Otto-von-Guericke Universität, Magdeburg
Fakultät für Informatik

Projektbericht

Visual Analytics of Participant Evolution in Longitudinal Cohort Study Data

Autor: Benedikt Mayer

Datum der Abgabe: 20.03.2018

Betreuung: Prof. Dr. habil. Bernhard Preim, Uli Niemann

Institut für Simulation und Grafik

Inhaltsverzeichnis

1 Motivation

Für die Auswertung epidemiologischer Langzeitstudien werden bisher häufig die Ergebnisse der einzelnen Studienzeitpunkte (Momente) separat betrachtet. Der Grund hierfür ist die Fluktuation im Lauf der Studie, sowohl im Bezug auf die Teilnehmer der Studie, welche zu einzelnen Momenten der Studie aus verschiedenen Gründen verhindert sein können, als auch im Bezug auf die erfassten Werte. So ist eine Anpassung bestimmter Messmethoden oder die Ablösung alter Techniken durch neue insbesondere bei epidemiologischen Studien üblich.
Dies erschwert die Untersuchung der Entwicklung der gesamten Kohorte (Entire Population, EP) sowie einzelner Teilgruppen (Subpopulations, SPs). Dennoch können die in den Veränderungen über die Zeit enthaltenen Informationen für Epidemiologen von großem Nutzen sein um einzelne SPs besser zu verstehen und um Risikopatienten, in diesem Fall bezüglich Steatosis hepatis (Fettleber), frühzeitig identifizieren zu können. Ansätze, um mit den beschriebenen Herausforderungen umzugehen, welche die Untersuchung longitudinaler epidemiologischer Daten mit sich bringt, wurden in der nachfolgend beschriebenen Anwendung „TemporalExplorer“ umgesetzt.
Der, beziehungsweise die Nutzerin des Tools wird dabei geschlechtsneutral als „der Nutzer“ oder „der Anwender“ bezeichnet. Der Datensatz, auf den sich diese Anwendung bezieht, stammt aus der „Study of Health in Pomerania“ (SHIP) [1]. Er besteht aus drei Momenten, SHIP-0, SHIP-1 und SHIP-2. Diese werden an einigen Stellen abkürzend als s0, s1 und s2 bezeichnet.

Die grundlegende Idee der vorgestellten Anwendung ist es, SPs in der Kohorte zu finden, deren Mitglieder einen ähnlich hohen Leberfettwert besitzen, die sich jedoch durch die Ausprägung bezüglich anderer Variablen als mrt_liverfat beschreiben lassen. Hierfür wird der Leberfettwert diskretisiert, sodass drei Risikoklassen entstehen. Die erste Klasse beinhaltet alle Probanden mit einem Leberfettwert zwischen 0 % und 10 %, die zweite alle Teilnehmer mit einem Leberfettwert zwischen 10 % und 25 % und die dritte alle mit einem Wert über 25 %. Zur Beschreibung der SPs werden Klassifikationsregeln verwendet, welche mit Hilfe des Weka HotSpot Algorithmus bezüglich der Variablen aus SHIP-2 gebildet werden [2]. Dieses Vorgehen ist an der Technik von Niemann et al. orientiert, bei denen jedoch ein Fokus auf der separaten Untersuchung der einzelnen Momente liegt [3].
In dieser Arbeit werden die gefundenen SPs hingegen mit einem Fokus auf

die Entwicklung über die Zeit untersucht.
Da durch jede gefundene Regel eine Subpopulation (SP) beschrieben wird, bezeichnen wir im Folgenden die durch eine Regel implizierte SP als die *zu der Regel gehörende SP*.

2 Verwandte Arbeiten

Ansätze für Untersuchungen der SHIP Daten, unterstützt durch Visual Analytics Techniken, wurden 2016 von Klemm vorgestellt. In [4] schlägt er Vorgehensweisen sowohl für die explorative als auch die Bestätigungsanalyse der erfassten Daten vor.
Gemeinsam mit Kollegen hat Klemm ebenfalls ein Konzept für eine vollständige Regressionsanalyse mit Hilfe einer 3D Heat Map erarbeitet [5].
Weiterhin wurde von Amendez et al. ein Ansatz, um mit Hilfe von Subspace Clustering nach plausiblen SPs zu suchen, in einem interaktiven Framework umgesetzt [6].
In der hier vorgestellten Arbeit werden die SPs basierend auf einem einzigen Zeitpunkt generiert. In der Arbeit von Krause et al. hingegen können Kohorten basierend auf Informationen erzeugt werden, die zu verschiedenen Zeitpunkten erfasst wurden. Sie schlagen ein interaktives Framework zur iterativen Erzeugung von Kohorten durch zeitabhängige Einschränkungen vor [7]. Durch eine visuelle Umsetzung wird auch Anwendern ohne Erfahrung im Bezug auf Datenbankabfragen eine Nutzung der Anwendung ermöglicht.
Wie bereits erwähnt dient als Grundlage dieser Arbeit der Ansatz für den InteractiveRuleMiner von Niemann et al. In ihrer Arbeit wurde ein Workflow für die Vorverarbeitung der Daten, das Data-Mining und die Untersuchung der Ergebnisse vorgeschlagen [3]. So wird auch eine Methode vorgestellt, um herauszufinden, wie die Mitglieder gefundener SPs auf die unterschiedlichen Risikoklassen verteilt sind. Das ist eine Betrachtung, welche in der hier vorgestellten Arbeit nur in geringem Umfang unterstützt wird.
Ein weiteres Problem, welches in dieser Arbeit mit grundlegenden Methoden gelöst wurde, sind fehlende Werte einzelner Probanden für bestimmte Variablen. Diese werden entweder direkt dargestellt, oder, wie zum Beispiel bei der Berechnung von Mittelwerten, ignoriert. In der Arbeit von Alemzadeh et al. [8] und Spratt et al. [9] wird sich dieser Frage in größerem Detail gewidmet und es werden fortgeschrittene Methoden zur Schätzung unbekannter Messwerte verwendet, mit einem Fokus auf multipler Imputation. Diese Technik

ermöglicht die Schätzung der fehlenden Werte basierend auf den Abhängigkeiten zwischen mehreren Variablen. Eine abgewandelte Technik zur Imputation wird auch in Abschnitt 5.5 verwendet.
Techniken für die Visualisierung von Assoziationsregeln wurden unter anderem von Hahsler et al. erarbeitet, von denen sich einige auch auf Klassifikationsregeln übertragen lassen. So wird in dieser Arbeit auch eine der in [10] vorgestellten Methode in Abschnitt 5.1 verwendet.

3 Grundlegende Begriffe

Bei der Bildung der HotSpot Regeln werden einige Bezeichnungen verwendet, die nachfolgend unter der Verwendung der Notation von Fürnkranz [11] erklärt werden. Dabei gehört jedes Mitglied der EP genau einer der drei oben genannten Risikoklassen an.
Fixiere unter dieser Voraussetzung eine Risikoklasse und eine Regel, deren Konsequenz die Zugehörigkeit zu der betrachteten Risikoklasse darstellt.
Es bezeichne P die Anzahl aller Probanden, die zu der Risikoklasse gehören und N die Anzahl aller Probanden, die nicht zu der Risikoklasse gehören. Außerdem sei $\hat{P}$ die Anzahl aller Probanden, die zu der Risikoklasse gehören und vom Antezedens der Regel abgedeckt werden (True Positives). Als Gegenstück dazu sei $\hat{N}$ die Anzahl aller Probanden, die nicht zu der Risikoklasse gehören, aber dennoch vom Antezedens der Regel abgedeckt werden (False Positives).
Aus diesen Definitionen ergibt sich, dass $P + N$ die Anzahl aller Probanden ist und $\hat{P} + \hat{N}$ die Anzahl aller Probanden, die vom Antezedens der Regel abgedeckt werden.
Mit diesen Bezeichnungen lassen sich die folgenden Qualitätsmaße definieren: $\mathit{Support} = \frac{\hat{P}}{P+N}$, $\mathit{Coverage} = \frac{\hat{P}+\hat{N}}{P+N}$, $\mathit{Confidence} = \frac{\hat{P}}{\hat{P}+\hat{N}}$, $\mathit{Lift} = \frac{\frac{\hat{P}}{\hat{P}+\hat{N}}}{\frac{P}{P+N}}$ und $\mathit{Odds\ Ratio} = \frac{\frac{\hat{P}}{P-\hat{P}}}{\frac{\hat{N}}{N-\hat{N}}}$. Dabei ist die Confidence ein häufig verwendetes Maß, um die Qualität einer Regel zu beurteilen. Sie wird auch als primäres Qualitätsmaß im HotSpot Algorithmus bei der Suche nach relevanten Regeln verwendet. Die Coverage gibt Auskunft über die Größe der SP und sowohl Support als auch Lift sind gängige Maße im Bereich von Assoziations- bzw. Klassifikationsregeln. Bei der Odds Ratio handelt es sich um ein oft im Bereich der Epidemiologie verwendetes Maß.

4 Kurzbeschreibung der Anwendung

Die Anwendung besteht aus verschiedenen Komponenten, die gemeinsam die Untersuchung von SPs ermöglichen, welche mit Hilfe des HotSpot-Algorithmus generiert wurden. Das allgemeine Konzept für die Vorgehensweise bei der Nutzung des Tools ist nachfolgend beschrieben.
Zunächst legt der Nutzer einige für die Bildung der Regeln relevante Parameter fest. Anschließend erhält er eine Übersicht über die generierten Regeln in einem Support-Confidence Plot. Einen detaillierten Überblick über die dargestellten Regeln findet der Nutzer außerdem unterhalb des generierten Plots in Form einer Tabelle, der Rule Overview Table. Die Einträge der Tabelle passen sich an die Auswahl bestimmter Regeln im Support-Confidence Plot an, sodass die ausgewählte Teilmenge separat untersucht werden kann. Dies dient insbesondere dazu, bei sich überlagernden Punkten im Plot eine exakte Selektion zu ermöglichen, sowie einen Nutzer, der die Verwendung reiner Tabellen bevorzugt, in seiner Arbeitsweise zu unterstützen.
Die nächste Komponente dient dazu, nachdem eine Regel ausgewählt wurde, mit geringem Aufwand abschätzen zu können, ob die gewählte Regel aus Expertensicht potentiell sinnvoll ist. Hierfür werden die SP-EP Barcharts verwendet, welche die Verteilung der SP und der EP bezüglich der in dem Antezedens der Regel verwendeten Variablen gegenüberstellen und eine äquivalente Behandlung von numerischen und kategorialen Variablen ermöglichen. Zeigt sich an diesem Punkt, dass die gewählte Regel zu abwegig ist, so kann der Nutzer eine andere auswählen oder sogar die Parameter für die Regelbildung anpassen um eine neue Menge von Regeln zu generieren.
Außerdem kann eine tabellarische Übersicht verwendet werden, um Auskunft über die Prädiktabilität der Zugehörigkeit von Probanden zu der gewählten SP zu erhalten. Dies geschieht in der Komponente Prediction Rule Table.
Nachfolgend kann die gefundene SP im näheren Detail untersucht werden. So werden durch den Nutzer zunächst bis zu acht numerische und acht kategoriale Variablen bestimmt, deren Entwicklung über die einzelnen Momente in zwei Plots veranschaulicht wird. Der Numerical Development Plot dient zur Veranschaulichung der numerischen, und der Categorical Development Plot zur Veranschaulichung der kategorialen Variablen. Für die numerischen Werte wurde hier außerdem eine Technik angewandt, welche die Werte von Variablen, die in bestimmten Momenten nicht erfasst wurden, mit Hilfe sogenannter Substitutionsvariablen approximiert.
Wird einer der Punkte in dem numerischen oder dem kategorialen Plot se-

lektiert, so erhält der Nutzer unter dem entsprechenden Plot genauere Informationen über die Entwicklung der SP bzw. der EP bezüglich der entsprechenden Variable. Im Fall einer numerischen Variable werden dabei Boxplots verwendet und im Fall einer kategorialen Variable stacked Barcharts.
Eine genaue Beschreibung der genannten Komponenten erfolgt im nächsten Abschnitt.
Sollten dem Anwender während der Nutzung des Tools bestimmte Begriffe oder Methoden unklar sein, so kann er eine entsprechende Erklärung im dafür vorgesehenen Reiter „Appendix“ nachlesen.

5 Bestandteile der Anwendung

Die einzelnen Komponenten des Tools greifen in ihrer Funktionalität eng ineinander. Um die resultierenden Zusammenhänge darzulegen, wird in diesem Abschnitt auf verschiedene Aspekte eingegangen. Dazu zählt, was die Voraussetzung für die Nutzung der einzelnen Komponenten ist, wie sie zu bedienen sind, welche Informationen sie darstellen und wie diese zu interpretieren sind. Außerdem wird erläutert, welche Usability Aspekte in den Aufbau der Anwendung eingeflossen sind, wie die Ergebnisse der einzelnen Module im Workflow weiterverwendet werden können, was bemerkenswerte Erkenntnisse sind und wo die Grenzen der jeweiligen Komponente liegen.

5.1 Support-Confidence Plot

In diesem Modul wird ein Support-Confidence Plot erstellt, welcher als Punkte alle in SHIP-2 gefundenen Regeln enthält. Der Nutzer besitzt die Möglichkeit, zunächst die folgenden Parameter festzulegen (Standardwerte in Klammern):
Maximum Rule Length (2): Wie viele Variablen maximal im Antezedens der Regel enthalten sein dürfen. Eine Verringerung dieses Parameters führt zu einer Verringerung der Berechnungszeit.
Risk Class (liverfat $\leq$ 10 %): Für welche Risikoklasse Regeln gesucht werden sollen. Standardmäßig ist die erste Klasse gewählt, da hierfür die Regelsuche den geringsten Aufwand besitzt, was ein schnelles Starten der Anwendung ermöglicht. Beim Start der Anwendung wird nämlich bereits ein Support-Confidence Plot für die Standardwerte der Parameter berechnet.
Minimum Support (0.3): Wie hoch der Support einer Regel mindestens sein

muss, damit die Regel in den Ergebnissen aufgeführt wird. Eine Verringerung des Wertes bewirkt eine Erhöhung der Berechnungszeit.
Minimal Confidence Gain (0.15): Um welchen Betrag die Confidence einer Regel bei Hinzunahme einer zusätzlichen Variable in das Antezedens mindestens steigen muss. Eine Verringerung des Wertes führt zu einer Erhöhung der Berechnungszeit.
Branching Factor (10): Wie viele Kindregeln bei der Suche nach Regeln generiert werden. Eine Erhöhung des Wertes bewirkt eine Erhöhung der Berechnungszeit.
Checkbox, ob Prediction Rules für die Größe der Punkte im Plot verwendet werden sollen (false): Prediction Rules (nähere Erklärung in Abschnitt 5.4) geben Auskunft über die Vorhersagbarkeit der Zugehörigkeit von Probanden zu der jeweils zur Regel gehörenden SP. Je besser die Vorhersagbarkeit ist, desto größer wird im Plot der Punkt dargestellt, der die entsprechende Regel repräsentiert. Die Auswahl dieser Option erhöht die Berechnungszeit drastisch.
Hat der Nutzer die Parameter nach seinen Wünschen angepasst, so wird durch Betätigen eines entsprechenden Buttons ein Plot der Regeln erstellt (vgl. Abbildung 1). Der Button wurde eingebaut, damit nicht bei jeder Parameteranpassung direkt die mitunter sehr zeitaufwändige Berechnung eines neuen Plots gestartet wird.
Im Plot selbst steht jeder Punkt für eine gefundene Regel und die beiden Achsen repräsentieren die Qualitätswerte Support und Confidence. Die Farbe gibt Auskunft über den Lift der Regeln und, falls „Prediction Rules“ ausgewählt wurden, beschreibt die Größe eines Punktes die Güte der Vorhersagbarkeit der zu der Regel gehörenden SP. Beim Hovern mit dem Mauszeiger über einem Punkt wird das Antezedens der entsprechenden Regel sowie ggf. ein Qualitätswert für die Vorhersagbarkeit angezeigt. Die Verwendung von Support und Confidence für die beiden Achsen und Lift für den Farbton der Punkte wurde von Hahsler et al. vorgeschlagen [10].
Mögliches Overplotting, zu dem es bei bestimmten Parameterkombinationen kommen kann (zum Beispiel bei niedrigem Minimum Support und niedrigem Minimal Confidence Gain), wurde durch semitransparente Darstellung der Punkte begegnet. Dieser Technik sind in stark besiedelten Regionen jedoch Grenzen gesetzt, da durch die Überlagerung zu vieler Punkte Opazität in der Darstellung erreicht wird.
Daher wird zusätzlich eine weitere Technik verwendet, um mit Overplotting umzugehen. So erfolgt die Auswahl bestimmter Punkte standardmäßig

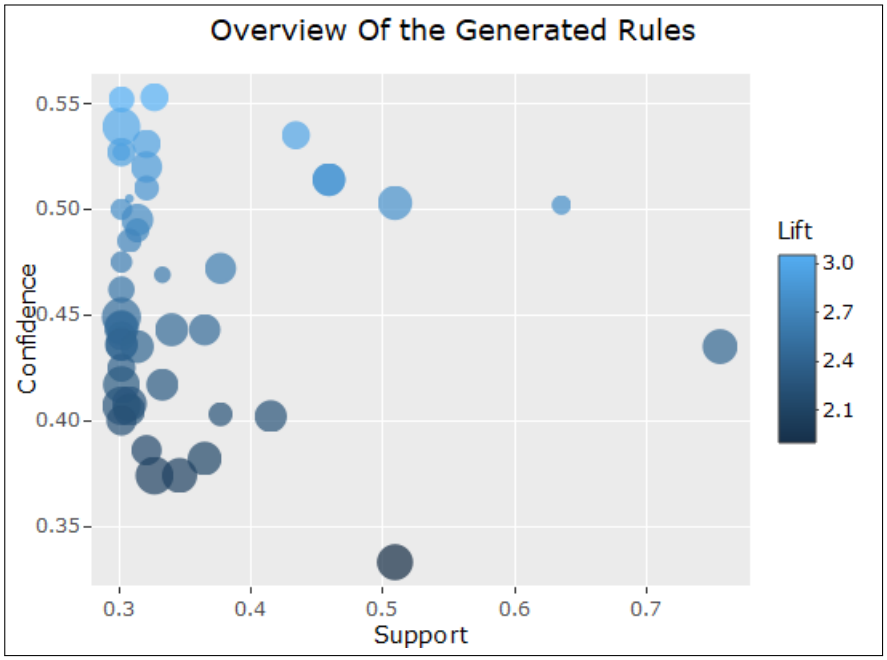


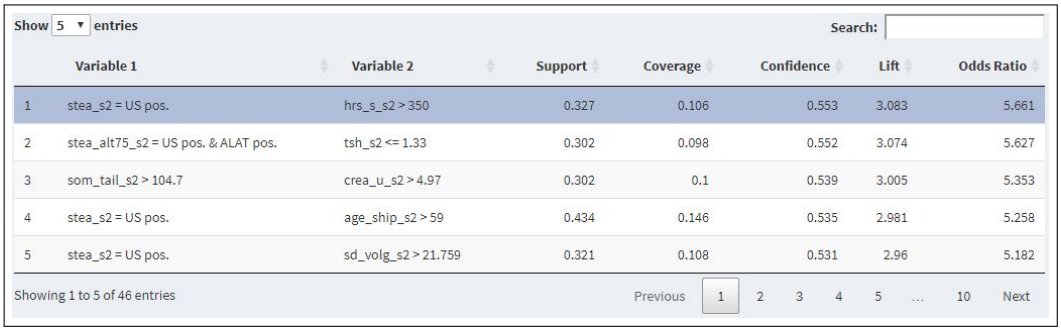
Abbildung 1: Support-Confidence Plot mit variablem Kreisdurchmesser. Einige gute Regeln im rechten und oberen Bereich. Die auf den Prediction Rules basierende Größe der Kreise ist scheinbar unabhängig von Support und Confidence.

durch Box Selection in Kombination mit einer Tabellenübersicht über die ausgewählten Punkte (nähere Erklärung in Abschnitt 5.2), in welcher der gewünschte Punkt selektiert wird. Auch wenn die Auswahl durch Box Selection bei der Selektion einzelner Punkte umständlicher ist als ein einfacher Klick, wurde dennoch diese Methode gewählt. Der Grund hierfür ist, dass sich gelegentlich mehrere Punkte für den Nutzer schwer erkennbar exakt überlagern und ein Klick nur den an oberster Stelle geplotteten Punkt selektieren würde. Mit Hilfe der Box Selection in Kombination mit der tabellarischen Übersicht lässt sich also unter der Berücksichtigung aller Regeln eine bestimmte davon eindeutig auswählen. Eine auf diese Weise gewählte Regel ist die Grundlage für die nachfolgenden Untersuchungen.
Für den Nutzer potentiell interessante Regeln sind im Allgemeinen diejenigen, welche einen hohen Support sowie eine hohe Confidence aufweisen, das heißt durch Kreise im oberen und im rechten Teil des Plots dargestellt werden. Diese können alle auf einmal mit Hilfe der Lasso Selection ausgewählt

werden, erreichbar über die Menüleiste oberhalb des Plots. Zudem besitzen gute Regeln einen hohen Lift, das bedeutet einen hellen Farbton (Farbskala jeweils in der Legende, vgl. Abbildung 1), sowie, falls Prediction Rules aktiviert sind, einen großen Kreisdurchmesser.

5.2 Rule Overview Table

Um Regeln in der Tabelle anzeigen zu können, muss zunächst ein Support-Confidence Plot erstellt werden. Wird in diesem Plot mit Hilfe der Box Selection eine Teilmenge der Regeln ausgewählt, so werden die zu der Teilmenge gehörenden Regeln in der Tabelle aufgelistet. Wurde nicht explizit eine Teilmenge bestimmt, so werden alle gefundenen Regeln aufgeführt (vgl. Abbildung 2).
Die Tabelle gibt Auskunft über das Antezendens sowie Support, Coverage,


Show 5 entries Search:

	Variable 1	Variable 2	Support	Coverage	Confidence	Lift	Odds Ratio
1	stea_s2 = US pos.	hrs_s_s2 > 350	0.327	0.106	0.553	3.083	5.661
2	stea_alt75_s2 = US pos. & ALAT pos.	tsh_s2 <= 1.33	0.302	0.098	0.552	3.074	5.627
3	som_tail_s2 > 104.7	crea_u_s2 > 4.97	0.302	0.1	0.539	3.005	5.353
4	stea_s2 = US pos.	age_ship_s2 > 59	0.434	0.146	0.535	2.981	5.258
5	stea_s2 = US pos.	sd_volg_s2 > 21.759	0.321	0.108	0.531	2.96	5.182

Showing 1 to 5 of 46 entries Previous 1 2 3 4 5 … 10 Next

Abbildung 2: Tabellarische Übersicht über in 5.1 ausgewählte Regeln. Die selektierte Regel (blau markiert) mit sehr guten Qualitätswerten.

Confidence, Lift und Odds Ratio einer Regel. Die initiale Sortierung erfolgt absteigend nach der Confidence, es kann jedoch ebenfalls nach beliebigen anderen Spalten sortiert werden. Um nach Regeln zu filtern, welche eine bestimmte Variable in ihrem Antezedens enthalten, befindet sich ein Suchfeld in der rechten oberen Ecke. Da die Suche Regular Expressions erkennt, können auch Begriffe aus den Ergebnissen ausgeschlossen werden. Möchte man zum Beispiel nur Regeln in der Tabelle anzeigen, welche nicht den Substring „stea“ in ihrem Antezedens führen, so kann man dies durch die Eingabe von *^((?!stea).)*$* im Suchfeld erreichen.
In der Tabelle kann maximal eine Regel ausgewählt werden. Standardmäßig

ist dies die oberste Regel in der Liste, was insbesondere dann nützlich ist, wenn nur eine Regel im Support-Confidence Plot ausgewählt wurde, da in diesem Fall keine zusätzliche Selektion in der Tabelle erfolgen muss.
Häufig im Antezedens enthaltene Variablen sind stea_s2 sowie stea_alt75_s2, was auf Grund von deren Bedeutung naheliegend ist. Zudem sind aber auch häufig Variablen wie som_tail_s2, som_bmi_s2 oder som_gew_s2 zu finden, die auf einen allgemein hohen Fettgehalt im Körper hinweisen. Dies deckt sich mit Erkenntnissen von Bedogni et al. [12]. Außerdem scheinen Blut- beziehungsweise Laborwerte wie tg_s_s2, hrs_s_s2, ggt_s_s2 und crea_u_s2 eine hohe Aussagekraft zu besitzen.
Eine mit Hilfe der Tabelle eindeutig bestimmte Regel repräsentiert wie eingangs erwähnt eine SP, die als Grundlage für die weiteren Untersuchungen verwendet wird.

5.3 SP-EP Barcharts

Zunächst soll jedoch abgeschätzt werden können, ob es sich bei der gewählten Regel um eine plausible Wahl handelt. Hierfür werden nach der Selektion einer Regel in 5.1 bzw. 5.2 Histogramme verwendet, um die Verteilung der zugehörigen SP sowie der EP gegenüberzustellen. Die Histogramme werden pro im Antezedens vorkommender Variable gemeinsam in einer stacked Barchart visualisiert (vgl. Abbildung 3). Barcharts werden von Zhang et al. unter traditionellen Techniken für die Visualisierung von Kohorten geführt [13], weshalb sie für den Nutzer bereits bekannt und damit leicht interpretierbar sein sollten. Außerdem besitzen die Histogramme maximal bis zu 15 Bins und es wird für numerische und kategoriale Variablen dieselbe Darstellung verwendet. Dies unterstützt den Nutzer zusätzlich dabei, sich mit geringem mentalen Aufwand ein Bild von der Verteilung der gewählten SP im Vergleich zu der Verteilung der EP im Bezug auf die Antezedens-Variablen zu machen.
Die Komponente ist somit nicht für eine detaillierte Analyse der selektierten SP konzipiert, sondern für einen ersten Eindruck über ihre Beziehung zu der EP. Im gesamten Workflow ist es daher möglich, dass der Nutzer zunächst einige Male zwischen der Auswahl einer Regel und dem Betrachten der zugehörigen SP-EP Barcharts wechselt, bis er eine SP gefunden hat, welche er tiefergehend analysieren möchte.

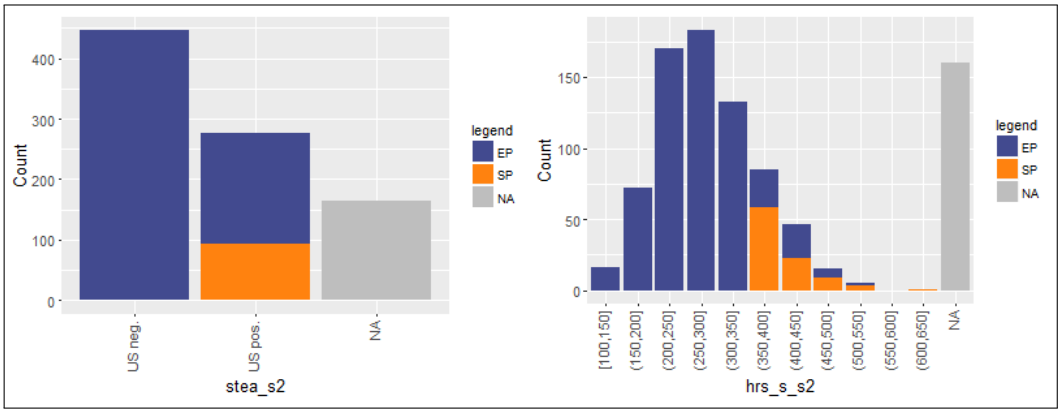


Abbildung 3: Stacked Barcharts des Antezedens einer in 5.2 ausgewählten Regel. Links zu sehen: Die Mitglieder der SP besitzen „US-pos“ als Ausprägung von stea_s2, was im Bezug auf alle Probanden die seltener vorkommende Ausprägung ist. Rechts zu sehen: Betrachtet man die Variable hrs_s_s2, so sind die meisten Probanden mit einem Wert von über 350 ein Mitglied der SP. Bei beiden Variablen konnten für eine ähnlich hohe Anzahl an Teilnehmern keine Werte ermittelt werden.

5.4 Prediction Rule Table

Für diese Komponente muss wie für die SP-EP Barcharts zunächst eine Regel mit Hilfe von 5.1 bzw. 5.2 ausgewählt werden.
Für die zu der Regel gehörende SP erlaubt die Tabelle allgemein gesprochen die Vorhersage der Zugehörigkeit eines Probanden zu der entsprechenden SP, ausgehend von den Informationen aus den vorherigen Momenten.
Im Speziellen bedeutet dies: Für die Bildung der Tabelle wird erneut ein HotSpot Algorithmus ausgeführt. Anstatt wie zu Beginn wird hier jedoch nicht der Leberfettwert als Zielvariable verwendet, sondern eine neue Variable, welche angibt, ob ein Proband zu der gewählten SP gehört oder nicht. Für die Bildung der Regeln werden dabei die Variablen verwendet, welche aus den Momenten SHIP-0 und SHIP-1 stammen. (Erinnerung: Die SPs werden basierend auf Variablen aus SHIP-2 bestimmt.) Die auf diese Weise gefundenen Regeln werden als *Prediction Rules* bezeichnet. Die Parameter für die Suche nach den Prediction Rules sind in diesem Fall nicht vom Nutzer anpassbar sondern auf bestimmte Werte festgelegt. Diese Entscheidung wurde getroffen, um die Komplexität der Anwendung in angemessenem Umfang zu halten. Sollte sich zeigen, dass der Nutzer jedoch an der Anpassung der Parameter

Interesse besitzt, so kann diese Funktionalität in zukünftigen Anwendungen eingebaut werden. Die aktuell festgelegten Werte sind eine Maximum Rule Length von 1, ein Minimal Support von 0.3, ein Minimal Confidence Gain von 0.15 und ein Branching Factor von 5.
Diese Werte werden auch zur Bestimmung der Kreisdurchmesser im Support-Confidence Plot genutzt, wenn dort die entsprechende Checkbox für die Verwendung von Prediction Rules ausgewählt wird. In diesem Fall wird für den Kreisdurchmesser die maximale Confidence aller gefundenen Prediction Rules verwendet. Dabei wird jedoch eine Mindestgröße garantiert, falls die maximale Confidence zu gering ist oder keine Prediction Rules gefunden wurden, um eine ausreichende Erkennbarkeit aller Regelpunkte zu gewährleisten. Die Mindestgröße entspricht einer maximalen Confidence von 0.1.
Die Übersichtstabelle der Prediction Rules ist von der selben Form wie die Rule Overview Table. Sie besitzt die gleichen Spalten sowie die gleiche Such- und Sortierungsfunktionalität (vgl. Abbildung 4).
Ist für den Nutzer wichtig, dass die Zugehörigkeit zu einer SP gut vor-

Search:

	Variable 1	Support	Coverage	Confidence	Lift	Odds Ratio
1	som_tail_s0 <= 71	0.371	0.143	0.795	2.59	8.769
2	som_tail_s1 <= 78.8	0.518	0.207	0.77	2.51	7.578
3	som_huef_s1 <= 95.8	0.316	0.138	0.705	2.296	5.393
4	hrs_s_s1 <= 200	0.324	0.144	0.688	2.239	4.966
5	som_huef_s0 <= 92.7	0.327	0.147	0.685	2.23	4.9

Showing 1 to 5 of 5 entries

Abbildung 4: Tabelle der Prediction Rules einer in 5.2 ausgewählten Regel mit guten Qualitätswerten.

hersagbar ist, so bietet sich mit dieser Tabelle die Möglichkeit, bereits vor der genaueren Analyse potentiell relevante von nicht relevanten Regeln zu trennen. Wurde eine SP gefunden, welche plausibel scheint, ist die Tabelle ebenfalls interessant. Lassen sich hier zu einer plausiblen SP noch gute Regeln finden um zu bestimmen, ob ein Proband zukünftig zu der SP gehören wird, so handelt es sich für den Analysten um eine sehr interessante SP.
In dieser Anwendung wurde keine gesonderte Visualisierung für die Ergebnisse verwendet, da zunächst geprüft werden soll, ob der Ansatz, für Regeln erneut nach Regeln zu suchen, für Nutzer generell interessant ist. Wenn sich

herausstellt, dass dies der Fall ist, so lässt sich in zukünftigen Anwendungen hierauf ein größerer Fokus legen.
Da der Branching Factor bei der Generierung der Prediction Rules auf 5 beschränkt wurde, ist die Anzahl der gefundenen Regeln gering. Dabei ist auffällig, dass zum Teil Confidence Werte erreicht werden, die zunächst erstaunlich hoch wirken. Dies hängt jedoch von der vorherzusagenden SP ab. Gehört sie zu einer Regel, die durch prägnante Werte wie zum Beispiel som_bmi_s2 geformt wird, für die meist zuverlässige Messungen aus den vorherigen Momenten vorliegen, so lassen sich eher aussagekräftige Prediction Rules finden.
Im Bezug auf den Support-Confidence Plot ist anzumerken, dass es für gut vorhersagbare Regeln keine räumlich vorherrschende Lage zu geben scheint. So können sowohl in Bereichen mit hohem Support und hoher Confidence gut vorhersagbare Regeln vorgefunden werden, als auch in Bereichen mit eher niedrigen Qualitätswerten.

5.5 Numerical Development Plot

Für die Berechnung des Plots in diesem Abschnitt muss zunächst eine SP in 5.1 bzw. 5.2 und mindestens eine numerische Variable aus einer Variablenübersicht ausgewählt werden. Die in der Übersicht aufgeführten Variablennamen sind *momentunabhängig*, das bedeutet, sie sind keinem bestimmten Moment zugeordnet (z.B. som_bmi). Zudem sind sie nach Kategorien sortiert, wie zum Beispiel Anthropometric, Blood, etc. Die standardmäßig ausgewählten Variablen sind som_bmi und mrt_liverfat. Außerdem ist die Anzahl maximal auswählbarer Variablen auf 8 beschränkt, um eine Überladung des Plots zu vermeiden. Unterstützt wird die Auswahl zudem durch eine Liste aller Variablen und zugehörigen Werte, welche die durchschnittliche Stärke der Abweichung der SP von der EP über alle Momente hinweg beschreiben. Die Berechnung der Stärke der Abweichung, $\overline{\mu_{SP}(v_i)}$, wird nachfolgend beschrieben.
Wurde auf diese Art mindestens eine numerische Variable ausgewählt, so wird deren Entwicklung über die Momente hinweg in einem Line Plot visualisiert. Die angetragenen Punkte stellen hierbei für eine Variable v (z.B. som_bmi) und einen Moment s_i, $i \in \{0, 1, 2\}$, (z.B. s_2) jeweils den folgenden Wert dar:

$$\overline{\mu_{SP}(v_i)} \coloneqq \frac{\mu_{SP}(v_i) - \mu_{EP}(v_i)}{\sigma_{EP}(v_i)} \tag{1}$$

wobei $\mu_{EP}(v_i)$ bzw. $\mu_{SP}(v_i)$ den Mittelwert der EP bzw. SP bezüglich der Variable v im Moment s_i bezeichnen und $\sigma_{EP}(v_i)$ die Standardabweichung der EP. $\overline{\mu_{SP}(v_i)}$ ist somit gewissermaßen eine normierte Version von $\mu_{SP}(v_i)$. Dadurch soll eine Vergleichbarkeit der Entwicklungen verschiedener Variablen ermöglicht werden.
Die Normierung ist wie folgt zu verstehen: Die Lage eines Punktes im Plot gibt Auskunft über die Differenz des Mittelwerts der SP und der EP gemessen in Standardabweichungen der EP bezüglich der entsprechenden Variable. Dies ermöglicht eine Einschätzung, wie sehr die Verteilung einer SP von der Verteilung der EP abweicht. Liegt zum Beispiel ein Wert $\overline{\mu_{SP}(v_i)}$ über 0, so bedeutet dies, dass die SP bezüglich Variable v im Moment s_i im Durchschnitt höhere Werte besitzt als die EP.
Anstatt $\overline{\mu_{SP}(v_i)}$ hätte auch ein anderes Maß zur Messung der Effektstärke verwendet werden können. Außerdem wäre eine Normierung auf Werte zwischen 0 und 1 möglich gewesen, indem man zum Beispiel durch die maximale auftretende Abweichung unter aller Probanden teilt. Das Problem hierbei wäre jedoch gewesen, dass starke Ausreißer den normierten Wert extrem beeinflusst und somit die Vergleichbarkeit erschwert hätten. Deshalb wurde die beschriebene Normalisierung mit Hilfe der Standardabweichung gewählt.
Im Plot werden die Werte in den einzelnen Momenten entweder durch einen kreisförmigen oder einen quadratischen Marker repräsentiert. Ein kreisförmiger Marker bedeutet, dass in dem entsprechenden Moment Messungen der zugehörigen Variable durchgeführt wurden. Ein quadratischer Marker wird verwendet, wenn eigentlich keine Werte der zugehörigen Variable in dem entsprechenden Moment vorliegen. In diesem Fall wird eine Substitutionstechnik verwendet, um den Wert für die Variable in diesem Moment zu schätzen.
Die Substitutionstechnik ähnelt der in der Statistik als Imputation bekannten Technik zur Approximation fehlender Werte. Diese wird unter Anderem von Alemzadeh et al. [8] und Spratt et al. [9] vorgestellt, jedoch wird sie häufig nur verwendet, um fehlende Werte einzelner Probanden zu schätzen. In dem hier vorliegenden Fall fehlen jedoch nicht nur einzelne Werte bestimmter Variablen, sondern die Variablen als solche sind zu bestimmten Zeitpunkten nicht verfügbar. Es findet sozusagen die Imputation nicht für einzelne Probanden, sondern für den normalisierten Mittelwert der Variablen selbst statt.
Für die Beschreibung der Technik wird nachfolgend ein mathematisch formeller Formulierungsstil verwendet.

Substitutionstechnik:

Es sei V_{all} die Menge aller Variablen, von denen in allen Momenten Messungen vorliegen (z.B. *som_bmi*). Für ein $w \in V_{all}$ bezeichne w_i die Werte von w in einem bestimmten Moment s_i (so wird z.B. *som_bmi_s2* als som_bmi_2 dargestellt).
Betrachte nun eine Variable v, für die in einem Moment s_l, $l \in \{0, 1, 2\}$, keine Messungen vorliegen (z.B. $mrt\_liverfat$ für $l = 0$, das heißt im Moment s_0). In diesem Fall existiert der Wert $\overline{\mu_{SP}(v_l)}$ nicht.
Gesucht ist also ein Wert $\widetilde{\mu_{SP}(v_l)}$, welcher als Ersatz für $\overline{\mu_{SP}(v_l)}$ in den Plot eingetragen werden kann.
Betrachte hierfür den zu s_l nächstgelegenen Moment s_k, in welchem eine Messung von v vorliegt und entsprechend $\overline{\mu_{SP}(v_k)}$ existiert (z.b. $k = 2$ für $mrt\_liverfat$).
Es seien p_j, $j \in \{1, ..., m\}$, die Mitglieder der SP und $p_j(v_i)$ der Wert eines Mitglieds der SP für die Variable v im Moment s_i, $i \in \{0, 1, 2\}$. Definiere außerdem $\overline{p_j(v_i)} \coloneqq \frac{p_j(v_i) - \mu_{EP}(v_i)}{\sigma_{EP}(v_i)}$ für $i \in \{0, 1, 2\}$ und $j \in \{1, ..., m\}$. (Diese Definition ist analog zu der des normierten Mittelwertes einer SP in 1.)
Berechne damit den Pearson-Korrelationskoeffizienten zwischen $\overline{p_j(v_k)}$ und $\overline{p_j(w_k)}$ für alle $w \in V_{all}$. Fixiere das w, welches den betragsmäßig größten Korrelationskoeffizienten liefert. ($\star$ für spätere Referenz)
Führe eine lineare Regression durch und erhalte daraus für $j \in \{1, ..., m\}$ die Gleichungen $\overline{p_j(v_k)} = a + b\,\overline{p_j(w_k)} + \xi$ mit $a, b \in \mathbb{R}$ und dem Fehlerterm ξ.
Verwende diesen Zusammenhang, um die Werte $\overline{p_j(v_l)}$, welche eigentlich nicht existieren, zu schätzen als:

$$\widetilde{\overline{p_j(v_l)}} = a + b\,\overline{p_j(w_l)}\ . \tag{2}$$

Diese Schätzung lässt sich nutzen, um den Wert $\overline{\mu_{SP}(v_l)}$ zu approximieren.
Betrachte hierfür zunächst den Moment s_k, in welchem die Werte $\overline{p_j(v_k)}$ existieren. In diesem Fall gilt offensichtlich $\mu_{SP}(v_k) = \frac{1}{m}\sum_{j=1}^{m} p_j(v_k)$ und damit:

$$\begin{aligned}\overline{\mu_{SP}(v_k)} &= \frac{\mu_{SP}(v_k) - \mu_{EP}(v_k)}{\sigma_{EP}(v_k)} = \frac{\frac{1}{m}\sum_{j=1}^{m} p_j(v_k) - \mu_{EP}(v_k)}{\sigma_{EP}(v_k)} \\ &= \frac{1}{m}\sum_{j=1}^{m} \frac{p_j(v_k) - \mu_{EP}(v_k)}{\sigma_{EP}(v_k)} = \frac{1}{m}\sum_{j=1}^{m} \overline{p_j(v_k)}\end{aligned}$$

Nutze diesen Zusammenhang nun im Moment s_l, in welchem laut (2) die Approximation $\widetilde{p_j(v_l)}$ für $\overline{p_j(v_l)}$ vorliegt, um

$$\widetilde{\mu_{SP}(v_l)} = \frac{1}{m} \sum_{j=1}^{m} \widetilde{p_j(v_l)}$$

zu schätzen.
Trage den Wert $\widetilde{\mu_{SP}(v_l)}$ anstatt des fehlenden $\overline{\mu_{SP}(v_l)}$ als normierten Mittelwert der SP für die Variable v im Moment s_l in den Plot ein. Für die Markergröße wird dabei der betragsmäßig maximale Korrelationskoeffizient aus ($\star$) verwendet (vgl. Abbildung 5).

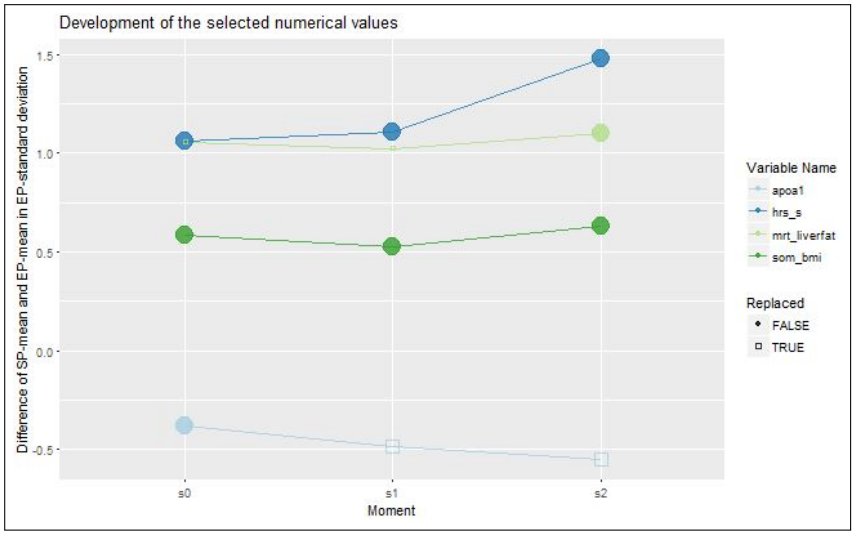


Abbildung 5: Numerical Development Plot bezüglich einer Regel für die zweite Risikoklasse. Alle Variablen außer apoa1 steigen tendenziell im Lauf der Zeit. Für apoa1 wurde eine möglicherweise sinnvolle Substitution durchgeführt, für mrt_liverfat höchstwahrscheinlich nicht (abzulesen an der Größe der quadratischen Marker, detaillierte Informationen durch Klick auf den entsprechenden Punkt).

Der Nutzer hat durch die angepasste Visualisierung von substituierten Werten sowie die Darstellung des Betrags des Korrelationskoeffizienten als Markergröße die Möglichkeit zu beurteilen, ob es sich um eine sinnvolle Substitution handelt. Dabei unterstützt ihn zusätzlich eine Textinformation unterhalb

des Plots, die bei dem Klick auf einen Punkt im Plot angibt, welche Variable im Fall einer Substitution verwendet wurde und wie hoch der Betrag des zugehörigen Korrelationskoeffizienten ist.
Für die Substitutionstechnik wurde generell ein linearer Zusammenhang angenommen. Dies ist nicht immer eine korrekte Annahme, jedoch hat der Nutzer die Möglichkeit, die Angemessenheit des Modells in fragwürdigen Fällen selbst einzuschätzen, da ihm sowohl die Substitutionsvariable als auch das Maß des linearen Zusammenhangs mitgeteilt wird. Dadurch wird eine Kombination der eher herkömmlichen Methode der linearen Regression und des neueren Ansatzes der Klassifikationsregeln kombiniert. Der Vorteil ist hierbei, dass sich durch die Beschränkung auf eine SP bei der Regression bessere lineare Modelle finden lassen, als wenn man diese für die EP sucht, da die Mitglieder der SP größere Ähnlichkeiten miteinander aufweisen.
Häufig besitzen Variablen, die im Antezedens der zur SP gehörenden Regel verwendet werden, eine hohe Abweichung von der EP in diesem Plot. Da für diese Variablen explizite Cutoff Werte für die Bildung der SP gegeben sind, ist dieser Umstand allerdings naheliegend. Eine bemerkenswertere Feststellung ist hingegen, dass bei Gruppen mit erhöhtem Leberfettwert häufig Werte, die den Alkoholkonsum der Probanden beschreiben, über den gesamten Verlauf der Studie kaum von dem Durchschnittswert der EP abweichen. Da sie auch fast nie in den Antezedenzien der generierten Regeln vorkommen, ist also die Vermutung naheliegend, dass sie keinen großen Einflussfaktor für die Steatosis hepatis darstellen. Dieser Zusammenhang lässt sich auch bei der Untersuchung der kategorialen Variablen in Abschnitt 5.7 feststellen. Um diese Erkenntnis zu verifizieren, müsste geprüft werden, auf welche Weise die zu Grunde liegenden Werte erfasst wurden. Beruhen sie lediglich auf Aussagen der Probanden, so besteht die Möglichkeit, dass die Angaben verfälscht sind. Ein Grund hierfür ist, dass manche Probanden unter Umständen falsche Angaben bezüglich ihres Alkoholkonsums machen, um eher den gesellschaftlichen Normen zu entsprechen.
Eine Besonderheit in der Darstellung ist, dass der normierte Mittelwert für age_ship über die Momente hinweg fallen kann. Dieser Umstand mag zunächst seltsam wirken, ist jedoch durchaus naheliegend. In den normierten Werten spiegelt sich lediglich die durchschnittliche Abweichung der SP von der EP wieder. So gibt es zum Beispiel eine SP, deren Durchschnittsalter zu den unterschiedlichen Studienzeitpunkten gerundet die folgenden Werte besitzt: s0 - 45, s1 - 51, s2 - 56. Das Durchschnittsalter der EP hingegen entwickelt sich wie folgt: s0 - 44, s1 - 50, s2 - 56. Von s1 auf s2 fällt entspre-

chend der normierte Durchschnittswert von age_ship für die beschriebene SP. Die Ursache hierfür liegt darin, dass die Erfassung der für SHIP-2 relevanten Daten in einem Zeitraum von 2008 bis 2012 stattgefunden hat. In der Anwendung wird jedoch vereinfachend angenommen, dass alle Messungen aus SHIP-2 zum selben Zeitpunkt vorgenommen wurden. Dieser Umstand könnte in zukünftigen Anwendungen differenzierter betrachtet werden.

Anmerkung: Die Substitutionstechnik könnte erweitert werden, indem man für die Substitution nicht nur die Variablen verwendet, welche in allen Momenten vorkommen, sondern auch jene, die lediglich in dem Moment vorkommen, in dem eine Messung fehlt und in einem Moment, in dem substituiert werden kann. Dies könnte genutzt werden, wenn Messmethoden durch neue, aber dennoch vergleichbare Methoden abgelöst werden, wobei die neuen Messungen ähnliche Werte wie bei der alten Methode ergeben, genauer gesagt, sie eine hohe Korrelation zueinander aufweisen. Bei einem häufig wechselnden Studienprotokoll wäre dies hilfreich, um zeitliche Entwicklungen besser nachvollziehen zu können.

5.6 Development Boxplots

Wird ein bestimmter Punkt im Numerical Development Plot angeklickt, so werden Boxplots angezeigt, welche die zeitliche Entwicklung der entsprechenden Variable veranschaulichen (vgl. Abbildung 6). Außerdem wird die Information angezeigt, ob die Variable zu mindestens einem Zeitpunkt substituiert wurde und, falls dies der Fall ist, durch welche Variable und mit welchem betragsmäßigen Korrelationskoeffizienten.
An dieser Stelle wären auch andere Visualisierungsformen denkbar gewesen, wie zum Beispiel parallele Koordinaten mit einer gesonderten Hervorhebung für die Mitglieder der SP. In dieser Darstellungsform hätte die Entwicklung einzelner Probanden untersucht werden können, was in den Boxplots nicht möglich ist, da keine Verbindung zwischen den Punkten verschiedener Momente existiert. Allerdings hätten in diesem Fall fortgeschrittene Techniken für den Umgang mit Overplotting verwendet werden müssen. Dahingegen erlaubt die Darstellung mit Hilfe der Boxplots eine abstraktere und damit einfacher nachzuvollziehende Gegenüberstellung von SP und EP und zudem das Ablesen in der Statistik häufig verwendeter Werte. So stehen die obere bzw. die untere Grenze einer Box für das obere bzw. das untere Quartil. Die Linie innerhalb des Plots steht für den Median und der obere bzw. der untere

Whisker für den 1.5-fachen Interquartilsabstand, ausgehend vom oberen bzw. unteren Quartil.
Für die Boxplots werden nur die Momente verwendet, in denen tatsächlich

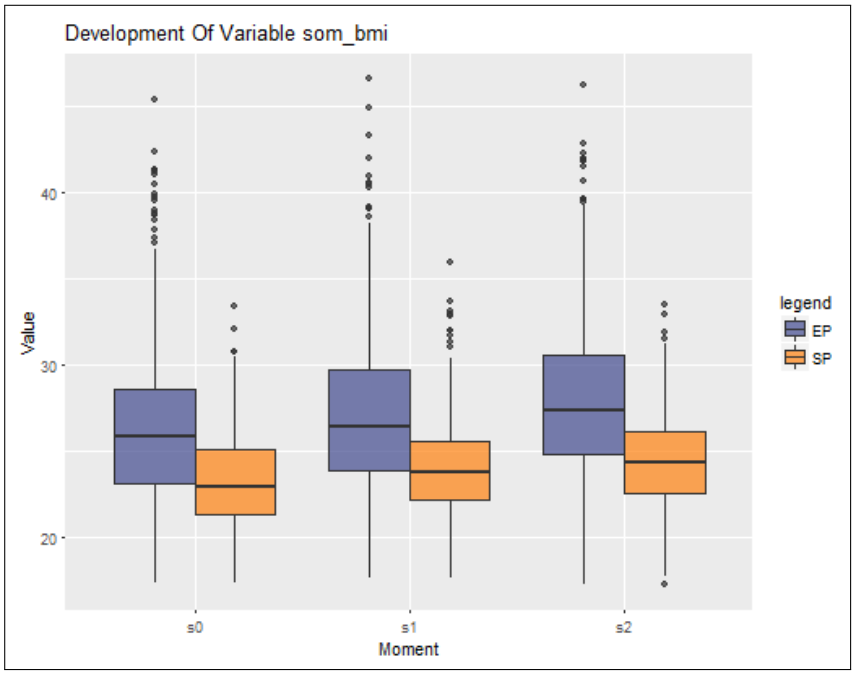


Abbildung 6: Development Boxplots für die Variable som_bmi für eine ausgewählte SP. Allgemeiner Anstieg der Werte sowohl bei der SP als auch bei der EP, die SP jedoch im Schnitt mit deutlich niedrigeren Werten.

Werte für die betrachtete Variable verfügbar sind.
Insgesamt ermöglicht die Boxplot-Übersicht einen noch tieferen Einblick in die Entwicklung der SP im Vergleich zur EP bezüglich der gewählten Variable. So gibt es zum Beispiel Variablen, die im Numerical Development Plot über die Momente hinweg ungefähr gleich bleiben, in den Development Boxplots hingegen steigen. In diesem Fall weist die EP ebenfalls eine steigende Entwicklung auf, wodurch sich der Anstieg der SP relativiert. Unter anderem kommt dies für die Variable som_bmi vor (vgl. Abbildung 6, in der die BMI-Werte der SP eigentlich ansteigen). Würde man nur die unnormierte Entwicklung der SP bezüglich ihrer BMI-Werte untersuchen, könnte man zu dem Schluss kommen, dass es sich um einen aussagekräftigen Anstieg der BMI-Werte handelt. Da der BMI für die EP mit steigendem Alter allerdings

ebenfalls zunimmt, ist davon auszugehen, dass dies schlicht eine altersbedingte Entwicklung ist. Durch die Normierung wird somit das Altern der Kohorte berücksichtigt.
Außerdem kann mit Hilfe der Boxplots abgeschätzt werden, wie die Mitglieder der SP auf die einzelnen Risikoklassen verteilt sind, indem die Variable mrt_liverfat zur näheren Untersuchung ausgewählt wird. Da die Visualisierung allerdings in Form von Boxplots stattfindet, wird die Verteilung auf die Risikoklassen nicht exakt wiedergegeben.

Das Modul unterstützt also den Nutzer dabei, in Kombination mit dem Numerical Development Plot die Charakteristika der SP im Bezug auf numerische Variablen, insbesondere im Hinblick auf Entwicklungen über die Zeit, besser zu verstehen und dadurch beurteilen zu können, ob es sich tatsächlich um eine plausible SP handelt.

5.7 Categorical Development Plot

Im Fall von kategorialen Variablen kommt ebenfalls eine Technik zur Anwendung, welche es ermöglichen soll, die Abweichung der Verteilung der SP von der Verteilung der EP zwischen verschiedenen Variablen zu vergleichen. Da die Variablen nicht numerisch sind, kann die Technik aus 5.5 nicht verwendet werden. Stattdessen wird auf den Kontingenzkoeffizienten *Cramérs V* zurückgegriffen. Außerdem wird hier auf eine Substitutionstechnik verzichtet.
Voraussetzung ist wie im obigen Fall, dass eine zu betrachtende SP bestimmt und mindestens eine kategoriale Variable in einer Übersicht selektiert wurde. Standardmäßig ist dies stea_alt75, eine Variable welche Auskunft über den Ultraschallbefund der Leber sowie den ALAT Wert gibt, wobei ein erhöhter ALAT Wert auf eine Lebererkrankung hindeutet. Unterstützt wird die Auswahl hierbei erneut durch eine Sortierung der Variablen nach bestimmten Kategorien und durch eine zusätzliche Liste, welche die durchschnittliche Stärke der Abweichung der SP von der EP über alle Momente hinweg für die einzelnen Variablen angibt.
Für alle Momente, in denen Messungen der ausgewählten Variablen durchgeführt wurden, wird im Plot der jeweilige Wert des Cramérs V der SP im Bezug zur EP dargestellt (vgl. Abbildung 7). Zur selben Variable gehörende Punkte sind mit Linien verbunden. Das Cramérs V berechnet sich wie folgt. Zunächst wird der χ^2-Koeffizient basierend auf einer Tabelle bestimmt, welche eine Spalte für die Mitglieder der SP besitzt und eine Spalte für die

Mitglieder der EP, die nicht Teil der SP sind. Die Anzahl der Zeilen der Tabelle entspricht der Anzahl an Ausprägungen der entsprechenden Variable. NAs werden dabei als eine mögliche Ausprägung betrachtet. In den einzelnen Zellen steht jeweils die Anzahl der Probanden welche der Gruppe angehören, die der aktuellen Spalte entspricht, und die Ausprägung bezüglich der betrachteten Variable besitzen, die der aktuellen Zeile entspricht.
Das Ergebnis wird weiterverwendet, um Cramérs V zu berechnen. Dafür wird die Formel

$$V = \sqrt{\frac{\chi^2}{n\,(k-1)}} \tag{3}$$

genutzt, wobei n die Anzahl aller Probanden und k das Minimum der Anzahl an Spalten und der Anzahl an Zeilen in der oben genannten Tabelle ist, in diesem Fall also immer 2. Einzige Ausnahme: Bei Variablen wie park_s0, für die alle Probanden der EP denselben Wert besitzen, ist $k = 1$.
Der Vorteil des Cramérs V gegenüber dem reinen χ^2-Koeffizienten ist, dass durch die Normierung in 3 der resultierende Wert auf einen Bereich zwischen 0 und 1 beschränkt wird und so die Verteilungen unterschiedlicher Variablen besser verglichen werden können.
Punkte, die weiter von 0 entfernt sind, zeigen an, dass die Verteilung der SP für die zugehörige Variable in diesem Moment stärker von der Verteilung der EP abweicht, wohingegen ein Wert nah bei 0 eine ähnliche Verteilung von SP und EP signalisiert.
Es gibt Fälle, in denen die SP sehr klein ist, wodurch die Berechnung des χ^2-Koeffizienten unter Umständen an Zuverlässigkeit verliert. Dadurch kann auch das Cramérs V an Aussagekraft verlieren. In diesen Fällen können jedoch die im nächsten Abschnitt vorgestellten Development Barcharts verwendet werden, um ein besseres Verständnis für die Verteilung der SP im Vergleich zur EP zu erlangen.
Ebenso wie der Numerical Development Plot dient diese Komponente dem Nutzer dazu, die gewählte SP besser zu verstehen.

Anmerkung zur Wahl der voreingestellten Variablen im Numerical Development Plot und Categorical Development Plot: Die Variablen wurden so gewählt, dass mit ihrer Hilfe schnell ein gutes Verständnis (Mental Map) von der Anwendung beim Nutzer entstehen kann. Es werden von diesen Variablen nämlich bereits alle der nachfolgenden Fragen bezüglich Spezialfällen beantwortet: Was passiert, wenn bei einer numerischen Variable nicht alle

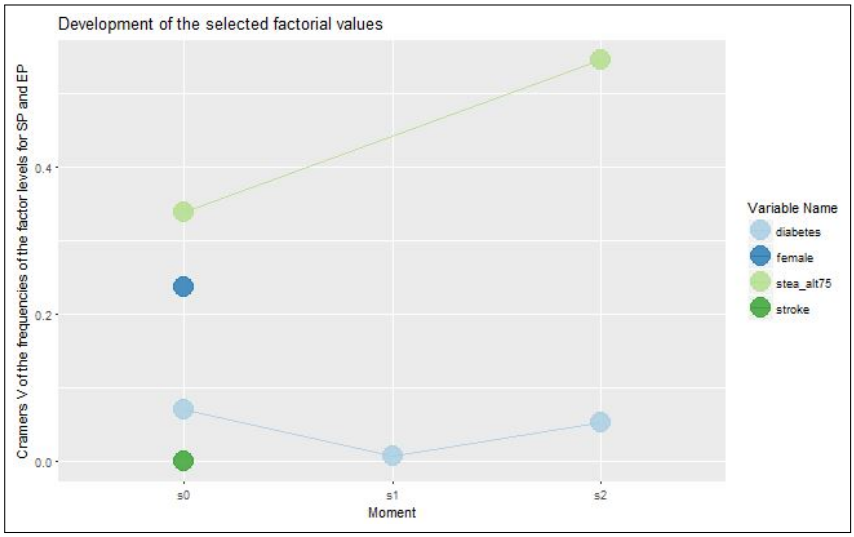


Abbildung 7: Factorial Development Plot einer ausgewählten SP für bestimmte Variablen. diabetes mit eher geringer, stea_alt75 hingegen mit starker Abweichung. Scheinbar hauptsächlich Probanden eines bestimmten Geschlechts in der SP, abzulesen an der verstärkten Abweichung in female.

Momente vorhanden sind? Wie sieht eine Kurve aus, die Substitutionsvariablen verwendet? Was geschieht, wenn bei einer kategorialen Variable nicht alle Werte vorhanden sind? Welche Folgen hat es, wenn in einem Plot in einem Moment von keiner Variable Messungen vorliegen?
Durch die Beantwortung dieser Fragen erlangt der Nutzer schnell ein Gefühl für die beiden Module, die für ihn zunächst unter Umständen ungewohnt sein mögen.

5.8 Development Barcharts

Diese Komponente dient analog zu den Development Boxplots dazu, die Entwicklungen einzelner kategorialer Variablen separat in größerem Detail zu betrachten. Hier wird nach der Selektion eines Punktes im Categorical Development Plot die Verteilung der Mitglieder der SP und der EP auf die einzelnen Ausprägungen der entsprechenden Variable für die jeweiligen Momente gegenübergestellt. Der Nutzer hat dabei die Wahl, die Verteilungen mit absoluten oder mit relativen Häufigkeiten zu vergleichen. Die erste Möglichkeit berücksichtigt die Größe der SP, die zweite hingegen ermöglicht einen

deutlich leichteren Vergleich der Verteilungen von EP und SP, insbesondere bei kleinen SPs (vgl. Abbildung 8).

Um eine bessere visuelle Abtrennung von SP und EP zu erhalten, besitzt

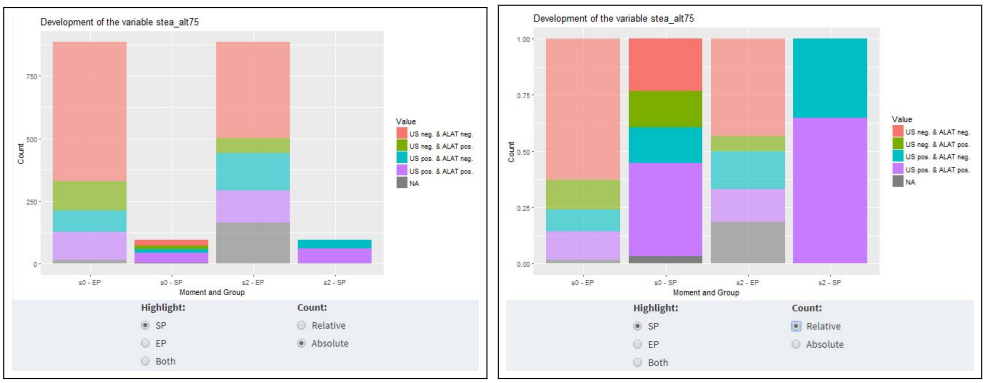


Abbildung 8: Development Barcharts für die Variable stea_alt75 in den Momenten s0 und s2 für eine ausgewählte SP. Der Fokus liegt jeweils auf der SP. Links werden absolute Häufigkeiten verwendet, rechts relative. In der Darstellung mit relativen Häufigkeiten ist eine klare Abweichung der Verteilung der SP von der Verteilung der EP zu erkennen, die sich von s0 auf s2 verstärkt.

der Nutzer außerdem die Option, den visuellen Fokus entweder auf die SP oder die EP zu legen, um diese hervorzuheben.
Mit dieser Darstellung lassen sich daher die Ergebnisse aus dem Categorical Development Plot verifizieren.

6 Umsetzung und Implementierung

6.1 Organisation und Arbeitsweise

In einem OneNote Dokument wurde die angestrebte Funktionalität der zu implementierenden Methoden zunächst in eigenen Abschnitten für die jeweiligen Komponenten in Worten beschrieben und bei hoher Komplexität zudem in Pseudocode formuliert. Während der tatsächlichen Implementierung, für die die Programmiersprache R verwendet wurde, wurde außerdem eine projektübergreifende To-do-Liste geführt. Deren einzelne Punkte verlinkten auf offene Probleme bzw. Aufgaben innerhalb der Abschnitte für die Umsetzung

der einzelnen Komponenten. Diese Verlinkung wurde ebenfalls an Stellen innerhalb der einzelnen Abschnitte verwendet, an denen die Funktionalität verschiedener Komponenten ineinander greift.
Zudem wurde eine online abrufbare grafische Übersicht des angestrebten Workflows für den eigenen Überblick und für den Austausch mit dem Betreuer erstellt.
Für die Darstellung des Tools als interaktive HTML-Seite wurde das R-Package RShiny verwendet.

6.2 Besondere Herausforderungen

Wie eingangs beschrieben bietet der SHIP-Datensatz als epidemiologische Langzeitstudie einige besondere Herausforderungen. Der Umgang damit wird nachfolgend noch einmal gesondert herausgestellt.
Zum einen sind bei Langzeitstudien Drop-Outs nicht zu verhindern, sprich, dass Teilnehmer im Lauf der Studie ausfallen. Das Problem ist im Fall dieses Projekts dahingehend abgemildert, dass lediglich die Probanden betrachtet wurden, bei denen die Ausprägung der Zielvariable mrt_liverfat_s2 bekannt ist. Dennoch ist es der Fall, dass für die übrigen Variablen teilweise in bestimmten Momenten keine Werte erfasst wurden. Diese Tatsache wurde an verschiedenen Stellen wie folgt berücksichtigt. In den SP-EP Barcharts existiert pro Plot ein Balken, der die Anzahl von NA Einträgen angibt. Im Numerical Development Plot wurden NA Einträge bei der Berechnung von Mittelwert und Standardabweichung ignoriert, ebenso in den Development Boxplots. Im Categorical Development Plot und in den Development Barcharts wurden sie als extra Variablenausprägung in die Visualisierung aufgenommen.
Eine weitere Herausforderung ist die Variabilität des Studienprotokolls, die verursacht wird durch die Hinzunahme neuer und das Ablösen alter Variablen über mehrere Studienzeitpunkte hinweg. Diesem Umstand wurde auf experimentelle Weise mit den Development Plots begegnet. So ist im Categorical Development Plot, in den Development Boxplots und den Development Barcharts erkennbar, wenn Variablen nicht in allen Momenten vorkommen. Im Numerical Development Plot wurde durch die Substitutionstechnik sogar eine Methode vorgestellt, um einen Ersatz für die fehlenden Werte zu finden.
Außerdem wurde dem Altern der Kohorte in der Anwendung auf zwei Arten begegnet. Zum einen durch die Normierung der Durchschnittswerte der SPs im Numerical und im Categorical Development Plot, welche die Alte-

rungseffekte relativieren und zum anderen durch die Development Boxplots und Barcharts, welche eine unnormierte Betrachtung der Entwicklung ermöglichen. Außerdem wurden die bezüglich der Zielvariable relevanten Regeln ausschließlich in einem Moment, SHIP-2, ermittelt.

6.3 Verworfene Ansätze

Ein Ansatz, welcher keine allzu befriedigende Ergebnisse lieferte, bestand darin, zwei Arten der Suche nach HotSpot Regeln zu vergleichen. Hierfür wurde zunächst der Datensatz nach den numerischen Variablen gefiltert. Für alle Variablen, die in mehr als einem Moment vorkommen, wurde jeweils die Differenz der Werte aus den verschiedenen Momenten gebildet und als zusätzliche Variable zu dem ursprünglichen Datensatz (mit Informationen über alle Momente) hinzugefügt. Darauf wurde der HotSpot Algorithmus angewandt und die Qualität der gefundenen Regeln mit der Qualität der Regeln verglichen, welche der Algorithmus, angewandt auf den ursprünglichen Datensatz ohne die Differenzwerte, lieferte.
Dabei war die Erwartung, dass die gefundenen Regeln aus dem Datensatz mit den zusätzlichen Differenzen mindestens genau so gut sein würden, wie die Regeln aus dem Datensatz ohne die Differenzen, da die Informationen aus dem zweiten komplett in dem ersten enthalten waren.
In der Tat wurde auch eine Verbesserung der Qualität der Regeln festgestellt, allerdings nur in einem sehr geringen Umfang. So erhöhte sich zum Beispiel die Confidence der besten Regel im Bezug auf die zweite Risikoklasse bei gleichbleibenden Parametern lediglich um 0.04, was auf Grund der deutlich höheren Anzahl an Variablen in diesem Fall lediglich probabilistische Gründe haben kann.
Die Differenzen-Technik wäre dann nützlich gewesen, wenn eine Person bereits über einen längeren Zeitraum unter Beobachtung steht und entsprechend Messungen zu verschiedenen Zeitpunkten vorliegen. Sie hätte die Möglichkeit berücksichtigt, dass für die Zugehörigkeit einer Person zu einer bestimmten Risikoklasse eventuell nicht primär bestimmte Werte zu einem festen Zeitpunkt relevant sind, sondern eine (möglicherweise drastische) Änderung bestimmter Werte ausschlaggebender ist.

Ein weiterer Ansatz war, ebenfalls eine Substitutionstechnik für den Categorical Development Plot zu entwickeln. Das ist auch geschehen, allerdings wurde das dafür verwendete Modell als zu abwegig vermutet. Mit mehr Auf-

wand ließe sich hierfür vermutlich ebenfalls ein angemessenes Modell entwickeln.

Ein Problem, welches während der Entwicklung der Anwendung auftrat, dann allerdings behoben werden konnte, war die benötigte Zeit für die Berechnung der Kreisdurchmesser im Support-Confidence Plot. Der ursprüngliche Ansatz war, die Kreisdurchmesser automatisch basierend auf der Qualität der besten Prediction Rule für die jeweilige SP zu bestimmen. Hierfür wurde zunächst pro SP der gesamte String-Output des HotSpot Algorithmus für die Prediction Rules geparsed, um anschließend nach der besten Regel zu suchen. Nachdem sich dieser Ansatz jedoch als zu zeitaufwändig herausstellte, wurden zwei Schritte unternommen, um damit umzugehen. Zum einen wurde die beschriebene Art der Bestimmung der Kreisdurchmesser optional gemacht. Zum anderen wurde die Eigenschaft der Implementierung des verwendeten HotSpot Algorithmus genutzt. Die im Ausgabe-String enthaltenen Regeln sind nämlich absteigend nach ihrer Confidence geordnet. So genügt bereits eine Substring-Suche nach der Confidence der ersten Regel im Output, wenn als Gütekriterium für die Bestimmung der Kreisdurchmesser die maximal mögliche Confidence der Prediction Rules der jeweiligen SP verwendet werden kann. Dies wurde als angemessener Kompromiss befunden, wenngleich die Confidence alleine nicht optimal zur Bestimmung der Qualität einer Regel ist. Die Technik wurde dennoch verwendet, da der Nutzer die Möglichkeit besitzt, durch Selektion einer SP die zugehörigen Prediction Rules in dem entsprechenden Modul (Abschnitt 5.4) auch basierend auf weiteren Qualitätsmaßen zu untersuchen.

7 Ausblick

Zukünftige Aufgaben umfassen unter anderem, die Anwendung mit weniger Restriktionen, verursacht durch verwendete Packages, umzusetzen. Dies bedeutet zwar zusätzlichen Aufwand, da mehr Funktionalität selbst implementiert werden muss, allerdings auch eine bessere Anpassbarkeit an die Ansprüche und Vorstellungen, welche an das Tool gerichtet werden.
Ein weiteres Ziel wäre eine tiefergehende Auseinandersetzung mit den Prediction Rules, falls diese bei dem Nutzer auf Interesse stoßen. So könnte man die bisher festgelegten Parameter für die Regelsuche anpassbar machen. Des Weiteren könnte man in diesem Fall eine Visualisierungstechnik vergleichbar

mit der im Support-Confidence Plot verwenden. Dies würde ein besseres und schnelleres Verständnis für die gefundenen Prediction Rules ermöglichen, als rein durch die Verwendung einer Tabelle wie bisher.
Zudem könnten für die Substitutionstechnik weitere Modelle als nur ein lineares verwendet werden. Weitet man diese auch auf kategoriale Variablen aus, so könnte man sogar für den Categorical Development Plot die Anwendung einer Substitutionstechnik ermöglichen.
Außerdem könnten die im Abschnitt Verwandte Arbeiten angesprochenen Methoden zur genaueren Untersuchung einzelner SPs und zum differenzierteren Umgang mit fehlenden Werten einzelner Probanden in die vorgestellte Anwendung eingebaut werden, um dem Nutzer eine noch präzisere Untersuchung der unterschiedlichen SPs zu ermöglichen.

Literatur

[1] H. Völzke, D. Alte, C. O. Schmidt, (...), R. Biffar, U. John und W. Hoffmann, Cohort Profile: The Study of Health in Pomerania, in *International Journal of Epidemiology*, 40 (2), 2011, pp.294-307.

[2] E. Frank, M. A. Hall und I.H. Witten, The WEKA Workbench, *Online Appendix for ‚Data Mining: Practical Machine Learning Tools and Techniques'*, Morgan Kaufmann, Fourth Edition, 2016.

[3] U. Niemann, H. Völzke, J. Kühn und M. Spiliopoulou, Learning and Inspecting Classification Rules from Longitudinal Epidemiological Data to Identify Predictive Features on Hepatic Steatosis, in *Expert Systems with Applications*, 41, 2014, pp.5405–5415.

[4] P. Klemm, Interactive Visual Analysis of Population Study Data, PhD thesis, Otto-von-Guericke University Magdeburg, 2016.

[5] P. Klemm, K. Lawonn, S. Glaßer, U. Niemann, K. Hegenscheid, H. Völzke und B. Preim, 3D Regression Heat Map Analysis of Population Study Data, in *IEEE Transactions on Visualization and Computer Graphics*, 22 (1), 2016, pp.81-90.

[6] S. Alemzadeh, T. Hielscher, U. Niemann, L. Cibulski, T. Ittermann, H. Völzke, M. Spiliopoulou und B. Preim, Subpopulation Discovery and Validation in Epidemiological Data, *EuroVis Workshop on Visual Analytics*, 2017.

[7] J. Krause, A. Perer, H. Stavropoulos, Supporting Iterative Cohort Construction with Visual Temporal Queries, in *IEEE Transactions on Visualization and Computer Graphics*, 22 (1), 2016, pp.91-100.

[8] S. Alemzadeh, U. Niemann, T. Ittermann, H. Völzke, D. Schneider, M. Spiliopoulou und B. Preim, Visual analytics of missing data in epidemiological cohort studies, in *Eurographics Workshop on Visual Computing for Biology and Medicine (VCBM)*, 2017, pp.43-52.

[9] M. Spratt, J. Carpenter, J. A. C. Sterne, J. B. Carlin, J. Heron, J. Henderson und K. Tilling, Strategies for Multiple Imputation in Longitudinal Studies, in *American Journal of Epidemiology*, 172 (4), 2010, pp.478–487.

[10] M. Hahsler, S. Chelluboina, K. Hornik und C. Buchta, The arules {R}-Package Ecosystem: Analyzing Interesting Patterns from Large Transaction Datasets, in *Journal of Machine Learning Research*, (12), 2011, pp.1977-1981.

[11] J. Fürnkranz, D. Gamberger und N. Lavrac, *Foundations of Rule Learning*, Berlin: Springer-Verlag, 2012, pp.135-169.

[12] G. Bedogni, S. Bellentani, L. Miglioli, F. Masutti, M. Passalacqua, A. Castiglione und C. Tiribelli, The Fatty Liver Index: a simple and accurate predictor of hepatic steatosis in the general population, *BMC Gastroenterology*, 6 (33), 2006, 7 pages.

[13] Z. Zhang, D. Gotz und Adam Perer, Iterative Cohort Analysis and Exploration, *Information Visualization*, 14 (4), 2015, p.293.