**	**Maß**
insurance	0	0,0029	1,00	0,0002
	1	0,0047	1,00	0,0019
	2	0,0539	0,20	0,0429
	3	0,2448	<0,01	0,3511
religion		0,0038	1,00	0,0007
martial_status		0,0027	1,00	0,0004
gender		0,0002	1,00	<0,0001
ethnicity_grouped		0,0030	1,00	<0,0001
age	0	0,0048	<0,01	0,3191
	1	0,0143	0,90	0,0615
	2	0,0683	0,53	0,0971
	3	0,1381	0,06	0,0783
spo2	0	0,2322	0,00	0,0021
	1	0,1808	<0,01	0,0016
	2	0,2712	<0,01	0,0072
	3	0,4470	<0,01	0,0059
pao2fioratio	0	0,0179	0,02	0,1995
	1	0,0190	0,76	0,1945
	2	0,1388	0,57	0,2374
	3	0,1295	0,19	0,2769
bicarbonate	0	0,0864	0,01	0,0772
	1	0,1750	0,05	0,0967
	2	0,1428	0,99	0,1179
	3	0,2500	1,0	0,0314
potassium	0	0,0346	<0,01	0,2460
	1	0,0367	0,16	0,0258
	2	0,2238	<0,01	0,0554
	3	0,1333	0,40	0,0398
sodium	0	0,0778	<0,01	0,0011
	1	0,0847	<0,01	0,0013
	2	0,2797	<0,01	0,0053
	3	0,2000	0,18	0,0030
glucose	0	0,0916	<0,01	0,1211
	1	0,0436	0,08	0,1294
	2	0,3061	<0,01	0,3957
	3	0,3493	<0,01	0,1838

Ergebnisse des KS-Tests

Für den KS-Test wurde das Signifikanzniveau auf unter 0,05 festgelegt, orientiert an dem Vorgehen von Baowaly et al. (21), s. Kapitel 2.2.3. Das bedeutet, wenn kein signifikantes Ergebnis vorliegt, konnte eine unterschiedliche Verteilung nicht nachgewiesen werden.

Werden zunächst die Ergebnisse des KS-Tests für die kategorischen Merkmale betrachtet, werden *p-Werte* identifiziert, die größer als 0,05 sind. Der *p-Wert* für *admission_type_1* und *insurance_0* liegt z. B. bei 1,00.

Vor allem für die kategorischen Merkmale liegen keine Unterschiede in der Verteilung vor. Während z. B. für das Merkmal *insurance* mit einem *p-Wert* von 1,00 kein signifikanter Unterschied in der Verteilung nachgewiesen werden konnte, wurde für *insurance_3* mit einem *p-Wert* von 0,01 ein signifikanter Unterschied festgestellt. Verglichen mit der prozentualen Verteilung der Häufigkeiten in Tabelle 20 bestätigen sich die Ergebnisse, in denen sich die prozentualen Anteile zwischen den beiden Datensätzen für *insurance* gleich sind und sich für *insurance_3* unterscheiden.

Für die numerischen Merkmale konnte in vielen Fällen eine signifikant unterschiedliche Verteilung nachgewiesen werden. Die Unterschiede in der Verteilung können möglicherweise mit den vorhandenen Fehlwerten in Verbindung gebracht werden, da in den numerischen Merkmalen verglichen mit den kategorischen Merkmalen unabhängig von den FA mehr Fehlwerte vorliegen.

Für den Test sind keine *Null*-Werte berücksichtigt worden und für beide Datensätze wurde die gleiche Anzahl an Werten bei einer unterschiedlichen Häufigkeitsverteilung verwendet. Dadurch wurde die Anzahl der Werte einer der verglichenen Merkmale reduziert und hat die Abweichung der Verteilungen beeinflusst. Für das Resultat von z. B. *bicarbonate_3 (p-Wert = 1,00)* ist zu beachten, dass hier eine kleine Fallzahl (realer DS, n=4) vorliegt. Hier können die Testergebnisse angezweifelt werden.

Ergebnisse der KL-Divergenz

Werden die Ergebnisse der KL-Divergenz untersucht, liegen für Resultate mit dem Wert 0 eine identische Verteilung der beiden Datensätze vor. Je größer die Distanz von 0 ist, desto mehr weichen die Verteilungen voneinander ab.

Für die meisten kategorischen Merkmale liegt die KL-Divergenz nahe 0, sodass für diese Merkmale von einer identischen Verteilung in beiden Datensätzen ausgegangen werden

kann. Für das kategorische Merkmal der *Religionszugehörigkeit (religion)* z. B. liegt das Maß der KL-Divergenz bei 0,0007. Für dieses Merkmal ist nach der KL-Divergenz die Verteilung nahezu identisch. Für Ausnahmen der kategorischen Merkmale, wie z. B. für die Merkmale *admission_type_3* (Maß = 0,4399)*, insurance_3* (Maß = 0,3511) ist die KL-Divergenz höher. Hier stimmt das Ergebnis mit dem signifikanten Ergebnis des KS-Tests und den Zahlen der Häufigkeitsverteilung überein, sodass davon ausgegangen werden kann, dass Unterschiede in der Verteilung vorliegen.

Werden alle Ergebnisse der KL-Divergenz mit denen der Häufigkeitsverteilung und des KS-Tests für die kategorischen Merkmale verglichen, liegen ebenfalls übereinstimmende Ergebnisse vor.

In den Ergebnissen der KL-Divergenz für die numerischen Merkmale liegen höhere Werte vor. Für das Merkmal *potassium* z. B. liegt das Maß bei 0,246. Hier resultiert aus der KL-Divergenz, dass sich die Verteilung unterscheidet. Wird das Ergebnis des KS-Tests (*p-Wert* <0,01) zusätzlich betrachtet, wird das Ergebnis bestätigt.

Nicht für alle numerischen Merkmale sind die Ergebnisse so eindeutig, sodass zwischen dem KS-Test und der KL-Divergenz Uneinigkeiten in den Ergebnissen vorliegen. Diese Differenzen können mit fehlenden Werten und dem damit verbundenen Vorgehen entstammen. Für die Berechnung der KL-Divergenz wurden Fehlwerte nicht berücksichtigt und eine unterschiedliche Anzahl der Werte in den beiden Datensätzen wurden auch hier für den verwendeten Test angeglichen. Die ermittelten Resultate der Vergleiche sind daher sowohl für den KS-Test als auch für die KL-Divergenz nicht vollumfänglich belastbar.

4.6.6 **Log-Cluster Metrik**

Für den Vergleich der gesamten Verteilung der Datensätze wurde nach dem Vorgehen von Goncalves et al. die in Kapitel 2.2.3 beschriebene *Log-Cluster Metrik* durchgeführt. Für die Clusteranalyse wurden für die kategorischen Merkmale *(admission_type, insurance, religion, marital_status, icustay_id, gender, ethnicity_grouped)* beider Datensätze 20 Cluster erstellt (45). In Abbildung 22 sind die Verteilungen der Datensätze in die verschiedenen Cluster in einem Balkendiagramm dargestellt. Die Anzahl der Zuordnungen in

die verschiedenen Cluster ist ähnlich. Zusätzlich wurde der U_C-Wert für den Vergleich der Clusterzuordnungen mit der von Goncalves et al. beschrieben Metrik berechnet (45). Das Resultat ist ein U_C-Wert von -8,15 berechnet.

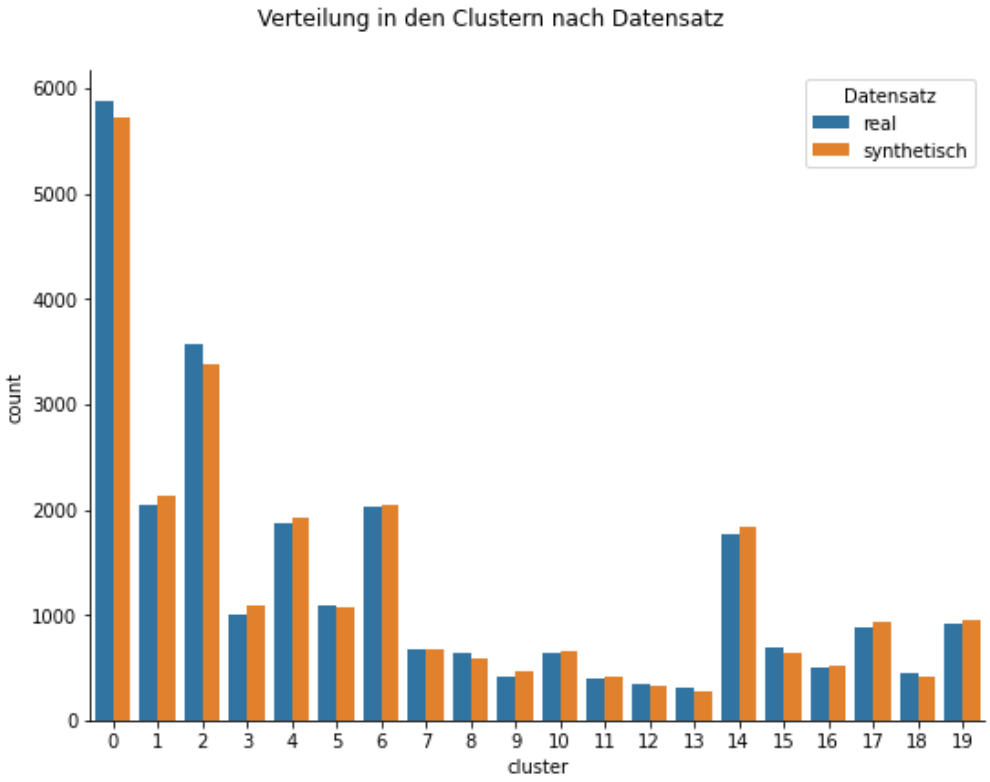


Abbildung 22: Verteilung in den Clustern nach Datensatz bei 20 Clustern

In einem weiteren Versuch wurde die optimale Clusteranzahl mit der *Ellenbogen-Methode*[35] ermittelt, dessen Ergebnis auf vier Cluster schloss. Der aus dieser Clusteranalyse resultierende U_C-Wert lag bei -10,57.

Die Ellenbogenkurve, s. Abbildung 29 und das Histogramm mit der Clusterverteilung, s. Abbildung 30, sind in Anhang C – Datenbeschreibung hinterlegt.

[35] Bei der Ellenbogenmethode wird eine Kurve zwischen der Summe der quadratischen Abstände jedes Datenpunktes und der Anzahl der Cluster erstellt, um die optimale Anzahl der Cluster für die Clusteranalyse zu identifizieren (112).

4.7 Prüfung der Privatheitswahrung

Um zu prüfen, ob Rückschlüsse aus den synthetischen Daten auf im realen Datensatz erfasste Personen möglich sind, wurden Funktionen erstellt. Zum einen wurde eine Funktion implementiert, mit der sich Äquivalenzklassen unter Berücksichtigung eines sensiblen Attributs zusammenfassen lassen, um zu prüfen, wie hoch die k-Anonymität bzw. ℓ-Diversity ist. Dies erfolgt, ohne dass die Merkmale in den Äquivalenzklassen, wie bei bekannten Anonymisierungstechniken z. B. verrauscht oder zusammengefasst werden, wie in Kapitel 2.3 beschrieben wurde.

Zum anderen wurde eine Funktion implementiert, die aus beiden Datensätzen vergleicht, wie viele Datensätze mit den gleichen Merkmalen vorliegen. Die zu prüfenden Merkmale werden dabei im Funktionsaufruf übergeben.

4.7.1 Prüfung der *k-Anonymität* und *ℓ-Diversity*

Um die Privatheitswahrung zu prüfen wurde eine Funktion erstellt, mit der die Höhe des *k-Wertes* der *k-Anonymität* bzw. die Höhe des *ℓ-Wertes* der *ℓ-Diversity* ermittelt werden kann. Dabei werden die Äquivalenzklassen erschlossen, ohne dass Merkmale verrauscht oder zusammengefasst werden.

Tabelle 22: k-Anonymität und l-Diversity nach Szenarien

			Realer DS		**Syn. DS**	
Sensibles Attribut	**Quasi-Id. 1**	**Quasi-Id. 2**	**k**	**ℓ**	**k**	**ℓ**
religion	admission_type	gender	331	11	252	15
marital_status	admission_type	gender	331	2	252	6
insurance	admission_type	gender	331	4	252	5
religion	admission_type	age	1	1	1	1
religion	admission_type	ethnicity_grouped	1	1	1	1
insurance	admission_type	ethnicity_grouped	1	1	1	1
insurance	marital_status	ethnicity_grouped	7	4	8	5
religion	ethnicity_grouped	gender	7	4	8	5
ethnicity_grouped	admission_type	gender	331	6	252	6

Wie in Kapitel 3.6.1 beschrieben, lassen sich die *k-Anonymität* und *ℓ-Diversity* für verschiedene Szenarien prüfen. In Tabelle 22 sind ausgewählte Szenarien anhand der als *Quasi-Identifikatoren* verwendeten kategorischen Merkmale zusammengefasst. Hierbei wird bei der Berücksichtigung des *sensiblen Attributs* variiert. Für die Prüfung der *k-Anonymität* ist das *sensible Attribut* nicht relevant, weshalb sie bei wechselnden *sensiblen Attributen* und gleichbleibenden *Quasi-Identifikatoren* ihren Wert nicht ändern.

Aus den Ergebnissen kann geschlossen werden, dass die *ℓ-Diversity* im synthetischen DS in den meisten geprüften Szenarien größer ist als im realen Datensatz. Werden die Szenarien betrachtet, deren Wert für die *k- Anonymität* 1 beträgt, kann der Wert für die *ℓ-Diversity* auch nur 1 betragen, da dann mindestens eine Äquivalenzklasse mit nur einem einzelnen Eintrag vorliegt. Da *ℓ* die Mindestzahl der unterschiedlichen Ausprägungen in den Äquivalenzklassen angibt, kann *ℓ* nie größer als *k* sein. Bei dem Szenario, in dem das *Alter in Jahren (age)* berücksichtigt wurde, kann der Wert von 1 mit den feingliedrigen Ausprägungen begründet werden. Hier ist die Vielfalt der möglichen Äquivalenzgruppen größer. Werden hingegen die Szenarien betrachtet, die mit dem Merkmal *Geschlecht (gender)* erstellt wurden, bei dem zwei Ausprägungen *(männlich* und *weiblich)* vorliegen, sind *k-Wert* und *ℓ-Wert* und die dadurch erstellten Äquivalenzgruppen größer. Wurde ein Szenario mit mehr als zwei *Quasi-Identifikatoren* geprüft, lagen der *k-Wert* und *ℓ-Wert* ebenfalls bei 1.

4.7.2 **Übereinstimmungen finden**

Für die Prüfung der übereinstimmenden Objekte sind die sieben Merkmale *age, gender, marital_status, ethnicity_grouped, insurance, religion* und *admission_type* als *Quasi-Identifikatoren* berücksichtigt worden. Dann wurde für jedes Objekt des synthetischen Datensatzes mit jedem Objekt aus dem realen DS ein paarweiser Vergleich durchgeführt. Während des Abgleichs wurde für jedes Objekt des synthetischen Datensatzes protokolliert, wie viele Merkmale des Objektes mit wie vielen Objekten aus dem realen DS übereinstimmen. Anschließend wurde derselbe Vorgang nur mit dem realen DS wiederholt, um eine Kontrollgruppe zu bilden, wie in Kapitel 3.6.2 näher beschrieben wurde.

Übereinstimmungen im Gesamtdatensatz

In Tabelle 23 werden die Ergebnisse aus den paarweisen Vergleichen des synthetischen Datensatzes mit denen des realen Datensatzes, sowie die aus dem realen Datensatz, zusammengefasst.

In der Tabelle wird zu der Anzahl der übereinstimmenden Merkmale jeweils der ermittelte Maximalwert *(Max.)* angegeben. Dieser sagt aus, wie oft die entsprechende Anzahl an Merkmalen für ein spezifisches einzelnes Objekt aus dem synthetischen bzw. geprüften Datensatz übereinstimmen. Idealerweise sollte dieser Wert entweder bei 0 liegen, oder einen möglichst hohen Wert erhalten. Entweder liegen dann keine Übereinstimmungen vor oder die Gruppe identischer Objekte ist möglichst groß. Beides erschwert eine Re-Identifizierung.

In der benachbarten Spalte ist die Anzahl der konfrontierten Objekte *(Kon.)* dokumentiert. Dieser Wert gibt an, für wie viele Objekte aus dem geprüften DS mindestens ein Objekt mit der entsprechenden Anzahl gleicher Merkmale im anderen gegengeprüften DS gefunden wurde. Enthält z. B. ein Objekt aus dem ersten Datensatz die Merkmale *Geschlecht = ‚weiblich'* und *Einweisungsgrund = ‚elektiv'*, gibt es mindestens ein Objekt im zweiten Datensatz, das über die gleichen zwei Merkmale verfügt.

In der Kontrollgruppe, dem Vergleich zwei identischer Datensätze, wurden wider der Erwartung, dass für jedes Objekt mindestens ein identisches Objekt vorliegt, weniger Objekte ermittelt. Für den vergleich sieben identischer Merkmale wurden z. B. nur 25.741 konfrontierte Objekte gefunden, dass sich mit den fehlenden Werten für das Merkmal Alter begründen lässt. Fehlt die Angabe zum Alter, kann nicht von einem übereinstimmenden Objekt ausgegangen werden.

In der dritten Spalte steht die Anzahl der identifizierten Eins-zu-Eins-Übereinstimmungen *(1:1)*. Für die Anzahl dieser Objekte aus dem geprüften Datensatz existiert genau ein Objekt mit den identischen geprüften Merkmalen in der Stichprobe des gegengeprüften Datensatzes.

Wird z. B. der Vergleich zwischen dem synthetischen und realen DS mit 7 Merkmalen betrachtet, ist erkennbar, dass mindestens ein Objekt im synthetischen DS existiert, zu dem 210 Objekte im realen DS mit identischen (geprüften) Merkmalen gefunden wurden. Bei dem Wert handelt es sich um das Maximum aus diesen Vergleichen. Das bedeutet, für die anderen Objekte ist die Anzahl der übereinstimmenden Objekte potentiell kleiner. Für 16.075 der 26.060 Objekte aus dem synthetischen Datensatz gibt es mindestens eine 7-fach-Übereinstimmung zu einem Objekt aus dem realen Datensatz. Die Differenz mit 38,31 % *(n=9.985)* zeigt die Objekte, die bei der Überprüfung nicht gefunden wurden und deren Kombination der Merkmale nur im synthetischen DS vorkommt. In 3.236 Fällen gibt es genau ein Objekt in beiden Datensätzen, das in allen sieben geprüften Merkmalen identische Werte vorweist.

Tabelle 23: Übereinstimmungen zwischen den synthetischen und realen DS und der Kontrollgruppe

Anzahl geprüfter Merkmale	**Verwendete Datensätze für die Überprüfung *(n=26.060)***					
	synthetisch : real			**real : real (Kontrollgruppe)**		
	Max.	**Kon.**	**1:1**	**Max.**	**Kon.**	**1:1**
1	14.058	26.060	0	14.570	26.060	0
2	12.137	26.060	0	12.403	26.060	0
3	10.441	26.060	0	10.447	26.060	0
4	7.620	26.060	0	7.627	26.060	0
5	3.899	26.042	14	3.989	26.057	7
6	1.048	25.594	421	1.048	25.741	358
7	210	16.075	3.236	210	25.077	7.586

Mit den in Tabelle 23 ermittelten Werten der *Eins-zu-Eins*-Übereinstimmungen wurden das Risiko und die Chance (Odds) berechnet, ob für ein Objekt eine *Eins-zu-Eins*-Übereinstimmung vorliegen kann. Das Risiko, ob für ein Objekt sieben *Quasi-Identifikatoren* im synthetischen und im realen DS übereinstimmen, liegt mit einem Verhältnis von 3.236 : 26.060 bei 12,42 %. In der Kontrollgruppe liegt das Risiko für eine *Eins-zu-Eins*-Übereinstimmung mit einem Verhältnis von 7.586 : 26.060 bei 29,11 %. Das aus diesen Werten ermittelte *Relative Risiko* beträgt 2,34 für die Kontrollgruppe. Das bedeutet, das Risiko für die Kontrollgruppe, dass ein identisches Objekt identifiziert wird, ist etwas mehr als zweimal so hoch, als für den Vergleich zwischen dem synthetischen und realen DS.

Die Chancen (Odds), dass im gleichen Beispiel eine *Eins-zu-Eins*-Übereinstimmung vorliegt, liegen für die Prüfung zwischen dem realen und synthetischen DS mit einem Verhältnis von 3.236 : 22.824 bei 14,18 % und für die Kontrollgruppe mit einem Verhältnis von 7586 : 18.474 bei 41,06 %. Das entspricht einer 2,90-fach höheren Chance *(Odds-Ratio)*, dass ein identisches Objekt im Vergleich zwischen dem realen DS (Kontrollgruppe) vorliegt, als im Vergleich zwischen dem synthetischen und realen DS.

Übereinstimmungen mit einem unabhängigen Datensatz

Um zu prüfen, ob bei der Synthese zufällig reale Objekte aus dem realen DS generiert wurden, sind in einem weiteren Versuch vor der Synthese 15,00 % *(n=3.910)* des Datensatzes zurückgehalten *(Split)* worden. Für Identifizierung der Übereinstimmungen wurde bei der Synthese die gleiche Anzahl *(n=3.910)* an Objekten aus dem anderen 85 % (*n=22.150*) des Gesamtdatensatzes generiert. Der Vergleich und die Erstellung der verwendeten Datensätze ist in Anhang C – Datenbeschreibung in Abbildung 31 schematisch dargestellt.

Die in Tabelle 24 aufgelisteten Ergebnisse der Überprüfung bringen hervor, dass auch im Abgleich mit dem unabhängigen Datensatz *Eins-zu-Eins*-Übereinstimmungen aller geprüften Merkmale vorliegen. Mit 575 *Eins-zu-Eins*-Übereinstimmungen liegt das Risiko bei 14,71% bzw. die Chance bei 17,43 %. Die Anzahl der konfrontierten Objekte für den Abgleich liegt bei 1.330. Das bedeutet, fast 66% der Objekte wurden bei die Überprüfung auf die Übereinstimmung mit allen *Quasi-Identifikatoren* (7 Merkmalen) nicht gefunden.

Tabelle 24: Übereinstimmungen synthetische Daten mit Daten-Split

Anzahl geprüfter Merkmale	Verwendete Datensätze für die Überprüfung *(n=3.910)*					
	synthetisch : real (Split)			real (Split) : real (Split) (Kontrollgruppe)		
	Max.	Kon.	1:1	Max.	Kon.	1:1
1	2.057	3.910	0	2.190	3.910	0
2	1.771	3.910	0	1.858	3.910	0
3	1.556	3.910	0	1.556	3.910	0
4	1.170	3.910	0	1.177	3.910	0
5	570	3.902	9	570	3.903	8
6	154	3.513	236	154	3.660	232
7	35	1.330	575	35	3.761	2270

Die prozentuale Angabe der nicht gefundenen Objekte ist im Vergleich zu dem Abgleich aus dem Gesamtdatensatz fast doppelt so hoch. Im Vergleich der synthetischen Daten mit dem unabhängigen Split wurden weniger übereinstimmende Objekte gefunden. Das Risiko liegt jedoch minimal höher (~ 2 %) als beim Abgleich der Gesamtdatensätze.

Ergebnisse von weiteren Berechnungen und den vorab erstellten Stichproben sind in

Tabelle 27 und der zugehörigen *Odds-Ratios* in Tabelle 28 in Anhang C – Datenbeschreibung zusammengefasst.

5 Diskussion und Ausblick

Im abschließenden Kapitel wird zunächst die Arbeit zusammengefasst und dann das Erreichen der in Kapitel 1.6 gesetzten Ziele diskutiert und reflektiert. Anschließend wird ein Ausblick auf zukünftige Perspektiven auf dem Gebiet synthetischer Daten gegeben.

5.1 Zusammenfassung

Das Ziel dieser Masterarbeit war die Generierung synthetischer medizinischer Forschungsdaten aus einem bestehenden Datensatz. Dabei lag der Fokus darauf, den generierten Datensatz auf die Datenqualität und eine mögliche Re-Identifizierung der im realen Datensatz enthaltenen Personen zu untersuchen.

Für dieses Vorhaben wurde zunächst der Begriff Datenqualität definiert. Dann sind mögliche Evaluierungstechniken für die Untersuchung der Datenqualität von synthetischen Daten mit einer Literaturrecherche ermittelt und beschrieben worden. Auch die Methoden zur Prüfung der Privatheitswahrung und die Verfahren und Anwendungen zur Erzeugung synthetischer Daten wurden dabei recherchiert.

Als nächstes wurden die Anforderungen an die für die Synthese verwendeten Daten, den synthetischen Datensatz und das verwendete Tool definiert. Diese konzentrieren sich darauf, einen synthetischen Datensatz zu erzeugen, der die gleichen statistischen Eigenschaften enthält, wie der ursprunggebende Datensatz, ohne Bezüge zu Personen aus dem realen Datensatz zu enthalten. Für die Identifikation eines geeigneten Tools wurde eine Nutzwertanalyse durchgeführt, welche zur Auswahl der Anwendung *DataSynthesizer* führte.

Um eine geeignete Datenextraktion aus dem als zugrundeliegendes Beispiel gewählten *MIMIC-III*-Datensatz zu erstellen, wurde aus einem bereits vorhandenen Teildatensatz eine Tabelle für die Synthese erstellt. Dafür wurden drei Tabellen verwendet, die durch

einen *Join* zusammengefasst wurden. Der Versuch die Synthese mit einer einfachen Verkettung der Tabellen durchzuführen, führte zu keinem brauchbaren Ergebnis. Deshalb wurde ein zweiter Pfad eröffnet, indem ein Längsschnitt der Daten erstellt wurde.

Nach der erfolgreichen Validierung dieses synthetischen Datensatzes ist dieser statistisch beschrieben und analysiert worden. Danach wurden der synthetische und der reale Datensatz hinsichtlich ihrer Datenqualität verglichen. Im KS-Test konnten für die kategorischen Merkmale keine signifikanten Unterschiede in der Wahrscheinlichkeitsverteilung nachgewiesen werden. Dieses Ergebnis wurde mittels der KL-Divergenz für alle kategorischen Merkmale ermittelt, indem darüber ein geringer Unterschied zwischen den Verteilungen nachgewiesen werden konnte. Auch bei der Prüfung der Zusammenhänge nach *Cramérs V* konnten für die kategorischen Merkmale nur wenige Differenzen nachgewiesen werden.

Für die numerischen Merkmale wurden sowohl signifikante als auch nicht signifikante Ergebnisse im KS-Test ermittelt. Im Vergleich zwischen den Resultaten aus dem KS-Test und der KL-Divergenz wurde der signifikante Nachweis einer unterschiedlichen Verteilung nur teilweise bestätigt. Andere Ergebnisse der KL-Divergenz widersprachen den Resultaten des KS-Tests. Aufgrund einer teilweise geringen Datendichte sind diese Ergebnisse nicht vollständig belastbar. Auch bei den paarweisen Vergleichen des Korrelationskoeffizienten nach *Pearson* lag eine Testunsicherheit aufgrund von fehlenden Werten vor. Hier entsprachen die Korrelationen teils dem Zufall bzw. könnten auch zufällig entstandene Scheinkorrelationen vorliegen. In der PKD lagen vor allem in den Vergleichen geringe Differenzen vor, in denen eine geringe Anzahl von Fehlwerten vorlag.

Als weitere Analyse zum Vergleich der Verteilung beider Datensätze wurde eine *Log-Cluster-Metrik* durchgeführt. Unter der Berücksichtigung von 20 bzw. 4 Clustern konnte eine ähnliche Anzahl der Zuordnungen der Objekte beider Datensätze identifiziert werden.

Im Anschluss an die Prüfung der Datenqualität ist der synthetische Datensatz auf die Bewahrung der Privatheit geprüft worden. Dafür wurde eine Funktion erstellt, welche

die *k-Anonymität* und *ℓ-Diversity* bestimmen kann. Hier konnte für die geprüften Szenarien für die synthetischen Daten ein höherer *ℓ-Wert* nachgewiesen werden als für die realen Daten, wobei der *k-Wert* in den realen Daten höher war.

In einer zweiten Funktion wurden die Übereinstimmungen von *Quasi-Identifikatoren* von verschiedenen Anzahlen kategorischer Merkmale aus beiden Datensätzen ermittelt. Durch die zusätzliche Erstellung einer Kontrollgruppe aus dem realen Datensatz konnte das *Relative Risiko* und das *Odds Ratio* berechnet werden, ob ein Objekt mit den berücksichtigten Merkmalen in beiden Datensätzen identisch ist. Das *Relative Risiko*, dass ein identisches Objekt identifiziert wird, ist in der Kontrollgruppe etwa zweimal so hoch. Die Chance ist für die Kontrollgruppe fast dreimal höher.

Um einen Abgleich der Übereinstimmungen mit einem unabhängigen Datensatz durchzuführen, wurde der Vergleich mit einem vor der Synthese zurückgehaltenen Datensatz wiederholt. Für den Abgleich auf Eins-zu-Eins-Übereinstimmungen wurden weniger Objekte berücksichtigt und das Risiko für eine Übereinstimmung war mit ∼2 % etwas höher als im Abgleich des synthetischen Datensatzes gegen den Gesamtdatensatz.

5.2 **Diskussion der Zielerreichung**

Ziel 1: Zusammenstellung qualitativer und quantitativer Anforderungen an Art, Umfang und Qualität der zu generierenden synthetischen Forschungsdaten.

Um die qualitativen und quantitativen Anforderungen an den synthetischen Datensatz zusammenzustellen, sind in Kapitel 2.5 die Anforderungen an die synthetischen Daten beschrieben und in der Anforderungsliste, Anhang A - Anforderungsliste (Tabelle 25) zusammengetragen worden. Vorab wurde in Kapitel 2.2.1 der Begriff Datenqualität mit der Schlussfolgerung definiert, dass die Datenqualität für jedes Vorhaben individuell definiert werden muss. Der generierte Datensatz sollte sich durch eine identische Datenmenge und Datenstruktur und eine möglichst ähnliche statistische Verteilung zwischen dem realen und synthetischen Datensatz auszeichnen. Außerdem sollte das Re-Identifi-

zierungs-Risiko, durch eine möglichst hoch erreichte *ℓ-Diversity*, im synthetischen Datensatz möglichst gering sein. Rückschlüsse zu Objekten aus dem realen Datensatz sollten nicht möglich sein.

Ziel 2: Identifikation und vergleichende Darstellung geeigneter Tools für die Generierung synthetischer Daten und Auswahl eines Favoriten nach den in Ziel 1 festgelegten Anforderungen.

In Kapitel 2.4.1 wurden mögliche Verfahren, wie das BN, das GAN und die Erstellung parametrischer und nicht parametrischer Modelle zur Erzeugung synthetischer Daten vorgestellt. Anschließend wurden vier Anwendungen zur Erzeugung synthetischer Daten mit ihren wichtigsten Eigenschaften in Kapitel 2.4.2 beschrieben.

Für die Auswahl einer dieser Anwendungen wurde eine objektive Entscheidung getroffen. Das Vorgehen der dafür durchgeführten Nutzwertanalyse wurde in Kapitel 3.3 beschrieben und das Ergebnis in Kapitel 4.1 präsentiert. Die Auswahl fiel auf die Anwendung *DataSynthesizer*, dessen Bereitstellung in Kapitel 3.4.5 erläutert wurde.

Ziel 3: Synthese eines neuen Datensatzes auf Basis der bereitgestellten Eingangsdaten aus der *MIMIC-III*-Datenbank.

Für die Generierung der synthetischen Daten ist ein Teildatensatz aus der *MIMIC-III*-Datenbank verwendet worden. Um einen Überblick über den verwendeten Teildatensatz zu erhalten, wurde dieser in einem Data Dictionary in Kapitel 4.2 und in Tabelle 26 beschrieben und zusammengefasst.

Die Synthese ist nach dem Vorgehen von McLachlan vorgenommen worden, das in Kapitel 3.1 schrittweise beschrieben wurde. Während des Vorgehens kam es bei der Validierung der Ergebnisdaten zu Unvollkommenheiten in der Datenstruktur des synthetischen Datensatzes, die in Kapitel 4.3 beschrieben wurden. Um die aufgetretenen Probleme des Eingangsdatensatzes zu lösen, wurde ein zum Erfolg führender Längsschnitt der Daten erstellt, dessen Generierung in Kapitel 4.4 erläutert wurde. Jedoch

lagen der Erstellung des Längsschnittes Einbußen des Datenumfangs zugrunde, die für den Verwendungszweck als akzeptabel angenommen wurden.

Ziel 4: Analyse, Beschreibung und Vergleich des ursprünglichen und des synthetischen Datensatzes in Hinblick auf die im Rahmen von Ziel 1 erarbeiteten Anforderungen an die Qualität der Daten.

Die Informationen und Evaluationsmetriken zur Prüfung der Datenqualität und zum Vergleich der statistischen Ähnlichkeit des realen Datensatzes und des synthetischen Datensatzes wurden in Kapitel 2.2.3 aufgelistet.

Für den Vergleich des synthetischen Datensatzes mit dem ursprunggebenden Datensatz wurden in Kapitel 4.6 die deskriptiven Statistiken, die *Paarweise Korrelations-Differenz*, der *Kolmogorov-Smirnov-Test,* die *Kullback-Leibler-Divergenz* und die *Log-Cluster Metrik* angewendet. Zusätzlich wurde eine Differenz des *Cramérs V*, analog zur *Paarweisen-Korrelations-Differenz*, für den Vergleich der kategorischen Merkmale erstellt und interpretiert. Nicht alle Tests eigneten sich aufgrund der fehlenden Werte für die Anwendung auf die verwendeten Daten, wie in Kapitel 3.2.3 erläutert wurde.

Aus den Analysen ging hervor, dass die statistischen Verteilungen der kategorischen Merkmale beider Datensätze nicht signifikant unterschiedlich sind bzw. nur geringe Unterschiede vorhanden sind. Für die Verteilung der numerischen Daten konnte mit den Tests aufgrund von fehlenden Werten keine zuverlässige Aussage getroffen werden. Jedoch wiesen die statistischen Kenngrößen, in Kapitel 4.5.2 (Tabelle 18) und Kapitel 4.6.2 (Tabelle 19) keine erheblichen Unterschiede nach. Die Häufigkeitsverteilung der einzelnen Merkmale unterscheidet sich etwas, sodass im synthetischen Datensatz in den Folgeaufenthalten eine höhere Anzahl der Häufigkeiten vorliegt. Die prozentuale Verteilung einiger Ausprägungen der kategorischen Merkmale in Tabelle 20 unterscheidet sich ebenfalls. Jedoch liegt für einen Großteil der Kategorien in beiden Datensätzen eine identische prozentuale Verteilung vor.

Zusätzlich wurden für die kategorischen Merkmale in Kapitel 4.6.6 die *Log-Cluster-Metrik* angewandt, um deren Verteilung zwischen den beiden Datensätzen zu vergleichen. Mit dem Histogramm in Abbildung 22 konnte eine ähnliche Verteilung der Objekte in die Cluster nachgewiesen werden. Zusätzlich wurde ein *UC*-Wert von -8,41 ermittelt, der sich wegen seiner Einordnung in den möglichen Wertebereich als schwer interpretierbar heurausstellte.

Ziel 5: Untersuchung der Güte des erstellten Modells in Bezug auf die Einhaltung des Datenschutzes (Ausschluss eines Personenbezugs) der in Ziel 3 generierten synthetischen Daten.

Auch die Methoden zur Bewahrung des Datenschutzes wurden in Kapitel 2.3 in Tabelle 4 zusammengestellt. Anschließend wurden gängige Anonymisierungs-Techniken und gängige damit verbundene Termini erläutert.

Für die Prüfung des Offenlegungsrisikos wurde, wie in Kapitel 3.6.1 beschrieben eine Funktion entwickelt, die sowohl die ℓ*-Diversity* als auch die *k-Anonymität* für verschiedene Szenarien bestimmen kann. Die Ergebnisse der Szenarien wurden in Kapitel 4.7.1 präsentiert. Hier lagen für alle Szenarien im synthetischen Datensatz höhere ℓ-Werte vor als für den realen Datensatz, während die *k-Werte* im realen Datensatz höher waren. Die höhere *k-Anonymität* lässt sich damit erklären, dass die Daten bereits einen Anonymisierungsprozess durchlaufen haben. Eine Privatheitswahrung der synthetischen Daten gegenüber den realen Daten konnte bei diesem Vorgehen nicht nachgewiesen oder bewertet werden.

In einer zweiten Funktion wurden Objekte ermittelt, deren kategorische Merkmale in beiden Datensätzen übereinstimmen, wie in Kapitel 4.7.2 beschrieben wurde. Das *Relative Risiko* für den Vergleich aus dem realen Datensatz (Kontrollgruppe) war mehr als doppelt (2,34-fach) so hoch. Das *Odds-Ratio* für die Identifikation eines identischen Objektes aus der Kontrollgruppe lag 2,90-fach höher.

Aus den Ergebnissen ließ sich schließen, dass die *Eins-zu-Eins*-Übereinstimmungen für die synthetischen Daten geringer sind, ein Offenlegungsrisiko konnte jedoch mit beiden erstellten Funktionen nicht ermittelt werden.

In einem weiteren Vergleich wurde ein synthetischer Datensatz mit einem vor der Synthese zurückgehaltenen Teildatensatz verglichen. Die Anteile der im Abgleich konfrontierten Objekte bei der Prüfung war kleiner, aber der prozentuale Anteil der *Eins-zu-Eins*-Übereinstimmungen war um etwa 2 % höher. Bei den *Eins-zu-Eins*-Übereinstimmungen zwischen den beiden voneinander unabhängigen Datensätzen könnte es sich jedoch um einen Zufall handeln. Da dieser Vergleich nur einmal durchgeführt wurde, besteht keine Referenz.

Hier sollte auch nicht außer Betracht gelassen werden, dass bei einer endlichen Anzahl an möglichen Kombinationen die Möglichkeit einer vollständigen Übereinstimmung aller Merkmale mit wachsender Probengröße zunimmt. Je mehr Objekte in einem synthetischen Datensatz vorliegen würden, desto mehr *7-fach*-Übereinstimmungen gäbe es mit dem ursprunggebenden Datensatz. Es kann nicht ausgeschlossen werden, dass bei einem ausreichend großem synthetischen Datensatz alle möglichen Kombinationen von Attributen vorhanden wären. Damit wären dann so viele Objekte mit identischen *Quasi-Identifikatoren* vorhanden, wie es Objekte im realen Datensatz gibt.

Auch zu berücksichtigen ist, dass die numerischen Parameter (Laborparameter) und die Anzahl der Folgeaufenthalte in der Prüfung nicht berücksichtigt wurden. Wird das gesamte Objekt betrachtet, würde sich die Vielfalt der einzelnen Objekte erhöhen. So können z. B. Objekte mit identischen *Quasi-Identifikatoren* vorliegen, deren Anzahl von Krankenhausaufenthalten und die darin enthaltenen Laborparameter unterschiedliche Werte enthalten.

5.3 Diskussion

In den folgenden Abschnitten werden die erarbeiteten Ergebnisse und die Methodik zur Erreichung der Ziele in dieser Arbeit reflektiert und diskutiert. Die Diskussion ist in fünf verschiedene Themen aufgeteilt.

5.3.1 Literatur

Bei der für die Grundlagenermittlung durchgeführten Literaturrecherche wurde eine Abfragesyntax erstellt, welche sowohl mit Freitext als auch mit MeSH-Termen suchte. Hierbei fiel auf, dass es für die verwendeten Begriffe *Datenqualität (Data Quality)* und *Medizinische Forschungsdaten (Medical research data)* keine MeSH-Terme vorhanden waren. Alternativ wurde für den Begriff *Data Quality* der MeSH-Term *Data accurancy* berücksichtigt, wobei dies nur einen Teil des ursprünglichen Begriffs einschließt. Anstatt *Medical research data* wurde der MeSH-Term *Big Data* in die Suchsyntax aufgenommen, wobei dieser Term dem eigentlichen Begriff übergeordnet ist. Durch die zusätzliche Freitextsuche führte die Literaturrecherche trotzdem zu dem gewünschten Ergebnis.

5.3.2 Verwendete Daten

Für die Synthese wurden drei der fünf Tabellen aus dem in Tabelle 25, Punkt 2 definierten Datensatz verwendet. In diesem verwendeten Teildatensatz waren mit den enthaltenen Tabellen Fehlwerte vorhanden. Während die Werte für die kategorischen Attribute vollständig erhoben wurden bzw. durch das Datenmanagement vervollständigt werden konnten, wurden für die Behandlungsfälle nicht alle Parameter erhoben. Ein Management im Umgang mit den fehlenden Werten war im Rahmen der Masterarbeit nicht vorgesehen und aufgrund der Struktur der verwendeten Daten nicht möglich.

Durch die Erstellung eines Längsschnittes, bei dem nur drei Tabellen berücksichtigt wurden, konnte ein synthetischer Datensatz erzeugt werden. Einschränkungen waren, dass im Längsschnitt weniger Objekte vorhanden waren. Durch die Verwendung des *Inner Joins* für die Verbindung der Tabellen entfielen viele der vorhandenen Objekte. Allerdings konnte durch die Verwendung verhindert werden, dass weitere Fehlwerte entstehen. Anstelle des *Inner Joins* kann ein *Outer Join* verwendet werden, um den Verlust

dieser Informationen zu verhindern. Der geringere Datenumfang war für den vorhandenen Verwendungszweck akzeptabel.

Für die Erstellung des Längsschnittes wurden die stündlich erfassten Parameter zu einer Messreihe je Krankenhausaufenthalt zusammengefasst. Bei einer Migration in das ursprüngliche Datenmodell würde zwischen allen Tabellen anschließend eine *1:1*-Beziehung bestehen. Dieser Arbeitsschritt war für die Erstellung des Längsschnittes notwendig, um die Anzahl der enthaltenen Attribute zu verkürzen. Außerdem konnte durch die Zusammenfassung die Anzahl der Fehlwerte reduziert werden. Hier ist jedoch auch zu erwähnen, dass für die erstellten Mittelwerte bereits zusammengefasste Werte verwendet wurden. Tang et al. haben bereits für die Erstellung der *Pivot*-Tabelle arithmetische Mittelwerte der Laborparameter aus den minütigen Messungen erstellt (91).

Die gewählte Lösung, aus den Daten einen Längsschnitt für die Synthese zu erstellen, stellte sich als aufwendig heraus. Außerdem sollte beachtet werden, dass diese Vorgehensweise sowohl in der Archivierung als auch in der Verarbeitung ressourcenintensiv ist, da Informationen in der Tabelle redundant erfasst wurden. So wurde z. B. auch für jedes Attribut im BN ein Knoten erstellt, wodurch sich die Berechnungszeit verlängerte.

Die Wunschanforderung, einen synthetischen, über mehrere Tabellen in Relation stehenden, Datensatz zu erzeugen, dessen Datenmodell sich nicht verändert, konnte mit den verwendeten Methoden nicht erfüllt werden.

5.3.3 **Verwendetes Tool**

Das ausgewählte Tool *DataSynthesizer* war geeignet, synthetische Daten zu generieren, welche die gleichen statistischen und formalen Eigenschaften aufweisen, wie der ursprunggebende Datensatz. Jedoch gab es vor allem Abweichungen zwischen den Verteilungen, wenn zu viele Fehlwerte vorlagen. Auch Ausreißer in den Eingangsdaten bewirkten eine stärkere Streuung bei der Synthese. Vor der Synthese sollte ein gründliches Datenmanagement durchgeführt werden, indem Ausreißer entfernt werden, um Streuungen in den synthetischen Daten zu verhindern.

Das Tool eignete sich jedoch weniger dazu, in Relation stehende Daten zu synthetisieren. Hier wurde bei der Entwicklung der Fokus auf die Synthese einer festgelegten Auswahl von Daten gelegt.

Die Zeitdaten und Altersangaben der synthetischen Objekte erhielten bei der Synthese keine chronologische bzw. aufsteigende Reihenfolge. Deshalb wurden die Zeitdaten bei der Analyse nicht weiter berücksichtigt. Die nicht plausiblen Altersangaben wurden bei der Analyse ignoriert. Der Grund dafür könnte darin liegen, dass durch das BN keine Zusammenhänge zwischen den Zeitdaten (Chronologie) erkannt wurden und die aufsteigende Zeitfolge somit nicht nachgebildet werden konnte. Ein möglicher Lösungsansatz wird in Kapitel 5.4 erwähnt. Anhand der fehlenden Plausibilitäten in den chronologischen Reihenfolgen könnte bei einem möglichen Angriff erkannt werden, dass der vorliegende Datensatz durch eine Synthese generiert wurde.

Die Benutzung der Software *DataSynthesizer* bestätigte die intuitive Bedienbarkeit, so dass keine lange Einarbeitung nötig war und zuverlässige Resultate erbracht wurden. Auch die Bereitstellung des Tools konnte sowohl problemlos über *WSL* als auch im *Jupyter Notebook* umgesetzt werden. Bei der Nutzung großer Datenmengen bzw. -strukturen erzeugte die Verwendung über die Webanwendung manchmal Fehler.

Die Bereitstellung der Tools *MedGAN* und *SynthEHR* war aufgrund nicht mehr verfügbarer Bibliotheken nicht möglich. Das vierte in der Nutzwertanalyse berücksichtigte Tool *SynthPop* konnte problemlos installiert werden. Jedoch fiel die Auswahl bei der Nutzwertanalyse nicht auf das Tool *SynthPop*, welches vor allem für das Kriterium einer „Intuitiven Bedienung“ weniger Punkte erhielt als das Tool *DataSynthesizer*.

5.3.4 **Datenqualität**

Das übergeordnete Ziel einen synthetischen Datensatz zu erzeugen, dessen Datenqualität und statistische Eigenschaften mit dem des ursprunggebenden Datensatzes übereinstimmen, konnte bis auf einige Vorbehalte erreicht werden. Mit den Vergleichstests, wie z. B. dem KS-Test und der KL-Divergenz, konnten in den statistischen Verteilungen vor

allem für die kategorischen Merkmale keine bzw. nur minimale Unterschiede nachgewiesen werden. Lagen für die Merkmale viele Fehlwerte vor, resultierten daraus in deren Häufigkeitsverteilungen Differenzen. Auch die Zuordnung in die Cluster bei der *Log-Cluster-Metrik* wurde eine ähnliche Verteilung der kategorischen Merkmale nachgewiesen.

Die Ergebnisse der Vergleichstests für die numerischen Merkmale lieferten zum Teil nicht vollständig belastbare und teilweise entgegengesetzte Ergebnisse. Aufgrund der Fehlwerte in den Datensätzen wurde die Anzahl der Werte in beiden Datensätzen für die Durchführung der Tests angeglichen. Dadurch wurden Werte zum Teil nicht für die Tests berücksichtigt und ließen stärkere Abweichungen in den Tests entstehen. Bei dem Vergleich der statistischen Kenngrößen der numerischen Merkmale konnte eine starke Übereinstimmung festgestellt werden. Jedoch konnte eine im Vergleich größere Streuung der synthetischen Daten festgestellt werden. Diese entstand wohlmöglich aufgrund der in den realen Daten vorhandenen Ausreißer.

Insgesamt waren für die Merkmale im synthetischen Datensatz eine größere Anzahl erfasster Werte vorhanden als im realen Datensatz. So konnten, durch die Erhöhung der Häufigkeiten, die paarweisen Korrelationen für den synthetischen Datensatz erstellt werden, die im Vergleich des realen Datensatzes nicht möglich waren oder Scheinkorrelationen berechnet wurden. Auch für andere statistische Tests könnten mit einer höheren Datendichte zuverlässigere Resultate erzielt werden. Obwohl die statistischen Verteilungen beider Datensätze ähnlich blieben, könnte eine höhere Anzahl der Häufigkeiten im synthetischen Datensatz auf eine Verzerrung der Eingangsdaten hinweisen.

Bei der *Log-Cluster Metrik* wurde unter der Berücksichtigung von 20 Clustern bzw. 4 Clustern eine ähnliche Verteilung in den Histogrammen der Objekte beider Datensätze in die verschiedenen Clustern festgestellt. Der ermittelte U_C-Wert von -8,41 bzw. -10,57 ist als Metrik in den möglichen Wertebereich schwer einzuordnen. Bei der angewendeten Formel kann ein Maximalwert von -0,6 bei unterschiedlichen und $-\infty$ bei genau identischen Datensätzen angenommen werden. Hinweise zur Interpretation finden sich in der Literatur nicht.

Die Qualität der verwendeten realen Daten (Eingangsdaten) war in Hinblick auf deren Vollständigkeit, Genauigkeit und Glaubwürdigkeit eingeschränkt. Beim Datenmanagement konnte für die Extremwerte nicht überall zweifelsfrei festgelegt werden, ob die Werte plausibel sein könnten. Mit den Ergebnissen aus der Prüfung der Datenqualität konnte belegt werden, dass bei der Aufbereitung der Eingangsdaten für die Synthese auf eine gute Datenqualität geachtet werden sollte. Diese bezieht sich vor allem auf die Genauigkeit, Vollständigkeit, Verständlichkeit, kompakte Darstellung und Interpretierbarkeit der Daten.

5.3.5 **Privatheit**

Im synthetischen Datensatz konnten keine Rückschlüsse über die *direkten Identifikatoren* (Identifikationsnummern) auf den realen Datensatz geschlossen werden, da diese durch einen neuen *Auto Increment* Index ersetzt wurden.

Die Ergebnisse der Analysefunktionen bestätigten, dass für den synthetischen Datensatz ein geringeres Risiko für eine *Eins-zu-Eins*-Übereinstimmung und eine höhere *ℓ-Diversity* vorlag. Dennoch konnte nicht eindeutig geklärt werden, ob ein Offenlegungsrisiko besteht und wie hoch es dann wäre. Das berechnete *Relative Risiko* bezieht sich ausschließlich darauf, ob ein identischer Datensatz in den Vergleichen vorliegt.

Der Informationsgehalt der synthetischen Daten konnte jedoch mit dem realen Datensatz verglichen werden und erhalten bleiben, ohne dass Daten für die De-Identifizierung zusammengefasst oder verrauscht wurden. Emam et al. entwickelten ein ausführliches Modell, um das Offenlegungsrisiko zu ermitteln (57). Die Umsetzung dieses Modells war in die Anforderungen nicht aufgenommen und wurde erst spät im Verlauf der Arbeit als eventuell geeignet identifiziert.

5.4 **Ausblick und Abschluss der Arbeit**

In dieser Arbeit wurden mögliche Methoden beschrieben, eingesetzt und entwickelt, um die Datenqualität synthetischer Daten zu prüfen und Informationen über eine mögliche

Re-Identifizierung zu erhalten. Auch Methoden für eine Generierung synthetischer Daten wurden identifiziert und beschrieben.

Aus den Ergebnisse dieser Masterarbeit lässt sich entnehmen, dass bei der Erstellung synthetischer Daten eine gute Datenqualität bei den Eingangsdaten vorausgesetzt werden muss. Auch Fehlwerte können die Datenqualität synthetischer Daten in Bezug auf die Häufigkeitsverteilung beeinträchtigen und die Belastbarkeit statistischer Tests verringern. Die Autoren Tucker et al. haben sich in ihrer Publikation auf einen möglichen Umgang mit fehlenden Werten bei der Generierung synthetischer Daten fokussiert. In ihrer Arbeit beschreiben sie ein Vorgehen, bei dem eine zusätzliche binäre Variable angelegt wird, die zu jedem Merkmal angibt, ob ein Fehlwertwert vorliegt. Die Variable wird als Knoten in das BN aufgenommen, sodass diese Information bei der Generierung zusätzlich berücksichtigt wird (13).

Der Vergleich bzw. die Ermittlung des Offenlegungs-Risikos zwischen dem ursprunggebenden Datensatz und dem synthetischen Datensatzes könnte in nachfolgenden Studien durch den Einsatz von *Data Mining*-Testverfahren vertiefend analysiert werden. Dazu könnten Verfahren wie das *k-Means-Clustering* oder *Kreuzvalidierungsverfahren* eingesetzt werden, in denen z. B. Ähnlichkeitsanalysen zwischen dem realen und synthetischen DS durchgeführt werden.

Mit den derzeit offen verfügbaren Tools ist es im Rahmen dieser Masterarbeit nicht gelungen, tabellenübergreifende Relationen in der Synthese zu berücksichtigen und nachzubilden. Hier besteht weiterer Forschungsbedarf, damit langfristig synthetische Daten, deren Objekte über mehrere Tabellen in Verbindung stehen und plausible Werte enthalten, ohne erheblichen Aufwand erzeugt werden können.

In den synthetischen Daten wurden Zeitreihen generiert, welche nicht in chronologischer Reihenfolge waren oder deren Zeitpunkte über längere Zeiträume variieren. Auch für die Altersangaben konnten durch das Tool keine Regeln erlernt werden, wodurch die Werte einen aufsteigenden Trend erhalten hätten.

Eine Möglichkeit, die fehlende Chronologie zwischen den Zeitdaten in den Objekten zu umgehen, wäre, die Angabe der Zeitdifferenz zwischen den Krankenhausaufenthalten in Tagen auszudrücken. Hier ist zu erwarten, dass das BN die Differenzwerte zwischen den unterschiedlichen Objekten verteilt, ohne einen Zusammenhang zum Datum der Erstaufnahme zu erhalten. Jedoch wären die Differenzen zwischen den Krankenhausaufenthalten innerhalb der Objekte plausibel.

Durch die Erhöhung der Komplexität bzw. der Elternknoten des Bayes'schen Netzes könnten komplexere Korrelationen zwischen den Knoten berücksichtigt werden. Dies wäre aber nur mit erheblich gesteigerter Rechenleistung möglich.

Um das Offenlegungsrisiko von synthetischen Daten zu berechnen, könnten die im Rahmen dieser Arbeit entworfenen Funktionen für die Prüfung der Privatheitswahrung weiterentwickelt werden. Diese sind grundsätzlich auch für andere Datensätze einsetzbar. Hier könnte an das von Emam et al. entwickelte Modell zur Ermittlung des Offenlegungsrisikos angeknüpft werden (57). In dem Modell werden neben der Prüfung von Übereinstimmungen zwischen den beiden Datensätzen weitere Komponenten eingeschlossen. Diese sind z. B. die Chance des Angreifers, neue Informationen zu erhalten (57).

Um auszuschließen, dass es sich bei der Anzahl der *Eins-zu-Eins*-Übereinstimmungen im Vergleich mit einem unabhängigen Datensatzes um einen Zufall handelt, könnte dieser Vergleich mehrfach durchgeführt werden. Dafür kann der Vergleich mit mehreren zufällig erstellen Splits und aus dem Restdatensatz generierten synthetischen Daten durchgeführt und verglichen werden.

Auch könnte die Funktion zur Ermittlung der *k-Anonymität* bzw. *ℓ-Diversity* um die Bestimmung der *t-Closeness* ergänzt werden, um diese für verschiedene Szenarien zusätzlich zu prüfen.

Für Forschungszwecke realitätsnahe, synthetische Daten zu nutzen, ist ein guter Weg, um die Privatsphäre ggf. betroffener Personen vor einer potenziellen Gefährdung oder Diskriminierung durch die Offenlegung personenbezogener Daten zu schützen.

Zusätzlich kann durch die Generierung synthetischer Daten ohne das Zusammenfassen oder Verrauschen der Daten ein höherer Informationsgehalt in den Daten erhalten bleiben, als es bei anonymisierten Daten mit herkömmlichen Techniken möglich wäre. Das bestätigt die zunehmende Bedeutung synthetischer Daten in retrospektiven Analysen, sowohl für den Erhalt einer hohen Datenqualität als auch für den Datenschutz und die Datenethik.

Literaturverzeichnis

1. Price WN, Cohen IG. Privacy in the Age of Medical Big Data. Nat Med. Januar 2019;25(1):37–43.

2. Khan R, Tao X, Anjum A, Sajjad H, Malik S ur R, Khan A, u. a. Privacy Preserving for Multiple Sensitive Attributes against Fingerprint Correlation Attack Satisfying c-Diversity. Wirel Commun Mob Comput. 28. Januar 2020;2020:e8416823.

3. Koch H, Schütze B, Spyra G, Welfer M. Datenschutzrechtliche Anforderungen an die medizinische Forschung unter Berücksichtigung der EU Datenschutz-Grundverordnung (DS-GVO) [Internet]. 2017 [zitiert 28. März 2021]. Verfügbar unter: https://www.gdd.de/arbeitskreise/datenschutz-und-datensicherheit-im-gesundheits-und-sozialwesen/materialien-und-links/datenschutzrechtliche-anforderungen-an-die-medizinische-forschung-unter-beru-cksichtigung-der-eu-datenschutz-grundverordnung/datenschutzrechtliche-anforderungen-an-die-medizinische-forschung-unter-beru-cksichtigung-der-eu-datenschutz-grundverordnung/view

4. Art. 4 DSGVO – Begriffsbestimmungen [Internet]. Datenschutz-Grundverordnung (DSGVO). [zitiert 28. März 2021]. Verfügbar unter: https://dsgvo-gesetz.de/art-4-dsgvo/

5. Burtscher B. Anonymisierung personenbezogener Daten - Ein branchenübergreifender Praxisleitfaden für Industrieunternehmen [Internet]. Berlin, Germany; 2003 Nov [zitiert 8. März 2021] S. 40. Verfügbar unter: https://bdi.eu/publikation/news/anonymisierung-personenbezogener-daten/

6. Datenschutz-Grundverordnung: DSGVO als übersichtliche Seite [Internet]. Verfügbar unter: https://dsgvo-gesetz.de/

7. Mustertext zur Patienteneinwilligung | Medizininformatik-Initiative [Internet]. [zitiert 23. April 2021]. Verfügbar unter: https://www.medizininformatik-initiative.de/de/mustertext-zur-patienteneinwilligung

8. Europäischer Gerichtshof. Urteil des EuGH vom 13. Mai 2014 in der Rechtssache C-131/12 [Internet]. [zitiert 23. April 2021]. Verfügbar unter: https://curia.europa.eu/juris/document/document.jsf;jsessionid=1D5BA68CCC1EAAAFEBAFC3A61DC46DD6?text=&docid=152065&pageIndex=0&doclang=DE&mode=lst&dir=&occ=first&part=1&cid=10521778

9. Datenschutzerklärung und ärztliche Aufbewahrungsfristen-p-9724.pdf [Internet]. [zitiert 31. März 2021]. Verfügbar unter: https://www.kvn.de/internet_media/Mitglieder/Beratung/IT+in+der+Arztpraxis/EU_Datenschutz_Grundverordnung/Datenschutzerkl%C3%A4rung_+%C3%84rztliche+Aufbewahrungsfristen-p-9724.pdf

10. Erwägungsgrund 26 - Keine Anwendung auf anonymisierte Daten [Internet]. Datenschutz-Grundverordnung (DSGVO). [zitiert 28. März 2021]. Verfügbar unter: https://dsgvo-gesetz.de/erwaegungsgruende/nr-26/

11. Sweeney L. k-anonymity: a model for protecting privacy. Int J Uncertain Fuzziness Knowl-Based Syst. Oktober 2002;10(05):557–70.

12. Ländervergleich Vereinigte Staaten von Amerika : Deutschland [Internet]. Laenderdaten.info. [zitiert 7. April 2021]. Verfügbar unter: https://www.laenderdaten.info/laendervergleich.php?country1=USA&country2=DEU

13. Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. Npj Digit Med. 9. November 2020;3(1):1–13.

14. Drechsler DJ, Jentzsch DN. Synthetische Daten - Innovationspotential und gesellschaftliche Herausforderungen [Internet]. Berlin: Stiftung Neue Verantwortung e. V.; 2018 S. 27. Verfügbar unter: https://www.stiftung-nv.de/sites/default/files/synthetische_daten.pdf

15. Foraker RE, Yu SC, Gupta A, Michelson AP, Pineda Soto JA, Colvin R, u. a. Spot the difference: comparing results of analyses from real patient data and synthetic derivatives. JAMIA Open. Dezember 2020;3(4):557–66.

16. Datenethikkommission der Bundesregierung, Bundesministerium des Innern, für Bau und Heimat. Gutachten der Datenethikkommission [Internet]. Berlin; 2019 S. 240. Verfügbar unter: www.datenethikkommission.de

17. Walonski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, u. a. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record | Journal of the American Medical Informatics Association | Oxford Academic. 30. August 207 n. Chr. [zitiert 26. April 2021]; Verfügbar unter: https://academic.oup.com/jamia/article/25/3/230/4098271

18. Azizi Z, Zheng C, Mosquera L, Pilote L, Emam KE. Can synthetic data be a proxy for real clinical trial data? A validation study. BMJ Open. 1. April 2021;11(4):e043497.

19. Rubin DB. Discussion Statistical Disclosure Limitation. 1993; Verfügbar unter: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf

20. Datenethikkommission [Internet]. Bundesministerium der Justiz und für Verbraucherschutz. [zitiert 5. April 2021]. Verfügbar unter: https://www.BMJV.de/DE/Themen/FokusThemen/Datenethikkommission/Datenethikkommission_node.html

21. Baowaly MK, Lin C-C, Liu C-L, Chen K-T. Synthesizing electronic health records using improved generative adversarial networks. J Am Med Inform Assoc. 1. März 2019;26(3):228–41.

22. Luber S. Was ist ein Generative Adversarial Network (GAN)? [Internet]. [zitiert 28. April 2021]. Verfügbar unter: https://www.bigdata-insider.de/was-ist-ein-generative-adversarial-network-gan-a-999817/

23. Dube K, Gallagher T. Approach and Method for Generating Realistic Synthetic Electronic Healthcare Records for Secondary Use. In: Gibbons J, MacCaull W, Herausgeber. Foundations of Health Information Engineering and Systems. Berlin, Heidelberg: Springer; 2014. S. 69–86. (Lecture Notes in Computer Science).

24. McLachlan S, Dube K, Gallagher T. Using the CareMap with Health Incidents Statistics for Generating the Realistic Synthetic Electronic Healthcare Record. In: 2016 IEEE International Conference on Healthcare Informatics (ICHI). 2016. S. 439–48.

25. Synthea [Internet]. syntheticheath; 2021 [zitiert 6. März 2021]. Verfügbar unter: https://github.com/synthetichealth/synthea

26. FHIR – Hl7wiki [Internet]. [zitiert 8. März 2021]. Verfügbar unter: http://wiki.hl7.de/index.php?title=FHIR

27. Weyer J, Roos M. Agentenbasierte Modellierung und Simulation. TATuP Z Für Tech Theor Prax. 14. Dezember 2017;26:11.

28. About MDClone - A Technology Company Powering Healthcare Organizations [Internet]. MDClone. [zitiert 26. April 2021]. Verfügbar unter: https://www.mdclone.com/about-mdclone

29. PySyft [Internet]. OpenMined; 2021 [zitiert 6. März 2021]. Verfügbar unter: https://github.com/OpenMined/PySyft

30. Welcome to Python.org [Internet]. Python.org. [zitiert 8. März 2021]. Verfügbar unter: https://www.python.org/

31. Table of contents [Internet]. Anonimatron. [zitiert 6. März 2021]. Verfügbar unter: https://realrolfje.github.io/anonimatron/documentation/

32. Screen Reject – Teilprojekt 3: Klinisches Data Warehouse zur Abstoßungsdiagnostik nach NTx [Internet]. [zitiert 14. Dezember 2020]. Verfügbar unter: http://screen-reject.f3.hs-hannover.de/

33. Hochschule Hannover. Smart Data Analytics : Schriften des Forschungsclusters Smart Data Analytics 2020 [Internet]. Angewandte Forschung für die Welt von morgen. Hochschule Hannover; 2020 [zitiert 9. März 2021]. Verfügbar unter: https://serwiss.bib.hs-hannover.de/1789

34. openehr.org/about/what_is_openehr [Internet]. [zitiert 5. April 2021]. Verfügbar unter: https://www.openehr.org/about/what_is_openehr

35. MIMIC-III Critical Care Database [Internet]. [zitiert 5. April 2021]. Verfügbar unter: https://mimic.physionet.org/about/mimic/

36. MeSH [Internet]. [zitiert 10. Juni 2021]. Verfügbar unter: https://www.dimdi.de/dynamic/de/klassifikationen/weitere-klassifikationen-und-standards/mesh/

37. PRISMA [Internet]. [zitiert 22. Juni 2021]. Verfügbar unter: http://prisma-statement.org/prismastatement/flowdiagram.aspx

38. Würthele VG. Datenqualitätsmetrik für Informationsprozesse: Datenqualitätsmanagement mittels ganzheitlicher Messung der Datenqualität [Internet] [Doctoral Thesis]. ETH Zurich; 2003 [zitiert 1. Mai 2021]. Verfügbar unter: https://www.research-collection.ethz.ch/handle/20.500.11850/147830

39. Wang RY, Strong DM. Beyond Accuracy: What Data Quality Means to Data Consumers. J Manag Inf Syst [Internet]. 11. Dezember 2015 [zitiert 1. Mai 2021]; Verfügbar unter: https://www.tandfonline.com/doi/abs/10.1080/07421222.1996.11518099

40. Held J. Datenqualität für TestdatenEine Nutzbarkeitsanalyse für Testdatensammlungen. [Nürnberg]: Technische Fakultät der Friedrich-Alexander-UniversitätErlangen-Nürnberg; 2016.

41. Hildebrand K, Gebauer M, Hinrichs H, Mielke M, Herausgeber. Daten- und Informationsqualität [Internet]. Wiesbaden: Springer Fachmedien Wiesbaden; 2018 [zitiert 6. Mai 2021]. Verfügbar unter: http://link.springer.com/10.1007/978-3-658-21994-9

42. Datenqualität - Definition [Internet]. DigitalWiki. 2015 [zitiert 4. Mai 2021]. Verfügbar unter: http://www.digitalwiki.de/datenqualitaet/

43. Helfert M. Planung und Messung der Datenqualität in Data-Warehouse-Systemen [Internet]. [Bamberg]: Universität St. Gallen - Hochschule für Wirtschafts-, Rechts- und Sozialwissenschaften (HSG); 2002. Verfügbar unter: http://dataquality.computing.dcu.ie/thesis/(German).Markus.Helfert.thesis.2002.pdf

44. Benaim AR, Almog R, Gorelik Y, Hochberg I, Nassar L, Mashiach T, u. a. Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies. JMIR Med Inform. 20. Februar 2020;8(2):e16492.

45. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. BMC Med Res Methodol. 7. Mai 2020;20(1):108.

46. Guan J, Li R, Yu S, Zhang X. A Method for Generating Synthetic Electronic Medical Record Text. IEEE/ACM Trans Comput Biol Bioinform. Februar 2021;18(1):173–82.

47. Yoon J, Drumright LN, van der Schaar M. Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). IEEE J Biomed Health Inform. August 2020;24(8):2378–88.

48. Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, Byrd JB, u. a. Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing. Circ Cardiovasc Qual Outcomes. 1. Juli 2019;12(7):e005122.

49. Hershey JR, Olsen PA. Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07. 2007. S. IV-317-IV–320.

50. Weisstein EW. Frobenius Norm [Internet]. Wolfram Research, Inc.; [zitiert 23. Juni 2021]. Verfügbar unter: https://mathworld.wolfram.com/FrobeniusNorm.html

51. What Is a Heatmap? [Internet]. Investopedia. [zitiert 11. September 2021]. Verfügbar unter: https://www.investopedia.com/terms/h/heatmap.asp

52. Haas PDH-D. Definition: Cluster [Internet]. https://wirtschaftslexikon.gabler.de/definition/cluster-30562. Springer Fachmedien Wiesbaden GmbH; [zitiert 30. Juli 2021]. Verfügbar unter: https://wirtschaftslexikon.gabler.de/definition/cluster-30562

53. Luber S. Was ist der k-Means-Algorithmus? [Internet]. [zitiert 22. Juni 2021]. Verfügbar unter: https://www.bigdata-insider.de/was-ist-der-k-means-algorithmus-a-734637/

54. Maschinelles Lernen mit Entscheidungsbaumverfahren – Artikelserie – Data Science Blog [Internet]. [zitiert 23. Juni 2021]. Verfügbar unter: https://data-science-blog.com/blog/2017/02/13/entscheidungsbaumverfahren-artikelserie/

55. Kolmogorov–Smirnov Test. In: The Concise Encyclopedia of Statistics [Internet]. New York, NY: Springer; 2008 [zitiert 19. Juli 2021]. S. 283–7. Verfügbar unter: https://doi.org/10.1007/978-0-387-32833-1_214

56. Wang S, Bonomi L, Dai W, Chen F, Cheung C, Bloss CS, u. a. Big Data Privacy in Biomedical Research. IEEE Trans Big Data. Juni 2020;6(2):296–308.

57. Emam KE, Mosquera L, Bass J. Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation. J Med Internet Res. 16. November 2020;22(11):e23139.

58. Olatunji IE, Rauch J, Katzensteiner M, Khosla M. A Review of Anonymization for Healthcare Data. ArXiv210406523 Cs [Internet]. 13. April 2021 [zitiert 4. Juni 2021]; Verfügbar unter: http://arxiv.org/abs/2104.06523

59. Vimercati S de C di, Foresti S. Quasi-Identifier. In: van Tilborg HCA, Jajodia S, Herausgeber. Encyclopedia of Cryptography and Security [Internet]. Boston, MA: Springer US; 2011 [zitiert 2. Juni 2021]. S. 1010–1. Verfügbar unter: https://doi.org/10.1007/978-1-4419-5906-5_763

60. Äquivalenzrelationen - lernen mit Serlo! [Internet]. Serlo. [zitiert 30. Juli 2021]. Verfügbar unter: https://de.serlo.org/mathe/56795/äquivalenzrelationen

61. Hauf D. K-Anonymity, l-Diversity and T-Closeness [Internet]. Karlsruhe: IPD Uni-Karlsruhe; Verfügbar unter: https://dbis.ipd.kit.edu/img/content/SS07Hauf_kAnonym.pdf

62. Gumz JD. Anonymisierung: Schutzziele und Techniken [Internet]. 2019. Verfügbar unter: https://cdn0.scrvt.com/fokus/784daae14fc72f91/bcebf7142066/Anonymisierung---Schutzziele-und-Techniken.pdf

63. Petrlic R, Sorge C. Datenschutz: Einführung in technischen Datenschutz, Datenschutzrecht und angewandte Kryptographie. Springer Fachmedien Wiesbaden; 2017. 178 S.

64. Templ M, Kowarik A, Meindl B, Alfons A, Ribatet M, Gussenbauer J. Package „simPop“ [Internet]. 2020. Verfügbar unter: https://cran.r-project.org/web/packages/simPop/simPop.pdf

65. Young J, Graham P, Penny R. Using Bayesian Networks to Create Synthetic Data. J Official Stat. 2009;Vol. 25(No. 4):549–67.

66. Directed acyclic graphs (DAGs) - BitcoinWiki [Internet]. [zitiert 30. Juli 2021]. Verfügbar unter: https://de.bitcoinwiki.org/wiki/Directed_acyclic_graphs_(DAGs)

67. Mikrodaten [Internet]. Statistisches Bundesamt. [zitiert 12. Juli 2021]. Verfügbar unter: https://www.destatis.de/DE/Service/Statistik-Campus/ESC/mikrodaten.html

68. Hindupur A. the-gan-zoo [Internet]. 2021 [zitiert 29. Juni 2021]. Verfügbar unter: https://github.com/hindupuravinash/the-gan-zoo

69. Nowok B, Raab GM, Dibben C. synthpop : Bespoke Creation of Synthetic Data in R. J Stat Softw [Internet]. 2016 [zitiert 6. Juli 2021];74(11). Verfügbar unter: http://www.jstatsoft.org/v74/i11/

70. GitHub: Where the world builds software [Internet]. GitHub. [zitiert 5. Juli 2021]. Verfügbar unter: https://github.com/

71. Python 3.0 Release [Internet]. Python.org. [zitiert 5. Juli 2021]. Verfügbar unter: https://www.python.org/download/releases/3.0/

72. Erste Normalform (1NF) | Normalisierung von Datenbanken [Internet]. Datenbanken - für Anfänger und Profis. [zitiert 5. Juli 2021]. Verfügbar unter: https://www.datenbanken-verstehen.de/datenmodellierung/normalisierung/erste-normalform/

73. DataResponsibly/DataSynthesizer [Internet]. Data, Responsibly; 2021 [zitiert 5. Juli 2021]. Verfügbar unter: https://github.com/DataResponsibly/DataSynthesizer

74. Ping H, Stoyanovich J, Howe B. DataSynthesizer: Privacy-Preserving Synthetic Datasets. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management [Internet]. Chicago IL USA: ACM; 2017 [zitiert 5. Juli 2021]. S. 1–5. Verfügbar unter: https://dl.acm.org/doi/10.1145/3085504.3091117

75. Data Responsibly [Internet]. [zitiert 9. August 2021]. Verfügbar unter: http://demo.dataresponsibly.com/synthesizer/

76. Choi E. mp2893/medgan [Internet]. 2021 [zitiert 6. Juli 2021]. Verfügbar unter: https://github.com/mp2893/medgan

77. Nguyen HT. Einführung in die Welt der Autoencoder – Data Science Blog [Internet]. [zitiert 5. Juli 2021]. Verfügbar unter: https://data-science-blog.com/blog/2020/04/01/einfuhrung-in-die-welt-der-autoencoder/

78. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. 19. Mai 2017;20.

79. Nowok B. synthpop: An R package for generating synthetic versions of sensitive microdata for statistical disclosure control. :9.

80. hazy/synthpop [Internet]. Hazy; 2021 [zitiert 7. Juli 2021]. Verfügbar unter: https://github.com/hazy/synthpop

81. Unsere Forschung - ZDIN [Internet]. [zitiert 20. Juli 2021]. Verfügbar unter: https://www.zdin.de/einblicke/unsere-forschung

82. GESUNDHEIT - ZDIN [Internet]. [zitiert 20. Juli 2021]. Verfügbar unter: https://www.zdin.de/zukunftslabore/gesundheit

83. Fundierte Entscheidungen treffen mit der Nutzwertanalyse [Internet]. Projekte leicht gemacht. 2015 [zitiert 22. Juli 2021]. Verfügbar unter: https://projekte-leicht-gemacht.de/blog/pm-methoden-erklaert/nutzwertanalyse/

84. McLachlan S. Realism in synthetic data generation [Internet] [Masterthesis]. [Palmerston North, New Zealand]: (MCSE, MCT, DipSysEng, GradDipInfSc, GradDipLaw, GradDipBus, MIITP, MBCS) School of Engineering and Advanced Technology Massey University; 2017. Verfügbar unter: https://mro.massey.ac.nz/handle/10179/11569

85. Luber S. Was ist eine Pivot-Tabelle? [Internet]. [zitiert 7. August 2021]. Verfügbar unter: https://www.bigdata-insider.de/was-ist-eine-pivot-tabelle-a-841640/

86. Johnson, Alistair, Pollard, Tom, Mark, Roger. MIMIC-III Clinical Database [Internet]. PhysioNet; 2015 [zitiert 16. April 2021]. Verfügbar unter: https://physionet.org/content/mimiciii/1.4/

87. About [Internet]. [zitiert 18. August 2021]. Verfügbar unter: https://physionet.org/about/

88. apakbin94/ICU72hReadmissionMIMICIII: Prediction of ICU Readmissions Using Data at Patient Discharge using MIMICIII Database [Internet]. GitHub. [zitiert 22. September 2021]. Verfügbar unter: https://github.com/apakbin94/ICU72hReadmissionMIMICIII

89. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, u. a. MIMIC-III, a freely accessible critical care database. Sci Data. 24. Mai 2016;3(1):160035.

90. Tang F, Xiao C, Wang F, Zhou J. Predictive Modeling in Urgent Care [Internet]. ILLIDAN Lab; 2020 [zitiert 5. August 2021]. Verfügbar unter: https://github.com/illidanlab/urgent-care-comparative

91. Tang F, Xiao C, Wang F, Zhou J. Predictive modeling in urgent care: a comparative study of machine learning approaches. JAMIA Open. 1. Juli 2018;1(1):87–98.

92. Bohinc T. Grundlagen des Projektmanagements: Methoden, Techniken und Tools für Projektleiter. GABAL Verlag GmbH; 2010. 187 S.

93. Paarweiser Vergleich Nutzwertanalyse incl Excel Vorlage [Internet]. Sixsigmablackbelt.de. 2016 [zitiert 29. Juli 2021]. Verfügbar unter: https://www.sixsigmablackbelt.de/paarweiser-vergleich/

94. Linux-Subsystem in Windows 10 installieren (Bash aktivieren) – so geht's [Internet]. GIGA. 2018 [zitiert 9. August 2021]. Verfügbar unter: https://www.giga.de/downloads/windows-10/tipps/windows-10-bash-aktivieren-installieren-und-oeffnen-so-geht-s/

95. Anaconda | Individual Edition [Internet]. Anaconda. [zitiert 16. Oktober 2021]. Verfügbar unter: https://www.anaconda.com/products/individual

96. developers T pip. pip: The PyPA recommended tool for installing Python packages. [Internet]. [zitiert 25. Oktober 2021]. Verfügbar unter: https://pip.pypa.io/

97. Luber S. Was ist Pandas? [Internet]. [zitiert 11. August 2021]. Verfügbar unter: https://www.bigdata-insider.de/was-ist-pandas-a-950229/

98. Luber S. Was ist NumPy? [Internet]. [zitiert 11. August 2021]. Verfügbar unter: https://www.bigdata-insider.de/was-ist-numpy-a-735687/

99. Luber S. Was ist Seaborn? [Internet]. [zitiert 9. August 2021]. Verfügbar unter: https://www.bigdata-insider.de/was-ist-seaborn-a-977098/

100. Numerisches Python: Einführung in Matplotlib [Internet]. [zitiert 12. August 2021]. Verfügbar unter: https://www.python-kurs.eu/matplotlib.php

101. Luber S. Was ist Scikit-learn? [Internet]. [zitiert 11. August 2021]. Verfügbar unter: https://www.bigdata-insider.de/was-ist-scikit-learn-a-756150/

102. Project Jupyter [Internet]. [zitiert 13. August 2021]. Verfügbar unter: https://www.jupyter.org

103. DataSynthesizer [Internet]. Data, Responsibly; 2021 [zitiert 3. September 2021]. Verfügbar unter: https://github.com/DataResponsibly/DataSynthesizer

104. Sibbertsen P, Lehne H. Statistik: Einführung für Wirtschafts- und Sozialwissenschaftler [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2021 [zitiert 16. Oktober 2021]. Verfügbar unter: https://link.springer.com/10.1007/978-3-662-62696-2

105. Kearney M. Cramér's V. 20. Dezember 2017; Verfügbar unter: https://www.researchgate.net/publication/307963787_Cramer%27s_V

106. IBM Docs [Internet]. 2021 [zitiert 11. September 2021]. Verfügbar unter: https://prod.ibmdocs-production-dal-6099123ce774e592a519d7c33db8265e-0000.us-south.containers.appdomain.cloud/docs/de/cognos-analytics/11.1.0?topic=terms-cramrs-v

107. Kreienbrock L, Pigeot I, Ahrens W. Epidemiologische Methoden [Internet]. Heidelberg: Spektrum Akademischer Verlag; 2012 [zitiert 16. Oktober 2021]. Verfügbar unter: http://link.springer.com/10.1007/978-3-8274-2334-4

108. TensorFlow [Internet]. [zitiert 20. August 2021]. Verfügbar unter: https://www.tensorflow.org/?hl=de

109. Python R. Python Virtual Environments: A Primer – Real Python [Internet]. [zitiert 20. August 2021]. Verfügbar unter: https://realpython.com/python-virtual-environments-a-primer/

110. MIMIC Code Repository [Internet]. MIT Laboratory for Computational Physiology; 2021 [zitiert 20. August 2021]. Verfügbar unter: https://github.com/MIT-LCP/mimic-code

111. Döring N, Bortz J. Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften | Lehrbuch Psychologie [Internet]. 5. Auflage. [zitiert 15. September 2021]. Verfügbar unter: https://lehrbuch-psychologie.springer.com/forschungsmethoden-und-evaluation-den-sozial-und-humanwissenschaften

112. K-Means Clustering: Techniken zum Finden der optimalen Cluster [Internet]. ICHI.PRO. [zitiert 17. Oktober 2021]. Verfügbar unter: https://ichi.pro/de/k-means-clustering-techniken-zum-finden-der-optimalen-cluster-139544899986683

Anhang A - Anforderungsliste

Tabelle 25: Anforderungstabelle

Anf.-ID	Art d. Anf.	Beschreibung	Werte/Daten/Erläuterungen M -Mussanforderung W - Wunschanforderung
1		**Allgemeiner Aufbau**	
1.1	M	MIMIC-III-DS (DS) verwenden	Extraktion aus dem MIMIC-III DS
1.2	W	Analoge Auswahl der Attribute (wie Arbeit aus AP 1.5 (Zukunftslabor))	Im Review beschriebene Attribute verwenden, siehe Punkt 2
1.3	W	Bestehendes Datenmodell erhalten	Im Datenmodell bestehende Beziehungen zwischen den Tabellen bleiben nach der Synthese erhalten
1.4	M	Bestehende Datentypen erhalten	Die im Realdatensatz (RD) definierten Datentypen sollen im synthetischen Datensatz (SD) identisch sein
1.5	M	Plausibilität der Daten	Einleuchtende, realistische Zusammenhänge der Daten
1.6	M	Integrität der Daten	Korrekter Ausgang der Datentypen
1.7	M	Eindeutige Vorgehensweise bei der Erstellung des Modells	Schrittweise Dokumentation der Synthese
1.8	W	Rekonstruierbarkeit(des erstellten Modells)	Mögliche Wiederverwendbarkeit auf den gleichen oder einen anderen DS
1.9	M	Daten in Erster Normalform	Atomare Aufteilung der Informationen
2		**Datenrextraktion aus MIMIC-III enthält:**	Vorhandene Attribute im verwendeten RD
2.1	W	age	Berechnet aus date of birth
2.2	W	ethnicity	Ethnische Herkunft
2.3	W	admission type	Art der Aufnahme
2.4	W	marital status	Familienstand
2.5	W	insurance	Versicherungstyp
2.6	W	religion	Religionszugehörigkeit
2.7	W	gender	Geschlecht
2.8	W	potassium score	Kaliumwert
2.9	W	arterial blood pressure	Arterieller Blutdruck
2.10	W	albumin score	Albuminwert

Anf.-ID	Art d. Anf.	Beschreibung	Werte/Daten/Erläuterungen M -Mussanforderung W - Wunschanforderung
2.11	W	blood urea nitrogen (BUN) score	Blutharnstoff
2.12	W	creatinine score	Kreatininwert
2.13	W	sodium score	Sodium
2.14	W	bicarbonate score	Biocarbonatwert
2.15	W	heart rate	Herzfrequenz
2.16	W	systolic blood pressure	systolischer Blutdruck
2.17	W	temperature	Körpertemperatur
2.18	W	respiratory rate	Atemfrequenz
2.19	W	spo2	Sauerstoffsättigung
2.20	W	glucose level	Blutzucker
2.22	W	physiology score (sapsii)	Maßzahl für den physiologischen Zustand eines Patienten
2.23	W	pao2fio2 score	Verhältnis zwischen dem arteriellen Sauerstoffpartialdruck (PaO2 in mmHg) und dem fraktioniert eingeatmeten Sauerstoff (FiO2 ausgedrückt als Bruchteil, nicht als Prozentsatz)
2.24	W	sirs	Systemic Inflammatory Response Syndrome
2.25	W	organ failure (sofa)	Maßzahl zur Beurteilung des Organversagens bei Sepsis
3		**Zielformat des SD**	
3.1	M	CSV	Ausgabeformat des SD als CSV
3.2	M	Speicherort der Daten	Durchführung der Synthese auf virtueller Maschine der Hochschule Hannover
4		**Statistischer Umfang des SD**	
4.1	M	Gleiche Anzahl zu generierender Objekte wie im RD	Identischer Datenumfang zwischen RD und SD (1:1)
4.2	M	Nahe Übereinstimmung der Deskription der erfassten Merkmale mit dem RD in Bezug auf:	Gleiche statistische Eigenschaften beider DS
4.2.1	M	Arithmetischen Mittelwert (MW)	Geringe Abweichung zwischen den MW der Merkmalen beider DS
4.2.2	M	Standardabweichung	ähnliche Streuung zwischen den Merkmalender Daten beider DS

Anf.-ID	Art d. Anf.	Beschreibung	Werte/Daten/Erläuterungen M -Mussanforderung W - Wunschanforderung
4.2.3	M	Minimum	ähnliches Minimum beider DS
4.2.4	M	Maximum	ähnliches Maximum beider DS
4.2.5	M	Median	ähnlicher Median beider DS
5		**Erfüllung des Datenschutzes**	
5.1	M	k-Anonymität	Erreichen eines möglichst hohen K-Wertes
5.2	M	l-Diversity	Erreichen einer l-Diversity
5.3	W	t-Closeness	Erreichen einer t-Closeness
6		**Datenqualität des SD**	
6.1	M	Teilweise SD ohne Offenlegungs-Risiko	Synthese der Direkten Identifikatoren, Quasi-Identifikatoren und sensiblen Attributen
6.2	W	Vollsynthetischer DS	Synthese aller Attribute
6.3	M	Paarweise Korrelations-Differenz (PKD)	Mindestens paarweiser Vergleich zwischen den Attributen beider DS
6.4	W	Kolmogorow-Smirnow-(KS)Test	Vergleich der Verteilung beider DS anhand des KS-Tests
6.5	W	Kullback-Leibler Divergenz (KLD)	Vergleich beider DS mit KLD
6.6	W	Log-Cluster-Metrik	Vergleich der DS mit Log-Cluster Metrik
6.7	W	Cross-Classification-Maß (2 Metriken: CrCl-RS, CrCl-SR	Detaillierte Analyse mit angegebenen Test
6.8	M	Anzahl vorhandener Objekte 1:1 zum Testdatensatz	Gleicher Datenumfang wie realer DS
6.9	W	Offene Zugänglichkeit ohne Verletzung des Datenschutzes	Frei verfügbarer DS ohne Datenschutzverletzung
7		**Anforderungen an das Tool**	
7.1	W	Integrierbarkeit	Einsetzbar in anderen Systemen wie z. B. SQL Server Data Tool (SSDT)
7.2	M	Open Source	Offener, konfigurierbarer Quellcode
7.3	M	Kostenfreie kommerzielle Verwendbarkeit	Keine erwerbbare Software
7.4	W	Unbegrenzte Anzahl an Daten generieren	Selbstbestimmung der Anzahl generierter synthetischer Objekte

Anf.-ID	Art d. Anf.	Beschreibung	Werte/Daten/Erläuterungen M -Mussanforderung W - Wunschanforderung
7.5	M	Generische Nutzbarkeit/anwendungsfallbezogene Konfigurierbarkeit	Anpassung an die Gegebenheiten anderer DS
7.6	W	Betriebssystem: Windows	Lauffähigkeit auf dem aktuellen Windows-Betriebssystem
7.7	W	Eingangsdaten: CSV	Eingabe der Daten zur Synthese als CSV-Format
7.8	W	Intuitive Bedienbarkeit	Selbsterklärende Steuerung des Tools
7.9	M	Mit aktuell verfügbaren Quellpaketen lauffähig	Vorhandensein für das Tool benötigter Programmbibliotheken
7.10	w	Kontinuierliche Weiterentwicklung	Regelmäßige Überarbeitung und Aktualisierung der verwendeten Software

Anhang B – Data Dictionary

Tabelle 26: Data Dictionary

Attribut	Datentyp	Ausprägungen/ Wertespezifikation	Beschreibung	PK	FK	NN	Index
admissions							
row_id	String		Einzigartiger Identifikator für Zeile	x		x	
subject_id	String		Verbindungsattribut zu *icustay_detail*		x	x	x
hadm_id	Integer		Identifikator für Krankenhausaufenthalt, Verbindung zu *icustay_detail*		x	x	x
admittime	datetime		Erfassungsdatum			x	
admission_type	Varchar()	Electice, emergency, newborn, urgent	Art der Einweisung			x	
insurance	Varchar()	Private, Medicare, Medicaid insurance, Self Pay, Government	Entlassungsstelle			x	
religion	Varchar()	Unknown, catholic, Protestant Quaker, buddist, other, jewish, jehova's witness, Greek orthodox, episcopalian, Christian scientist, Methodist, Unitarian-universalist,	Religionszugehörigkeit			x	

Attribut	Datentyp	Ausprägungen/ Wertespezifikation	Beschreibung	PK	FK	NN	Index
		Hindu, Hebrew, 7th day Adventist, muslim Baptist, Romanian, east orth, lutheran					
marital_status	Varchar()	Divorced, life partner, married, separated, single, unknown, widowed	Familienstand			x	
icustay_detail							
subject_id	Integer		Verbindungsattribut zu admissions		x		
hadm_id	Integer		Verbindungsattribut zu admissions und pivoted_bg		x		
icustay_id	Integer		Verbindungsattribut zu pivoted_lab/pivoted_vital	x			
gender	Varchar()	"M" "F"	Geschlecht			x	
age	Integer		Alter des Patienten				
admittime	datetime		Erfassungsdatum			x	
ethnicity_grouped	Varchar()	"asian", "white", "unknown", "black", "other", "hispanic", "native"	Ethnische Herkunft - gruppierte Darstellung aus der Variable *"ethnicity"*				

Attribut	Datentyp	Ausprägungen/ Wertespezifikation	Beschreibung	PK	FK	NN	Index
pivoted_bg							
hadm_id	String		Verbindungsattribut (df_admissions)		x		
icustay_id	String		Verbindungsattribut zu df_pivoted_lab/df_pivoted_vital		x		
charttime	timestamp		Zeitpunkt des Eintegerrages			x	
spo2	float	%	Sauerstoffsättigung				
fio2	float	%	Inspiratorische Sauerstoffkonzentration				
pao2fio2ratio	float		PaO2/FiO2 Ratio: Verhältnis zwischen dem arteriellen Sauerstoffpartialdruck (PaO2 in mmHg) und dem fraktioniert eingeatmeten Sauerstoff (FiO2 ausgedrückt als Bruchteil, nicht als Prozentsatz)				
bicarbonate	float	mmol/l	Bicarbonatwert				
potassium	float	mmol/l	Kaliumwert				
sodium	float	mmol/l	Natriumwert				
glucose	float	mg/dl	Blutzucker				
pivoted_lab							
icustay_id	String		Identifikator, Verbindungsattribut zu icusty_detail		x		
hadm_id	String		Identifikator, Verbindungsattribut zu icusty_detail, admissions		x		
subject_id	String		Identifikator, Verbindungsattribut zu icusty_detail, admissions		x		
charttime			Erfassungszeit				

Attribut	Datentyp	Ausprägungen/ Wertespezifikation	Beschreibung	PK	FK	NN	Index
albumin	float	g/dl	Albuminwert				
bicarbonate	float	mmol/l	Bicarbonatwert				
creatinine	float	mg/dl	Kreatininwert				
glucose	float	mg/dl	Blutzuckerwert				
potassium	float	mmol/l	Kaliumwert				
sodium	float	mmol/l	Natriumwert				
bun	float	mg/dl	Blutharnstoff				
pivoted_vital							
icustay_id	String		Identifikator, Verbindungsattribut zu admissions		x		
charttime			Erfassungszeit				
heartrate	float	Schläge/min	Herzfrequenz				
sysbp	float	mmHg	Systolischer Blutdruck				
diasbp	float	mmHg	Diastolischer Blutdruck				
resprate	float	Atemzüge/min	Atemfrequenz				
tempc	float	°C	Körpertemperatur				
spo2	float	%	Sauerstoffsättigung				
glucose	float	mg/dl	Blutzucker				

Anhang C – Datenbeschreibung

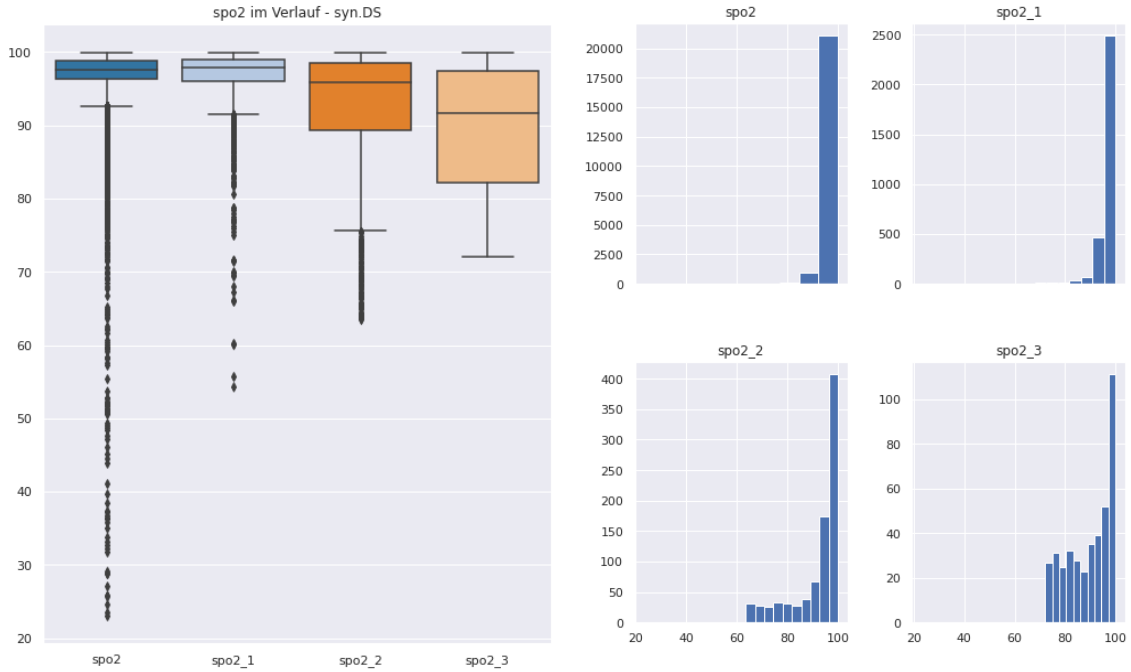


Abbildung 23: Boxplots und Histogramme zur Sauerstoffsättigung (spo2) - synthetischer DS

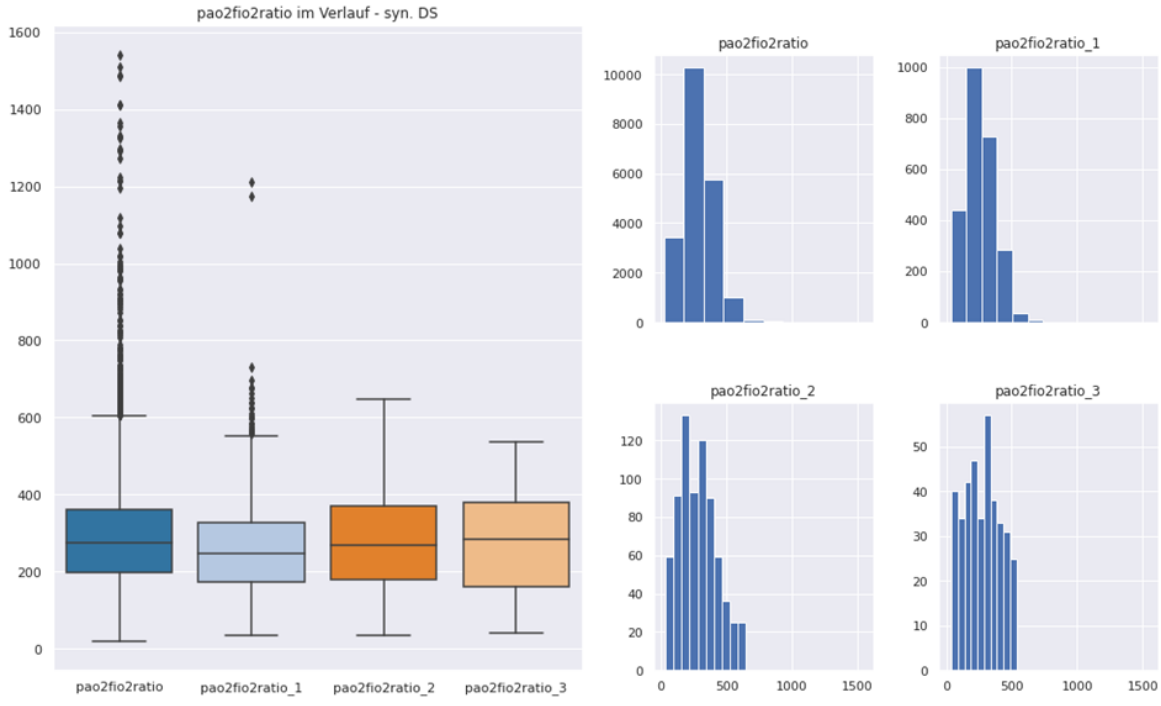


Abbildung 24: Boxplots und Histogramme zum PaO^2/FiO^2-Ratio - synthetischer DS

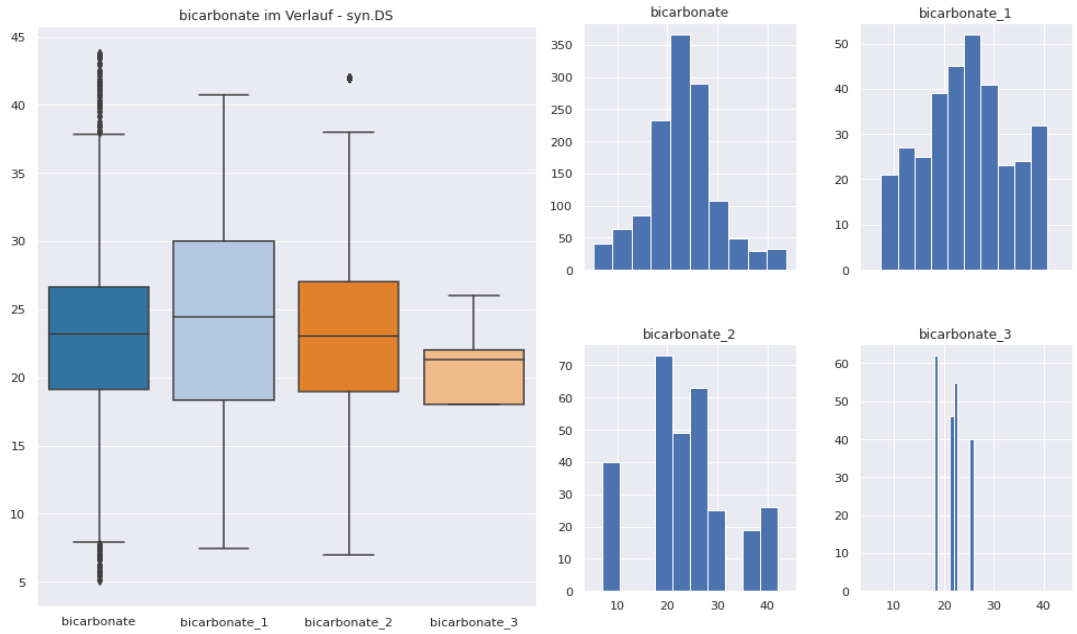


Abbildung 25: Boxplots und Histogramme zu Bicarbonat - synthetischer Datensatz

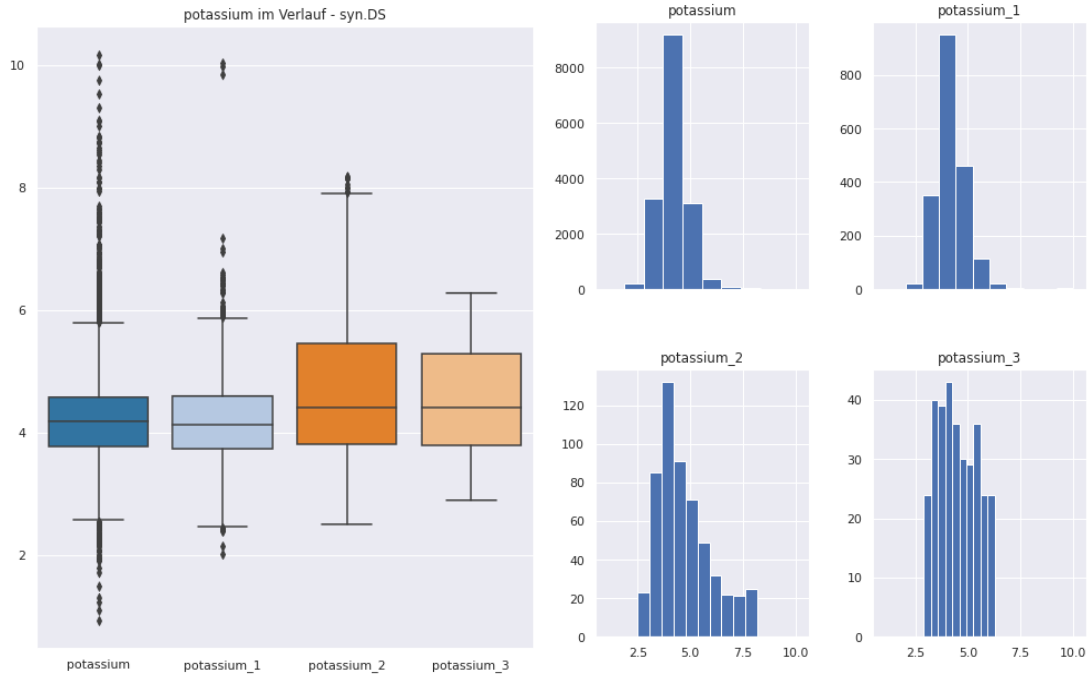


Abbildung 26: Boxplots und Histogramme zu Kalium (potassium)- synthetischer Datensatz

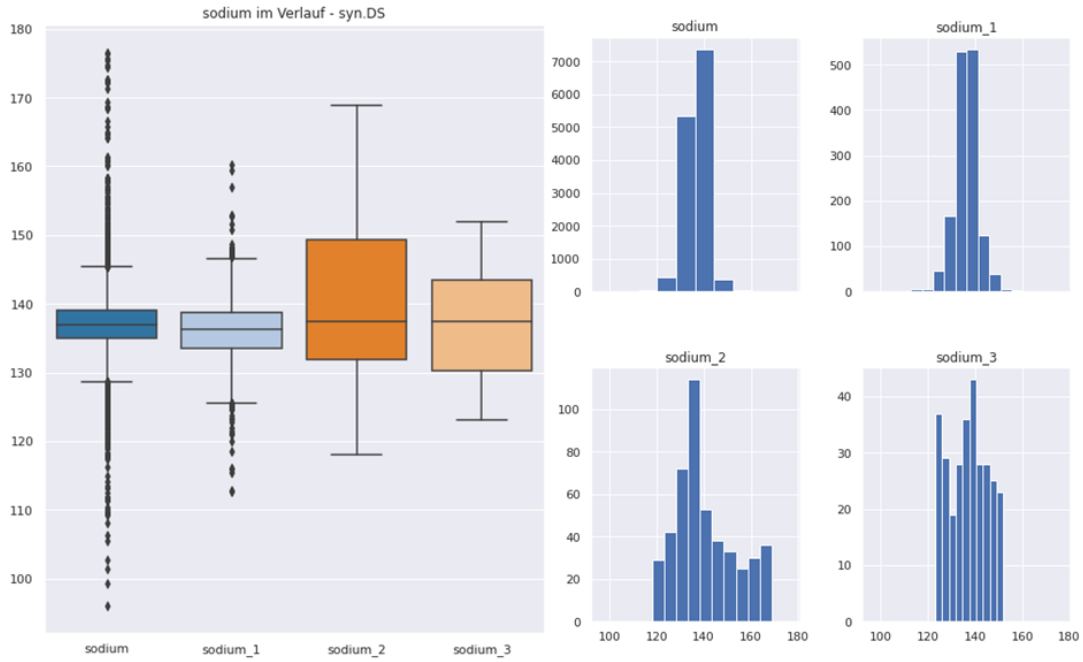


Abbildung 27: Boxplots und Histogramme zu Natrium (sodium) – synthetischer DS

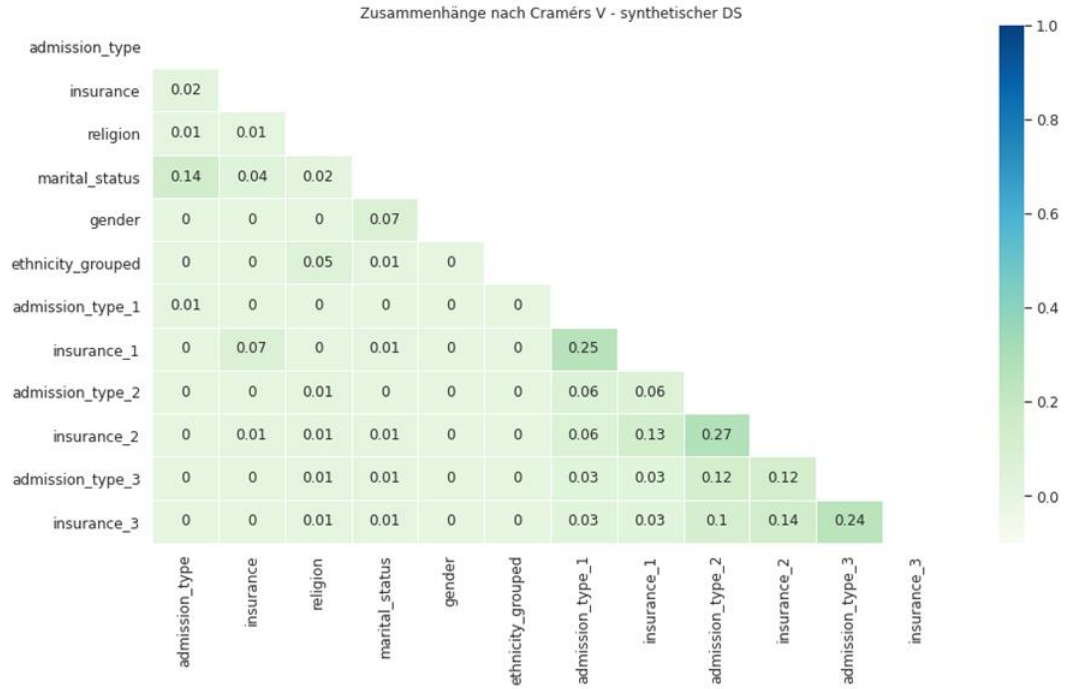


Abbildung 28: Zusammenhänge nach Cramérs V - synthetischer DS

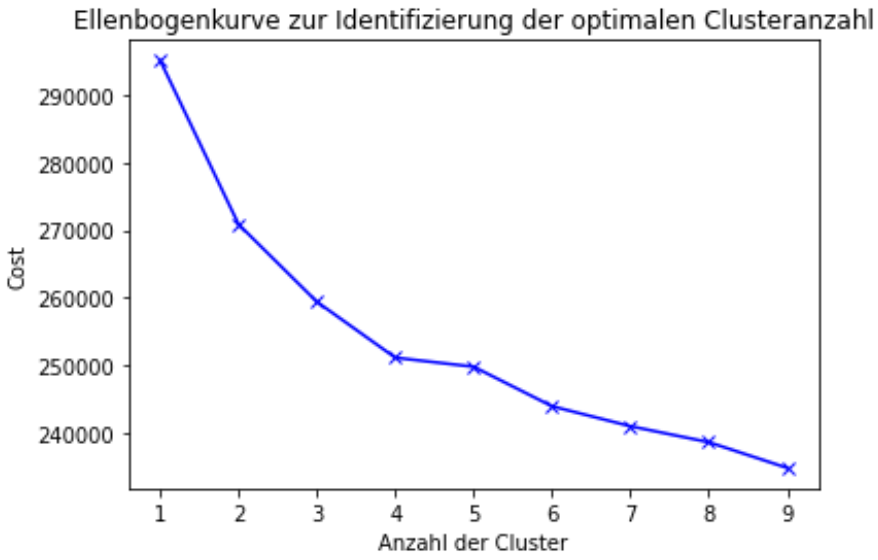


Abbildung 29: Ellenbogenkurve zur Ermittlung der optimalen Clusteranzahl

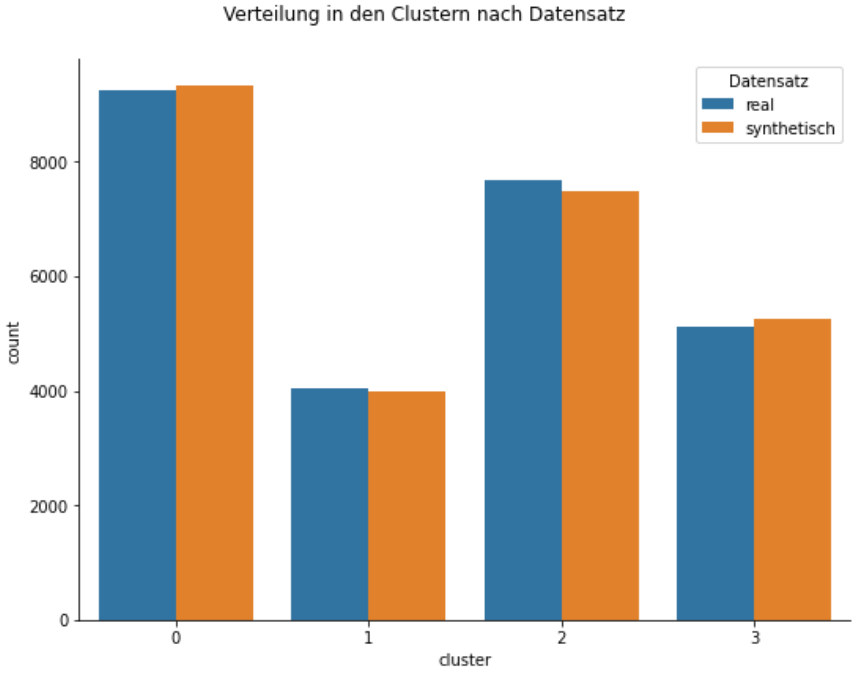
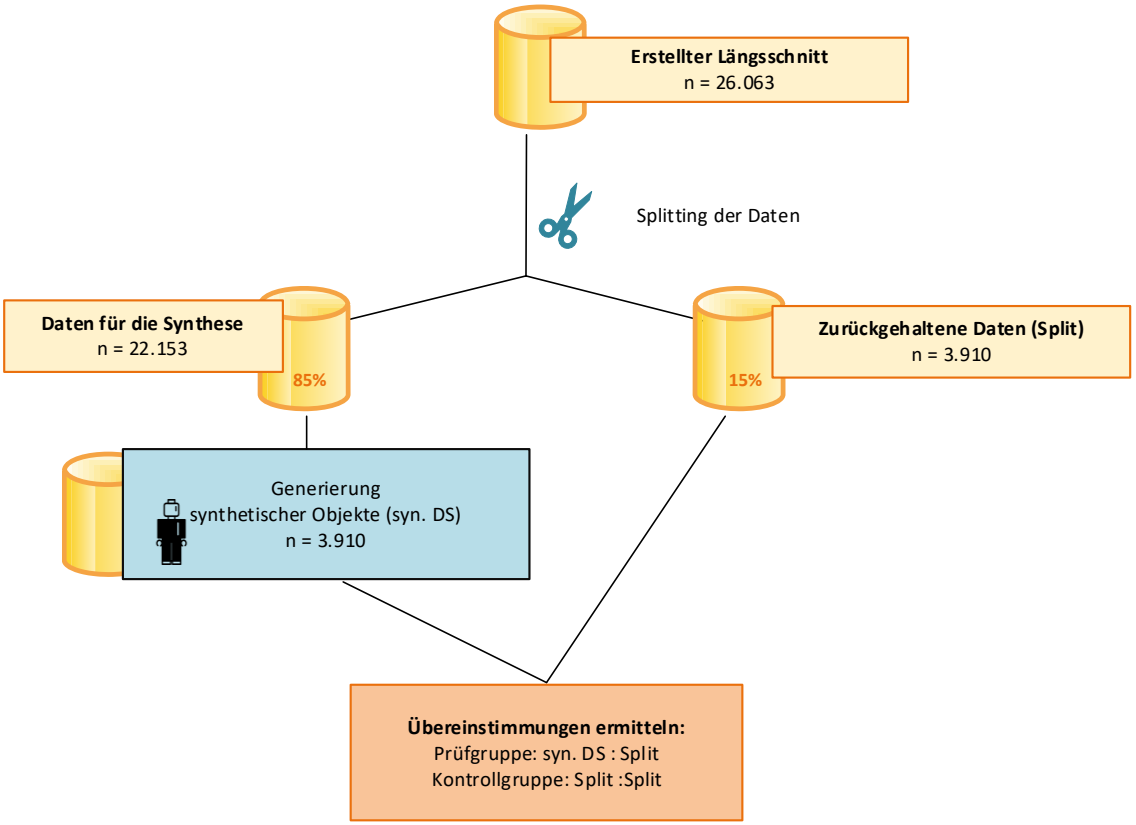


Abbildung 30: Verteilung bei vier Clustern



Bildquelle: https://www.flaticon.com/de/kostenloses-icon/roboter-im-anzug_28695

Abbildung 31: Vergleich des unabhängigen Datensatzes

Tabelle 27: Übereinstimmungen zwischen Stichproben aus realen und synthetischen DS

Anz. gepr. Merk-male	***Anzahl der Objekte in den Stichproben***											
	2.500			***5.000***			***10.000***			***26.060***		
	Max.	**Kon.**	**1:1**	**Max.**	**Kon.**	**1:1**	**Max**	**Kon.**	**1:1**	**Max.**	**Kon.**	**1:1**
1	1.345	2.500	0	2.613	5.000	0	5.278	10.000	0	14.058	26.060	0
2	1.173	2.500	0	2.349	5.000	0	4.667	10.000	0	12.137	26.060	0
3	1.003	2.500	0	2.016	5.000	0	3.996	10.000	0	10.441	26.060	0
4	744	2.500	0	1.494	5.000	0	2.970	10.000	0	7.620	26.060	0
5	405	2.488	14	796	4.988	13	1.580	9.985	10	3.899	26.042	14
6	119	2.200	180	190	4.650	236	385	9.602	333	1.048	25.594	421
7	22	678	320	38	1.908	744	85	4.798	1.448	210	16.075	3.236

Tabelle 28: Odds-Ratio und Relatives Risiko aus den Vergleichen

Stichprobe *synthetisch – real* *Prüfgruppe(PG)*				Stichprobe *real – real* *Kontrollgruppe (KG)*			Relatives Risiko		Odds Ratio	
n	**1 : 1**	**Risiko (%)**	**Odds (%)**	**1 : 1**	**Risiko (%)**	**Odds (%)**	**PG : KG**	**KG : PG**	**PG : KG**	**KG : PG**
2.500	320	12,80	14,68	491	19,64	24,44	0,65	1,53	0,60	1,66
5.000	744	14,88	17,48	1.160	23,20	30,21	0,64	1,56	0,58	1,73
10.000	1.448	14,48	16,93	2.450	24,50	32,45	0,59	1,69	0,52	1,92
26.060	3.236	12,42	14,18	7.586	29,11	41,06	0,43	2,34	0,35	2,90
3.910 (Split)	575	14,71	17,24	2.270	58,06	138,41	0,25	3,95	0,12	8,03

Anhang D – DVD

I. Masterarbeit (PDF)

II. Erstellte Jupyter Notebooks

   a. Datenauszug erstellen

   b. Datenbeschreibung

   c. Längsschnitt

   d. Datenbereinigung synthetischer Datensatz

   e. Datenbeschreibung – synthetischer Datensatz

   f. Datenqualität - PKD_KS_KLD

   g. Datenqualität – Log-Cluster Metrik

   h. Privatheit – k-anonymity_and_l-diversity

   i. Privatheit – find matches

III. Online-Quellen