miracum

Medical Informatics in Research and Care in University Medicine

FROM DATA TO KNOWLEDGE – STRATIFIZIERTE SUBGRUPPEN FÜR DIE ENTWICKLUNG VON PRÄDIKTIONSMODELLEN

Erfahrungen, Erkenntnisse und Berichte 2018 bis 2022

2. ÜBERARBEITETE AUFLAGE

USE CASE **2**

Herausgeber
Steering Board des MIRACUM-Konsortiums

Artdirection
W.A.S.

Illustrationen
Nina Eggemann

Druck
Druckhaus Sportflieger

Printed in Germany

Daniela Zöller

Harald Binder

Fotos: Sommer

Vorwort

Das MIRACUM-Konsortium hat für die Aufbau- und Vernetzungsphase der Medizininformatik-Initiative (MII) drei Use Cases definiert und darüber Anwendungsszenarien beschrieben, welche auf den Basiskomponenten der MIRACUM Datenintegrationszentren (DIZ) aufsetzen und so deren Nutzen durch ergänzende innovative IT-Lösungen belegen. In Use Case 2 „From Data to Knowledge – Stratifizierte Subgruppen für die Entwicklung von Prädiktionsmodellen“ wurden mit Techniken des maschinellen Lernens (insbesondere mit Deep Learning) valide Vorhersagemodelle auf Grundlage einer große Fülle an Daten entwickelt. Der schrittweise inhaltliche Ausbau der DIZ an den MIRACUM Standorten bot dabei eine solide Datenbasis, um Patientenkohorten anhand klinischer Parameter, Biomarker und molekularer/genomischer Untersuchungen in Subgruppen zu stratifizieren. Das Konsortium hatte sich auch zur Aufgabe gemacht, entstehende Vorhersagemodelle schnellstmöglich mittels Smart-Apps in den Klinikalltag zurückzuspielen, um Ärzt:innen in ihren diagnostischen und therapeutischen Entscheidungen zu unterstützen. Der klinische Fokus lag hierbei zunächst auf medizinische Fragestellungen aus dem Bereich Asthma/COPD und auf Hirntumoren. In dieser Broschüre stellen wir unsere Ziele, bisher etablierte Konzepte und die derzeitigen Umsetzungsergebnisse vor. Auch bedanken wir uns hiermit bei unserem Use Case 2-Team für die letzten fünf Jahre sehr engagierter und fruchtvoller Zusammenarbeit.

Daniela Zöller
Harald Binder
Use Case 2 Koordinatoren

Hans-Ulrich Prokosch
Konsortialleiter MIRACUM

Verantwortliche

Prof. Dr. Harald Binder

Er hat in Regensburg und an der University of California, Irvine, Psychologie und Mathematical Behavioral Sciences studiert und promovierte 2006 an der LMU München. Nach seiner Postdoc-Zeit in Freiburg und einer Professur an der Universitätsmedizin Mainz trat er 2017 die Professur für Medizinische Biometrie und Statistik am Uniklinikum Freiburg an, verbunden mit der Leitung des gleichnamigen Instituts. Sein Forschungsschwerpunkt ist die Verknüpfung statistischer Techniken mit Ansätzen des Deep Learning für molekularen und klinischen Daten.

Prof. Dr. Till Acker

Er hat Medizin in Freiburg, London, San Diego und Kapstadt studiert. Seit 2008 ist er Direktor des Instituts für Neuropathologie an der Justus-Liebig-Universität Gießen, seit 2015 medizinischer Forschungsdekan und seit 2019 Vorsitzender der Deutschen Gesellschaft für Neuropathologie und Neuroanatomie (DGNN). Sein Forschungsschwerpunkt liegt auf der molekularen und funktionellen Kartierung der Rolle des Mikromilieus in der Regulation von „Hallmarks of Cancer". Klinisch ist seine Gruppe an der Aufdeckung morphomolekularer Biomarker durch angewandte KI für die Differentialdiagnose und Prognose bei Hirntumoren interessiert.

Dr. Daniela Zöller

Sie hat an der Hochschule Koblenz Biomathematik studiert und 2017 an der Universitätsmedizin Mainz im Bereich Biostatistik promoviert. Nach einer 3-jährigen Postdoc-Zeit im Bereich Knowledge Discovery and Synthesis des Instituts Medizinische Biometrie und Statistik des Universitätsklinikums Freiburg, hat sie 2021 die Leitung des Bereichs Medical Data Science übernommen. Ihr Forschungsschwerpunkt liegt im Bereich der Statistischen Methodenentwicklung für Routine- und Registerdaten und für verteilte Analysen unter Datenschutzbeschränkungen.

Prof. Dr. Harald Renz

Seit 1999 ist er Direktor des Instituts für Laboratoriumsmedizin am Universitätsklinikum Gießen und Marburg (UKGM) und Professor der Philipps-Universität Marburg. Seit 2010 ist Prof. Renz stellvertretender Sprecher des UKGM Lung Center. Sein Forschungsinteresse gilt der Pathogenese von Allergien und Asthma. In den Jahren 2012/13 war er Gastprofessor und Fulbright-Stipendiat an der Harvard Medical School in Boston. Seit 2015 ist er CMO des UKGM, Standort Marburg. Von 2010 bis 2016 war er Präsident der Deutschen Gesellschaft für Allergologie und Klinische Immunologie (DGAKI) und seit 2018 ist er Vizepräsident der Deutschen Gesellschaft für Laboratoriumsmedizin. Derzeit hält er Gastprofessuren in Moskau, Russland und Moshi, Tansania.

Fotos: Sommer; JLU / Rolf K. Wegst; UK Gießen und Marburg GmbH

Team

Daniel Amsel
Christian Bruns
Alejandro Cosa Linan
Hildegard Dohmen
Frederike Euchner
Kiana Farhadyar
Patrick Fischer
Denis Gebele
Timm Greulich
Julian Gründner
Christian Haverkamp
Petar Horki
Nattika Jugkaeo
Kapil Karki
Saskia Kiefer
Stefan Lenz
Anke Lux
Adeline Makoudjou
Sebastian Mate
Michael Maxheim
Attila Nemeth
Michael Neumaier
Michelle Pfaffenlehner
Alan Race
Stephan Ringshandl
Miriam Rößner
Jannik Schaaf
Jan Scheer
Stefanie Schild
Sebastian Schindler
Bernd Schmeck
Kai Schmid
Jannik Sehring
Christian Seidemann
Susanne Seuchter
Ghislain Sofack
Amir Tabatabaei
Dativa Tibyampansha
Dennis Toddenroth
Frederik Trinkmann
Abishaa Vengadeswaran
Johannes Wolf
Jochen Zohner

Von der Theorie zur Praxis: Der Nutzen von medizinischen Big Data für Vorhersagemodelle

Das Gesundheitswesen produziert gigantische Datenmengen mit vielen bislang unbekannten Informationen. Statistische Werkzeuge, wie z.B. Deep Learning, können aus diesen ungehobenen Schätzen potenziell präzise Vorhersagen über Krankheitsverläufe oder Therapieoptionen in praxisrelevanten Situationen ermitteln. Ziel des Use Case 2 ist es, relevante Datenmuster zu identifizieren und zu validen Vorhersagemodellen zu kombinieren.

Aufgrund individueller Unterschiede, beispielsweise in der Ansprechrate auf Medikamente, ist oftmals die richtige Therapiekombination noch unbekannt und zahlreiche Krankheitsbilder in Deutschland werden noch immer nach Trial-and-Error therapiert. Ein System, das für alle Beteiligten gleichermaßen unbefriedigend ist: Patient:innen haben oftmals mit unerwünschten Nebenwirkungen von Medikamenten zu kämpfen, die letztendlich nur suboptimal wirken; Ärzt:innen wollen den Patient:innen schnelle und wirkungsvolle Therapien anbieten; und auch für die Pharmaindustrie wird es zum Problem, sich den Vorwürfen ausgesetzt zu sehen, unwirksame Mittel zu vertreiben. Dazu kommt, dass dieses Vorgehen tatsächlich auch volkswirtschaftlich eigentlich nicht zu verantworten ist, da die Ressourcen des Gesundheitswesens nicht optimal genutzt werden. Aus diesem Grund ist es wünschenswert, Patient:innen, anhand von Biomarkern und -signaturen in therapeutisch relevante Subgruppen (Endotypen) einteilen zu können, um sie effizienter und individueller zu behandeln.

Noch zu viel Theorie für die Praxis

Im klinischen Kontext ist eine immer größere Menge an Patientendaten elektronisch verfügbar. Sowohl klinische als auch Labordaten, Röntgenaufnahmen und individuelle Krankheitsbilder sind dabei oft sogar im Zeitverlauf vorhanden, so dass prinzipiell hochinformative Profile vorliegen. Um jedoch

» Wir möchten im Use Case 2 für mindestens zwei große Krankheitsbilder Vorhersagemodelle mit Deep Learning entwickeln, trainieren und evaluieren. Und wir möchten zeigen, wie unsere Ergebnisse als innovative IT-Lösungen die Mediziner bei konkreten Entscheidungen unterstützen. «

Harald Binder, 2018

Hirntumore im Visier

Hirntumore werden als besonders belastend empfunden und gehen mit einer hohen Morbidität und Mortalität sowie hohen sozioökonomischen Kosten einher. Bei Kindern und Jugendlichen sind Hirntumore der zweithäufigste Krebstyp. Bei Erwachsenen ist der häufigste Hirntumor das Glioblastom mit einer infausten Prognose und medianen Überlebenszeit von 12-15 Monaten und einer der schlechtesten Fünf-Jahres-Überlebensraten unter allen Krebserkrankungen. Bis heute ist die korrekte Diagnose von Hirntumoren herausfordernd, auch bei erfahrenen Neuropathologen, was die Notwendigkeit herausstreicht, unsere Fähigkeit zur Diagnose und adäquaten Therapie von Hirntumoren zu verbessern.

umsetzbares Wissen zu generieren, müssen aus diesen Daten mit Verfahren der Statistik und der künstlichen Intelligenz Muster identifiziert werden, welche für die Behandlung von Patient:innen relevant sind.

Aus solchen Mustern können Diagnose- und Vorhersagemodelle entwickelt werden, die zurück in den Routineeinsatz übertragen werden müssen. Trotz der Fortschritte bei der Entwicklung derartiger Modelle in den letzten Jahren, ist es immer noch eine Herausforderung, auch tatsächlich den Kreis bis in die klinische Routine zu schließen. So markiert in vielen Fällen schon die Bewertung der prinzipiellen Einsetzbarkeit von Vorhersagemodellen bereits den Abschluss solcher Forschungsprojekte. Nur selten wird versucht, die entwickelten Modelle auch tatsächlich in der Krankenversorgung zum Einsatz zu bringen. Um das Potenzial wirklich auszuschöpfen, ist es wichtig, diesen Schritt zu gehen und Algorithmen und Tools für die tägliche Entscheidungsunterstützung einzusetzen und zu verbreiten.

Die Förderinitiative MII hat mehr Welten in unseren Kliniken zusammengeführt und zu Gesprächen und gemeinsamen Zielen gebracht, als nur zwei, drei methodische Disziplinen. Hans-Ulrich Prokosch (Universität Erlangen-Nürnberg) und Till Acker (Universität Gießen)

Im Use Case 2 wird MIRACUM deshalb nicht nur Vorhersagemodelle entwickeln, sondern diese in die klinischen Versorgungsprozesse integrieren. Ziel ist es, für mindestens zwei große Krankheitsbilder zu demonstrieren, wie Vorhersagemodelle entwickelt, trainiert und evaluiert werden können, und wie diese in innovative IT-Lösungen überführt werden können, die Ärzt:innen bei konkreten Entscheidungen unterstützen.

MIRACUM hat mit Asthma/COPD (Chronisch obstruktive Lungenerkrankung) und Neuroonkologie den thematischen Schwerpunkt auf zwei medizinische Bereiche gelegt, die gute Beispiele für die Integration heterogener Datenquellen zur Identifizierung prognostisch relevanter Subgruppen darstellen. Obwohl dies zwei unterschiedliche medizinische Spezialgebiete sind, erfordern sie einen gemeinsamen methodischen Kern, welcher in den Datenintegrationszentren (DIZ) der MIRACUM-Standorte etabliert werden wird. Beide medizinischen Fragen profitieren über die Maße von den Datenkombinationen mehrerer Kliniken. Für beide Indikationen gilt, je größer die Fallzahlen, desto besser lassen sich klinisch relevante Muster identifizieren und bewerten. So ist z.B. die Existenz bestimmter Patienten-Untergruppen oft bereits aus Forschungsdatensätzen bekannt, doch noch ist unklar, wie solche Untergruppen basierend auf Routinedaten erkannt werden können und ob Patient:innen zuverlässig als Mitglied bestimmter Untergruppen identifiziert werden können.

Asthma & COPD

Asthma und COPD gehören zu den häufigsten chronisch-entzündlichen Lungenerkrankungen. Jüngste Forschungsergebnis-

Künstliche Intelligenz für Prädiktion
Die Erstellung von Prognosen oder prädiktiven Vorhersagen spielt in vielen Bereichen des täglichen Lebens eine Rolle. So benutzen E-Commerce-Anbieter Techniken der künstlichen Intelligenz/ des Deep Learnings um das Kaufverhalten von Kunden auf Basis der Historie besuchter Webseiten vorherzusagen. In ähnlicher Weise können biomedizinische Parameter als „Marker" bzw. „Signaturen" verwendet werden, um individuell auf bevorstehende Krankheitsveränderungen hinzuweisen bzw. abhängig vom Krankheitsbild bzw. -verlauf individuell passende und maßgeschneiderte therapeutische Maßnahmen zu ergreifen.

se weisen auf eine große Heterogenität der zellulären und molekularen Entzündungssignalwege bei Patientenuntergruppen hin, welche als „Endotypen" bezeichnet werden. Die Endotypisierung von Asthma- und COPD-Patient:innen entwickelt sich so zu einer vorrangigen Aufgabe, da darauf aufbauende gezielte Therapien die Möglichkeit bieten, strategische Mediatoren in diesen Entzündungsnetzwerken selektiv und spezifisch zu beeinflussen. Eine Präzisionsmedizin bei Asthma und COPD basiert auf der Charakterisierung solcher Endotypen unter Verwendung von Biomarkern. Eine besondere Herausforderung besteht darin, dass eine adäquate Charakterisierung der Patient:innen die Integration longitudinaler Daten erfordert, die verschiedenen Krankheitszuständen entsprechen. Leider ist auch die longitudinale sektorübergreifende Datenspeicherung und Datenintegration ein Gebiet, in dem sich Deutschland bislang nicht als Vorreiter auszeichnete.

Eine weitere Herausforderung in diesem Zusammenhang ist die Entwicklung von Biomarker-Panels, die Endotypen präzise widerspiegeln. Eine wichtige Rolle kommt dabei Eosinophilen im Blut zu, wobei gerade Fragen bei nicht-eosinophilem Asthma und COPD aufgeworfen wurden. Neuere klinische und experimentelle Daten legen nahe, dass diese Patientenpopulation nicht nur größer als erwartet ist, sondern auch innerhalb eines Endotyps eine breite Heterogenität besteht. Was erklärt, dass eine erfolgreiche

» Wir wissen, dass das One-size-fits-all-Modell in der Therapie oft nicht funktioniert. Die großen Fallzahlen helfen uns, klinisch relevante Muster besser identifizieren und bewerten zu können. «

Harald Renz

Therapie dieser Patient:innen von Variablen abhängen kann, die wir bisher nicht immer im Griff haben. Für uns ergibt sich daraus auch gezielte Therapien für diese wichtigen Untergruppen von Asthma-/COPD-Patient:innen mit im Auge zu haben.

Um also klinisch relevante Asthma- und COPD-Endotypen besser zu definieren, haben wir den MIRACUM-DIZ-Datensatz um Elemente von Asthma- und COPD-Patient:innen erweitert, wobei insbesondere Parameter der Lungenfunktionsmessung zu nennen sind. Alle Universitätskliniken, die an MIRACUM teilnehmen, bieten bereits bei Erwachsenen- und/ oder Kinderasthma-/ COPD-Patient:innen eine (spezialisierte) Betreuung an. Insgesamt erwarten wir mehr als 3.500 Fälle pro Jahr. Hier zeigt sich der Vorteil einer großen Zahl von Zentren, um eine Unterklassifizierung von Patientengruppen zu ermöglichen. Die Integration heterogener Datenquellen ermöglicht dabei eine tiefe Immunphänotypisierung.

Als Ergebnis werden wir die erste Plattform entwickeln, die die klinische Beratung von Asthma-/ COPD-Patienten basierend auf klinischem Clustering ermöglicht. Dies wird nicht nur eine wichtige Basis für die Auswahl und Implementierung von Behandlungsoptionen bieten, sondern langfristig auch zu einer Risikobewertung und Vorhersage der Patient:innen hinsichtlich ihrer klinischen Ursache führen, z.B. eine Entwicklung von Asthma Exazerbation als Haupttreiber der Schwere der Erkrankung.

COPD: Unterschiede zwischen Patient:innen mit und ohne Alpha-1-Antitrypsin-Mangel

Mittlerweile haben sich durch die enge Zusammenarbeit mit Klinikern an den Standorten Freiburg und Marburg mehrere wissenschaftliche Fragestellungen ergeben, die kurz davor sind, publiziert zu werden. Ein spezieller Fokus liegt dabei auf COPD-Patient:innen mit einer Erbkrankheit, Alpha-1-Antitrypsin-Mangel (AATM), der bewirkt, dass die Leberzellen das Enzym Alpha-1-Antitrypsin fehlerhaft oder in zu geringer Menge bilden oder freisetzen. Dieser Enzymmangel kann dann bereits in jungen Jahren eine COPD auslösen, wodurch sich solche Patient:innen

» Prädiktionsmodelle benötigen strukturierte Daten, die wir innerhalb von MIRACUM liefern können. Dies ist eine essentielle Voraussetzung, um diagnostische und therapeutische Ansätze in der Präzisionsmedizin mit Hilfe von KI optimieren zu können. «

Till Acker

grundlegend von anderen COPD-Patient:innen unterscheiden, welche zumeist älter sind und eine starke Rauchhistorie haben. In der ersten Fragestellung vergleichen wir die Ergebnisse innerhalb des MIRACUM-Konsortiums mit denen des COSYCONET Registers. Die Daten des COSYCONET Registers sind im Gegensatz zu den MIRACUM Daten speziell zu Forschungszwecken erhoben worden und bilden dadurch einen Goldstandard. Wir konnten die wichtigsten Ergebnisse reproduzieren: Beispielsweise ist bekannt, dass Patient:innen mit AATM ein erhöhtes Risiko für Lungenemphysemen und Bronchiektasien im Vergleich mit Patient:innen ohne AATM aufweisen, was sich auch in den MIRACUM Analysen gezeigt hat. Umgekehrt konnte wie erwartet kein Unterschied beim Auftreten von Lipoprotein oder Diabetes beobachtet werden. Dies zeigt, dass die Routinedaten grundsätzlich plausibel sind und somit, bei vorsichtiger Interpretation, in der Forschung verwendet werden können. Des Weiteren haben wir erste Hinweise auf das große Potential der Routinedaten zur Unterstützung der Gewinnung neuer medizinische Kenntnisse gefunden: Der Anteil an Pneumokokkeninfektionen und an Pseudomonasinfektionen war bei den Patient:innen mit AATM im Verhältnis höher als bei den Patient:innen ohne AATM. Auch wenn diese Ergebnisse aktuell noch als Work-In-Progress zu sehen sind, könnte diese Beobachtung einen Hinweis auf unterschiedliche Infektionsprofile und damit unterschiedliche klinische Behandlungsfokusierungen liefern.

Asthma & COPD: Der Einfluss von chronischen Lungenerkrankungen auf den Verlauf von hospitalisierten Patient:innen mit COVID-19

Im Zuge der COVID-19 Pandemie haben sich neue Fragestellungen entwickelt, für die Use Case 2 bereits wichtige Vorarbeiten geleistet hatte: Wie wird der Verlauf der COVID-19 Erkrankung im Krankenhaus durch chronische Lungenerkrankungen wie Asthma und COPD beeinflusst? Um dies zu adressieren, wurde der Use Case 2 Datensatz für Asthma und COPD um COVID-19 relevante Parameter wie Beatmungsdauer und Krankenhausmortalität erweitert.

Hirntumore im Visier

Das zweite medizinische Anwendungsfeld liegt im Bereich der Hirntumore. Sie haben einen direkten und zutiefst beeinträchtigenden Einfluss auf die kognitiven Funktionen, die geistigen Fähigkeiten und die Persönlichkeit der Patient:innen. Neben diesen hohen Morbiditätsraten sind sie mit einer hohen Mortalität und hohen sozioökonomischen Kosten verbunden. Eine präzise Tumordiagnostik ist daher von zentraler Bedeutung für die richtige Behandlung der Patient:innen.

Bis heute ist die korrekte Diagnose von Hirntumoren herausfordernd, was die Notwendigkeit hervorhebt, unsere Fähigkeit zur Diagnose und adäquaten Therapie von Hirntumoren zu verbessern. Moderne Hochdurchsatz-Analysen von großen Hirntumor-Kohorten haben gezeigt, dass DNA-Methylierung als robuste Methode zur Klassifizierung verschiedener Tumoreinheiten und zur Entdeckung neuartiger, molekular unterschiedlicher Tumorsubtypen verwendet werden kann. Hierbei werden pro Hirntumor parallel bis zu 850.000 DNA-Methylierungsstellen untersucht und analysiert. In Zusammenarbeit mit dem Deutschen Krebsforschungszentrum (DKFZ) in Heidelberg kombinieren wir darauf aufbauend DNA-methylierungsbasierte integrative Hirntumordiagnosen mit klinischen und longitudinalen Schlüsselparametern, um den Einfluss der molekularen Hirntumor-Klassifikation auf die Diagnostik und Behandlung von neuroonkologischen Patient:innen zu untersuchen.

Mit 7.000 Neuerkrankungen pro Jahr in Deutschland gehören Hirntumore zu den selteneren Tumorerkrankungen, was die Identifizierung von Subtypen weiter erschwert. Die Größe des MIRACUM-Konsortiums ist hierbei für unsere Forschung ein unschätzbarer Vorteil. Zu den wichtigsten Fragen gehört die Untersuchung neuartiger Tumor-Untergruppen und ihrer klinischen Verläufe. Für die klinische Anwendung wird schließlich entscheidend sein, ob methylierungsbasierte Hirntumorklassen Ärzt:innen dabei unterstützen können, Patient:innen, die wahrscheinlich innerhalb der nächsten 12 Monate versterben werden, zuverlässiger zu identifizieren, um sie besser palliativ versorgen zu können.

Durch Analyse dieser hochkomplexen Daten mit Techniken des Deep Learning werden Schlüsselfragen zu den Auswirkungen der verfeinerten, integrativen histomolekularen Diagnostik auf die Therapie und den Outcome von Hirntumorpatient:innen beantwortet. Da auch in anderen medizinischen Bereichen verstärkt molekulare Messungen aus Hochdurchsatzverfahren (omics), wie z.B. zur DNA-Methylierung oder NGS (next generation sequencing) Verfahren, verwendet werden, betrachten wir die beschriebenen Analysen dementsprechend als beispielhaft und zentral für weitere zukünftige Anwendungen.

Foto: iStock/luchschen; Baier

Internationale Kooperationen auf dem Weg zum verteilten maschinellen Lernen

Valide Prädiktionsmodelle, wie sie in Use Case 2 entwickelt werden sollen, benötigen große Datenmengen, welche oftmals nur durch die Kombination diverser Datenquellen erstellt werden können. Die Herausforderung besteht im Extrahieren relevanter Subgruppen aus heterogenen Datenquellen – dafür kooperiert MIRACUM jetzt auch mit Pionieren aus Newcastle.

» Eine zentrale Datenbank wäre ein sehr prominenter Angriffspunkt und sollte womöglich schon aus Sicherheitsgründen vermieden werden. «

Neue Techniken des maschinellen Lernens sollen komplexere Mechanismen und Muster identifizieren, um Empfehlungen für individuellere Behandlungen von Patient:innen geben zu können. Um neue Erkenntnisse zu generieren, sind diese Verfahren aber auf große, freiwillig gespendete Datenmengen (Stichwort „Big Data“) angewiesen. Durch die Erarbeitung eines Datenschutzkonzeptes an den einzelnen Standorten des MIRACUM-Konsortiums ist es möglich, die sensiblen Patientendaten unterschiedlicher Standorte mithilfe verteilten Rechnens mit DataSHIELD zu analysieren und Prädiktionsmodelle zu entwickeln.

Die Analyse soll zu den Daten

Der große Datenschatz, der derzeit verteilt auf die zehn Universitätskliniken im MIRACUM-Konsortium liegt, soll nicht in einen zentralen Speicher zusammengeführt, sondern verteilt über alle Standorte für Auswertungen genutzt werden. So kann die notwendige kritische Masse an Daten für die Beantwortung bestimmter Fragestellungen zustande kommen. Das Motto und datenschutzrechtliche Grundkonzept des MIRACUM-Konsortiums lautet daher „Wir wollen nicht die Daten zur Analyse bringen, sondern die Analysen zu den Daten.“ D.h. Analysen werden dezentral durchgeführt, ohne dass Kliniken die Patientendaten selbst herausgeben müssen. Viele der Algorithmen zur Datenauswertung lassen sich so umformulieren, dass sie auch auf verteilte Daten angewendet werden können.

Zwischen- und Endergebnisse dieser Analysen lassen keine Rückschlüsse auf einzelne Patient:innen zu. Dabei werden die teilnehmenden Institutionen (Krankenhäuser) zu einem Analysenetzwerk verbunden, so dass Wissenschaftler:innen über die Plattform Analysen in den teilnehmenden Krankenhäusern durchführen können, ohne Individualdaten der Patient:innen zu sehen.

VERTRAUENSWÜRDIGE NETZE

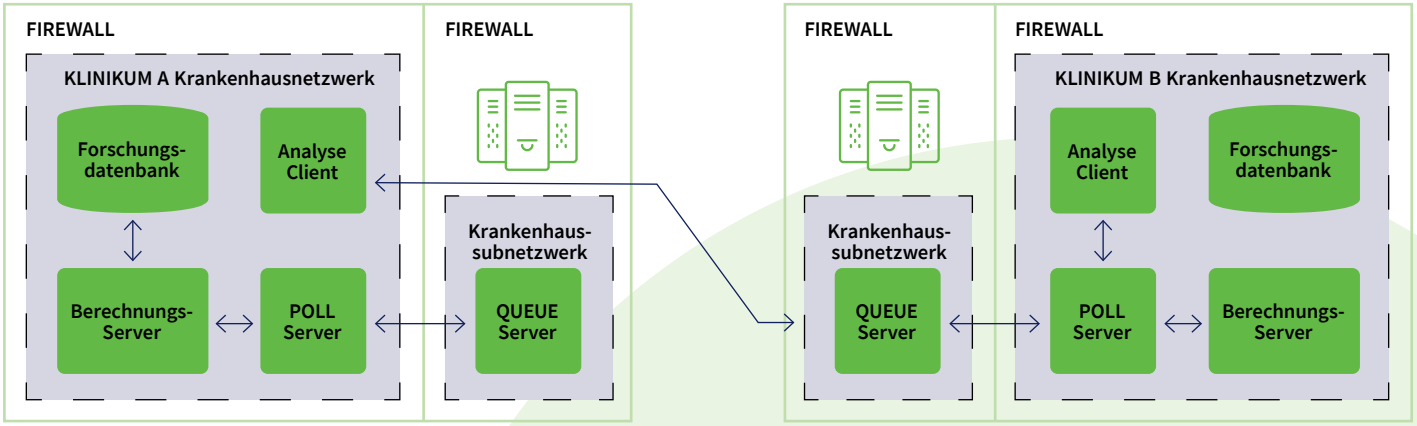


Abbildung 1: **Dieser Queue-Poll-Mechanismus** erlaubt es Krankenhäusern einfach einem Analysenetzwerk beizutreten, ohne die bestehenden Sicherheitsstandards der Institutionen absenken zu müssen. Auf diese Weise werden im MIRACUM-Konsortium mehrere Standorte sicher zu einem vertrauenswürdigen Netzwerk verbunden.

Das Projekt DataSHIELD

Um solche Analysen durchzuführen, muss eine entsprechend sichere Infrastruktur zwischen den Kliniken aufgebaut werden. Eine Forschergruppe aus England hat auf diesem Gebiet bereits Pionierarbeit geleistet und die Open-Source-Software „DataSHIELD“ entwickelt (Gaye et al. 2014, Budin-Ljøsne et al. 2015). Diese Software bietet bereits verschiedene Verfahren an, die zum statistischen Handwerkszeug gehören. Angefangen bei der Berechnung einfacher Kennzahlen, wie Durchschnittswerten oder Häufigkeiten, bis hin zu komplexeren Regressionsmodellen. Zusätzlich zu diesen fertigen Analyseverfahren bietet DataSHIELD aber auch eine flexible Infrastruktur, um neue Arten von Analysen zu entwickeln, die dann auf vernetzte, aber über mehrere Standorte verteilte Datenbestände angewendet werden können. Hier setzen die Use Case 2 Kernentwickler:innen aus Freiburg (Biometrie: Maschinelles Lernen) und Erlangen (Medizininformatik: sichere IT-Infrastruktur) an und bringen ihre Expertise ein.

Erweiterung des DataSHIELD-Werkzeugkastens

Um die Analysen zu den Datenbeständen in einem Netzwerk an die einzelnen Standorte zu bringen, nutzt das Original DataSHIELD-Konzept einen Zugriff auf das jeweilige Klinikumsnetzwerk von außen, um die Analysen dann im lokalen Netzwerk bereit zu stellen. Dies würde aus Sicherheitsgründen in einem deutschen Universitätsklinikum durch die vor dem Klinikumsnetzwerk eingerichtete Firewall abgewiesen. Dieses Problem wurde vom MIRACUM Use Case 2 Team gelöst, indem eine DataSHIELD Erweiterung konzipiert und implementiert wurde, welche diesen Analysen-Verteilungs-/Zugriffsweg umdreht (siehe Abbildung 1).

Die mit einem zentralen Client erstellten Analysen werden zunächst außerhalb in einer Warteschlange (Queue) abgelegt. Innerhalb des Netzwerks fragt ein Prozess diese Warteschleife ständig ab, um zu prüfen, ob dort eine neue auszuführende Analyse bereitgestellt wurde. Eine dort abgelegte Analyse wird dann von diesem Prozess mittels eines, über die Firewall erlaubten Prozesses, von außen hinein geholt (Poll).

Vertrauenswürdige Netzwerke generieren

Dieser Queue-Poll-Mechanismus erlaubt es Krankenhäusern, einfach einem Analysenetzwerk beizutreten, ohne die bestehenden Sicherheitsstandards der Institutionen absenken zu müssen. Weiterhin erfasst der Mechanismus jede Anfrage und bietet die Möglichkeit, nur bestimmte Analyse Clients (aus anderen Institutionen) eine Anfrage zu erlauben. Auf diese Weise werden im MIRACUM-Konsortium mehrere Standorte sicher zu einem vertrauenswürdigen Netzwerk verbunden.

Die Community der DataSHIELD-Entwickler:innen

Um diese Erweiterung in die DataSHIELD-Community zu bringen, wurde bereits im Mai 2018 ein erstes Abstimmungstreffen mit dem Kernentwicklerteam (Prof. Paul Burton, Dr. Olly Butters und Dr. Rebecca Wilson) initiiert. Dieses nahm den Konzeptvorschlag sofort positiv auf, und reagierte mit einer freundlichen Aufnahme in die DataSHIELD-Entwickler:innen-Community. Nach dieser ersten Abstimmung konnten die MIRACUM-Entwi-

DATASHIELD ANSATZ INNERHALB DES MIRACUM PROJEKTS

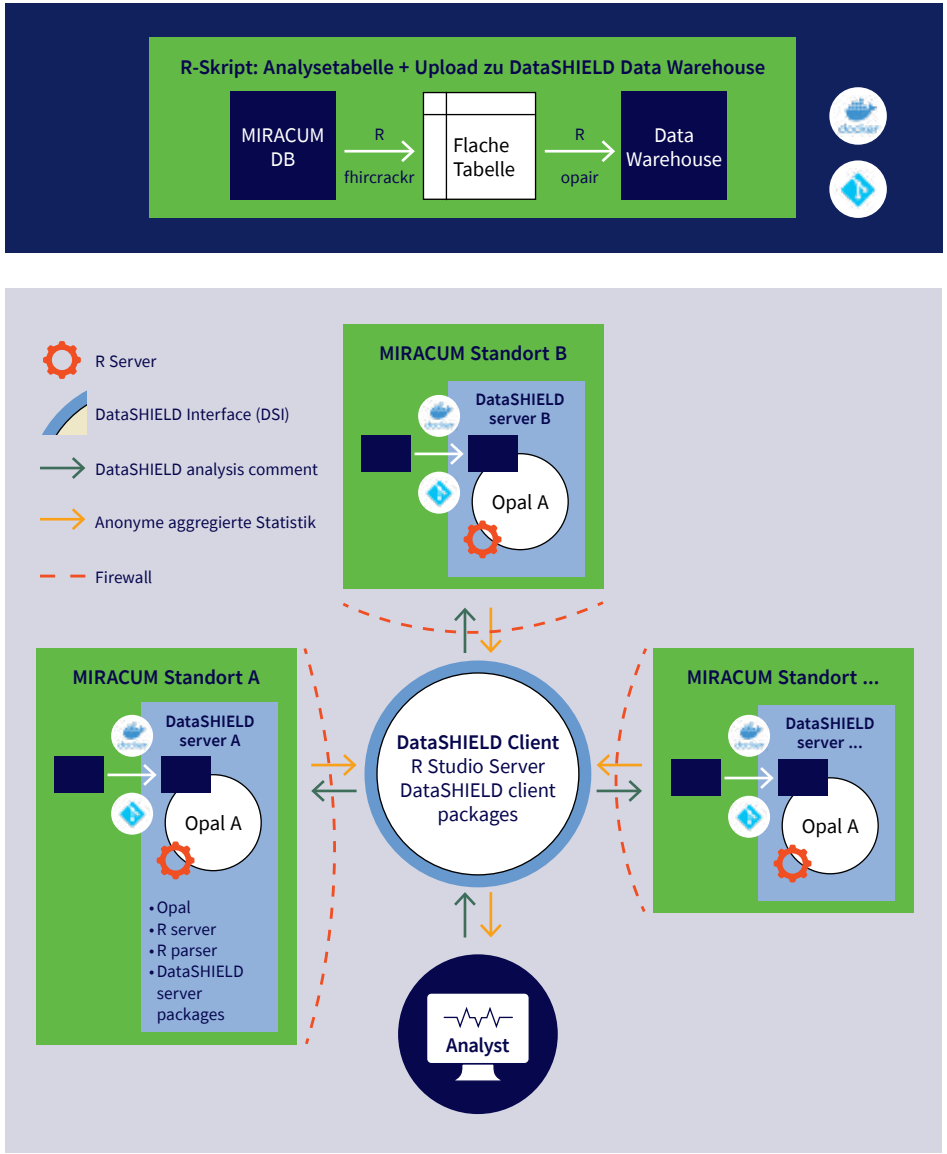


Abbildung 2: Der MIRACUM DataSHIELD-Ansatz ist ein automatisierter Prozess.

» Beim MIRACUM DataSHIELD-Ansatz handelt es sich um einen automatisierten Prozess, wobei die Daten harmonisiert im ersten Schritt von allen Standorten auf die lokalen Opal Datawarehouse Server geladen werden müssen. Der Analyst hat hierbei nur Zugriff auf DataSHIELD-Funktionen und aggregierte Daten. «

cklungen schon im November 2018 beim internationalen DataSHIELD-Workshop in Newcastle (UK) präsentiert werden. Es ergab sich auch ein Einblick in weitere EU-Projekte, die ihrerseits Gesundheitsdaten dezentral analysieren wollen. Europaweit gibt es mehrere Studien, die DataSHIELD einsetzen und mit Ihrem Feedback helfen, die Software zu verbessern und auch weitere Gruppen, die neue Funktionalitäten beisteuern. Das von der EU und vom kanadischen Gesundheitsministerium geförderte und 2019 angelaufenen EU-CAN-Connect-Projekt arbeitet daran, mittels DataSHIELD eine Datenplattform nach den Richtlinien FAIR-Data-Initiative aufzubauen.

TMF Task Force Verteilte Analysen

Das Thema „Verteilte Analysen“ spielt für die gesamte Medizininformatik-Initiative (MII) eine große Rolle. 2020 wurden zwei eintägige Workshops zu diesem Thema veranstaltet, bei der die verschiedenen Konsortien der MII bisherige Lösungensansätze präsentiert und Vor- und Nachteile vergleichend diskutiert wurden. Neben dem Ansatz DataSHIELD von MIRACUM sind insbesondere der Personal Health Train und Secure Multi Party Computation zu nennen. Im Nachgang des Workshops wurde von der TMF die Task Force Verteilte Analysen gegründet, in der Use Case 2 von Harald Binder und Daniela Zöller aus Freiburg vertreten sind. Hier wurden die Ergebnisse zusammengefasst und eine Stellungnahme zum aktuellen Stand verfasst. DataSHIELD wurde dabei als einzige Lösung identifiziert, die aktuell bereits für diverse Einsatzzwecke verwendet werden kann. Aus diesem Grund wird innerhalb der MII nun allgemein empfohlen, DataSHIELD an allen Standorten zur Verfügung zu stellen. Seitdem haben sich bereits einzelne Standorte anderer Konsortien an Use Case 2 gewandt, um die von MIRACUM entwickelte Queue-Poll-Lösung zu nutzen. Allerdings ist ein weiteres Ergebnis der Task Force, dass DataSHIELD in einigen Bereichen aktuell nicht einsatzbar ist, wie z.B. beim Record-Linkage oder bei der Zusammenführung sehr kleiner Fallzahlen. Wir haben darauf hin Kontakt mit dem Kernentwicklerteam in England aufgenommen, die großes Interesse an einer Kooperation mit den anderen Ansätzen be-

MIRACUM-Analyse-Methoden helfen Wissenschaftler:innen bei der Auswertung komplexer Datensätze.

kundet haben. Es gibt also Bewegung rund um DataSHIELD, zu der auch die deutsche Entwicklungsinitiative durch MIRACUM-Elan beigetragen hat.

Der im MIRACUM verwendete DataSHIELD-Ansatz wird in Abbildung 2 dargestellt. Dabei handelt es sich um einen automatisierten Prozess, wobei die Daten harmonisiert im ersten Schritt von allen Standorten auf die lokalen Opal Datawarehouse Server geladen werden müssen. Der Analyst hat hierbei nur Zugriff auf DataSHIELD Funktionen und aggregierte Daten (siehe Abbildung 2).

Strukturelle Vorarbeiten

Der Use Case 2 hat die strukturellen Vorarbeiten für die Beantwortung vieler medizinischer Fragestellungen geleistet. Dazu haben wir DataSHIELD erfolgreich an allen Standorten installiert, getestet und die Funktionalität (u.a. für maschinelles Lernen und insbesondere Verfahren des Deep Learning) erweitert. Für verschiedene Fragestellungen haben wir ein R-Skript vorbereitet, welches die Daten entsprechend der wissenschaftlichen Vorgaben für die Analyse aufbereitet und in das lokale DataSHIELD-Data Warehouse lädt. Die Skripte werden mit Docker Containern an die Standorte verteilt, wo Mitarbeiter:innen nach der Ausführung die Daten nochmals kontrollieren und abschließend für die Analyse freigeben können. Um sicher zu stellen, dass die Daten den Anforderungen genügen, wurde die Datenqualität kontrolliert, bevor mit den eigentlichen Analysen gestartet wurde.

Fokus: chronische Lungenerkrankungen und Hirntumore

Gemäß dem Motto „Learning by doing“ wird der Fokus zuerst auf die zwei bereits vorgestellten Erkrankungsgruppen Asthma/COPD und Hirntumore gelegt, um die neuen Techniken zu entwickeln und zu erproben. Diese beiden Projekte stellen die Beteiligten vor verschiedene Herausforderungen. In Bezug auf Daten von Asthma/COPD-Patient:innen hat man es vor allem mit Daten aus dem klinischen Routinebetrieb zu tun. Hier ist eine enge Zusammenarbeit zwischen Klinikern, Biostatistikern und Medizininformatikern notwendig, um relevante Fragestellungen, dafür geeignete Auswertungsverfahren und Daten zu identifizieren und aus den Krankenhausinformationssystemen so zu extrahieren.

Freiburger Biometrie entwickelt neues Werkzeug

Für die Patient:innen mit Hirntumoren liegen ebenfalls umfangreiche genetische Daten vor. Dabei gilt es, aus dieser unüberblickbaren Menge genetischer Merkmale Hinweise auf diejenigen Eigenschaften herauszufinden, die einen Einfluss auf die Schwere der Erkrankung bzw. die Heilungschancen haben könnten. Diese Hinweise können dann in Laborexperimenten weiter untersucht werden, um die genauen Wirkmechanismen zu erforschen. Dafür entwickelte die Freiburger Biometrie ein neues Werkzeug zum Finden von Einflussfaktoren für DataSHIELD, das seit 2019 an Hirntumordaten erprobt werden kann.

Diese selbst gestellten Herausforderungen dienen auch der Motivation, um eine nachhaltige und flexible Dateninfrastruktur aufzubauen, damit Forschung und Patient:innen von dem verstreuten Datenschatz profitieren. So wird darauf hingearbeitet, Werkzeuge bereit zu stellen, damit Ärzt:innen und Biolog:innen Krankheiten genauer verstehen können und individuellere Behandlungen ermöglicht werden, ohne dass dabei der Schutz der Gesundheitsdaten vernachlässigt werden muss.

Fotos: iStock (demaio; janiecbros)

DataShield Literatur

Budin-Ljøsne I, Burton P, Isaeva J, Gaye A, Turner A, Murtagh MJ, et al.: DataSHIELD: an ethically robust solution to multiple-site individual-level data analysis. Public Health Genomics. 2015;18(2):87-96.

Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, Minion J et al.: DataSHIELD: taking the analysis to the data, not the data to the analysis. Int J Epidemiol. 2014 Dec;43(6):1929-44.

Die strukturellen Vorarbeiten sind abgeschlossen

Use Case 2 des MIRACUM-Konsortiums ist angetreten für chronische Lungenerkrankungen, wie Asthma oder COPD, patientenrelevante Verbesserungen durch valide Prädiktionsmodelle schnell in den Praxisalltag zu bringen. Über die Mühen und den Lohn dieser Grundlagenarbeit gab Prof. Dr. Harald Renz, Direktor des Marburger Instituts für Laboratoriumsmedizin, Pathobiochemie und Molekulare Diagnostik des Universitätsklinikums Gießen und Marburg GmbH, Auskunft.

INTERVIEW MIT Prof. Dr. Harald Renz
TEXT Claudia Dirks

Prof. Dr. Harald Renz,
Universitätsklinikum
Gießen-Marburg

Lassen Sie uns am Ende dieser ersten Förderperiode Bilanz ziehen. Wie steht es um den Use Case 2 des MIRACUM-Konsortiums, From Data to Knowledge – stratifizierte Subgruppen für die Entwicklung von Prädiktionsmodellen? Sind Sie so weit gekommen, wie Sie sich das mit Ihrem Team vor vier Jahren erhofft hatten?

Leider sogar bei weitem nicht. Der Use Case 2 konnte erst richtig an Fahrt aufnehmen, nachdem eine gewisse Grundstruktur durch die einzelnen Datenintegrationszentren (DIZ) an den einzelnen Standorten aufgebaut war, und implementiert werden konnte. Wir mussten also erst einmal die Voraussetzung schaffen, überhaupt mit Daten arbeiten zu können. Insofern waren wir, wie alle anderen klinischen Anwendungsprojekte, davon abhängig, dass es a) eine arbeitsfähige DIZ-Struktur gibt und b) diese dann in Breite und Tiefe das Datenmaterial aus den Krankenhausinformationssystemen zugespielt bekommen, die in der Wissenschaftslandschaft verarbeitet werden können.

Das klingt nach doch sehr viel Vorarbeit, die erst einmal geleistet werden musste, ohne überhaupt einen Patienten zu Gesicht bekommen zu haben?

Ja, da haben wir schon gemerkt, dass wir dicke Bretter bohren. Die technischen Herausforderungen, dies alles zum Laufen zu bringen, sind gewaltig gewesen. Das hatte mit der MII erst einmal gar nichts zu tun. Das ist den technischen, datenschutzrechtlichen und den ethischen Grundvoraussetzungen geschuldet, in denen wir hierzulande navigieren.

» Ich habe mir vor fünf Jahren den Weg weniger mühsam vorgestellt. Doch es ist enorm, was geleistet wurde. Wir haben funktionsfähige DIZ und es steht ein robuster Werkzeugkasten zur Verfügung, um klinische Daten in dieses DIZ-Umfeld zu exportieren und zu integrieren. «

Können Sie mal ein konkretes Beispiel aus Ihrer eigenen Erfahrungswelt nennen?

Es war beispielsweise eine Herkules-Aufgabe, die Labordaten hier in Marburg aus der KIS-, LIS-Landschaft ins Marburger DIZ einzuspielen. Wir sind zurückgegangen bis 2006 – und dieser Retro-Export ist gerade erst abgeschlossen.

Jetzt kommt das nächste Problem. Wenn Sie Labordaten über Standorte hinweg miteinander vergleichen wollen, dann brauchen Sie deren Harmonisierung und eine standardisierte Erhebung der Laborwerte über die Standorte hinweg.

… und auch damit ist ja die Arbeit auch noch nicht getan

Absolut richtig. Dann kommt auch noch der Abgleich mit der Erwartungshaltung der Kliniker. Es ist leider gar nicht so, dass alle relevanten Fälle der Kliniken genommen werden, um diese dann wissenschaftlich auswerten zu können. Die Auswertung findet nämlich am jeweiligen DIZ statt. Das bedeutet, diese aggregierten Daten erstellen quasi ein Super-Aggregat – was aber dennoch etwas anderes ist, als 1.000 Individualdaten zu haben, die in einen Datensatz zusammengeführt werden. Sagen wir so: Die Lernkurve für Methodiker, wie für Kliniker ist eine sehr steile in den vergangenen Jahren gewesen.

Daraus kann ja auch viel Gutes entstehen?

Ja, absolut. Es hat tatsächlich dazu geführt, dass untereinander ein besseres Verständnis für die jeweils andere Position entstanden ist. Jeder Spezialbereich hat nun einmal seine eigene Sprache. Wenn ein Methodiker über Statistische Modelle oder Künstliche Intelligenz redet, ist das meist etwas völlig anderes als das, was der Kliniker darunter versteht. Und andersrum ist es ebenso. Da ein gemeinsames Verständnis zu entwickeln, war die eigentlich Herausforderung. Aber genau so wurden die Use Cases ja auch angelegt.

Hätte sich die MII aus heutiger Sicht mehr an internationalen Vorbildern orientieren sollen oder ist diese Näherung ein Prozess, durch den jedes Land einmal gehen muss, um ein gemeinsames Verständnis von Krankenversorgung zu entwickeln?

Im internationalen Vergleich hat Deutschland sicherlich einen etwas schwereren Stand, allein was die Nutzbarkeit der Daten für die Forschung angeht. Da haben wir großen Aufholbedarf, allein, weil die juristischen Hürden so hoch sind, was die Nutzbarkeit medizinischer Daten anbelangt. Das mag seine (guten) Gründe haben, aber es erschwert dennoch die Nutzung im wissenschaftlichen Umfeld.

Welches Land hat denn aus Ihrer Sicht einen strategischen Vorteil hinsichtlich der Datenbereitstellung?

Nehmen wir das Beispiel USA. Die Daten gehören nicht den Patienten, sie gehören demjenigen, der die Daten erhoben hat. Die dürfen dann auch mit den Daten arbeiten. Das hat große Vorteile, wenn Sie an wissenschaftliche Auswertung und Verwertung denken. Und wiederum eher negative Auswirkungen auf den Daten- bzw. Patientenschutz hat. Das muss immer gegeneinander abgewogen werden – und da entscheiden wir in Deutschland in den allermeisten Fällen zugunsten des Datenschutzes. Das – und der Föderalismus – erschweren die klinische Forschung hierzulande nun einmal.

Sie klingen etwas ernüchtert? Oder anders: Ist es für Sie absolut folgerichtig, dass es eine zweite Förderperiode geben wird, in der dann aus Sicht des Use Case 2 endlich gearbeitet werden kann?

(Lacht.) Ich habe mir vor fünf Jahren den Weg weniger mühsam vorgestellt – und nicht verbunden mit so viel Kernarbeit. Das hat mich doch auch überrascht.

Wenn ich jedoch zurückschaue, ist es enorm, was geleistet wurde. Wir haben an den meisten Standorten, funktionsfähige DIZ, es steht ein robuster Werkzeugkasten zur Verfügung, um klinische Daten in dieses DIZ-Umfeld zu exportieren und zu integrieren. Labordaten sind drin, Lungenfunktionsdaten kommen als nächste. Das ist jetzt eben ein notwendiger Prozess, der aber einen sehr guten Weg bereitet.

Welche zwei dringendsten offenen Baustellen sehen Sie denn konkret vor sich?

Wenn wir jetzt beim Use Case 2 bleiben, dann möchten wir individuelle Krankheitsverläufe abbilden können. Unsere Patient:innen laborieren über Jahrzehnte – nicht nur mit ihren Krankenhäusern –, sondern auch mit niedergelassenen Ärzt:innen an ihren Krankheiten herum. Wir können also quasi Verläufe nur exakt nachzeichnen, wenn wir die sektorübergreifende Erfassung anstoßen.

Asthma-Patient:innen kommen heute kaum noch in die Klinik wegen ihres Asthmas. Nur ca. 1 Prozent kommt in Ausnahmefällen überhaupt ins Krankenhaus. 99% werden ausschließlich in der Praxis behandelt – wir müssen also die Daten EINES Patienten aus ganz verschiedenen Orten zusammenführen, um individuelle Verläufe nachzeichnen zu können.

» Ich freu‘ mich sehr darüber, dass die Begeisterung der Kliniker nicht nachgelassen hat. Alle haben in der vergangenen Zeit begriffen, wie sinnvoll diese Arbeit ist. «

Die andere große Herausforderung ist die Verknüpfung unterschiedlicher Datensätze miteinander.

Was aus diesen vergangenen Jahren kommt denn heute schon bei Ihren Patient:innen an und auch bei Ihren Behandelnden an? Arbeiten die heute mit mehr Daten als noch vor zwei Jahren?

Leider ist davon noch nichts wirklich bei den Patient:innen angekommen. Deswegen stellen wir für Modul 3 der kommenden Förderperiode jetzt einen konsortienübergreifenden Antrag zu COPD und Asthma, um eben genau solche Vorhersagemodelle zu entwickeln, die den Patient:innen tatsächlich auf dem Weg mit seiner Krankheit helfen zurechtzukommen.

Ich möchte als Patient wissen, ab wann komme ich in ein erhöhtes Risiko? Also bevor ich keine Luft mehr bekomme und in eine schwierige gesundheitliche Situation gerate. Dazu gehören Umwelt- oder Wetterdaten oder überhaupt Umweltexpositionen, die für Lungenpatienten besonders schwierig werden können. Das sind relevante Informationen.

Können die Patient:innen auch selbst etwas beitragen, um Daten aus ihrem Alltag einzubringen?

Ja, das ist ein gutes Beispiel. Es gibt ja mittlerweile immer mehr Möglichkeiten, über Biosensoren, Biosignale in Echtzeit abzurufen. In unseren Fällen sind das EKG, Atemmuster oder die Sauerstoffsättigung im Blut besonders interessant – das sind alles Daten, die die Patient:innen selbst generieren können.

Unsere Herausforderung an dieser Stelle ist es, diese Daten zugänglich zu machen und sinnvoll in den jeweiligen Patientenpfad zu integrieren, um Patient:in wie Mediziner:in Informationen über die aktuelle und individuelle Risikokonstellation zur Seite zu stellen. Das hätte dann tatsächlich einen unglaublichen Mehrwert für alle Beteiligten mit unmittelbarer Auswirkung auf eine verbesserte Patientenversorgung.

Die Patient:innen sollen also eher präventiv unterstützt werden, so dass es gar nicht mehr zur Hospitalisierung kommt?

Genau. Was wir erreichen wollen, ist die Verbesserung der Lebensqualität der Patient:innen; sie sollen sich fitter und besser fühlen. Gleichzeitig wollen wir eine Hospitalisierung verhindern, die bei dieser Erkrankung zu großen Teilen die sozio-ökonomischen Kosten treibt, obwohl sie in vielen Fällen vermeidbar wäre. Oder, wenn Krankenhaus, dann doch wenigstens so kurz wie möglich – und natürlich wollen wir als oberstes Ziel die Sterblichkeit reduzieren.

Jetzt haben Sie innerhalb der MII eine lange, produktive Kennenlernphase hinter sich. Schauen wir in die Zukunft, soll jetzt auch der nicht-universitäre Bereich mit eingebunden werden. Wie wird der Niedergelassene Bereich denn idealerweise mitgenommen?

Konkret für unseren Bereich sehe ich da ganz gelassen in die Zukunft. Wir haben sechs große Spezialpraxen für den Antrag mit an Board, um im Sinne eines Proof-of-Concept dieses Miteinander einmal durchzuspielen. Wenn das funktioniert, wird es auch für andere eine starke vertrauensbildende Maßnahme sein und ein gutes Signal, dass sowas produktive für alle Beteiligten sein kann.

Diese sechs Spezialpraxen sind Teil des Förderantrags für die kommenden Jahre?

Ja, genau. Diese Praxen sind Partner in unserem Antrag für das Modul 3, das unmittelbar darauf aufsetzt, was wir mit unserem Use Case 2 in MIRACUM gemacht haben.

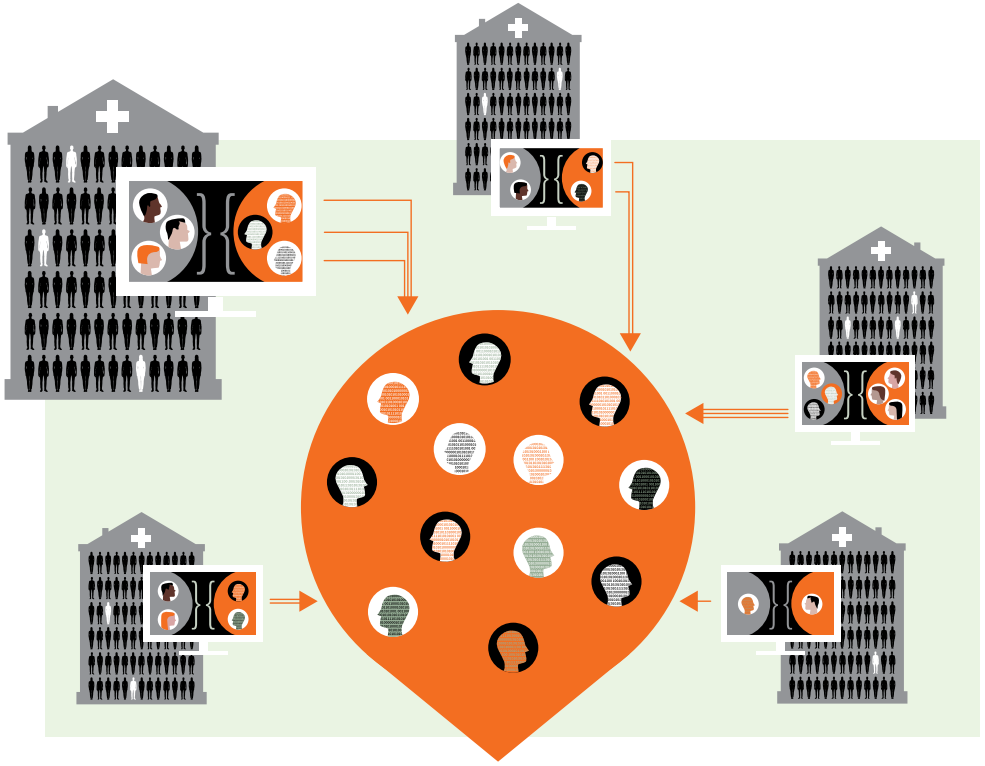
Worüber freuen Sie sich in der Rückschau auf die vergangenen Jahre in Use Case 2 denn besonders?

Ich freu‘ mich sehr darüber, dass die Begeisterung der Kliniker nicht nachgelassen hat! Das ist das Wichtigste. Alle haben in der vergangenen Zeit begriffen, wie sinnvoll diese Arbeit ist und sind über die Jahre motiviert und konstruktiv geblieben. Und natürlich haben wir eine handvoll Unterprojekte, die wir jetzt veröffentlichen können – was für uns Forscher:innen ja auch einen großen Wert hat.

Worauf werden Sie beim Miteinander im nächsten Durchgang Ihr besonderes Augenmerk legen?

Das Erwartungsmanagement der unterschiedlichen Professionen muss früh genug auf den Tisch gebracht und gemeinsam eingeordnet werden. Hier gilt es wirklich, von Anfang an klar zu sein und mit einer Erwartungshaltung zu operieren, die der Realität entspricht.

Fotos: iStock (mi-viri); UK Gießen-Marburg



Die virtuellen Patient:innen

Rigide Datenschutzbestimmungen machen in Deutschland der Forschungslandschaft oftmals schwer zu schaffen. Um dennoch mit eigenen Daten forschen zu können, prüft das Team des Use Case 2, ob synthetische Daten Teil der Lösung werden können.

Der Schutz von Gesundheitsdaten ist wichtig, da sie personenbezogene Daten beinhalten, die sowohl Patient:innen als auch deren Familienangehörige betreffen. Datenschutzbestimmungen bilden andererseits jedoch ein kritisches Hindernis für den Zugang zu Patientendaten, die zur Verbesserung der Gesundheitsversorgung genutzt werden könnten. Dieses Problem adressieren wir z. B. auch im Use Case 2 des MIRACUM-Konsortiums, dessen Ziel es ja ist, personalisierte Vorhersagemodelle für die medizinische Behandlung zu erstellen. Dazu sollen die Daten mehrerer MIRACUM Standorte gemeinsam analysiert werden. Aktuell ist es allerdings noch so, dass diese aus Datenschutzgründen nicht zusammengeführt werden dürfen. Im Rahmen der technischen Lösungen, die wir zum Umgang mit dieser Problematik entwickeln, wollen wir zunächst Daten chronischer Lungenerkrankungen (Asthma und COPD) analysieren und damit im MIRACUM Projekt wissenschaftliche Fragestellungen auf Daten von Patient:innen mit Hirntumoren untersuchen.

Synthetische Daten sind eine Lösung, die zur Überwindung von Datenschutzhürden genutzt werden können. Synthetische Daten sind Daten, die keine echten Daten enthalten, sondern lediglich allgemeine Merkmale und statistische Beziehungen echter Daten nachbilden. Für die Datennutzung in der Forschung bedeutet dies, dass pro Standort virtuelle Patientendaten erstellt werden, die nicht an die Daten eines einzenen Patienten gebunden sind. Solche Daten können dann gemeinsam genutzt werden und ermöglichen den Einsatz statistischer Standardanalysen, oder auch Techniken der künstlichen Intelligenz.

Generative Modelle

Für die Erzeugung synthetischer Daten aus realen Daten ist eine Art statistischer oder maschineller Lernansatz erforderlich. Konkret verwenden wir generative Modelle, welche die systematische und zufällige Variabilität der Originaldaten abbilden, indem ein Modell der statistischen Verteilung der Daten erstellt wird. Tiefe Techniken des Deep Learnings können sich sehr flexibel unterschiedlichen Datenverteilungen annähern. In MIRACUM verwenden wir unter anderem einen der populärsten generativen Ansätze, sogenannte Variational Autoencoder (VAE).

Wie funktioniert ein Variational Autoencoder?

Die Funktionsweise eines Variational Autoencoder kann gut an einem Beispiel erklärt werden: Wir wollen wissen, „wie krank“ ein:e Patient:in ist. Im Idealfall haben wir für alle Patient:innen ein oder zwei abstrakte Werte, die den Grad der Erkrankung repräsentieren. Damit könnten wir Daten für virtuelle Patient:innen generieren, indem wir plausible Werte für eine oder zwei Dimensionen auswählen.

In der realen Welt werden Krankheiten jedoch durch viele verschiedene beobachtete Symptommuster gekennzeichnet. Deshalb wollen wir von Mustern, die in realen Messungen, Beobachtungen oder Symptomen zu sehen sind, auf eine kleine Anzahl dieser niedrigdimensionalen Werte abbilden. Diese latenten Merkmale werden im Grunde nie direkt beobachtet, also treffen wir einige plausible Annahmen, um ein Muster zu finden. Wir nehmen zum Beispiel an, dass wir, wenn wir den Krankheitswert aller Individuen in einem Häufigkeitsdiagramm aufzeichnen,

eine Normalverteilung oder Gaußsche Glockenkurve haben. Gemäß dieser Annahme können wir dann aus den Daten die Transformation lernen, die Messungen, Beobachtungen und Symptome in die latenten Krankheitswerte transformiert.

Für den nächsten Schritt stelle man sich vor, dass ein:e Mediziner:in Messungen, Beobachtungen und Symptome von Patient:innen mit bestimmten Erkrankungen mit anderen Wissenschaftler:innen diskutieren möchte, natürlich ohne den Datenschutz zu verletzen. Mit den latenten Krankheitswerten für verschiedene Individuen ist es möglich, eine weitere Transformation vom Krankheitswert zurück zu passenden Messungen, Symptomen oder Diagnosen zu lernen. Damit können schließlich plausibel gewählte latente Krankheitswerte eines virtuellen Patient:innen zurücktransformiert werden, so dass sie wie plausible Patientendaten aussehen.

Kiana Farhadyar (M. Sc.),
Universität Freiburg

Herausforderungen bei der Verwendung Variational Autoencoder für klinische Daten

Der Variational Autoencoder hat eine eingebaute Normalverteilungsannahme. Wir wollen Variational Autoencoder jedoch auch für Daten verwenden, bei denen die ursprünglichen Messungen eine ganz andere Form haben könnten. Es gibt zum Beispiel einige Laborwerte, welche ihr Minimum bei Null haben, im Durchschnitt um 50 liegen und sich aber bis zu Werten im Tausenderbereich erstrecken können, sodass sie nicht einer Normalverteilung folgen. Ein anderes Beispiel sind Messungen, die sich zwischen zwei Gruppen von Personen unterscheiden können, wie es z. B. in Daten zur Lungenfunktion der Fall ist. Ein typischer Messwert in diesen Daten ist die Vitalkapazität, welche bei Frauen und Männern aufgrund der verschiedenen Körpergrößen unterschiedlich ist. Wenn man nun die Verteilung aufzeichnet, ist sie daher nicht mehr glockenförmig, sondern eher eine Kombination aus zwei Glockenkurven, welche durch eine Kerbe getrennt wären. Da wir realistische virtuelle Patient:innen haben wollen, müssen wir in der Lage sein, plausible Werte für derartige Verteilungen zu erzeugen. Wir müssen in den erwähnten Labordaten die kleineren Werte für Frauen im Vergleich zu Männern in der Vitalkapazität in unseren virtuellen Patient:innen wiederfinden, um sie für weitere Forschungen nutzbar zu machen. Leider haben wir festgestellt, dass Variational Autoencoder dazu neigen, Daten zu erzeugen, die einer Glockenkurve folgen, auch wenn die Eingabe anders aussieht. Daher mussten wir eine neue Lösung entwickeln, welche diese Herausforderung angeht.

Unsere Methode

Das Wichtige an den neuen Methoden zur synthetischen Datenerzeugung ist, dass sie nicht auf eine einzige Art von Daten zugeschnitten sein sollten. Es ist zum Beispiel keine gute Lösung, wenn wir nun einige andere Annahmen treffen, welche wiederum die Ergebnisse bei tatsächlich glockenförmigen Daten verschlechtern würden. Bei der Methode, die wir entwickeln, ändern wir also nicht die Annahmen, sondern versuchen, Transformationen zu finden, die schwierige Datenverteilungen bei Bedarf in glockenförmige Verteilungen umwandeln können.

Um die jeweils beste Transformation für die Daten zu finden, benutzen wir Maßzahlen, die anzeigen, ob deren Verteilung glockenförmig ist oder ob es sich um eine andere Verteilung handelt. Nach diesen Maßzahlen kann man die besten Parameter für die Transformation in eine glockenförmige Verteilung finden. Durch die Anwendung dieser Transformationen, kombiniert mit der Verwendung des Variational Autoencoders mit Normalverteilungsannahme und der Verwendung von Rücktransformationen erhalten wir nun synthetische Daten für anspruchsvollere Arten von Verteilungen. Auf diese Weise können wir realistische Daten virtueller Patient:innen für viele verschiedene Anwendungen erstellen.

Diese Methode führen wir unter dem Namen Pre-transformation Variational Autoencoder (PTVA). Das zugehörige Manuskript befindet sich zur Zeit im Gutachterprozess, kann bei Interesse aber gerne zur Verfügung gestellt werden (Kontakt: Kiana Farhadyar, Universität Freiburg).

Die virtuellen Patient:innen in DataSHIELD

Die Generierung virtueller Patientendaten muss verteilt über verschiedene MIRACUM Standorte durchgeführt werden. Dafür und allgemein zur Durchführung gemeinsamer, standortübergreifender Analysen benutzen wir die Software DataSHIELD als Infrastruktur. Das MIRACUM-Team steht in regelmäßigem Austausch mit dem Kernentwicklerteam an der Universität in Newcastle.

Die in DataSHIELD verfügbaren Verfahren konnten wir auch dafür benutzen, um eigene Analyseansätze umzusetzen. Im letzten Jahr konnten wir die Generierung synthetische Daten mittels generativer Modelle bereits an einem anderen Verfahren, den so genannten Deep Boltzmann Machines, in DataSHIELD erproben. Es ist durchaus möglich den DataSHIELD-Werkzeugkasten noch um VAEs zu erweitern, um qualitativ hochwertige synthetische Daten für diverse Verteilungen von Daten erstellen zu können.

Foto: Universität Freiburg

» Diese Resultate zeigen die Bedeutung der standortübergreifenden Nutzbarmachung klinischer Routine- und hochdimensionaler Patientendaten. «

Wo wir jetzt stehen

Im „UC2 – From Data To Knowlege“ haben die MIRACUM Standorte in den letzten Jahren an diversen Studien zu den Themen Asthma/COPD und Neuroonkologie gearbeitet. Das Hauptziel der UC2-Studien besteht darin, durch die gemeinsame Analyse der im MIRACUM-Konsortium verfügbaren Routinedaten medizinisch nützliche Ergebnisse zu erzielen. Die Verwendung von Routinedaten für die Forschung zählt zur sekundär Nutzung und die Daten stehen bzgl. der Datenqualität, -verfügbarkeit und unbekannten Selektionsprozessen bei Patient:innen vor großen Herausforderungen, sodass in der Fachliteratur beständig über die Nützlichkeit diskutiert wird. Ein sekundäres Ziel von UC2 ist es deshalb, die Verlässlichkeit der Routinedaten zu untersuchen. Im Folgenden präsentieren wir eine Auswahl an durchgeführten Studien.

Der Standort Freiburg führt aktuell unter der Leitung von Dr. Sebastian Fähndrich eine Studie mit retrospektiven Daten von Individuen mit Chronisch Obstruktiven Lungenerkrankungen (COPD) durch. Ziel dieser Studie ist es, kardiovaskuläre Komorbiditäten zwischen Individuen mit der zusätzlichen Prädisposition Alpha-1-Antitrypsin-Mangel (AATM) und Individuen ohne AATM zu vergleichen. Frühere Untersuchungen weisen darauf hin, dass Individuen mit COPD und AATM ein niedrigeres kardiovaskuläres Risikoprofil aufweisen als Individuen mit COPD und ohne AATM. Für die Analysen haben wir Routinedaten aus den Datenintegrationszentren von 6 der 10 Universitätskliniken des MIRACUM-Konsortiums verwendet und konnten einige neuartige Erkenntnisse erzielen. In den Analysen können signifikante Unterschiede bei kardiovaskulären Komorbiditäten, Blutlipiden und Glukoseparametern zwischen den beiden Gruppen beobachtet werden. Darüber hinaus können wir abschätzen, welchen Einfluss die Biomarker high sensitivity-Troponin und N-terminales pro B-Type natriuretic peptide (NT-proBNP) auf die Vorhersage der Krankenhaussterblichkeit bei Personen mit AATM haben. Diese vorläufigen Ergebnisse aus Freiburg stimmen mit bereits publizierten Ergebnissen überein, womit die Routinedaten grundsätzlich plausibel sind und bei vorsichtiger Interpretation in der Forschung verwendet werden können.

Am Standort Marburg wurden drei Projekte realisiert. Neben einem „COPD“- und „Pneumonie“-Projekt wurden im Projekt „Asthma“ aus Routinedaten, welche im Rahmen von MIRACUM der Auswertung zugänglich waren, Parameter ermittelt, welche einen Einfluss auf die Dauer des Krankenhausaufenthaltes haben. Dabei flossen Daten aus fünf universitären Standorten ein. Berücksichtigt wurden diverse Laborparameter, Alter, Geschlecht, selektive Komorbiditäten sowie die jeweiligen Entlassungsgründe. Relevante Parameter wurden in Modellen eingefügt, so dass der Einfluss

jedes Einzelparameters bestimmt werden konnte. Hier zeigten sich Unterschiede in Bezug auf die Modellwahl. Dabei lieferte das nicht-lineare Modell zusätzlich zu dem linearen Modell wertvolle Erkenntnisse. Darüber hinaus wurde analysiert, inwieweit die Ergebnisse durch unvollständige Datensätze von Patienten beeinflusst werden. Diese Überlegungen sind essenziell, um einschätzen zu können, ob die gewonnenen Erkenntnisse auf alle Patienten:innen verallgemeinert werden können.

Im Bereich der Neuroonkologie arbeiten die MIRACUM Standorte unter der Federführung von Prof. Dr. Till Acker aus Gießen zusammen mit der BMBF geförderten Nachwuchsforschergruppe AI-RON daran, die Nützlichkeit von in der Routine erhobenen hochdimensionalen Omics- und Bilddaten zur Anlernung von neuronalen Netzwerken unter Verwendung von Swarm-Learning für die Prognoseabschätzung bei Hirntumorpatienten zu evaluieren. Eine erste Anwendung ist Erstellung eines verteilten datenschutzkonformen Hirntumorentitäten spezifischen t-distributed Stochastic Neighbor Embedding (t-SNE) zur Typisierung von Hirntumoren anhand von DNA-Methylierungsdaten, der erstmals erfolgreich vor einigen Wochen zusammen mit den DIZ der Standorte Frankfurt und Marburg umgesetzt wurde und eine ebenso gute True Positive Rate wie ein lokales Pendant zeigt. Das etablierte System ist flexibel erweiterbar um weitere interessierte Standorte bzw. Institute. Aktuell wird daran gearbeitet mit weiteren Standorten eine CopyNumberVariation (CNV) mit datenschutzkonformen Swarm-Learning zu etablieren, da CNVs mit der Prognose bei Hirntumorpatienten korrelieren und wichtige Information für den Behandlungsablauf liefern. In Pilotstudien konnte zudem bereits ein Deep Learning Algorithmus entwickelt werden, welcher in der Lage ist anhand von digitalisierten histopathologischen Schnitten, eine Tumor-Subtypisierung vorzunehmen, welche sonst nur mit Hilfe teurer und zeitaufwändiger Methylierungsarrays erfolgen kann. Diese Resultate zeigen die Bedeutung der standortübergreifenden Nutzbarmachung klinischer Routine- und hochdimensionaler Patientendaten.

Zum Weiterlesen ...

Publikation, die im Use Case 2 entstanden sind:

1. Farhadyar K, Bonofiglio F, Zoeller D, Binder H. Adapting deep generative approaches for getting synthetic data with realistic marginal distributions. export.arXiv.org > stat > arXiv:2105.06907; submitted 14.05.2021

2. Lenz S, Hess M, Binder H. Deep generative models in DataSHIELD. BMC Med Res Methodol 21, 64 (2021). https://doi.org/10.1186/s12874-021-01237-6

3. Gruendner J, Wolf N, Tögel L, Haller F, Prokosch HU, Christoph J. Integrating Genomics and Clinical Data for Statistical Analysis by Using GEnome MINIng (GEMINI) and Fast Healthcare Interoperability Resources (FHIR): System Design and Implementation. JMIR 2020;22:e19879. Doi: 10.2196/19879

4. Jaravine V, Balmford J, Metzger P, Boerries M, Binder H, Boeker M. Annotation of Human Exome Gene Variants with Consensus Pathogenicity. Genes 11 (9), 1076 (2020). https://doi.org/10.3390/genes11091076

5. 5. Gruendner J, Schwachhofer T, Sippl P, Wolf N, Erpenbeck M, Gulden C, Kapsner LA, Zierk J, Mate S, Stürzl M, Croner R, Prokosch HU, Toddenroth D. KETOS: Clinical decision support and machine learning as a service – A training and deployment platform based on Docker, OMOP-CDM, and FHIR Web Services. PLoS ONE 14(10): e0223010. https://doi.org/10.1371/journal.pone.0223010

6. Gruendner J, Prokosch HU, Schindler S, Lenz S, Binder H. A Queue-Poll Extension and DataSHIELD: Standardised, Monitored, Indirect and Secure Access to Sensitive Data. Stud Health Technol Inform. 2019;258:115-119

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

MITGLIED DER