Entwicklung einer webbasierten Anwendung zur interaktiven Analyse klinischer Daten

Rebecca Cristina Fröhlich

Eine Arbeit präsentiert für den Abschluss

Bachelor of Science

Algorithmische Bioinformatik

Heinrich Heine Universität Düsseldorf

Deutschland

5. Juli 2021

Danksagung

Ich möchte mich bei Prof. Dr. Klau dafür bedanken, dass Sie es mir ermöglicht haben, meine Bachelorarbeit an Ihrem Lehrstuhl zu schreiben. Vielen Dank auch an Prof. Dr. Dilthey, dass Sie sich bereit erklärt haben, die Zweitkorrektur zu übernehmen. Ein großes Dankeschön geht an Philipp Spohr für die Beantwortung all meiner Fragen und die großartige Unterstützung während der gesamten Arbeitszeit.

Zusammenfassung

Die Erkenntnisse, die aus gesammelten Daten gewonnen werden können, spielen in der medizinischen Forschung eine große Rolle. In einer großen Menge an Zahlen und anderen Messwerten Zusammenhänge und Muster zu erkennen, ist jedoch eine Herausforderung. Damit die Analyse größerer Datenmengen effektiv erfolgen kann, ist eine geeignete Darstellung der Daten notwendig. Medizinische Daten lassen sich gut analysieren, wenn diese im zeitlichen Verlauf visualisiert werden. Selten werden dabei Daten gemeinsam analysiert, die an verschiedenen Stellen erhoben werden und sich in ihrem Format grundsätzlich unterscheiden. Das Ziel dieser Arbeit ist die gleichzeitige graphische Darstellung von mehreren Datentypen und die Anwendung einer Principal Component Analysis auf einen Teil der Daten. Die Kombination vieler Daten hat zur Folge, dass eine große Menge an Informationen an einem Ort zur Verfügung steht. Somit besteht die Möglichkeit, dass Zusammenhänge erkannt werden, die bei der einzelnen Betrachtung der Daten verborgen geblieben wären. Gleiches gilt für die Durchführung einer Principal Component Analysis.

Inhaltsverzeichnis

1 Einleitung

Die Analyse von Daten ist eine Möglichkeit, um in der Forschung zu neuen Erkenntnissen zu gelangen. Es gibt vielfältige Methoden Daten zu analysieren. Dazu zählen unter anderem die Visualisierung der Daten und die Anwendung einer statistischen Analyse. Visuell dargestellte Daten können deren Analyse erleichtern, da Cluster oder Muster und somit auch Zusammenhänge oder Unterschiede klarer erkennbar sind als dies bei der Betrachtung einer Tabelle der Fall wäre. Statistische Analysen sind von Bedeutung, da diese Zusammenhänge hervorheben können, die bei der bloßen Betrachtung der Daten nicht zu erkennen sind.

Das Ziel der Bachelorarbeit ist die Bereitstellung dieser beiden Analysemethoden in Form einer Webanwendung. Die betrachteten Daten stammen von Leukämiepatienten, die eine Stammzellentransplantation erhalten haben. Intention der Webanwendung ist es, die Untersuchung dieser Daten zu unterstützen und so die Gewinnung möglicher neuer Erkenntnisse zu befähigen. Die Visualisierungen der Webanwendung zeichnen sich besonders dadurch aus, dass eine Kombination von verschiedenen Datentypen wie klinischen Daten, Mikrobiomdaten und Medikationsdaten der Patienten möglich ist. Als statistische Analyse wird die Principal Component Analysis zur Verfügung gestellt. In allen Bereichen ist die Anwendung möglichst interaktiv gestaltet, um den Nutzern eine individuelle Verwendung zu ermöglichen.

2 Hintergrund

In diesem Abschnitt wird auf den wissenschaftlichen Hintergrund des Projektes eingegangen. Dies schließt sowohl eine Beschreibung der betrachteten Erkrankung als auch einen Überblick über die Therapiemethode der Knochenmarktransplantation mit ein. Des Weiteren werden der Ursprung und die jeweilige Verwendung der Eingabedaten thematisiert. Zuletzt erfolgt eine Erläuterung einiger die Webanwendung betreffenden Aspekte, inklusive der Anforderungen, die an die Webanwendung gestellt wurden und bei der Umsetzung beachtet werden mussten.

2.1 Medizinischer Hintergrund

Die im Rahmen der Bachelorarbeit betrachteten Daten stammen von Patienten mit einer Leukämieerkrankung. Leukämie ist eine Form von Krebs, bei der sich die weißen Blutkörperchen (Leukozyten) unkontrolliert vervielfältigen. Es können auch Zellen betroffen sein, die sich erst noch zu Leukozyten entwickeln [1]. Leukozyten werden unterteilt in Granulozyten, Monozyten und Lymphozyten [2]. Granulozyten und Monozyten entwickeln sich aus myeloischen Zellen. Lymphozyten entstehen aus lymphatischen Zellen. [1] Bei einer Leukämieerkrankung wird unterschieden zwischen akuten und chronischen Verläufen sowie zwischen myeloischen und lymphatischen Leukämien. Daraus ergeben sich die vier häufigsten Leukämieformen: Akute myeloische Leukämie, akute lymphatische Leukämie, chronische myeloische Leukämie und chronische lymphatische Leukämie. [2] Die Ursache einer Leukämieerkrankung ist in den meisten Fällen nicht bekannt [3].

Eine Therapiemöglichkeit der Leukämie ist die Knochenmarktransplantation. Mittels einer anderen Therapieform, wie einer Chemotherapie, werden zuvor möglichst alle malignen Zellen zerstört. Danach kann ein Patient eine Knochenmarktransplantation erhalten. Bei der Knochenmarktransplantation werden einem Patienten Knochenmarkzellen zugeführt, die die Bildung gesunder Zellen fördern sollen. Es wird unterschieden zwischen einer allogenen und einer autologen Transplantation. Für eine allogene Knochenmarktransplantation muss ein passender Spender gefunden werden. Gesucht wird nach einer Person, deren HLA-Merkmale (Humane Leukozyten-Antigene) mit den HLA-Merkmalen der erkrankten Person übereinstimmen, um die Wahrscheinlichkeit einer GvHD (Graft-versus-Host-Disease) zu minimieren. [3] Die Suche nach einem geeigneten Spender kann problematisch sein, da es etwa 7000 HLA-Merkmale

[1] *https://www.kompetenznetz-leukaemie.de/content/patienten/leukaemien/e8895/index_ger.html#leukaemie_e13974, zuletzt aufgerufen: 03.07.2021*

[2] *https://www.krebsdaten.de/Krebs/DE/Content/Krebsarten/Leukaemien/leukaemien_inhalt.html, zuletzt aufgerufen: 03.07.2021*

[3] *https://www.blutstammzellspende.ch/de/hintergrundinformationen/hla-merkmale-und-vererbung, zuletzt aufgerufen: 03.07.2021*

gibt. [4] Bei einer autologen Knochenmarktransplantation wird dem Patienten in einer Remissionsphase das Knochenmark entnommen und zu einem späteren Zeitpunkt wieder zugeführt. Bei dieser Form der Knochenmarktransplantation kann es passieren, dass das wieder zugeführte Knochenmark noch maligne Zellen enthält. [4]

2.2 Medizinische Daten - Input der Anwendung

Als Eingabedaten fungieren die im Folgenden näher beschriebenen klinischen Daten, Mikrobiomdaten, Medikationsdaten und Patienten-Charakteristika. Alle Daten stammen von Patienten, die aufgrund einer Leukämieerkrankung eine Knochenmarktransplantation erhalten haben. Die Daten wurden im Universitätsklinikum Düsseldorf erhoben und im Anschluss pseudonymisiert. Dabei wurde jedem Patienten eine Identifikationsnummer (ID) zugewiesen. Die Identifikationsnummern der Patienten in den unterschiedlichen Eingabedaten korrespondieren miteinander, sodass ein Patient in allen Daten immer dieselbe Identifikationsnummer hat. Insgesamt liegen derzeit die Daten von zweiunddreißig Patienten vor. Die vergebenen Identifikationsnummern wurden allerdings nicht fortlaufend von eins bis zweiunddreißig vergeben, sondern liegen verteilt im Bereich von vier bis zweiundvierzig. Zusätzlich zu der Identifikationsnummer wird allen Daten ein Zeitpunkt zugeordnet. Dieser wird in Tagen relativ zum Transplantationstag (Tag 0) angegeben. In den folgenden vier Unterabschnitten werden die spezifischen Eigenschaften der verschiedenen Daten genauer erläutert.

2.2.1 Klinische Daten

Unter klinischen Daten sind in diesem Fall alle Parameter zu verstehen, die im Rahmen einer medizinischen Untersuchung erhoben wurden und in Form eines Diagramms dargestellt werden können. Dazu zählen Blutwerte wie Leukozyten, die sich auf einer quantitativen Skala veranschaulichen lassen. Genauso ist das Stattfinden bestimmter Untersuchungen ist in diesen Daten vermerkt. Unter Verwendung einer nominalen Skala können beispielsweise Materialentnahmen im zeitlichen Verlauf visualisiert werden. Klinische Parameter werden auch als Analyten bezeichnet. Der Wert eines Analyten wird innerhalb der Tabelle eines Laborbereichs immer bezogen auf einen Patienten und einen Zeitpunkt bestimmt.

Insgesamt werden über die fünf betrachteten Bereiche - Mikrobiologie, Serologie, Transfusion, Virologie und Zentrallabor - hinweg 1176 Analyten erfasst. Diese sind jedoch nicht eindeutig, da es namentlich identische Analyten (mit sich unterscheidenden Werten) in mehreren Laborbereichen geben kann. Ein Beispiel dafür ist die Materialentnahme, die es sowohl im Bereich der Serologie und Virologie als auch im Bereich der Mikrobiologie gibt. In der Anwendung

[4] *https://www.krebsgesellschaft.de/onko-internetportal/basis-informationen-krebs/therapieformen/stammzelltransplantation.html, zuletzt aufgerufen: 03.07.2021*

erfolgt die Unterscheidung namentlich identischer Analyten durch das Hinzufügen des zugehörigen Bereichs in Form einer Abkürzung. Bei Verwendung der gegebenen Ordnerstruktur sind diese wie folgt zuzuordnen: Mikrobiologie - mibi, Serologie - sero, Transfusion - transfusion, Virologie - viro und Zentrallabor - zentral. Durch eine Veränderung der Ordnernamen der Eingabedaten können jedoch auch andere Abkürzungen genutzt werden. Wie dies funktioniert wird in 4.2.1 erläutert. Nicht jeder Analyt ist auch für jeden Patienten vertreten. Stattdessen kann es vorkommen, dass sehr spezifische Analyten nur für eine Teilmenge von Patienten zugeordnete Werte haben.

Die Visualisierung innerhalb der Webanwendung erfolgt standardmäßig als Punktdiagramm (Scatterplot), kann aber auch in Form eines Liniendiagramms stattfinden. Die x-Achse fungiert als Zeitachse, die Zeit wird in Tagen relativ zum Transplantationszeitpunkt angegeben, und die y-Achse verhält sich abhängig von der Wahl des Datentyps entweder quantitativ oder nominal. Farblich codiert werden in dieser Darstellung die gewählten Patienten.

2.2.2 Mikrobiomdaten

Die Mikrobiomdaten enthalten Informationen über die Zusammensetzung von Stuhlproben. Bei diesen Informationen handelt es sich um die Ausgabe einer Sequenzierungssoftware (zum Beispiel: kraken2), welche die durch die Stuhlproben gegebenen Daten verarbeitet und so aufbereitet, dass diese graphisch darstellbar sind. Jede Probe ist einem Patienten und einem Zeitpunkt zugeordnet. Für jede Probe liefert die Sequenzierungssoftware Daten, welche für verschiedene taxonomische Stufen Auskunft über die jeweilig vorhandenen Taxone und deren zugehörige Read-Anzahlen geben. Unter taxonomischen Stufen verstehen sich Domäne, Reich, Stamm, Klasse, Ordnung, Familie, Gattung und Art. Je nach Stufe sind zudem unterschiedliche Anzahlen an Subleveln vorhanden.

In der Webanwendung werden die Mikrobiomdaten als gestapelte Säulendiagramme (Barplots) dargestellt. Hierfür müssen zuerst ein Patient sowie eine taxonomische Stufe festgelegt werden. Die x-Achse wird als Zeitachse (in Tagen relativ zum Transplantationstag) verwendet und die Read-Anzahlen werden in normalisierter Form auf der y-Achse dargestellt. Dazu wird für jede Säule die Summe aller Read-Anzahlen als hundert Prozent betrachtet. Die Höhe jedes Abschnitts einer Säule ergibt sich aus dem zugehörigen Anteil (bezogen auf die hundert Prozent) der Read-Anzahlen. Die Visualisierung der Taxone erfolgt mithilfe verschiedener Farben. Informationen wie beispielsweise die absoluten Read-Anzahlen werden im Tooltip angezeigt. Unter Tooltip ist eine Funktion zu verstehen, die in die graphischen Darstellungen integriert ist. Schwebt der Mauszeiger über einen Punkt oder einen farbigen Bereich eines Graphen, so erscheint ein Kästchen mit weiteren Informationen über den gerade betrachteten Ausschnitt des Graphen.

2.2.3 Medikationsdaten

Die Medikationsdaten beinhalten alle Medikationsgaben der Patienten. Dies schließt unter anderem Informationen über den gegebenen Wirkstoff, die Dosierung und die Darreichungsform ein. Außerdem wird jedes Medikament einer der folgenden fünf Kategorien zugeordnet: Antiinfektiva, Immunglobuline, Konditionierung, Zytostatikum und Sonstiges.

Dargestellt werden die Medikationsdaten für jeweils einen Patienten als Punktdiagramm (Scatterplot). Wie auch in den anderen Visualisierungen wird die x-Achse als Zeitachse genutzt. Auf der y-Achse werden die verschiedenen Wirkstoffe in alphabetischer Reihenfolge aufgelistet. Somit hat die y-Achse eine nominale Skalierung. Die zuvor beschriebenen Kategorien werden durch die farbliche Komponente unterschieden. Alle weiteren Informationen werden mithilfe der Tooltip-Funktion zur Verfügung gestellt.

2.2.4 Patienten-Charakteristika

Als Patienten-Charakteristika werden jene Daten bezeichnet, die weitere Eigenschaften der Patienten beinhalten. Unter anderem wird in diesen Daten die Erkrankung sowie die Knochenmarktransplantation eines Patienten mithilfe einiger Anhaltspunkte beschrieben. Ein Beispiel dafür sind die Grunderkrankung, die Art und Anzahl der Vortherapien des Patienten und die Donor Charakteristik. Für alle Patienten werden die identischen Parameter mit individuellen Werten versehen und somit Vergleiche ermöglicht.

Die Besonderheit dieser Daten ist, dass diese nicht in graphischer Form dargestellt werden, sondern nur über den Tooltip zu sehen sind. Da ein Tooltip nur für graphische Darstellungen verwendbar ist, werden die Informationen der Patienten-Charakteristika in der Datentypen kombinierenden Abbildung in den Tooltip der klinischen Daten integriert. Ähnliche Daten werden auch bei der Principal Component Analysis für die Farbgebung der einzelnen Punkte im Diagramm verwendet. Zusätzlich werden auch wieder alle Daten im Tooltip dargestellt. An dieser Stelle wird außerdem zwischen zeitabhängigen und zeitunabhängigen Daten unterschieden. Für zeitunabhängige Charakteristika ändert sich an der Darstellung nichts. Diese werden genauso wie auch in der Abbildung der klinischen Daten angezeigt. Zeitabhängige Charakteristika können je nach gewähltem PCA Datum unterschiedliche Werte für einen Patienten haben.

2.3 Aspekte der Webanwendung

Stand der Anwendung zu Beginn der Bachelorarbeit Zu Beginn der Bachelorarbeit existierte bereits ein Teil der jetzigen Anwendung. Alle im Rahmen der Bachelorarbeit vorgenommenen Implementierungen bauen darauf auf. Der genaue Verlauf der Entstehung der Anwendung kann dem GitLab-Repository, welches den Source-Code enthält, entnommen werden (https://gitlab.cs.uni-duesseldorf.de/albi/medreactor).

Implementierung als Webanwendung Die Implementierung in Form einer Webanwendung ist sinnvoll, da eine Webseite von jedem elektronischem Gerät mit Internetzugang aufgerufen werden kann. Somit ist für die Verwendung der Webseite weder ein spezielles Gerät, noch die Installation von Software notwendig. Aus diesem Grund muss die Anwendung auch nicht an unterschiedliche Betriebssysteme angepasst werden.

Anforderungen an die Webanwendung An die Entwicklung der Webanwendung wurden einige Anforderungen gestellt, die sich aus den Vorstellungen der Verwendung der fertigen Anwendung ergaben. Da sich diese Vorstellungen stetig weiterentwickeln können, können auch die Anforderungen kontinuierlich um kleinere Details erweitert werden. Das Grundgerüst bleibt dabei jedoch bestehen. Dieses sowie andere wichtige Erweiterungen werden in diesem Abschnitt beschrieben. Inwiefern die gestellten Anforderungen erfüllt wurden, wird in Abschnitt 5 diskutiert.

Die wichtigste Anforderung ist die Möglichkeit, verschiedene Datentypen kombinieren zu können. Datentypen beschreibt in diesem Fall die in den Abschnitten 2.2.1 bis 2.2.4 beschriebenen Daten. Die ursprüngliche Idee war es, klinische Daten in Kombination mit Mikrobiomdaten darstellen zu können. Das heißt, beide Datentypen sollen in einem Graph visualisiert werden, die klinischen Daten als Punkt- oder Liniendiagramm und die Mikrobiomdaten als Säulendiagramm. Ergänzt wurde diese Anforderung durch das gleichzeitige Anzeigen der Medikationsdaten. Diese werden aber nicht in demselben Graphen, sondern in einem eigenen Koordinatensystem unter dem kombinierten Punkt- und Säulendiagramm dargestellt. Erstellt man jedoch mehrere Punkt- und Säulendiagramme, so wird die identische Anzahl an Koordinatensystemen mit Medikationsdaten angezeigt. Diese Anzeige erfolgt alternierend. Zu jedem Graph mit klinischen Daten und Mikrobiomdaten gehört eine Darstellung der Medikationsdaten.

Eine weitere Anforderung ergibt sich durch die geforderte Interaktivität der Webanwendung. Um die Anwendung sinnvoll nutzen zu können, ist es von großer Bedeutung, dass die verfügbaren Funktionen frei kombiniert werden können. Aus dieser Anforderung ergibt sich die Option, dass die Visualisierungen sowohl der Mikrobiomdaten als auch der Medikationsdaten je nach Wunsch entweder angezeigt oder ausgeblendet werden können. Im Rahmen der Anforderung der Interaktivität sind Wahlmöglichkeiten von Parametern ebenso wie Filtermöglichkeiten besonders erwünscht.

Die nächsten beiden Anforderungen haben eine Gemeinsamkeit. Für beide ist die Programmierung eines weiteren Tabs innerhalb der Anwendung sinnvoll. Zusätzliche Tabs sind hilfreich, um die Abgrenzung der verschiedenen Optionen der Tabs zu verdeutlichen. Zudem ist so immer klar, welche Einstellungen zu welcher Visualisierungsmöglichkeit gehören. Der erste Tab ist für eine Principal Component Analysis (PCA) bestimmt. Diese soll ebenso interaktiv

zu bedienen sein, wie die kombinatorische Darstellung der verschiedenen Datentypen. Im Optimalfall kann in diesem Tab auch die Entwicklung verschiedener Datenpunkte im zeitlichen Verlauf angezeigt werden. In dem zweiten Tab soll es möglich sein mehrere klinische Parameter gleichzeitig für die y-Achse auszuwählen, um den Vergleich dieser zu erleichtern.

Eine andere Anforderung zielt auf ein vereinfachtes Verständnis der Namen der klinischen Parameter ab. Diese sind in den Dateien meist abgekürzt. Weniger bekannte Abkürzungen können schwierig zu verstehen sein. Somit könnte die Wahl des korrekten Datentyps für die y-Achse (quantitativ oder nominal) erschwert sein. Um dem entgegen zu wirken, sollen nicht nur die Abkürzungen der Parameter in dem Auswahlmenü zur Verfügung stehen, sondern auch deren Erklärungen.

Als letzte Anforderung sollen alle Eingabedaten auf die Korrektheit ihrer Spaltenüberschriften hin überprüft werden, um mögliche Fehlerquellen zu minimieren. Zudem soll sichergestellt werden, dass die klinischen Daten geladen wurden. Alle anderen Daten sind optional. Für diese muss zuerst getestet werden, welche tatsächlich hochgeladen wurden. Im Anschluss können die Spaltenüberschriften aller vorhandenen Daten geprüft werden.

3 Methodik

In diesem Kapitel werden verschiedene technische Aspekte thematisiert. Dazu gehört die Frage, worin eine gute Datenvisualisierung besteht. Weiterhin werden die verwendeten Darstellungswerkzeuge erörtert. Der letzte Punkt ist die Erläuterung der Principal Component Analysis in Bezug auf den mathematischen Hintergrund und die Verwendung in der Implementierung der Webanwendung.

3.1 Datenvisualisierung

Datenvisualisierung wird definiert als die graphische Darstellung von computergenerierten oder anderweitig erfassten und digitalisierten Daten [4, S.477]. In der Webanwendung werden alle Daten in zweidimensionalen Koordinatensystemen visualisiert. Zusätzlich wird in allen Visualisierungen die Farbkomponente zur Darstellung einer weiteren Variable genutzt. Die Eingabedaten der PCA können mehr als zwei Dimensionen haben. Das Ergebnis wird jedoch immer zweidimensional visualisiert. Als ein Teilgebiet der Datenvisualisierung wird die Informationsvisualisierung betrachtet. Innerhalb dieser Unterteilung, getroffen von Card et al. [5], wird Informationsvisualisierung durch Nazemi et al. als „computer-basierte, interaktive visuelle Repräsentation von abstrakten Daten zur Stärkung der Kognition“ [4, S.479] definiert. Nazemi et al. nennen drei Hauptkriterien, die bei der Informationsvisualisierung eine Rolle spielen. Dazu zählen die zu visualisierenden Daten, die Visualisierung und das Ziel der Visualisierung [4, S.480].

Die in der Webanwendung darstellbaren Daten werden durch mehr als einen Datentyp repräsentiert. So gibt es sowohl quantitative, als auch nominale Daten, die es zu visualisieren gilt. Für nominale Daten kann es vorkommen, dass die y-Achse der Visualisierung der klinischen Daten zu klein ist, sodass die Werte der y-Achse übereinander geschrieben werden und unleserlich sind. In diesem Fall muss die y-Achse verlängert werden, damit die Darstellung interpretierbar ist. Das Ziel aller Visualisierungen der Webanwendung ist die größtmögliche Interaktivität. Mithilfe dieser können auch die Praktikabilität und somit die Interpretationsmöglichkeit beeinflusst werden. Ein Beispiel ist die Filtermöglichkeit der Medikationsdaten, durch die kleinere, übersichtlichere Graphen entstehen können.

Eine weitere wichtige Komponente ist die Wahl der Farben einer Darstellung. Diese sollen insbesondere bei der Verwendung von mehreren Farben in einem Säulendiagramm voneinander abgrenzbar sein. Problematisch ist die Abgrenzung der Farben für Personen, deren Farbwahrnehmung eingeschränkt ist. In dem Fall ist es von Vorteil, wenn aufeinanderfolgende Farben auch als Graustufen möglichst große Kontraste aufweisen.

3.2 Verwendete Darstellungswerkzeuge

Dieser Abschnitt dient der Erläuterung zweier Werkzeuge, die für die Gestaltung der graphischen Benutzeroberfläche verwendet wurden. Vega-Lite wird für die Darstellung der Eingabedaten verwendet und befindet sich somit hinter allen in der Webanwendung erstellbaren Koordinatensystemen. Für das Design aller technischen Elemente der graphischen Benutzeroberfläche, beispielsweise Knöpfe oder Tabellen, wird Material-UI genutzt.

3.2.1 Vega-Lite

Die Plots der Webanwendung werden mithilfe der Grammatik Vega-Lite erzeugt. Vega-Lite dient der Entwicklung von Visualisierungen mit interaktiven Elementen [6]. Der Code für einen Plot wird im JSON Format geschrieben und kann in eine JavaScript Datei integriert werden. Durch die Möglichkeit der Verwendung von Funktionen innerhalb des JSON Codes ist die Benutzung von Kontrollstrukturen wie Schleifen und if-Abfragen realisierbar. Diese werden dann in JavaScript geschrieben. Einfache if-Abfragen sind auch mit Vega-Lite alleine umsetzbar.[5]

Um eine elementare Graphik zu erstellen, erfordert Vega-Lite die Spezifikation der folgenden Punkte: data, mark und encoding. Data enthält die Eingabedaten der Graphik. Mark gibt an, welchen Typ beziehungsweise welche Form die Darstellung in der Graphik hat. Beispiele für mark sind Punkt-, Linien- oder Säulendiagramme. Encoding spezifiziert die auf der x-Achse und y-Achse dargestellten Datenfelder/ Spalten der Eingabedaten, sowie den Datentyp. Der Datentyp kann beispielsweise quantitativ, nominal, ordinal oder temporal sein. Über die encoding Einstellungen können auch die Farbgebung und der Tooltip angepasst werden. Weiterhin ist die Festlegung der Breite und Höhe eines Koordinatensystems möglich. Außerdem kann hier die Anwendbarkeit der Zoom-Funktion aktiviert werden. Für die Entwicklung elaborierter Graphiken ist die Verwendung der Funktionen layer und concat von Bedeutung.[6] Mithilfe von layer lassen sich verschiedene graphische Darstellungen gleichzeitig in einem Koordinatensystem plotten. Dies wird in der Webanwendung für die Anzeige der klinischen Daten in Kombination mit den Mikrobiomdaten verwendet. Concat existiert in zwei unterschiedlichen Ausprägungen: vconcat und hconcat. Das v steht für vertikal, während horizontal mit h abgekürzt wird. Gemeint ist mit der jeweiligen Richtung, ob zwei durch concat verknüpfte Koordinatensysteme nebeneinander (horizontal) oder untereinander (vertikal) angezeigt werden. Die Verknüpfung der Koordinatensysteme der Medikationsdaten und klinischen Daten erfolgt durch vconcat. Über die Funktion mit dem Namen resolve kann festgelegt werden, dass sich beispielsweise durch layer kombinierte Graphiken zwar ein Koordinatensystem teilen, die Ach-

[5] *https://vega.github.io/vega-lite/docs/condition.html, zuletzt aufgerufen: 03.07.2021*

[6] *https://vega.github.io/vega-lite/docs/composition.html, zuletzt aufgerufen: 03.07.2021*

sen und Farbgebungen aber unabhängig voneinander sein können.

Die Dokumentation von Vega-Lite ist unter der folgenden URL zu finden: https://vega.github.io/vega-lite/docs/. Dort werden alle möglichen Funktionen mit ihren jeweiligen Optionen erläutert. Zudem gibt es eine große Menge an Beispielgraphiken mit dem jeweils zugehörigen Code im JSON Format.

3.2.2 Material-UI

Jedes auf der Webseite zu sehende, technische Element abgesehen von allen Koordinatensystemen, dem Dateneingabefeld und der Funktion, die die Reihenfolge der Medviewer-Graphiken verändern kann, ist die Implementierung einer Material-UI Komponente. Material-UI ist eine Open Source JavaScript Bibliothek,[7] die speziell für das Design von React-Anwendungen entwickelt wurde und auf Material Design basiert.[8]

Die umfangreiche Dokumentation von Material-UI in Form von Beispielen und Anwendungsschnittstellen (APIs) ermöglicht eine unkomplizierte Benutzung unterschiedlichster Komponenten.[9] Am häufigsten verwendet werden in der Webanwendung die Komponenten Autocomplete, Button, Checkbox, Select, Slider und TextField. Mit diesen Komponenten kann der Benutzer interagieren und so auf die Darstellung der Webanwendung einwirken. Die Pop-Up Fenster, die unter anderem über Fehlermeldungen Auskunft geben und die Tabelle, die die Principal Component Analysis Komponenten auflistet, werden auch mithilfe von Material-UI erstellt. Außerdem werden Design technische Komponenten wie Box, Divider, FormControl und Paper verwendet. Es ist nicht unüblich, dass für die Implementierung einer Komponente mehrere andere Komponenten verwendet werden. Gute Beispiele dafür sind Select, Dialog und Table. Jede Komponente hat Optionen, die verändert werden können, um die Komponente den individuellen Vorgaben einer Anwendung anzupassen. Diese Optionen beziehen sich auf das Design und die Funktion einer Komponente. Für weitere Designmöglichkeiten können Material-UI Komponenten mit CSS kombiniert werden.

3.3 Principal Component Analysis

Die verwendete statistische Analyse in der Webanwendung ist die Principal Component Analysis (PCA). Das Grundprinzip der Principal Component Analysis wurde erstmals 1901 von Pearson [7] beschrieben. Ziel der PCA ist die Reduktion von Dimensionen bei möglichst geringem Datenverlust. Die Anzahl der eingegebenen Dimensionen ist wie auch die Anzahl der ausgege-

[7] *https://github.com/mui-org/material-ui, zuletzt aufgerufen: 03.07.2021*

[8] *https://material-ui.com/company/about/, zuletzt aufgerufen: 03.07.2021*

[9] *https://material-ui.com/, zuletzt aufgerufen: 03.07.2021*

benen Dimensionen individuell festlegbar. Im Rahmen der Webanwendung ist die Anzahl der zu Beginn betrachteten Dimensionen variabel. Ausgegeben werden jedoch nach Anwendung der PCA immer zwei Principal Components. Es erfolgt also immer eine Reduktion auf zwei Dimensionen. Die mathematische Darstellung einer Principal Component erfolgt, ebenfalls nach Pearson, in Form einer Linearkombination:

$$z = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + ... + a_nx_n$$

$z, x_0, x_1, x_2, x_3, ..., x_n$ als Variablen, $a_0, a_1, a_2, a_3, ..., a_n$ als Konstanten (in der Anwendung die Analyten) [7].

Für die Berechnung der Principal Component Analysis in der Webanwendung wird eine existierende JavaScript Bibliothek genutzt.[10] Diese ermittelt die beiden Principal Components mithilfe von Eigenvektoren und Eigenwerten. Für jeden Patienten muss pro Analyt a_i $(i = 1, ..., n)$ mindestens ein Wert vorhanden sein. Es besteht die Möglichkeit, dass für einen Patienten und einen Analyt mehrere Werte zur Verfügung stehen. In diesem Fall wird der Durchschnitt aller vorhandenen Werte berechnet und im weiteren Verlauf als einziger Datenwert genutzt. Bevor die Berechnung der PCA beginnt, werden für jeden gewählten Analyt die Datenwerte normalisiert. Dazu werden zuerst das Minimum und Maximum jedes Analyten bestimmt. Darauffolgend wird jeder Wert w mithilfe der folgenden Gleichung normalisiert:

$$w_{normalisiert} = \frac{w - Minimum}{Maximum - Minimum}$$

Jeder in der PCA betrachtete Wert hat dann einen Wert zwischen null und eins.

Die erklärte Varianz gibt an, welchen Anteil der betrachteten Daten eine Principal Component repräsentiert. Berechnet wird diese für eine Principal Component p durch die folgende Gleichung:

$$Var_p = (\frac{Eigenwert_p}{\sum \forall Eigenwerte}) * 100$$

[10] *https://github.com/bitanath/pca, zuletzt aufgerufen: 03.07.2021*

4 Ergebnisse

Dieser Abschnitt behandelt die Ergebnisse des Projektes. Da ein großer Teil des Resultats durch die Webanwendung selbst repräsentiert wird, befasst sich dieser Abschnitt hauptsächlich mit der Verwendung der Anwendung. Dazu zählt einmal das Starten der Anwendung. Des Weiteren wird auf die Datenformate eingegangen, welche notwendig sind, damit das Hochladen von Eingabedaten problemlos ablaufen kann. Beschrieben werden außerdem die Funktionen der Anwendung.

4.1 Starten der Anwendung

Um die Webanwendung verwenden zu können, muss diese zuerst gestartet werden. Dafür gibt es zwei Optionen: Die Anwendung kann online als Webseite abgerufen werden oder die Anwendung wird als Localhost auf dem eigenen Computer gestartet. In beiden Fällen erfolgt das Hochladen von Eingabedaten und das tatsächliche Verwenden der Anwendung in einem Webbrowser.

Online Um die Anwendung online abrufen zu können, muss eine Internetverbindung vorhanden sein. Es ist erforderlich, dass die IP-Adresse des genutzten Gerätes Teil des Universitätsnetzwerks ist. Für die Verwendung der Webseite auf dem Universitätsgelände muss nur das durch die Universität zur Verfügung gestellte Netzwerk genutzt werden, um die Webseite zu laden. Um die Webseite außerhalb der Universität aufrufen zu können, ist es notwendig die IP-Adresse des Gerätes mithilfe einer VPN-Verbindung zu verändern. So kann die Anwesenheit in der Universität simuliert werden. Dem Gerät wird eine IP-Adresse des Universitätsnetzwerks zugeordnet, obwohl es sich tatsächlich nicht in diesem befindet. Auf diese Weise kann die Webseite von jedem Standort aus genutzt werden. Die URL der Webseite lautet: https://medreactor.cs.uni-duesseldorf.de/. Diese kann in jedem beliebigen Webbrowser geöffnet werden. Es wird das Menü angezeigt, welches zum Auswählen und Hochladen der Eingabedaten genutzt werden kann (siehe Abbildung 1).

Als Localhost Zu Entwicklungszwecken kann die Anwendung auch als Localhost gestartet werden. Wie dies funktioniert kann der README.md Datei des zur folgenden URL gehörenden Repositorys entnommen werden: https://gitlab.cs.uni-duesseldorf.de/albi/medreactor. In diesem befindet sich auch der zugehörige Source-Code. Da dieses Repository nicht explizit für die Bachelorarbeit erstellt wurde, können darin weiterhin Veränderungen durch andere Personen vorgenommen werden. Deshalb wurde der Stand des Repositorys zur Abgabe der Bachelorarbeit mithilfe eines Digital Object Indentifier permanent festgelegt [8].

4.2 Geforderte Datenformate

In Abschnitt 2.2 wurde der Inhalt der verschiedenen Datentypen beschrieben. In diesem Abschnitt wird auf die technischen Voraussetzungen eingegangen, die von den Eingabedaten

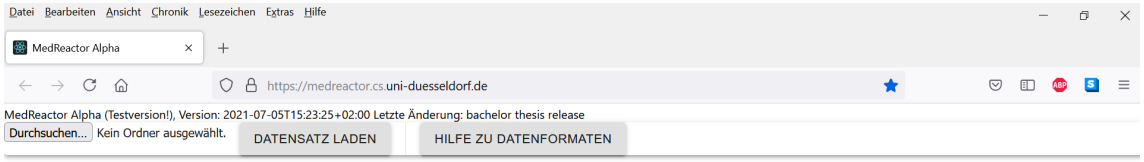

Abbildung 1: Ansicht des Browsers direkt nach dem Aufrufen von https://medreactor.cs.uni-duesseldorf.de/

erfüllt werden müssen, damit diese von Programm korrekt verarbeitet werden können. Zum einen ist dafür die Ordnerstruktur der Eingabedaten relevant, zum anderen müssen sich die richtigen Daten in den korrespondierenden Dateien befinden. Für diese Überprüfung sind der Dateiname und die Spaltenüberschriften (Schlüssel) von Bedeutung. Eine ähnliche, kürzere Beschreibung in englischer Sprache ist in der README.md Datei des GitLab Verzeichnisses mit dem Source-Code unter der folgenden URL zu finden: https://gitlab.cs.uni-duesseldorf.de/albi/medreactor. Derselbe Text ist auch hinter dem ‚Hilfe zu Datenformaten' - Knopf in der Webanwendung hinterlegt (siehe Abschnitt 4.3.1).

Das Hochladen der Daten funktioniert unabhängig von der Methode, die zum Starten der Anwendung verwendet wird, auf identische Art und Weise. Dies gilt auch für alle folgenden Funktionen. Sobald die Startseite geladen wurde gibt es keine Unterschiede mehr.

Für Teilabschnitte 4.2.2 bis 4.2.5 gilt: die geforderten Schlüssel und somit auch die Spalten können in jeder gewünschten Reihenfolge angegeben werden. Definiert als Schlüssel werden die Spalten, die definitiv vorhanden sein müssen. Zusätzliche Spalten sind erlaubt, werden jedoch vom Programm nicht berücksichtigt.
Die für jede Datei geforderten Schlüssel können der README.md Datei entnommen werden (https://gitlab.cs.uni-duesseldorf.de/albi/medreactor).

4.2.1 Ordnerstruktur der Eingabedaten

Die Startseite der Webanwendung (Abbildung 1) verfügt über einen ‚Durchsuchen...' - Knopf, der das Ordnermenü des verwendeten Computers öffnet. An dieser Stelle kann ein Ordner zum Hochladen ausgewählt werden. Da die Wahl einzelner Dateien nicht möglich ist, müssen

sich alle gewünschten Dateien in diesem Ordner befinden. Jede Datei im Upload-Ordner muss die Dateiendung .csv haben. Notwendig für die Verwendung der Anwendung sind grundsätzlich nur die klinischen Daten. Da sich diese jedoch in denselben Dateien wie die Patienten-Charakteristika befinden (für weitere Erläuterungen siehe Abschnitt 4.2.2), sind auch diese Daten unabdingbar. Optional hingegen sind sowohl die Mikrobiomdaten und Medikationsdaten als auch die Charakteristikadaten der PCA. Werden diese jedoch nicht hochgeladen, so können auch die dazugehörigen Funktionen nicht verwendet werden.

Wie in 2.2.1 bereits beschrieben, werden die Namen der unterschiedlichen Bereiche, welchen die Analyten der klinischen Daten entstammen, in Form einer Abkürzung den Analytnamen hinzugefügt, um auch zwischen Analyten, die denselben Namen haben, aber in verschiedenen Laboren untersucht wurden, differenzieren zu können. Das Format ist das Folgende: Analytname (Laborname). Die Abkürzungen, die die Labore unterscheiden, sind dabei mithilfe der Ordnernamen der Eingabedaten beeinflussbar. Da die Kombination aus Analyt- und Laborname benutzt wird, um die Daten eines in einem späteren Schritt gewählten Analyten aus der Gesamtmenge der Daten herauszufiltern, ist es notwendig, die im Folgenden beschriebene Ordnerstruktur einzuhalten. Sei hier einmal eine Beispielstruktur dargestellt, die alle vorhandenen Datentypen enthält.

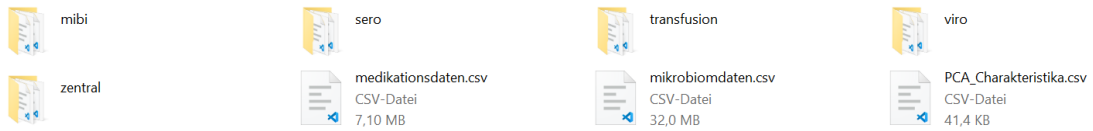

Abbildung 2: Alle dargestellten Ordner und Dateien befinden sich im Upload-Ordner. Jeder der fünf Ordner enthält die Dateien mit den dem Laborzweig zugehörigen klinischen Daten. In diesem Beispiel wurden die in 2.2.1 erklärten Laborabkürzungen verwendet.

Zur Veränderung der Labornamen muss nur der entsprechende Ordner umbenannt werden. Wird beispielsweise mibi in mikrobiologie geändert, so wird der Analyt X aus diesem Ordner als X (mikrobiologie), statt als X (mibi) angezeigt werden. Es können beliebig viele Ordner mit klinischen Daten im Upload-Ordner vorhanden sein. Diese sollten jedoch nur in einem Ordner ‚klinische Daten' zusammengefasst werden, wenn in der Anwendung keine Differenzierung in die verschiedenen Laborbereiche erwünscht ist. Ein solcher Ordner würde bewirken, dass alle Analyten wie folgt dargestellt werden: Analytname (klinische Daten). Aus einem ähnlichen Grund sollten klinische Daten auch in einen Ordner sortiert werden, wenn nur die Daten eines Laborbereichs hochgeladen werden. Angenommen, es sind nur die Mikrobiologiedaten gewünscht, so müssen sich diese trotzdem in einem Ordner (zum Beispiel: mibi), wie auch in Abbildung 2 gezeigt, befinden. Dies hat gegenüber einer Sortierung ohne Subordner im Upload-Ordner den Vorteil, dass Analyten ansonsten für jeden Patienten einzeln vorkommen (nach folgendem Prinzip: Analytname (ID.csv)) und aus diesem Grund auch nur für jeweils den zugehörigen Patienten angezeigt werden können.

4.2.2 Klinische Daten und Patienten-Charakteristika

Da die Informationen der Patienten-Charakteristika im Tooltip der Graphik der klinischen Daten dargestellt werden (siehe 2.2.4), wurden diese beiden Datentypen in jeweils einer Datei zusammengefasst. Aus diesem Grund haben diese Dateien eine größere Anzahl an Spalten. Wird im Folgenden von klinischen Daten gesprochen, so schließt dies die Patienten-Charakteristika mit ein. In 2.2 wurde bereits beschrieben, dass jeder Patient bei der Pseudonymisierung eine Identifikationsnummer zugeordnet bekommt. Für jeden Patienten gibt es pro Laborgebiet eine Eingabedatei. Bei fünf Laborbereichen, wie in Abbildung 2, gibt es somit für jeden Patienten fünf Dateien, eine in jedem Unterordner. Jede der Dateien muss nach dem folgenden Schema benannt werden: Nummer.csv. Dies ist notwendig für die Sortierung der Dateien im Programm. Jede Datei, deren Name mit einer Zahl beginnt wird wie eine Datei mit klinischen Daten behandelt. Deshalb müssen alle Dateinamen, deren Dateien klinische Daten enthalten, mit einer Nummer anfangen. Im Umkehrschluss sollten auch nur Dateien mit klinischen Daten eine Zahl im Namen haben. Wichtig ist, dass die Zahl nicht als Wort ausgeschrieben wird. Im Normalfall korrespondiert die Zahl im Dateinamen mit der Identifikationsnummer des Patienten. Die Anzahl der möglichen Dateien, die klinische Daten enthalten dürfen, ist nicht begrenzt. Die Spalten dieser Dateien sind tab-separiert. Beispiele für enthaltene Spalten sind: Patienten ID, Datum und Messwerte.

4.2.3 Mikrobiomdaten

Die Mikrobiomdaten aller Patienten werden in einer Datei gespeichert. Daraus resultierend ist es nicht erlaubt, mehr als eine Datei dieser Art im Upload-Ordner zu haben. Die Mikrobiomdatendatei muss das Wort Mikrobiomdaten im Dateinamen enthalten und die Endung .csv haben. mikrobiomdaten wird ebenfalls zugelassen. Gegengleich zu 4.2.2 muss der Dateiname mit einem Buchstaben beginnen. Die Datei ist Komma-separiert. Tabelle 1 gibt einen exemplarischen Überblick über den Inhalt.

PatID	Datum	Read-Anzahl	Level	Taxon
1	0	255	D	Bacteria
1	-10	1	D	Bacteria
2	5	312	C	Firmicutes
...	...	...	...	...

Tabelle 1: Die hier dargestellten Daten sind frei erfunden und sollen nur einen Überblick vermitteln.

Wie in der Tabelle 1 gezeigt, werden die Level durch das Sequenzierungsprogramm kraken2 mit dem jeweiligen Anfangsbuchstaben der taxonomischen Stufe in englischer Sprache abgekürzt. In der Webanwendung werden die Anfangsbuchstaben in die Langform der taxonomischen Stufen übersetzt. Eine Übersicht der Stufen, die von dem Programm erkannt und übersetzt werden, befindet sich in Tabelle 4 im Anhang. Alle dort nicht aufgeführten Abkür-

zungen können von der Anwendung nicht transformiert werden und erscheinen so, wie sie in den Eingabedaten stehen, auch in der graphischen Benutzeroberfläche.

4.2.4 Medikationsdaten

Genau wie die Mikrobiomdaten werden auch die Medikationsdaten aller Patienten in einer Datei zusammengefasst. Dementsprechend ist auch von dieser Dateiart nur eine Datei im Upload-Ordner erlaubt. Der Dateiname muss das Wort Medikationsdaten enthalten und mit einem Buchstaben beginnen (siehe 4.2.2). Das Wort mit kleinem Anfangsbuchstaben ist auch erlaubt. Die Datei ist Komma-separiert und hat die Dateiendung .csv.

Dosierung	Wirkstoff	Darreichungsform	Datum	PatID	Zuordnung
1-0-1	Valaciclovir	p.o.	0.0	1	Antiinfektiva
1-0-1	Valaciclovir	p.o.	-5.0	1	Antiinfektiva
1-0-1	Valaciclovir	p.o.	10.0	2	Antiinfektiva
...	...	...	...	...	...

Tabelle 2: Die hier dargestellten Daten sind frei erfunden und sollen nur einen Überblick vermitteln.

4.2.5 PCA Charakteristika

Diese Tabelle beinhaltet die Patienten Charakteristika, die für die Farbgebung innerhalb des Koordinatensystems mit den PCA Ergebnissen genutzt werden. Von dieser Tabellenart kann nur eine gleichzeitig hochgeladen werden. Das im Dateinamen zwingend notwendige Wort lautet Charakteristika (bzw. charakteristika). Die Datei ist Komma-separiert und hat dementsprechend die Dateiendung .csv.

4.3 Beschreibung der interaktiven Funktionen der Webanwendung

In diesem Abschnitt erfolgt eine Auflistung und kurze Beschreibung aller durch die Webanwendung bereitgestellten Funktionen. Diese werden aufgeteilt in grundlegende Funktionen, Features des Medviewers und die Principal Component Analysis.

4.3.1 Grundlegende Funktionen

Hochladen von Daten Wie schon in Abschnitt 4.2.1 erwähnt, erfolgt das Hochladen von Daten durch das Drücken des ‚Durchsuchen...' - Knopfs und das anschließende Auswählen eines Ordners. Wurde der entsprechende Ordner gefunden und ausgewählt, öffnet sich ein kleines Fenster, in dem die Upload-Absicht bestätigt werden muss, um die Daten tatsächlich hochzuladen. Durch die Betätigung des ‚Datensatz laden' - Knopfs wird der Upload-Prozess ausgelöst. Abhängig von der Anzahl der Dateien im Upload-Ordner kann dies einige Sekunden dauern. Nach dem erfolgreichen Hochladen der Daten wird das Menü des Medviewer-Tabs angezeigt (siehe Abbildung 3).

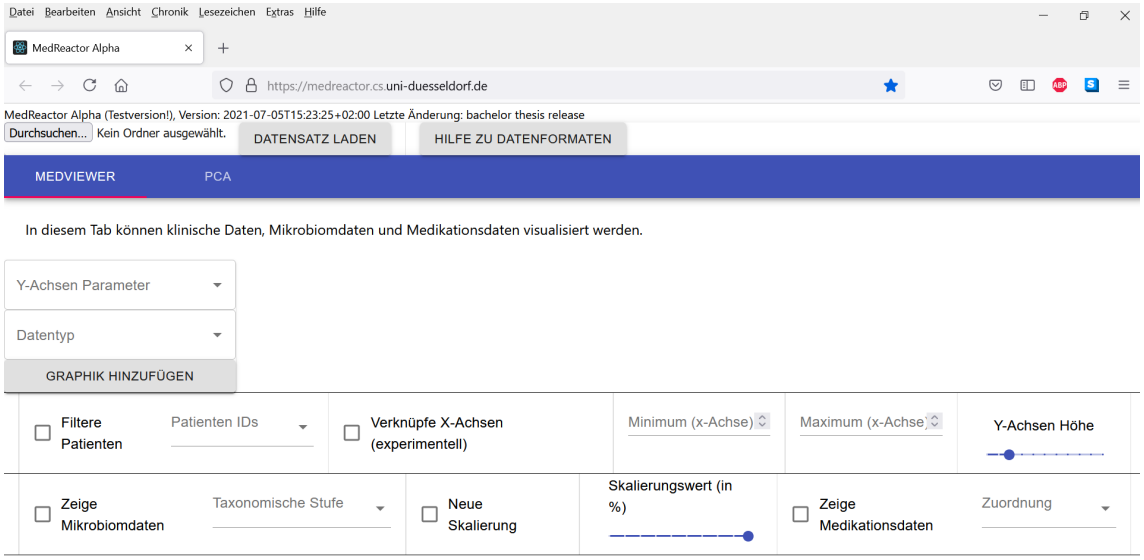

Abbildung 3: Ansicht des Medviewer-Tabs nachdem alle verfügbaren Daten erfolgreich geladen wurden. Liegen die Mikrobiomdaten oder die Medikationsdaten nicht vor, so werden die jeweils zugehörigen Menüpunkte deaktiviert (siehe Abbildung 8).

Um einen anderen Datensatz zu verwenden, nachdem ein Datensatz bereits hochgeladen wurde, kann entweder die Webseite neu geladen werden oder es wird direkt auf den ‚Durchsuchen...' - Knopf geklickt. Wird direkt ein neuer Datensatz hochgeladen ohne die Webseite zu aktualisieren, so erscheint nach der Betätigung des ‚Datensatz laden' - Knopfs ein Fenster, welches eine Bestätigung zum Verwerfen des aktuellen Datensatzes fordert, bevor die neuen Daten geladen werden. Bei beiden Methoden werden alle aktuellen Daten, Einstellungen und Graphen verworfen.

Hilfe zu den Datenformaten Werden Daten ausgewählt, die nicht mit dem geforderten Format übereinstimmen, so kann es zu Fehlermeldungen kommen. Das erforderliche Format ist im gesamten Abschnitt 4.2 ausführlich beschrieben. Zusätzlich ist der Text aus der README.md Datei des GitLab Repository (siehe auch Abschnitt 4.2) zum Thema Datenformate in der Webanwendung hinterlegt. Dieser wird durch das Betätigen des ‚Hilfe zu Datenformaten' - Knopfs angezeigt. Fehler, die den Ladeprozess eines Datensatzes abbrechen, sind:

- mindestens eine Datei, deren Endung nicht .csv entspricht
- mehr als eine Datei deren Name das Wort Mikrobiomdaten, Medikationsdaten oder Charakteristika enthält
- mindestens eine Datei, die aufgrund ihres Namens nicht zugeordnet werden kann
- mindestens eine Datei mit klinischen Daten, die nicht über alle erforderlichen Schlüssel verfügt

In diesen Fälle wird das Menü des Medviewer-Tabs (Abbildung 3) nicht angezeigt, da der Ladeprozess nicht abgeschlossen wird. Zusätzlich gibt es noch drei weitere Meldungen, die beim Hochladen von Daten angezeigt werden können. Fehlt entweder in der Mikrobiomdatendatei, der Medikationsdatendatei oder der PCA Charakteristikadatei einer der zwingend notwendigen Schlüssel, so wird dies durch eine Fehlermeldung auf dem Bildschirm angegeben. Die Konsequenz ist in diesem Fall aber nicht das Abbrechen des Ladeprozesses. Stattdessen werden die zugehörigen Optionen im zugehörigen Menü deaktiviert.

Tabs Die Webanwendung verfügt über zwei Tabs. Im Medviewer-Tab können alle Datentypen (klinische Daten inklusive Patienten-Charakteristika, Mikrobiomdaten und Medikationsdaten) visuell dargestellt werden. Dieser Tab wird immer nach dem Laden eines Datensatzes angezeigt. Der zweite Tab kann genutzt werden, um eine Principal Component Analysis durchzuführen. Die einzelnen Möglichkeiten der beiden Tabs werden im Folgenden genauer erläutert. Welcher Tab gerade angezeigt wird, kann zum einen am Inhalt erkannt werden. Zum anderen ist die Schrift des angezeigten Tabnamens weiß anstelle von grau und zusätzlich ist der Name des Tabs rot unterstrichen. Der Wechsel erfolgt durch das Klicken auf den gewünschten Tabnamen.

4.3.2 Tab: Medviewer

Der Tab mit dem Namen Medviewer ist bestimmt für die kombinierte Visualisierung aller gegebenen Datentypen. Wichtig ist dabei, dass das Erstellen jedes Graphen und die verschiedenen Anpassungsmöglichkeiten interaktiv nutzbar sind. Der folgende Abschnitt beschäftigt sich mit der Erstellung eines Graphen im Allgemeinen und den Optionen zur Anpassung des Graphen. Einige der Einstellungsoptionen beziehen sich auf einen bestimmten Datentyp.

Erstellen einer Graphik Um den vollen Umfang der Einstellungsmöglichkeiten ausschöpfen zu können, muss zuerst eine Graphik mit einem klinischen Parameter erstellt werden. Notwendig dafür ist die Wahl eines Y-Achsen Parameters. Diese erfolgt über ein Autocomplete-Feld[11] mit demselben Namen (Y-Achsen Parameter) am linken Rand der Webanwendungsseite. Durch das Auswählen des Feldes öffnet sich ein Dropdown-Menü, welches über alle Parameter verfügt, die in den Eingabedaten in der Spalte ABKU enthalten sind. Wie in 4.2.1 erklärt, werden die Analytnamen durch den zugehörigen Labornamen, der in Klammern dahinter angezeigt wird, ergänzt. Dies dient zum einen der eindeutigen Zuordnung der Analyten. Zum anderen wird so die Auswahl eines Analyten erleichtert. Der Vorteil eines Autocomplete-Felds an dieser Stelle ist die Suchfunktion des Feldes. Alle Analyten werden in alphabetischer Reihenfolge im Dropdown-Menü dargestellt. Durch das Eingeben von Teilen von Analyt- oder Labornamen wird die Liste gefiltert und das Finden eines spezifischen Analyten wird erleichtert. So können

[11] *https://material-ui.com/components/autocomplete/, zuletzt aufgerufen: 03.07.2021*

auch nur Parameter eines Labors angezeigt werden.

Nachdem ein Analyt gewählt wurde, muss der Datentyp festgelegt werden, der die Darstellung der Daten auf der Y-Achse beeinflusst. Zur Auswahl stehen die Datentypen quantitativ und nominal. Quantitativ ist der empfohlene Datentyp für Parameter, bei denen Daten aus Zahlwerten bestehen. Nominal ist eine gute Wahl für Daten, die wenige unterschiedliche Werte haben. Ein Beispiel dafür wären + und - oder positiv und negativ als Datenwerte. Wichtig zu erwähnen ist, dass mit dem Datentyp quantitativ nur Werte in der Graphik darstellbar sind, die aus Ziffern und potentiell einem Punkt für Dezimalzahlen bestehen. Werte mit anderen Symbolen werden für diesen Datentyp nicht angezeigt.

Wenn auch der Datentyp festgelegt wurde, kann die Graphik mithilfe des ‚Graphik hinzufügen' - Knopfs erstellt werden. Diese erscheint dann unter den bereits vorhandenen Menüpunkten. Pro Plot kann je ein Analyt für die Y-Achse gewählt werden. Es können beliebig viele Plots mit ebenso vielen unterschiedlichen Analyten untereinander dargestellt werden. Beim Hinzufügen eines neuen Graphen, wenn vorher mindestens ein anderer Graph erstellt wurde, erscheint dieser unter allen bisher visualisierten Plots. Die Reihenfolge verschiedener Graphiken kann verändert werden, indem die sechs Punkte in der linken oberen Ecke einer Graphik festgehalten werden und die Graphik so an die gewünschte Stelle gezogen wird. Der ‚Graphik entfernen' - Knopf löscht die zum Knopf gehörige Graphik. Durch Aktivieren des Kästchens mit der Beschreibung ‚Zeichne Linie' werden zusammengehörende Punkte (alle Punkte eines Patienten) durch eine Linie verbunden. Durch das erneute Bestätigen des Kästchens wird zur Punktdarstellung zurückgekehrt.

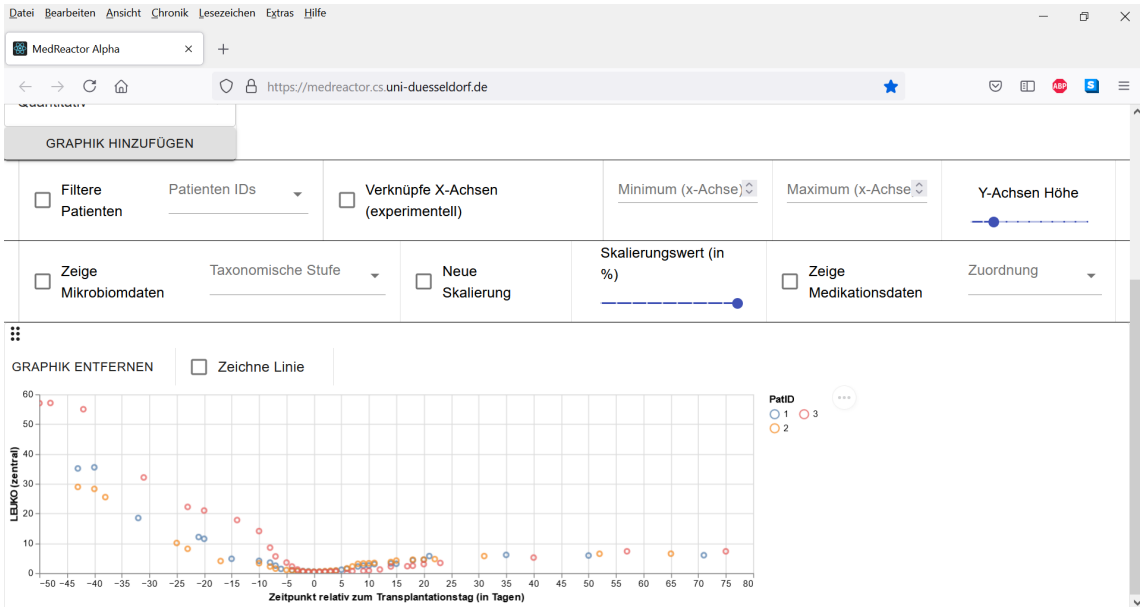

Abbildung 4: Ansicht einer Graphik mit frei erfundenen Daten direkt nach der Erstellung bevor weitere Einstellungen vorgenommen wurden. Zu jeder Graphik gehören eine Legende, der ‚Graphik entfernen' - Knopf, die sechs Punkte über diesem und das Kontrollkästchen ‚Zeichne Linie'.

Patienten ID Filter Unter Patienten ID Filter ist das Kontrollkästchen mit der Beschriftung ‚Filtere Patienten' sowie das Auswahlfeld mit der Bezeichnung ‚Patienten IDs' zu verstehen. Diese Einstellung bezieht sich wie alle anderen Optionen, die sich zwischen der Parameter und Datentyp Auswahl und dem ersten Graphen befinden, auch auf alle erstellten Graphen.

Ist das Kästchen deaktiviert, so wird keine Filterung vorgenommen. Es werden dann alle verfügbaren Patienten in der Legende angezeigt und die zugehörigen Datenpunkte geplottet. Verfügbare Patienten bedeutet, dass für alle Patienten, für die mindestens ein Datenwert für den gewählten Analyt in den Eingabedaten existiert, mindestens ein Punkt in dem Graphen zu sehen ist. Deshalb ist jeder dieser Patienten dann auch in der Legende vertreten. Für häufig analysierte Parameter wie Leukozyten oder C-reaktives Protein sind Werte für alle 32 Patienten vorhanden. Weniger oft gemessene Parameter haben unter Umständen nicht für alle Patienten mindestens einen Datenpunkt. In diesen Fälle sind nur die Patienten mit Daten im Graphen und der Legende vertreten.

Wird das Kontrollkästchen aktiviert, während das Auswahlfeld noch leer ist, verschwinden zunächst die bereits erstellten Graphen, sodass nur noch die Achsen der Koordinatensysteme zu sehen sind. Über das Auswahlfeld kann eine beliebige Anzahl an Patienten Identifikationsnummern gewählt werden. Gibt es für diese für den jeweiligen Y-Achsen Parameter Daten, so erscheinen diese im zugehörigen Graphen. Sowohl die Mikrobiomdaten als auch die Medikationsdaten können nur für jeweils einen Patienten gleichzeitig visualisiert werden. Sobald mehr als eine Identifikationsnummer im Filter gewählt wird, wird daher nur die Darstellung des klinischen Parameters für alle gewählten Patienten angezeigt.

Optionen für Mikrobiomdaten Zu den Optionen für die Mikrobiomdaten zählen das Kontrollkästchen ‚Zeige Mikrobiomdaten', das zugehörige Auswahlfeld mit der Beschriftung ‚Taxonomische Stufe', das Kontrollkästchen ‚Neue Skalierung' und der Schieberegler ‚Skalierungswert (in %)'. Um die Mikrobiomdaten eines Patienten zu visualisieren muss das Kästchen ‚Zeige Mikrobiomdaten' aktiviert werden. Außerdem muss mindestens eine taxonomische Stufe aus der gegebenen Liste ausgewählt werden. Die Visualisierung kann nur erfolgen, wenn genau ein Patient im Patientenfilter festgelegt wurde.

Graphisch dargestellt werden die Mikrobiomdaten als Säulendiagramme, auch Barplots genannt. Die Säulen werden im selben Koordinatensystem angezeigt, wie auch die klinischen Daten. Beide Darstellungen teilen sich eine x-Achse. Eine Säule steht an der Stelle der x-Achse, die mit dem Zeitpunkt der Probe, die durch die Säule visualisiert wird, korrespondiert. Die y-Achse des Säulendiagramms ist unabhängig von der y-Achse der klinischen Daten. Aus diesem Grund gibt es in diesem kombinierten Diagramm zwei y-Achsen. Die linke y-Achse verändert sich nicht und bezieht sich weiterhin auf den dargestellten klinischen Parameter. Auf der rechten Seite des Koordinatensystems wird eine weitere y-Achse hinzugefügt, die dem Säulendia-

gramm zugeordnet ist. Die Einteilung der neuen y-Achse ist immer gleich: sie beginnt bei null und endet bei eins. Dargestellt werden die normalisierten Read-Anteile einer Probe für die gewählte taxonomische Stufe. Eine Säule besteht somit aus mehreren kleineren Abschnitten, die aufeinander gestapelt die angezeigte Säule ergeben. Jeder Bereich einer Säule steht für den normalisierten Read-Anteil eines Taxons. Je größer ein Bereich ist, desto größer ist auch der normalisierte Read-Anteil des Taxons.

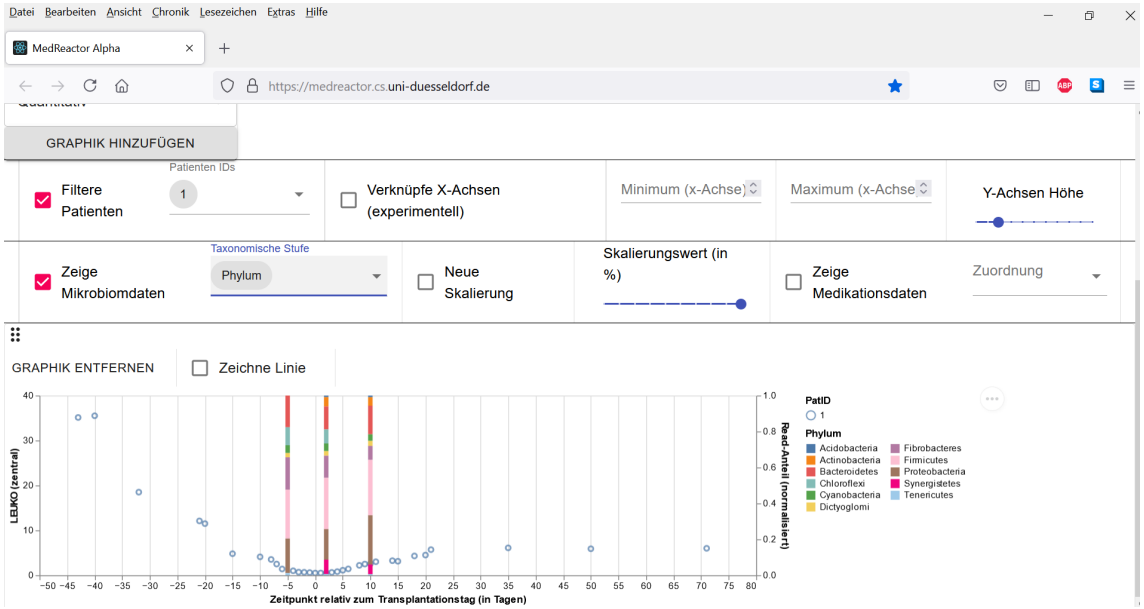

Abbildung 5: Ansicht der klinischen Daten (Leukozyten) und Mikrobiomdaten für den Patienten mit der ID 1 (frei erfundene Daten)

Die verschiedenen Bereiche sind durch Farben voneinander unterscheidbar. Insgesamt werden zwanzig Farben verwendet. Sobald alle Farben einmal genutzt wurden, wird in derselben Reihenfolge wieder mit der ersten Farbe begonnen. Werden taxonomische Stufen mit vielen Taxonen dargestellt, so kann eine Farbe in einer Säule mehrfach vorkommen. In diesem Fall kann der Tooltip bei der Identifizierung der Taxone helfen. Bei den für die Säulendiagramme verwendeten Farben handelt es sich zu einem Großteil um die Farbpalette tableau20[12]. Drei der 20 Farben wurden jedoch ausgetauscht, um möglichst wenige Farben zu verwenden, die sich sehr ähnlich sehen. Dabei wurden auch die beiden Grautöne der Farbpalette ersetzt. Außerdem wurde die Reihenfolge der Farben so verändert, dass der Kontrast von in der Legende aufeinander folgenden Farben möglichst groß ist. Im Source-Code können die derzeit verwendeten Farben durch jede Kombination an Farben, die in hexadezimaler Form angegeben werden, ersetzt werden. Die Reihenfolge der Farben in der Legende kann so festgelegt werden. Bei der Verwendung in den Säulendiagrammen ist es jedoch häufig der Fall, dass nicht alle in der Legende aufgelisteten Taxone in einer Säule sichtbar sind. Somit verschiebt sich die Reihenfolge der Farben je nach Zusammensetzung einer Säule. Aus diesem Grund kann nicht

[12] *https://vega.github.io/vega/docs/schemes/, zuletzt aufgerufen: 03.07.2021*

gewährleistet werden, dass sich die einzelnen Abschnitte einer Säule durch den größtmöglichen Kontrast voneinander abheben. Dies kann die Praktikabilität der Darstellung für Personen mit einer eingeschränkten Farbwahrnehmung beeinträchtigen, da unter Umständen nicht direkt erkennbar ist, wo ein Abschnitt einer Säule beginnt oder endet. Diese Informationen sind aber auch über den Tooltip ablesbar, sodass die Verwendung der Anwendung trotzdem möglich ist.

Das Auswahlfeld, über welches die taxonomische Stufe ausgewählt wird, erlaubt die Auswahl mehrerer taxonomischer Stufen. Wird mehr als eine taxonomische Stufe gewählt, so vergrößert sich die Anzahl der Säulen im Koordinatensystem. Grundsätzlich gilt: die Anzahl der Säulen entspricht der Anzahl der gewählten taxonomischen Stufen multipliziert mit der Anzahl der Proben, die für den betrachteten Patienten vorhanden sind. Von dieser Regel gibt es zwei Ausnahmen. Für ein Sublevel einer taxonomischen Stufe kann es vorkommen, dass in einer Probe keine Daten für dieses Sublevel vorhanden sind. In diesem Fall bleibt die Stelle in der Visualisierung, an der die Säule hätte angezeigt werden sollen, leer. Die zweite Ausnahme betrifft die Anzahl der gewählten taxonomischen Stufen. Es können maximal zehn taxonomische Stufen gleichzeitig betrachtet werden. Werden mehr als zehn Stufen ausgewählt, so werden die ersten zehn unverändert angezeigt. Alle weiteren Stufen bleiben unberücksichtigt. Damit sich die einzelnen Säulen nicht gegenseitig verdecken, wird die Position auf der x-Achse angepasst. Dies führt dazu, dass alle Säulen einer Probe meist nur gesehen werden können, wenn die Darstellung entsprechend vergrößert wird. Die Reihenfolge der Säulen ist ebenso wie die Reihenfolge der einzelnen Abschnitte der Legenden abhängig von der Auswahlreihenfolge der taxonomischen Stufen.

Durch die Aktivierung des Kästchens ‚Neue Skalierung' wird die Berechnung der Read-Anteile verändert. Der eingestellte Wert des Schiebereglers rechts vom Kontrollkästchen fungiert dabei als Grenzwert. Es sind Grenzwerte von mindestens 0.1 Prozent bis maximal 1.0 Prozent möglich. Wählbar sind alle Werte in diesem Intervall, die sich ohne Rest durch 0.1 dividieren lassen. Wird das Kästchen aktiviert, so wird zuerst die ursprüngliche normalisierte Darstellung berechnet. Alle Taxone bzw. Farbbereiche, die einen normalisierten Read-Anteil haben, der kleiner ist als der gewählte Grenzwert, werden in einer Kategorie zusammengefasst. Diese Kategorie bekommt den Namen ‚Anderes'. Der Vorteil dieser Neuskalierung ist, dass durch die Zusammenfassung vieler kleiner Taxone, weniger kleine Farbbereiche in einer Säule dargestellt werden. Dies hat eine verbesserte Übersichtlichkeit zur Folge. In der Visualisierung hat die Kategorie ‚Anderes' immer die Farbe grau, sodass nach der Aktivierung der Neuskalierung 21 Farben zur Darstellung der Barplots verwendet werden. Das grau wird jedoch nicht in die Reihenfolge der anderen zwanzig Farben mitaufgenommen. Somit kann jede Säule maximal einen grauen Bereich haben, da in den zwanzig Farben zur besseren Abgrenzung von der durch die Neuskalierung entstehenden Kategorie kein grau enthalten ist. Auch steht ‚Anderes' als Taxon immer am Ende der Legende, während alle übrigen Taxone weiterhin alphabetisch

aufgelistet werden.

Optionen für Medikationsdaten Für die Anzeige der Medikationsdaten ist die Auswahl von exakt einem Patienten notwendig. Wird zusätzlich das Kontrollkästchen ‚Zeige Medikationsdaten' aktiviert, erscheint unter jeder bereits erstellten Graphik ein weiteres Koordinatensystem. Wie bereits in Abschnitt 2.2.3 beschrieben, werden hier alle in den Daten vermerkten Wirkstoffe in Abhängigkeit der bekannten Zeitachse (x-Achse) dargestellt. Direkt nach dem Erscheinen des neuen Koordinatensystems ist die x-Achse so eingestellt, dass alle existierenden Datenpunkte zu sehen sind. Durch das Verschieben der x-Achse eines Plots - dies kann entweder die Medikationsdatendarstellung oder die Visualisierung der klinischen Daten sein - passt sich die x-Achse des jeweils anderen Plots an die gerade verschobene x-Achse an.

Jeder Wirkstoff ist einer von fünf Kategorien zugeordnet. Die Kategorien werden in der Legende mit ihrer zugehörigen Farbe aufgelistet. Das Feld ‚Zuordnung' stellt alle in den Daten gegebenen Zuordnungen zur Auswahl. Die möglichen Zuordnungen sind: Antiinfektiva, Immunglobuline, Konditionierung, Zytostatikum und Sonstiges. Über das Auswahlfeld kann jede Kombination der Zuordnungskategorien gewählt werden. Sobald mindestens eine Kategorie ausgewählt wird, werden die Medikationsdaten gefiltert. Es werden dann nur noch Wirkstoffe visualisiert, die einer der gewählten Kategorien zugeordnet werden. Es gibt Patienten, die keinen Wirkstoff erhalten haben, der den Immunglobulinen zugeordnet wird. Wird in diesem Fall nur die Kategorie Immunglobuline ausgewählt, bleibt das Medikationskoordinatensystem bis auf die Achsen leer.

Die Darstellung der Daten erfolgt in Form eines Punktdiagramms. Da jeder Punkt die Form eines Quadrats hat, ähnelt die Visualisierung bei genügender Verkleinerung einem Balkendiagramm. Es ist möglich, die Darstellung Medikationsdaten gleichzeitig mit der Darstellung der Mikrobiomdaten anzuzeigen.

Optionen der erstellten Plots Dieser Abschnitt beschäftigt sich mit den Einstellungsmöglichkeiten Verknüpfe x-Achsen (experimentell), Minimum (x-Achse), Maximum (x-Achse) und Y-Achsen Höhe. Prinzipiell ist die Verschiebung innerhalb eines Koordinatensystems durch das Ziehen der Achse in die gewünschte Richtung möglich. Die y-Achse ist jedoch in allen Darstellungen fixiert. Deshalb kann jeweils nur die x-Achse verschoben werden. Auch das rein oder raus zoomen verändert nur die angezeigten Werte auf der x-Achse. Durch einen Doppelklick auf ein Koordinatensystem werden alle Verschiebungen rückgängig gemacht. Werden zu dem Zeitpunkt auch Medikationsdaten angezeigt, so werden beide Darstellungen zurückgesetzt, unabhängig davon, auf welche von beiden der Doppelklick erfolgt. Soll die Höhe der y-Achse verändert werden, so kann dies mithilfe des Schiebereglers erfolgen. Die neue Höhe wird auf alle erstellten Graphiken mit klinischen Parametern angewandt. Auf die Koordinatensysteme mit den Medikationsdaten hat diese Einstellung keine Auswirkungen. Alle folgenden Optionen

beziehen sich immer auf beide Koordinatensysteme, das mit den klinischen Daten und das mit den Medikationsdaten.

Sollen sich nicht nur die x-Achsen einer Medikationsdatengraphik und einer Graphik mit klinischen Daten synchron verhalten, sondern auch die x-Achsen mehrerer Graphiken mit klinischen Daten, so kann das Kästchen ‚Verknüpfe x-Achsen (experimentell)' aktiviert werden. Dies führt die gewünschte Aktion aus, jedoch ist diese sehr langsam und führt zu Verzögerungen in der Verschiebung der einzelnen Achsen. Weiterhin ist die Synchronisierung nicht ganz exakt. Mithilfe der Optionen ‚Minimum (x-Achse)' und ‚Maximum (x-Achse)' können der äußerste Wert links und der äußerste Wert rechts der x-Achse festgelegt werden. Diese Werte gelten dann für alle erstellten Graphiken. Wird nur einer der beiden Werte spezifiziert, so wird der andere automatisch auf null gesetzt. Zu beachten ist an dieser Stelle, dass der Wert des Minimums kleiner als der des Maximums sein sollte, um eine sinnvolle Darstellung zu erhalten. Auch diese Beschränkung der x-Achse ist nicht hundertprozentig genau. Zu empfehlen sind daher Zahlen, die sich ohne Rest durch zehn dividieren lassen, da in diesen Fällen die Festlegung besser übereinstimmt. Um zu der initialen Darstellung ohne festgelegten Minimum- und Maximumwert zurückzukehren, müssen die Einträge beider Textfelder gelöscht werden.

Die Option ‚Verknüpfe x-Achsen' kann nicht gleichzeitig mit der Festlegung des Minimums und Maximums genutzt werden. Sollte versucht werden, beide Optionen gleichzeitig zu verwenden, so wird die Option ‚Verknüpfe x-Achsen' bevorzugt werden.

4.3.3 Tab: Principal Component Analysis

Im Tab, der die Überschrift PCA trägt, kann eine Principal Component Analysis durchgeführt werden. Der folgende Abschnitt erläutert welche Schritte erfolgen müssen, um eine PCA zu erstellen und welche Optionen für die nachfolgende Betrachtung des Ergebnisses zur Verfügung stehen.

PCA erstellen Zur Erstellung einer PCA müssen ein Datum und alle Komponenten, die in die Berechnung miteinbezogen werden sollen, gewählt werden. Das Datum wird über das erste Autocomplete-Feld gewählt. Dementsprechend ist eine Suche innerhalb der zum Feld gehörenden Liste möglich. Zusätzlich zum Datum wird über den darüber liegenden Schieberegler die zu berücksichtigende Zeitspanne festgelegt. Der voreingestellte Wert beträgt fünf. Es sind Werte von null bis zehn wählbar. Mithilfe von Datum und Schiebereglerwert wird eine Zeitspanne festgelegt. Beispielsweise ist die Zeitspanne für das Datum null und den Schiebereglerwert fünf die Zeit von Tag minus fünf bis Tag plus fünf.

Für eine Liste von Analyten, die zu den klinischen Daten gehören und die als quantitative Parameter, die für jeden der 32 Patienten mindestens einen Wert haben, charakterisiert wurden,

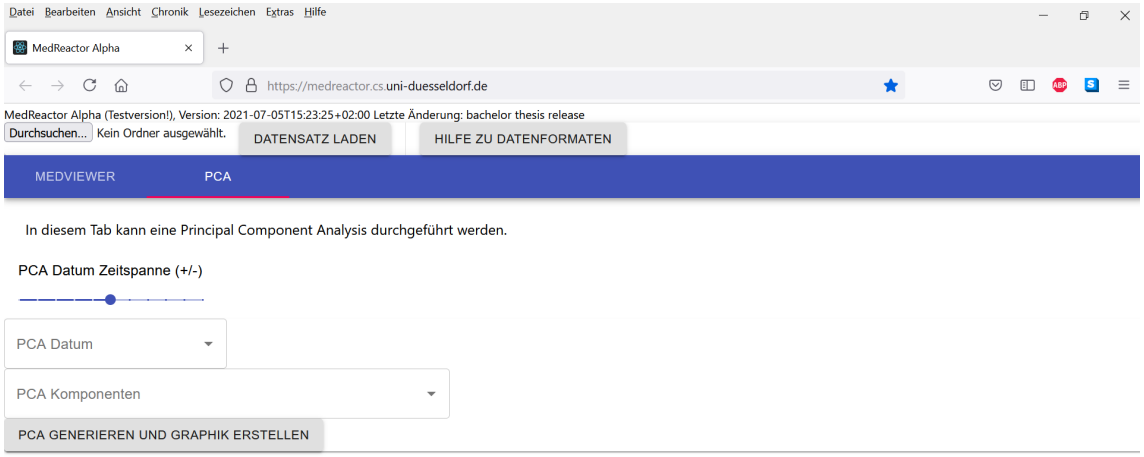

Abbildung 6: Ansicht des PCA-Tabs nach dem Laden der Eingabedaten

wird nun eine Überprüfung durchgeführt. Getestet wird, ob für einen Analyt für jeden Patienten innerhalb der gewählten Zeitspanne mindestens ein Wert vorliegt. Berücksichtigt werden an dieser Stelle nur Patienten für die klinische Daten hochgeladen wurden. Alle Parameter, für die sich diese Überprüfung als korrekt erweist, werden im zweiten Autocomplete-Feld aufgelistet. Dementsprechend verändern sich die wählbaren Parameter dieses Feldes durch die Veränderung der Zeitspanne, die sich aus Datum und dem Wert des Schiebereglers zusammensetzt. Deshalb muss das Datum zuerst gewählt werden, da ansonsten keine Parameter als PCA Komponenten zur Verfügung stehen. Gleichermaßen gilt, dass die PCA Komponenten neu ausgewählt werden müssen, wenn das Datum geändert oder der Schieberegler verschoben wird. Um die PCA Komponenten auszuwählen, können entweder alle gewünschten Analyten einzeln angeklickt werden oder es wird die Option ‚Alle verfügbaren Komponenten auswählen', welche ganz oben in der Liste der möglichen Analyten steht, gewählt. Diese soll das Anklicken einer langen Liste von Analyten ersparen, wenn alle Analyten berücksichtigt werden sollen. Es können keine Parameter mehr als einmal in die PCA einfließen. Werden zusätzlich zu der Option ‚Alle verfügbare Komponenten auswählen' auch noch ein oder mehrere Analyten ausgewählt, so werden die einzeln gewählten Analyten nicht berücksichtigt, da diese bereits in der alles auswählen Option mit inbegriffen sind.

Die Berechnung der PCA erfolgt mithilfe einer bereits existierenden JavaScript Bibliothek.[13] Diese setzt eine Komponentenanzahl von mindestens zwei voraus. Deshalb müssen für die Erstellung einer PCA mindestens zwei der Analyten gewählt werden. Werden in der Liste der möglichen Komponenten nicht mindestens zwei Analyten bereitgestellt, so existiert auch die

[13] *https://github.com/bitanath/pca, zuletzt aufgerufen: 03.07.2021*

Option ‚Alle verfügbaren Komponenten auswählen' nicht und es kann für die gewählte Zeitspanne keine PCA berechnet werden.

Sind alle Einstellungen erfolgt, so kann die PCA über den ‚PCA generieren und Graphik erstellen' - Knopf verwirklicht werden. Es können mehrere PCAs untereinander dargestellt werden. Um eine dieser PCAs wieder zu löschen, muss auf den zugehörigen ‚PCA entfernen' - Knopf geklickt werden. Der jeweils zugehörige Knopf befindet sich über der graphischen Darstellung.

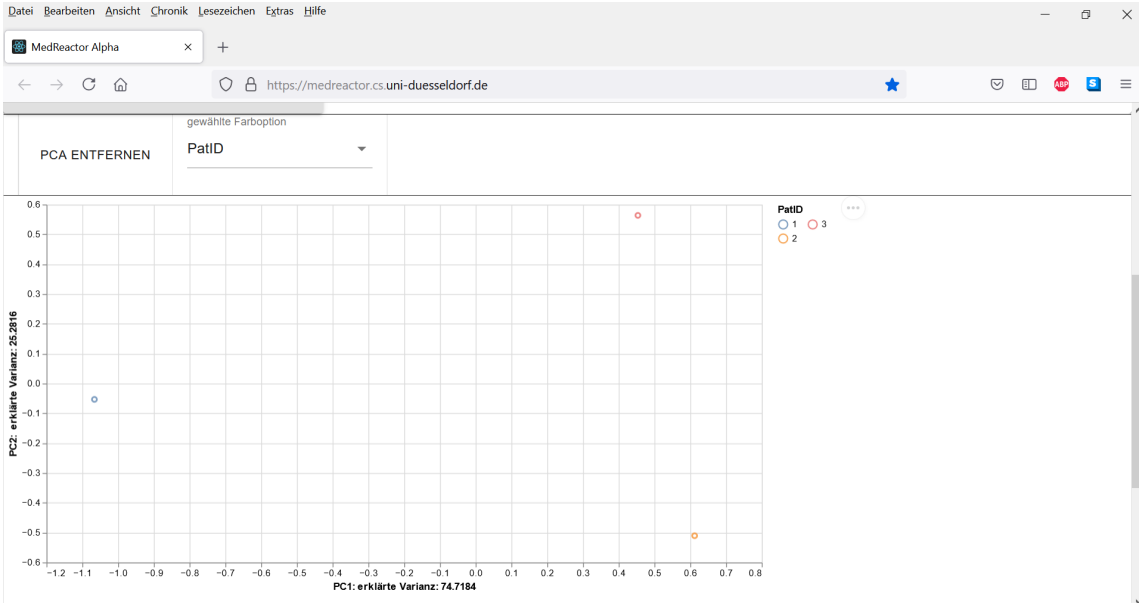

Abbildung 7: Darstellung der Ergebnisse einer PCA, die verwendeten Daten sind frei erfunden

Farboptionen Die PCA wird visualisiert als ein Punktdiagramm mit genauso vielen Punkten, wie Patienten für die Berechnung berücksichtigt wurden. Jeder Punkt hat zusätzlich zu seiner Position in einem zweidimensionalen Koordinatensystem eine Farbe. Direkt nach der Erstellung einer PCA sind die unterschiedlichen Farben den Patienten Identifikationsnummern zugeordnet. Über dem Koordinatensystem befindet sich eine Auswahlmöglichkeit mit dem Namen ‚gewählte Farboption', über die die Zuordnung der Punkte im Koordinatensystem zu einer Farbe verändert werden kann. Dafür werden die Daten der PCA Charakteristika Tabelle verwendet. Wurde diese nicht hochgeladen, so kann die Farboption nicht verändert werden. Es wird zwischen zeitunabhängigen und zeitabhängigen Charakteristika unterschieden. Zeitunabhängige Charakteristika sind zum Beispiel die Grunderkrankung und die Art der Vortherapie eines Patienten. Diese beziehen sich nur auf den zugehörigen Patienten und verändern sich im Zeitverlauf nicht. Wiederum zeitabhängig sind unter anderem folgende Charakteristika: GvHD und Rezidiv. Diese haben einen Starttag. Das Charakteristikum Parenterale Ernährung hat zudem einen Endpunkt. Für alle Tage, die innerhalb dieses Intervalls bzw. nach dem Starttag liegen, gilt das betrachtete Charakteristikum. Da bei der Erstellung der PCA eine Zeitspanne

statt eines einzelnen Tages betrachtet wird - außer der Schieberegler wird auf null eingestellt - könnte es passieren, dass sich die Zeitspanne der PCA mit der Zeitspanne des Charakteristikums nur teilweise überschneidet. Um diesem Problem entgegenzuwirken, wird bei zeitabhängigen Charakteristika nur der Tag, welcher als Datum für die PCA gewählt wurde, berücksichtigt. Die Zeitspanne der PCA ist somit für die Wahl der Farboption nicht relevant. Wird die gewählte Farboption geändert, so werden alle Punkte ihrem Wert der gewählten Tabellenspalte gemäß farblich angepasst. Wird ein zeitabhängiges Charakteristikum gewählt, wird außerdem das Datum berücksichtigt. Auch die Legende wird bei einer Veränderung der Farboption entsprechend aktualisiert.

PC Tabelle & Achsenbeschriftung (Varianz) Das Ergebnis der PCA wird in einem zweidimensionalen Koordinatensystem dargestellt. Die x-Achse symbolisiert die erste Principal Component (PC 1), die y-Achse steht für die zweite Principal Component (PC 2). Jede für die Erstellung der PCA gewählte Komponente hat eine Auswirkung auf die beiden Principal Components. Die genaue Zusammensetzung wurde in Abschnitt 3.3 erläutert. Statt der Darstellung als eine lange Gleichung wird die Komposition der Principal Components in Tabellenform angegeben. Zu jeder graphischen Darstellung einer PCA gehört eine Tabelle mit den folgenden vier Spalten: PCA Komponenten, Bezeichnungen, PC 1 und PC 2.

PCA Komponenten	Bezeichnungen	PC 1	PC 2
Leuko (zentral)	Leukozyten	0.4512	0.3902
HB (zentral)	Hämoglobin	-0.2945	0.0147
...	...	...	...

Tabelle 3: Die hier dargestellten Daten sind frei erfunden und sollen nur das Format der Tabelle veranschaulichen.

In der Webanwendung befinden sich links neben den Überschriften der Spalten (PCA Komponenten, PC 1 und PC 2) kleine Pfeile. Durch das Klicken auf einen der beiden Pfeile neben den Spalten PC 1 und PC 2 wird die Tabelle den Werten der gewählten Spalte entsprechend aufsteigend sortiert. Ein erneutes Klicken auf denselben Pfeil bewirkt die Sortierung in absteigender Reihenfolge. Danach wird wieder die aufsteigende Reihenfolge dargestellt. Sobald der Pfeil einer Spalte geklickt wird, werden alle anderen Pfeile zurückgesetzt, sodass wieder mit der aufsteigenden Reihenfolge begonnen wird. Der Pfeil, der zur Spalte PCA Komponenten gehört, stellt eine Besonderheit dar. Dieser sortiert die Spalten nicht auf- oder absteigend, sondern bringt die Tabelle wieder in den Zustand, den sie nach Erstellung der PCA hatte. PCA Komponenten erscheinen in der Tabelle in der Reihenfolge, in der sie ausgewählt wurden.

Die erklärte Varianz jeder Principal Component (siehe Abschnitt 3.3) wird in der Achsenbeschriftung vermerkt.

5 Diskussion

Zuvor wurde dargestellt, welche Anforderungen die Webanwendung erfüllen soll. In diesem Abschnitt wird diskutiert, inwiefern diese eingehalten wurden und welche Funktionen noch ergänzt werden können.

Die Kombination der verschiedenen Datentypen (klinische Daten, Mikrobiomdaten und Medikationsdaten) funktioniert, inklusive der Interaktivitätsfeatures, den Anforderungen entsprechend. Als eine Einschränkung der Interaktivität kann gesehen werden, dass die klinischen Daten die Basis der Anwendung bilden und deshalb nicht optional sind. Die zugehörige graphische Darstellung kann nicht ausgeblendet werden. Da diese die Darstellung der anderen Datentypen nicht beeinträchtigt, ist dies als minimale Einschränkung zu gewichten.

Eine weitere Anforderung war die Implementierung mehrerer Tabs. Tatsächlich verfügt die Anwendung derzeit über zwei Tabs (Kombination der Datentypen und PCA). Die Prinicipal Component Analysis ist in einem eigenständigen Tab implementiert. In diesem könnte noch die Darstellung der Datenpunkte im zeitlichen Verlauf ergänzt werden. Der dritte diskutierte Tab sollte die Möglichkeit bieten, mehrere Analyten in einem Koordinatensystem darzustellen. Dies ist derzeit nicht möglich. Ein kritischer Punkt an dieser Stelle sind die unterschiedlichen Einheiten der Analyten, sowie die Kombination von nominalen und quantitativen Analyten auf einer y-Achse. Weiterhin können aber verschiedene Analyten in mehreren Koordinatensystemen untereinander dargestellt werden.

Ergänzt werden muss außerdem die Umwandlung der Kurzformen der Analyten in die vollständigen Namen der Analyten. Die letzte genannte Anforderung betrifft die Überprüfung der Eingabedaten. Es wird überprüft, ob alle benötigten Schlüssel vorhanden sind. Die Daten in den Dateien selbst werden aber nicht geprüft.

Zusammenfassend lässt sich sagen, dass die meisten Anforderungen erfüllt wurden, es aber dennoch weitere Verbesserungsmöglichkeiten gibt.

Literatur

[1] Christian Prinz und Folker Schneller. In: *Basiswissen Innere Medizin*. Springer, Berlin, Heidelberg, 2012. Kap. Leukämien. URL: https://doi.org/10.1007/978-3-642-12377-1_16.

[2] Erich Gutenberg. In: A. V. Hoffbrand und J. E. Pettit. *Grundlagen der Hämatologie*. Steinkopff, Heidelberg, 1986. Kap. Leukozyten. URL: https://doi.org/10.1007/978-3-662-11915-0_6.

[3] Robert Koch-Institut und Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V. *Krebs in Deutschland 2015/2016*. 2019. Kap. 3.31 Leukämie. URL: http://dx.doi.org/10.25646/5977.3.

[4] Kawa Nazemi und Lukas Kaupp und Dirk Burkhardt und Nicola Below. *Praxishandbuch Forschungsdatenmanagement*. Markus Putnings und Heike Neuroth und Janna Neumann, De Gruyter Saur, 2021. Kap. 5.4 - Datenvisualisierung. URL: https://doi.org/10.1515/9783110657807-026.

[5] Stuart K. Card und Jock D. Mackinlay und Ben Shneiderman. *Readings in information visualization: using vision to think*. 1999.

[6] *Vega-Lite*. URL: https://vega.github.io/vega-lite/. (zuletzt aufgerufen: 03.07.2021).

[7] Karl Pearson. “On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. Philosophical Magazine Series 6 Volume 2 (Issue 11 1901), S. 559–572. URL: https://doi.org/10.1080/14786440109462720.

[8] Rebecca Fröhlich. *Bachelor Thesis Release*. Juli 2021. DOI: 10.5281/zenodo.5070462. URL: https://doi.org/10.5281/zenodo.5070462.

A Anhang

Abkürzung	Taxonomische Stufe (englisch)	Taxonomische Stufe (deutsch)
C	Class	Klasse
D	Domain	Domäne
F	Family	Familie
G	Genus	Gattung
K	Kingdom	Reich
O	Order	Ordnung
P	Phylum	Stamm
S	Species	Art
R	Root	-
U	Unclassified	-

Tabelle 4: Steht eine Ziffer hinter einem Abkürzungsbuchstaben, so handelt sich um ein Sublevel der taxonomischen Stufe, z.B.: D1, Domain Sublevel 1.

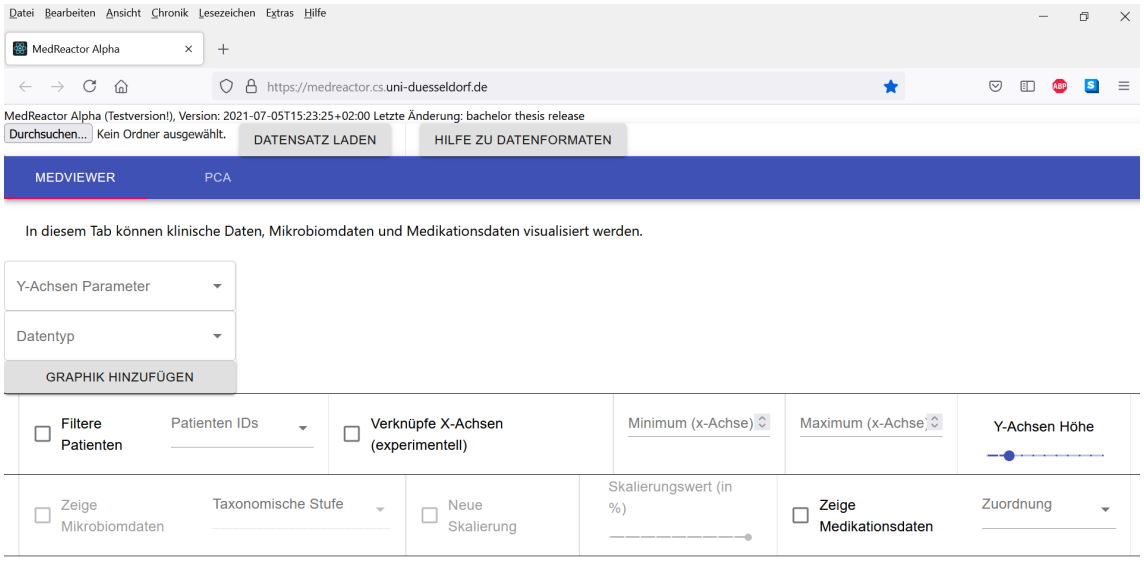

Abbildung 8: Werden keine Mikrobiomdaten hochgeladen, so sind die ersten vier Optionen in der zweiten Zeile nicht verfügbar (in dieser Abbildung zu sehen). Den Medikationsdaten werden die letzten beiden Menüpunkte der zweiten Zeile zugeordnet. Fehlen beide Dateien, so ist die komplette zweite Zeile deaktiviert. Die erste Zeile der Einstellungen kann immer verwendet werden.

Tabellenverzeichnis

Abbildungsverzeichnis