Aus der Klinik für Neurologie
(Direktor: Prof. Dr. Daniela Berg)
im Universitätsklinikum Schleswig-Holstein, Campus Kiel
an der Christian-Albrechts-Universität zu Kiel

Fehleranalyse zur Beurteilung der Qualität automatisierter FreeSurfer Segmentierungen von MRT Aufnahmen in einer neurogeriatrischen Kohorte

Inauguraldissertation
zur
Erlangung der Doktorwürde der Medizin
der Medizinischen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von
Maximilian James Kress
aus Kassel
Kiel 2020

1. Berichterstatter/in: Prof. Dr. Walter Maetzler, Klinik für Neurologie

2. Berichterstatter/in: Prof. Dr. Olav Jansen, Klinik für Radiologie und Neuroradiologie

Tag der mündlichen Prüfung: 02.07.2021

Zum Druck genehmigt, Kiel, den 03.03.2021

gez.: Prof. Dr. Ralf Baron

(Vorsitzender der Prüfungskommission)

Inhaltsverzeichnis

1. Einleitung

Unser Gesundheitssystem wird mit einer zunehmend alternden Gesellschaft vor neue Herausforderungen gestellt. Neben dem Anstieg von alters-assoziierten Erkrankungen ergibt sich eine erhöhte Wahrscheinlichkeit, im Alter an mehreren Erkrankungen gleichzeitig zu leiden (sogenannte Multimorbidität) (Fabbri et al., 2015). Eine schottische Querschnittsstudie (Barnett et al., 2012) zeigte, dass in der Altersgruppe von 65-84 Jahren 65% und in der Gruppe >84 Jahren 82% multimorbide erkrankt sind. Beachtlich ist auch, dass bereits im Alter von 45-64 Jahren 30% Multimorbidität aufweisen (Barnett et al., 2012). Diese Untersuchungen machen deutlich, dass eine stetige Verbesserung der Diagnosestrategien von Multimorbidität dringend notwendig ist (M. Tinetti, 2016; M. E. Tinetti, Bogardus, & Agostini, 2004). Auf dem Gebiet der Gehirnanatomie und –pathologie bieten bildgebende Verfahren, wie die strukturelle Magnetresonanztomographie, Möglichkeiten, krankheitsbedingte oder –begleitende Abweichungen des zentralen Nervensystems zu quantifizieren. Neuere Verfahren in der Bildverarbeitung, die z.B. in der Software FreeSurfer genutzt werden (siehe 2.2), erlauben eine automatisierte Detektion von Gehirnstrukturen, mit dem Nutzen beispielsweise Volumenanalysen dieser Strukturen, zu diagnostischen Zwecken, durchführen zu können. Diese Techniken sind jedoch in der Regel an Normalkollektiven entwickelt worden und deren Verwendbarkeit bei multimorbiden Patienten ist daher möglicherweise eingeschränkt. Diese Arbeit untersucht daher die Anwendbarkeit der Software FreeSurfer bei Magnetresonanz-Tomographien (MRT) in einer multimorbiden Kohorte mit neurologischen Erkrankungen. Dazu wird ein für diesen Zweck entwickeltes Bewertungssystem vorgestellt und angewandt.

1.1 Magnetresonanz-Tomographie

In vielen Bereichen der modernen klinischen Medizin ist die MRT ein etabliertes Verfahren zur primären Diagnostik, Determinierung akutmedizinischer Behandlungsstrategien sowie zur Verlaufs- und Langzeitkontrolle.

Zusammengesetzt aus den altgriechischen Wörtern „τομή“ (= Schnitt) und „γράφειν“ (=schreiben) bezeichnet die Tomographie eine Form der Schnittbildgebung. Anders als die Computertomographie (CT) nutzt die MRT keine ionisierende Röntgenstrahlung, sondern ein starkes Magnetfeld und Hochfrequenzimpulse im UKW-Radiowellenbereich.

Wie im alten Namen „Kernspintomographie“ erkennbar, basiert dieses Bildgebungsverfahren auf dem Drehimpuls des Atomkerns, dem Kernspin. Eine Modellvorstellung besteht darin, dass ein geladenes Teilchen wie das Proton, das den Kern eines Wasserstoffatoms bildet, sich um seine eigene Achse dreht und so ein Magnetfeld erzeugt.

Atomkerne mit ungerader Protonen- oder Neutronenzahl besitzen ein resultierendes magnetisches Moment, welches man mit einem geeigneten Instrument messen kann. In klinischen MRTs werden in der Regel nur Wasserstoffkerne gemessen. In einem externen Magnetfeld richten sich die Kernspins entlang der Feldlinien aus. Das dazu notwendige Grundmagnetfeld wird bei klinischen Geräten mit einer supraleitenden Magnetspule erzeugt, die mit flüssigem Helium gekühlt wird (Pabst, 2013; Weishaupt, Köchli, & Marincek, 2014). Die magnetische Flussdichte (B0), oder auch Magnetfeldstärke, wird in der Einheit „Tesla" (T) angegeben. Für klinische Zwecke werden Scanner mit 1,5 bis 7 Tesla verwendet. Möglich sind auch höhere Feldstärken. Allerdings eignen sich Geräte mit Magnetfeldstärken von bis zu 21,1 Tesla nur noch beispielsweise zur Molekular- oder Kleintierforschung (Schepkin et al., 2012). Im Vergleich dazu schwankt laut der amerikanischen „National Oceanic and Atmospheric Administration" (NOAA) das Erdmagnetfeld bei einer Feldstärke lediglich zwischen 30µT und 60µT.

Im Inneren des MRT laufen die Feldlinien des supraleitenden Magneten parallel zur Z-Achse. Im Grundzustand erfolgt die Ausrichtung der Rotationsachsen der Wasserstoffkerne mehr entlang als entgegen der Feldlinien. Somit richtet sich auch der entstehende Vektor ihres Magnetfeldes entlang der Feldlinien des Hauptmagneten aus. Um eine Aufnahmen eines gewünschten Körperteils und bestimmter Gewebetypen zu erhalten, nutzt man hochfrequente (HF) Radioimpulswellen, die die Drehachsen der spinnenden Protonen auslenken. Ein auf diese Weise angeregtes Messobjekt im Magnetfeld hat Eigenschaften, die man als T1- und T2-Zeit bezeichnet. Wenn die Protonen von diesem angeregten Zustand (hohes Energielevel) wieder in den Grundzustand zurückfallen, wird messbare Energie an die Umgebung abgegeben. Die T1-Zeit (oder auch T1 Relaxation) gibt die Zeit an, in der sich 63% Ursprungsmagnetisierung wieder entlang der Feldlinien des Hauptmagneten, eingestellt hat (Pabst, 2013). Die T1-Zeit ist für jedes Gewebe spezifisch, wodurch ein T1 Kontrast auf den MRT Bildern möglich wird (Pabst, 2013). Beispielsweise benötigt bei B0 = 3Tesla Fett ca. 390ms, wohingegen Liquor über 4000ms braucht (Bojorquez et al., 2017). Regt man ein Gewebe mehrfach an, ohne dass die Spins in der Zwischenzeit vollständig in den Grundzustand zurückkehren konnten, dann hängt die Anregung der Spins von der T1-Zeit des Gewebes ab. Ein so gemessenes Bild wird als T1-gewichtetes Bild oder kurz T1-Bild bezeichnet. Der zeitliche Abstand aufeinanderfolgender Anregungen wird als Repetitionszeit (TR) bezeichnet. T1 gewichtete Aufnahmen des Gehirns weisen meist einen guten Kontrast zwischen Grauer und Weißer Substanz auf und können so zur Segmentierung verschiedener Hirnregionen genutzt werden.

Kurz nach Anregung durch den Frequenzimpuls kreiseln die Protonen phasengleich, treten allerdings untereinander in Wechselwirkung und verlieren dadurch ihre Synchronität. Messbar ist hier ein schwächer werdender Protonenvektor. Diese Dephasierung wird als T2-

Relaxation bezeichnet und weist ebenfalls gewebsspezifische Zeiten (T2-Zeit) auf. So benötigt die Graue Substanz beispielsweise ca. 90ms und die Weiße Substanz ca. 70ms (Bojorquez et al., 2017). Die longitudinale (T1) und die transversale (T2) Relaxation laufen voneinander unabhängig ab (Lehmann, 2013).

Bei T2-gewichteten Bildern ist der Beobachtungszeitpunkt nach der Frequenzanregung (Echozeit = TE) entscheidend. Wählt man die TE so, dass die kreiselnden Protonen von schnell dephasierenden Geweben bereits zerstreut sind, liegt das Bildsignal schwerpunktmäßig auf Gewebearten, in denen die Homogenität der synchronen Wasserstoffprotonen lange erhalten bleibt. Hell erscheinen hier meistens Flüssigkeiten, wie z.B. Liquor, bei gleichzeitig etwas dunkleren fettreichen Geweben (Thomalla et al., 2009).

Um nicht nur ein gemitteltes Signal der Gewebe im MRT, sondern ein Schnittbild zu erhalten, muss man die einzelnen Schichten voneinander unterscheiden und die Bildpunkte (Voxel) dreidimensional im Raum anordnen können. Dazu werden mittels weiterer Spulen innerhalb des Hauptmagneten Gradienten von wenigen mT/m erzeugt, welche eine Ortskodierung entlang der X-, Y- und der Z-Achse möglich machen (Lehmann, 2013; Pabst, 2013; Weishaupt et al., 2014).

T1-gewichtete Bilder dienen vor allem zur strukturellen Analyse von Körpergeweben. In der Neuroradiologie sind sie Grundlage für die Beurteilung von Veränderung der Hirnanatomie (z.B. Entwicklungsanomalien, Raumforderungen, cerebrovaskuläre Erkrankungen, u.v.m.). Mit T2-gewichteten Bildern lässt sich freies und/oder gewebsständiges Wasser darstellen, wie es beispielsweise bei der Fragestellung „Zyste vs. Tumor“ oder bei der Detektion von ödematösen Gewebe in der Umgebung eines Schlaganfalls notwendig sein kann (Thomalla et al., 2009). Darüber hinaus existieren auch Verfahren, die die Funktion ausgewählter Hirnregionen anhand ihrer Stoffwechselleistung darstellen können. Mittels diffusionsgewichteter Sequenzen ist es möglich, die Bahnen der weißen Substanz nachzuverfolgen oder mit sogenannten T2* Sequenzen Eisenablagerungen nach einem Blutungsereignis sichtbar zu machen. Mit T1- und T2-Sequenzen seien die zwei wichtigsten und grundlegendsten Wichtungen von MRT Bildern erklärt. Neben T1- und T2 Wichtungen gibt es noch eine Vielzahl anderer Möglichkeiten, MRT Aufnahmen anzufertigen. Da diese nicht Teil dieser Dissertation sind, werden sie nicht weiter erklärt.

Im Hinblick auf die Aufnahmetechnik unterscheidet man zweidimensionale (2D) von dreidimensionalen (3D) Sequenzen. Für 2D Aufnahmen werden nacheinander die einzelnen Schichten des Messobjekts mit schichtselektiven HF Impulsen angeregt und gemessen. Fügt man alle Einzelbilder hintereinander zusammen, entsteht das Gesamtbild im klassischen Sinne eines Schichtbildes. Die Dicke einer Bildschicht hängt von der Breite der HF Impulse ab und liegt oft zwischen 2mm und 4mm (Schick, 2006).

Bei 3D Aufnahmen hingegen wird ein gesamtes Volumen mittels HF Impuls angeregt und gemessen. Die Ortskodierung erfolgt über phasenverschiebende Gradienten, die die Dicke der Schichten im untersuchten Volumen bestimmen. Bei 3D Aufnahmen ist die Schichtdicke meist wesentlich geringer (z.B. 1mm Schichtdicke) als bei 2D Aufnahmen und die Bildschichten liegen einander lückenlos an (Schick, 2006). Dies bietet den Vorteil, dass 3D Aufnahmen eine höhere Auflösung aufweisen und sich ein aus allen Raumrichtungen betrachtbares Bild rekonstruieren lässt. 3D Sequenzen ermöglichen so die Detektion kleinster Läsionen in einem untersuchten Gewebe und stellen die Grundlage zur Volumenbestimmung von gemessenen Gewebetypen dar. Dies ist von entscheidender Bedeutung bei der Untersuchung von Gehirnen.

Für diese Arbeit wurde eine T1 3D MP RAGE Sequenz benutzt. 1990 von John P. Mugler und James R. Brookeman vorgestellt, leitet sich die Bezeichnung „MP RAGE“ von dem englischen „**m**agnetization-**p**repared **ra**pid **g**radient-**e**cho“ ab (Mugler & Brookeman, 1990). Charakteristisch für diese Sequenz sind eine Vormagnetisierung des Messobjekts, kurze TR Zeiten und kleine Auslenkwinkel der Spin-Drehachsen. Mit dem besonderen Schema der MP RAGE Aufnahme lassen sich gut kontrastierte Bilder erzeugen. Besonders zur Beurteilung von Gehirnen liefert diese Aufnahmetechnik eine gute Unterscheidbarkeit von Grauer und Weißer Substanz. Diese Doktorarbeit beschäftigt sich mit der automatisierten Auswertung ebendieser MP RAGE Aufnahme mit dem frei verfügbaren Programm FreeSurfer.

1.2 Automatische Hirnsegmentierung

Nicht nur das Erheben und klinische Befunden von MRT Bildern des Gehirns (cMRT) durch einen Neuroradiologen hat an Wichtigkeit gewonnen, sondern auch die quantitative algorithmusgestützte Analyse von Bildern. Die automatisierte Hirnsegmentierung ist ein Analyseansatz, bei dem Bildpunkte oder Bildregionen eines Bildes jeweils einer anatomischen Struktur zugeordnet werden („gelabelt“). Das „Label“ und die Anzahl der möglichen anatomischen Strukturen hängen dabei von dem verwendeten anatomischen Atlas ab. Im Weiteren wird der Begriff „Labels“ für ebendiese anatomischen Regionen benutzt. So lassen sich Informationen zu Struktur, Funktion, Konnektivität oder Volumen einzelner Hirnareale gewinnen. Eine manuelle Segmentierung einer vollständigen Hirnaufnahme, unter Betrachtung der Intensitätsstufen in einem T1 Bild und einem Abgleich mit Hirnatlanten, bedarf selbst für erfahrene Bildbetrachter äußerst viel Zeit (Tage) und sehr guter neuroanatomischer Kenntnisse (Destrieux, Fischl, Dale, & Halgren, 2010; Fischl, 2012). So wurde die Notwendigkeit von vollautomatischen Segmentierungsverfahren erkannt und durch zunehmende Rechenleistung von Computern auch möglich (Helms, 2016). Bereits 2000 gab es Ansätze zur technischen Umsetzung einer automatisierten Hirnsegmentierung (Ashburner & Friston, 2000). Weit verbreitet ist ein Verfahren, welches in mehreren Schritten

eine komplette Segmentierung des Gehirns liefert und in dem frei verfügbaren Programm „FreeSurfer“ zusammengefasst ist (Fischl, 2012; Fischl et al., 2002). FreeSurfer, wurde am Athinoula A. Martinos Center for Biomedical Imaging im Rahmen des “Human Connectome Project” entwickelt. Das Programm, wie man es in der heutigen Version verwenden kann, ist das Resultat der seit zwei Jahrzehnten andauernden Zusammenführung einer großen Anzahl von Projekten und unterliegt einer kontinuierlichen Weiterentwicklung (Fischl, 2012). Die Zuordnung der Labels basiert einerseits auf gewebsspezifischen Intensitätswerten der Voxel und andererseits auf der räumlichen Anordnung der Voxel im Abgleich mit zugrundeliegenden Hirnatlanten. Eine alleinige Zuordnung nach Intensitätswerten ist wegen der bereits erwähnten hohen Variabilität und der daraus folgenden großen Überschneidung von Werten unterschiedlicher Gewebetypen nicht möglich (Destrieux et al., 2010; Fischl, 2012; Fischl et al., 2002). Erst wenn einer Position im Raum ein anatomischer Wert zugeordnet werden kann, sinkt die Anzahl der für diese Position möglichen Labels und eine Differenzierung u.a. nach Intensität kann erfolgen (Destrieux et al., 2010; Fischl et al., 2002, 2004). In dem zu segmentierenden T1-gewichteten Bild werden zunächst alle Bestandteile, die nicht Gehirn abbilden, entfernt (Jenkinson, Pechaud, & Smith, 2005; Smith, 2002). Dieses Gehirn wird anschließend in einem virtuellen Raum zur Analyse standardisiert ausgerichtet. Dies ist von besonderer Bedeutung, da die exakte Ausrichtung dringend notwendig ist, um das Gehirn mit standardisierten Atlantenbildern in Übereinstimmung zu bringen (Ashburner & Friston, 2000). Nun werden den Voxeln anhand ihrer Position im Raum Vorwahrscheinlichkeiten für anatomische Labels zugeordnet. Diese Vorwahrscheinlichkeiten werden von Informationen aus Häufigkeitshistogrammen, wie oft einem Voxel an dieser Position ein bestimmtes Label zugeordnet wurde, ergänzt (Fischl et al., 2002). Darüber hinaus stehen manche neuroanatomische Strukturen in einer vorhersehbaren räumlichen Beziehung zueinander. Fischl nennt an dieser Stelle als Beispiel die Anordnung von Amygdala und Hippocampus. Die Amygdala wird immer anterior und superior zum Hippocampus zu finden sein. Solche speziellen Beziehungen werden, verschlüsselt in regionalen Markov Random Fields (MRF), in die Karte aus den bereits modifizierten Vorwahrscheinlichkeiten integriert und es entsteht, unter Einbeziehung der Intensitätswerte der Voxel, die finale Segmentierung (Fischl, 2012; Fischl et al., 2002). MRFs ermöglichen, dass sich die Segmentierung besser individuellen anatomischen Gegebenheiten anpassen kann und führt zu einer Kantenglättung der Segmente (Fischl et al., 2002).

Anfängliches Ziel dieses sehr großen Projektes war das Erstellen von Oberflächenmodellen zur Rekonstruktion des zerebralen Kortex mit den verfügbaren Labels „Graue Substanz“, „Weiße Substanz“ und „Zerebrospinalflüssigkeit“. Mittlerweile sind beispielsweise in der „wmparc“-Segmentierung 182 verschiedene Labels enthalten, die u.a. die subkortikale weiße Masse unterteilen. Der benannte Atlas ist eine Zusammenfassung von 3 unterschiedlichen

Atlanten und ihren manuellen Segmentierungsprojekten, die ihrerseits Vorschläge zur Definition der anatomischen Grenzen von Strukturen liefen. Gerade die genaue Definition von anatomischen Strukturen (Beispiel: Grenze zwischen Gyrus und Sulcus) ist für manuelle Hirnsegmentierungen von großer Bedeutung, da diese manuelle Vorarbeit die Grundlage für die Algorithmus-basierte Segmentierung darstellt. Nur mit klaren anatomischen Definitionen sind manuelle Ergebnisse verschiedener Individuen vergleichbar und reproduzierbar. Desikan et al. (2006) liefern einen an 40 Probanden erprobten Atlas mit 34 kortikalen Labels (Desikan et al., 2006). In FreeSurfer ist dieser als *Desikan-Killiany-Atlas* bekannt. Er deckt eine Altersspanne von 19-87 Jahren ab (Desikan et al., 2006). Darüber hinaus sind in dem Atlas auch Daten von Gehirnen von Alzheimer-Patienten vertreten (n=10), um auch Aussagen über Gehirne, die von Atrophie betroffen sind, zuzulassen.

Der zweite zur „wmparc"-Segmentierung gehörende Atlas ist der *Destrieux-Atlas* (Destrieux et al., 2010; Fischl et al., 2004). Hier liegt eine vom *Desikan-Killiany-Atlas* abweichende Definition der Gyrus-Sulcus-Grenze vor und die Kohorte besteht ausschließlich aus jungen und gesunden Probanden (n=24, 18-33 Jahre) (Destrieux et al., 2010). 2009 wurden am *Destrieux-Atlas* Anpassungen vorgenommen, die unter anderem eine genauere Unterteilung des bis dahin definierten Labels „Corpus_Callosum" in Subgruppen (anterior, mid-anterior, central, mid-posterior, posterior) beinhaltet (Vogt, Berger, & Derbyshire, 2003; Vogt, Vogt, & Laureys, 2006).

Der dritte in der „wmparc"-Segmentierung enthaltene Atlas stellt eine Modifikation des *Desikan-Killiany-Atlasses* dar, der Anpassungen enthält, um die „Definition der [anatomischen] Regionen so konstant und so eindeutig wie möglich" zu machen und um die anatomischen Grenzen verlässlich an den automatischen FreeSurfer Algorithmus anzupassen (Klein & Tourville, 2012). Das daraus entstandene *Desikan-Killiany-Tourville* Protokoll wurde an 40 MRT Datensätzen angewandt, die aus einer frei verfügbaren Hirn MRT- /manuellen Hirnsegmentierungsdatenbank, bekannt als *Mindboggle-101*, bezogen werden konnten (Klein & Tourville, 2012). In FreeSurfer wird für diesen Atlas die Abkürzung *DKT40* benutzt und es sei angemerkt, dass 31 der 40 Probanden in ihren Zwanzigern, 4 in ihren Dreißigern und eine Person in den Sechzigern waren (Klein & Tourville, 2012). Die genannten Atlanten sind im FreeSurfer Programm enthalten und können mit dem „recon-all" Skript auf zu segmentierende T1 MRT Aufnahmen angewendet werden.

Auf dieser Grundlage bietet FreeSurfer eine einfache Möglichkeit zur volumetrischen Analyse von MRT Bildern. Da sich mittels der oben beschriebenen MRT Aufnahmetechniken zur Raumorientierung das Volumen eines einzelnen Voxel festlegen lässt, kann man durch Aufsummieren aller Voxel desselben Labels, unter Berücksichtigung derer Einzelvolumina, einen Rückschluss auf das Volumen der entsprechenden anatomischen Struktur ziehen.

1.3 Klinische Relevanz der automatischen Hirnsegmentierung

Besonders die Behandlung, Kontrolle und klinische Erforschung von Krankheitsbildern aus der Neurologie ist auf die Verbindung zur Radiologie angewiesen. Auch wenn die Betrachtung und Befundung von cMRTs durch einen Neuroradiologen ein stark standardisiertes Vorgehen aufweist, bleibt sie dennoch ein subjektives Verfahren, sodass gerade hier das Tool der Algorithmus-basierten Volumetrie zusätzliche und quantifizierbare Informationen liefern kann. Die neuroradiologische Betrachtung von Volumina beschäftigt sich vor allem mit der sichtbaren Volumenabnahme von Geweben (Atrophie) und deren Verteilungsmuster. Die Beobachtung, Erfassung und Interpretation von eventuell dazu passenden neurokognitiven, motorischen oder psychischen Symptomen leistet in ihrer Gesamtheit die Neurologie. Somit wird die Erforschung und klinische Einordnung von Atrophien gemeinsame Schnittstelle von Radiologie und Neurologie, wobei je nach betrachtetem Krankheitsbild offen bleibt, ob es sich bei einer Atrophie um dessen Ursache oder Folge handeln könnte. Anhand der Multiplen Sklerose (MS), Morbus Alzheimer/Alzheimer Demenz (AD) oder anderen dementiellen Erkrankungen soll deutlich werden, wie wichtig dieser interdisziplinäre Schulterschluss in der modernen Medizin und Forschung ist. Bei der AD stellt sich der Neurologie das Problem, dass Frühstadien eine Vielzahl von leichten neurologischen Symptomen aufweisen können und die Krankheit daher oft lange Zeit unerkannt bleibt (Masters et al., 2015). Symptome, die der AD vorausgehen charakterisieren sich in einer anfänglichen milden kognitiven Einschränkung (eng. *mild cognitve impairment* = MCI). Da sich bestimmte Atrophiemuster gut spezifischen neurologischen Symptomen zuordnen lassen, hat die Neuroradiologie ihren festen Bestandteil bei der AD-Diagnostik (Frisoni, Fox, Jack, Scheltens, & Thompson, 2010). So lässt sich anhand der Atrophie des entorhinalen Cortex und des Hippocampus ein Progress vom Stadium des MCI hin zur AD vorhersagen, da sich die Atrophieraten dieser Strukturen ca. 5,5 Jahre vor der AD Diagnose, sowie mitunter kurz vor dem Eintritt des Patienten in das MCI Stadium, von den Atrophieraten der Normalbevölkerung unterscheiden, und diese mit der Konzentration von pathognomonischen tau-Proteinen im Hirnwasser korrelieren. (Frisoni et al., 2010). Darüber hinaus ist die Bildgebung von essenzieller Bedeutung die AD differenzialdiagnostisch beispielsweise von einer vaskulären Demenz zu unterscheiden (Frisoni et al., 2010). Die Grenzen dieser Früherkennung liegen laut Frisoni im Design herkömmlicher AD MRT Studien. Diese schließen in der Regel Probanden mit zerebrovaskulären Erkrankungen und Ereignissen aus, sodass es „zur Beurteilung der potentiell unabhängigen oder synergistischen Beteiligung von neurodegenerativem AD und zerebrovaskulären Veränderungen am MCI weiterer Studien zur Untersuchung der grauen und weißen Substanz an einer repräsentativen Kohorte bedürfe“ (Frisoni et al., 2010).

Um die Erfassung, Quantifizierung und zentrumsübergreifende Vergleichbarkeit von Atrophie bei MS Patienten zu ermöglichen, empfehlen Marciniewicz und Kollegen (Marciniewicz, Bladowska, Podgórski, & Sąsiadek, 2019), standardisierte MRT Protokolle zu etablieren und neue Verfahren zu validieren.

So lässt sich Marciniewiczs Aufruf an die MS-Bildgebung, zusammen mit der von Frisoni beschriebenen Notwendigkeit nach Forschung an repräsentativen Studien, auf die Erforschung von vielen größeren geriatrischen Kohorten übertragen. Sollten sich für gerade diese Kohorten, welche MRT Diagnostik ohnehin oft in Anspruch nehmen, standardisierte MRT Datensätze in Form eines „Geriatrie-Protokolls“ in den klinischen Alltag etablieren, ließe sich innerhalb kürzester Zeit, im Sinne der explorativen klinischen Forschung, eine Datenbank von sofort auswertbaren und vergleichbaren Informationen erschaffen.

Damit wird klar, wie wichtig eine valide und reliable automatisierte Analyse von zerebralen MRT Bildern von multimorbiden Patienten mit neurologischen Erkrankungen ist. Wie aber bereits erwähnt, wurden bei der Entwicklung und Testung des oben beschrieben FreeSurfer Algorithmus Hirnatlanten benutzt, die zum überwiegenden Anteil an jungen und hirngesunden Probanden entwickelt worden sind (Desikan et al., 2006; Destrieux et al., 2010; Fischl et al., 2004; Klein & Tourville, 2012). Nur in einer der 3 Studien wurden Probanden, die an einer AD erkrankt waren, eingeschlossen, aber solche, die einen zerebralen Infarkt erlitten hatten oder bei denen ein Hirntumorleiden bekannt war, ausgeschlossen (Desikan et al., 2006).

Trotz das FreeSurfer an einer Vielzahl von relativ hirngesunden Menschen entwickelt wurde, ist die Anwendung auf Bilder von Patienten mit Schlaganfall dennoch möglich, wie beispielsweise volumetrische Untersuchungen des Thalamus und Hippocampus nach Schlaganfall zeigen (Brodtmann et al., 2012). Allerdings ist zu beachten, dass mit zunehmendem Strukturschaden oder sogar gänzlichem Verlust von Hirnstruktur die Kompensationsmechanismen der FreeSurfer Segmentierungssoftware schnell ausgeschöpft sind, welches sich in erheblichen Segmentierungsfehlern niederschlägt (Gajawelli et al., 2011). Schon alleine wegen der Häufigkeit von zerebrovaskulären Läsionen in der älteren Bevölkerung würde die automatisierte Evaluation von MRT Bildern von Patienten mit derartigen Läsionen ein enormer Fortschritt in der objektiven quantiativen MRT Analyse darstellen. Laut den Ergebnissen der Studie zur Gesundheit Erwachsener in Deutschland (DEGS1) aus 2013 liegt die Prävalenz von Schlaganfällen in der Altersgruppe 60–69 Jahre für Frauen bei 3,1% und für Männer bei 5,4% (Busch, Schienkiewitz, Nowossadeck, & Gößwald, 2013). In der Altersgruppe 70-79 Jahre sind es für Frauen 6,3% und für Männer sogar 8,3% (Busch et al., 2013).

1.4 System zur Bewertung von automatischen MRT Segmentierungen bei multimorbiden Patienten mit neurologischen Erkrankungen

Um bei multimorbiden Patienten mit neurologischen Erkrankungen trotz starker Schwankungen der Segmentierungsqualität Volumenanalysen durchführen zu können, muss eine systematische Einordung der Segmentierungsqualität durchgeführt werden. Dies wurde in dieser Doktorarbeit durchgeführt und wird im Folgenden detailliert vorgestellt. Diese systematische Einordnung basiert im ersten Schritt auf der deskriptiven Erfassung der Fehler, die innerhalb einer Segmentierung entstehen. In einem zweiten Schritt wird die Schwere der Fehler (leicht, mittel, schwer) anhand festgelegter Regeln definiert. Aus der Summe und Kombination der verschiedenen Fehlerstufen ergibt sich die Gesamtbewertung einer Segmentierung, wiederum dreistufig in GUT, MITTEL und SCHLECHT. Die Idee der systematischen Einordnung wurde aus der Literatur entnommen (siehe z.B. (Reuter et al., 2015; Rosen et al., 2018)). Die Segmentierungen der Kategorie GUT soll keine größeren Fehler aufweisen und somit definitiv in spätere Volumenanalysen eingeschlossen werden können. Die Kategorie SCHLECHT umfasst solche Segmentierungsergebnisse, die für spätere Analysen inakzeptable Fehler und Volumenabweichungen einzelner Labels aufweisen. Die Kategorie MITTEL weist eine mäßige Fehlerlast auf, die die Segmentierung in ihrer Integrität nicht global stört und daher optional für weitere Analysen verwendet werden kann. Auch 5-stufige systematischen Einordnungen sind in der Literatur beschrieben, allerdings konnten bereits Rosen et al. (2018) zeigen, dass diese im Vergleich zu 3-stufigen Einordnungen zu einer Reduktion der Interrater-Reliabilität führen.

Aus diesen Aspekten ergeben sich folgende Hypothesen und Überlegungen, die dann weiter bearbeitet werden.

1.5 Hypothesen und wichtigste Arbeitsschritte

1) Das hier verwendete dreistufige Bewertungssystem für die Beurteilung der Segmentierungsqualität T1-gewichteter cMRT-Bilder bei multimorbiden Patienten mit neurologischen Erkrankungen weist eine akzeptable Interrater-Reliabilität (Kappa) auf. Die Interpretation des Kappa orientiert sich an den von Landis und Koch (1977) vorgestellten Abstufungen.

2) Die Qualität der Segmentierung von FreeSurfer beeinflusst das Volumen der Hirnlappen (hier: Frontal- Temporal-, Parietal- und Okzipitallappen, Basalganglien und Kleinhirn) nicht systematisch. Dazu wird das Volumen der Hirnlappen zwischen den drei Qualitätsstufen verglichen. Da eine frühere Studie keinen systematischen Einfluss der Fehler auf das Volumen der Areale gefunden hat (McCarthy et al., 2015), gehen wir davon aus, dass sich auch bei unserer Population keine systematischen Volumenfehler finden.

3) Bei schlechterer Qualität der Segmentierung zeigt sich eine größere Streuung des Volumens. Um diesen Zusammenhang zu untersuchen wird die Homogenität der Varianz für die drei Qualitätsstufen überprüft. Fehler in der Segmentierung stellen sich als zu viel oder zu wenig von FreeSurfer markierte anatomischen Strukturen dar, die auf Pathologien oder technische Fehler zurückzuführen sind und es wird vermutet, dass die Variabilität der Volumina segmentierter Areale bei schlechterer Qualität höher ist, da dabei entweder größere Gebiete ausgelassen oder fälschlicherweise hinzugenommen werden. Dies erfolgt auf Ebene der sechs bereits oben beschriebenen Hirnlappen.

Eine explorative Untersuchung der zehn am häufigsten von Fehlern betroffenen FreeSurfer Segmente soll sich anschließen. Diese Subregionen werden auf Volumenunterschiede zwischen den drei Qualitätsstufen überprüft und die Verteilung der Varianz genauer betrachtet.

2. Methoden

2.1 Studienbeschreibung

Die Daten für die vorliegende Dissertation wurden im Rahmen der ComOn-Studie (Cognitive and Motor Interactions in the Older Population) erhoben (Geritz et al., 2020). Es liegt ein positives Votum der Ethikkommission vor (AZ D427/17). Ziel dieser Multicenter-Studie ist die quantitative Evaluation von Kognition und Motorik bei 1000 geriatrischen Patienten. Im Rahmen eines umfassenden Assessments werden dabei sowohl standardisierte neuropsychologischer Testverfahren für die Erfassung kognitiver Aspekte eingesetzt, als auch sensorbasierte Gang- und Gleichgewichtsmessungen. Außerdem wird eine große Anzahl von Fragebögen eingesetzt, die für den geriatrischen Patienten relevante Thematiken abdeckt. Teilnehmende Studienzentren finden sich in Deutschland, Italien, Portugal, Österreich und Brasilien. Die Konzeption und Durchführung der Studie erfolgt im Sinne der Internationalen Klassifikation der Funktionsfähigkeit, Behinderung und Gesundheit (ICF-Modell) der Weltgesundheitsorganisation (Ewert & Stucki, 2007). Im Rahmen des ICF-Modells werden Patientinnen und Patienten unter biologischen, psychologischen und sozialen Aspekten betrachtet. Um diesem Anspruch Rechnung zu tragen und insbesondere die alltagsrelevanten Fähigkeiten des geriatrischen Patientenklientels zu erfassen (Maetzler, Grond, & Jacobs, 2016) werden in der ComOn-Studie ein multiprofessionelles Team und verschiedene Messmethoden einbezogen. Für die Dissertation wurden Daten von Patienten berücksichtigt, die neben der Mobilitäts- und kognitiven Untersuchung auch ein cMRT erhalten haben.

2.2 Studienteilnehmer

Die ComOn Studie richtet sich an geriatrische Patientinnen und Patienten (Alter >70 Jahre), mit und ohne chronische neurologische Erkrankung. Des Weiteren können auch jüngere Patientinnen und Patienten teilnehmen, sofern sie das Kriterium der Multimorbidität erfüllen. Dabei werden ausschließlich Patientinnen und Patienten rekrutiert, die in der Lage sind, mindestens 10 Sekunden selbstständig zu stehen und mindestens 3 Meter selbstständig zu gehen (Hilfsmittel wie z.B. Gehstöcke sind erlaubt). Ausschlusskriterien für die Studienteilnahme sind Bewusstseinsstörungen (klinisch beurteilt), mehr als 2 Stürze in der vorherigen Woche, 5 oder weniger Punkte im MoCa (Montreal Cognitive Assessment) Test, Medikamenten- oder Drogenmissbrauch (außer Nikotin) und eine Sehschärfe von weniger als 60%. Für die hier relevante cMRT-Messung galten des Weiteren folgende Ausschlusskriterien: Herzschrittmacher, Defibrillatoren, Medikamentenpumpen, DBS-Systeme, Metallsplitter oder Gefäßclips aus ferromagnetischem Material, Temporäre Cava-Filter, Cochlea-Implantate oder Klaustrophobie.

Die Daten für die vorliegende Dissertation wurden am Universitätsklinikum Schleswig-Holstein, Campus Kiel, im Rahmen stationärer Aufenthalte von Patientinnen und Patienten auf der Neurogeriatrischen Station A2, der Inneren geriatrischen Station A2, sowie den neurologischen Allgemeinstationen N2 und N3 erhoben. Für diese Arbeit erhielten 53 freiwillige Probanden, im Rahmen ihres stationären Aufenthaltes, eine MRT. Das durchschnittliche Alter lag bei 78.3 Jahren (Standardabweichung: 6.2 Jahre). 25 von 53 Probanden waren weiblich (47.2%). Da es Ziel dieser Arbeit war, auch die Brauchbarkeit von MRT Aufnahmen mit schlechter Qualität zu bewerten, wurden alle diese MRT Aufnahmen für die Studie verwendet (8 als Trainingsdatensatz, 45 als Analysedatensatz). Für weitere Informationen siehe Tabelle 6.

2.3 Bildgebende Verfahren in der ComOn-Studie

Die Kopf-MRT-Datensätze der ComOn Studie werden in der Klinik für Neuroradiologie im UKSH, Campus Kiel erhoben. Sie dienen sowohl Forschungszwecken, als auch der Befundung durch einen Neuroradiologen bezüglich klinischer Fragestellungen.

2.3.1 Scanner

Es wurde ein „Achieva 3T TX“ (Philips Healthcare, Best, Niederlande) MRT Gerät mit einer magnetischen Flussdichte von B0 = 3Tesla und eine 32-Kanal-Kopfspule verwendet. Betrieben wurde es mit den Software Versionen 5.3, später 5.4.

2.3.2 MRT Protokoll

Das spezifisch für diese Untersuchung zusammengestellte „Kieler Geriatrie-Protokoll“ beinhaltet sechs unterschiedliche Sequenzen. Planungsbilder (sogenannte Localizer) dienen zunächst der groben Orientierung des Kopfes des Probanden im Raum. Anhand dieser sehr kurzen Sequenz wird die Raumausrichtung der anschließenden Sequenzen geplant. Ein „field mapping“ bestimmt vorab die Verzerrung des Magnetfeldes, um magnetfeldbedingte Artefakte zu vermeiden.
Die 3D T1 MP RAGE Sequenz orientiert sich am ADNI-Protokoll Version 2.6 Philips. Sie nutzt einen Inversion-Vorpuls mit der Inversionzeit TI=900ms, eine Repetitionszeit TR=6.7ms und einer Echozeit TE=3.1ms mit einem Flipwinkel von 9°. Es werden 170 transversale Schichten mit einer Matrix von 244 x 230 und einer Auflösung von 1.1x1.1 x 1.2mm^3 und SENSE Faktor 1.8 in 5:34 Minuten aufgenommen. Die Bilder werden auf eine Matrix von 256 x 256 rekonstruiert, so dass die resultierenden Bilder eine Auflösung von 1.05 x 1.05 x 1,2mm^3 haben.

Es handelt sich hierbei um eine Aufnahme, die die Basis für eine vollautomatische Segmentierung darstellt und den für FreeSurfer empfohlenen Einstellungen folgt (FreeSurferWiki, 2009).

Außerdem sind in dem Protokoll noch eine T2 (DRIVE), T2 (FLAIR), T2* und DTI 32R iso Sequenz enthalten, die jedoch in dieser Arbeit nicht verwendet werden.

Tabelle 1 fasst die Sequenzen des Protokolls zusammen.

Tabelle 1: Sequenzen des Kieler MRT Geriatrie-Protokolls

Name	**Schichten**	**Voxel Maße (mm)**	**Matrix**	**TR (ms)**	**TE (ms)**
T1 MPRAGE	170	1.05 x 1.05 x 1.2	256 x 256	6.64	3.09
T2 DRIVE 2mm	57	0.43 x 0.43 x 2.0	512 x 512	5501.9	80.0
T2 FLAIR 2mm	60	0.43 x 0.43 x 2.0	528 x 528	12000.0	160.0
T2*	38	0.73 x 0.73 x 4.0	288 × 288	661	20
DTI 32R iso	60 Schichten x 34 Volumen	1.75 x 1.75 x 2.0	128 x 128	6354.76	74.5

2.3.3 FreeSurfer Segmentierung

Die Segmentierung der T1 gewichteten Bilder wurde mit dem frei verfügbaren Programm FreeSurfer in der Version Stable v6.0 („freesurfer-x86_64-unknown-linux-gnu-stable6-20170118“) durchgeführt. Benutzt wurde das „recon-all“ Skript in den Standarteinstellungen, um eine Segmentierung nach dem oben beschriebenen „wmparc“ Atlas zu generieren (Desikan et al., 2006; Destrieux et al., 2010; Fischl et al., 2004; Klein & Tourville, 2012). Die Aufteilung der Regionen zu Hirnlappen ist unter 2.5 (Volumenanalyse) genauer beschrieben.

2.4 Bewertung einer Segmentierung

Zur visuellen Kontrolle der Segmentierungsergebnisse wurde das frei verfügbare Programm ITK-SNAP (http://www.itksnap.org/pmwiki/pmwiki.php?n=Downloads.SNAP3) in der Version 3.6.0 verwendet (Yushkevich et al., 2006). Das im Rahmen dieser Doktorarbeit ausgearbeitete Vorgehen zur Bewertung erfolgt in 3 Schritten:

Bestimmung der Fehlerart ⟶ Bewertung des Fehlers ⟶ Bewertung der Segmentierung

Nach diesen Schritten untergliedern sich die folgenden Abschnitte. Die Tabellen und Abbildungen dieses Abschnittes sind dem Tutorial entnommen, welches vom Doktoranden selbständig und in Rücksprache mit einem erfahrenen Neuroradiologen (PD Dr. Christian Riedel) und einem Bildverarbeitungsexperten (Oliver Granert) erstellt wurde und einem potentiellen Rater diese Methode anhand von Beispielen erklärt.

Tabelle 2: Mögliche Fehlerarten und ihre Beschreibung

Symbol	Bedeutung	Erklärung
+	leicht zu viel	über Segmentgrenze minimal (eine Schicht) hinausgehend, andere Strukturen sind diesem Label nicht fälschlicherweise zugeordnet
++	moderat zu viel	über Segmentgrenze deutlich (mehr als eine Schicht, aber max. +1/3 des Gesamtvolumens der betroffenen Struktur) hinaus, fälschlicherweise ist kleinen Teilen anderer Strukturen dieses Label zugeordnet.
+++	extrem zu viel	über Segmentgrenze deutlich (mehr als die Hälfte des Gesamtvolumens der betroffenen Struktur) hinaus, fälschlicherweise ist großen Teilen anderer Strukturen dieses Label zugeordnet
-	leicht zu wenig	Segment an den Rändern minimal (eine Schicht) nicht erfasst
--	moderat zu wenig	Segment um mehr als eine Schicht bis max. weniger als1/3 des Gesamtvolumens der betroffenen Struktur zu wenig erfasst

---	extrem zu wenig	Segment um weniger als die Hälfte des Gesamtvolumens der betroffenen Struktur oder kaum erfasst
*	Raumforderung	Raumforderung/Läsion liegt inmitten eines Segments und ist in allen Raumrichtungen fälschlicherweise als dieses markiert

2.4.1 Schritt 1: Bestimmung der Fehlerart

Die Beschaffenheit eines Fehlers innerhalb einer Segmentierung kann nur wenige Varianten annehmen. Entweder wird eine anatomische Struktur zu groß oder zu klein einer Atlasregion (weiter im Tutorial als „Label“ bezeichnet) zugeordnet oder inmitten einer solchen Zuordnung liegt eine anatomische Aussparung, beispielsweise ein Gefäß oder ein pathologischer Substanzverlust, die dem entsprechenden Label nicht hätte zugeordnet werden dürfen. Diese Abweichungen können entlang einer Richtung im Raum verschiedene Ausmaße annehmen. Im hier vorgestellten Bewertungssystem werden 3 Schweregrade unterschieden. Tabelle 2 zeigt das detaillierte Regelwerk nachdem die Fehlerart bestimmt und in einer Tabelle, in der alle Labels des Probanden gelistet sind, notiert wird.

Tabelle 3: Anatomische Raumrichtungen

fro	frontal	nach vorne
occ	occipital	nach hinten
cra	cranial	nach oben
cau	caudal	nach unten
lat	lateral	nach außen
med	medial	nach innen

In ebendieser Tabelle wird neben der Fehlerart auch die Fehlerrichtung im Raum angegeben, um die fragliche Stelle später einfacher wiederfinden zu können. Diese Notiz folgt den in Tabelle 3 angeführten Abkürzungen. Sie orientieren sich an den herkömmlichen anatomischen Raumrichtungen, wie sie, in ähnlicher Form, z.B. bei (Schünke, Schulte, Schumacher, Voll, & Wesker, 2018) gefunden werden kann.

2.4.2 Schritt 2: Bewertung des Fehlers

Die in Schritt 1 erkannten Fehler werden einer von 3 Kategorien zugeordnet. Hierfür gibt es wiederum feste Regeln. Markiert wird der Fehler in der oben benannten Tabelle mit den Farben Grün (leichter Fehler), Gelb (mittlerer Fehler) oder Rot (schwerer Fehler). Während

im Tutorial und in den Bewertungstabellen ausschließlich mit farblichen Markierungen gearbeitet wird, verzichtet diese Dissertation auf die farblichen Markierungen im Text und ersetzt sie, zur besseren Übersicht im gedruckten Format, mit typographischen Hervorhebungen.

a) Leichter Fehler (*grün = kursiv)*: Diese Fehler haben keinen sichtbar relevanten Einfluss auf die Segmentierung. Die Volumenabweichungen der entsprechenden Labels sind augenscheinlich marginal. Dem Fehler liegt offensichtlich keine Pathologie zu Grunde.
b) Mittlerer Fehler (*gelb = kursiv + unterstrichen*): Für diese Kategorie liegt eine mit dem bloßen Auge sichtbare deutliche Volumenabweichung des Labels vor. Als Ursache für diesen Segmentierungsfehler lässt sich eine Pathologie nicht ausschließen.
c) Schwere Fehler (***rot = kursiv + unterstrichen + fett***): Hier ist die Segmentierung des Labels komplex gestört und die fehlerhafte Stelle betrifft angrenzende Labels deutlich. Dem Fehler kann darüber hinaus eine offensichtliche Pathologie zu Grunde liegen

Tabelle 4 fasst Schritt 2 nochmal zusammen.

Tabelle 4: Schema zur Bewertung eins Fehlers

Farbe	Bedeutung	Erklärung
grün	leichter Fehler	wenig relevant, offensichtlich keine Pathologie
gelb	mittlerer Fehler	deutliche Volumenabweichung, Pathologie nicht ausgeschlossen
rot	schwerer Fehler	Segmentierung komplex gestört, weitere Segmente mitbetroffen, offensichtliche Pathologie

2.4.3 Schritt 3: Bewertung der Segmentierung

Dieser Schritt führt zur finalen Gesamtbewertung der Segmentierung. Auch hier existieren 3 Kategorien. An dieser Stelle soll anhand von Beispielen, die sich in ähnlicher Form auch im oben bereits erwähnten Tutorial wiederfinden, deutlich werden, nach welchen Regeln eine Gesamtbewertung durchgeführt wurde. Die verwendeten Kategorien sind GUT, MITTEL und SCHLECHT.

a) GUT

Die Bewertung GUT wird vergeben, wenn sich keine Fehler in einer Segmentierung ergeben. Außerdem wird GUT vergeben, wenn überwiegend *leichte* Fehler vorliegen, aber keine ***schweren***. Dies soll an 2 Beispielen verdeutlicht werden.

Beispiel 1 (GUT)

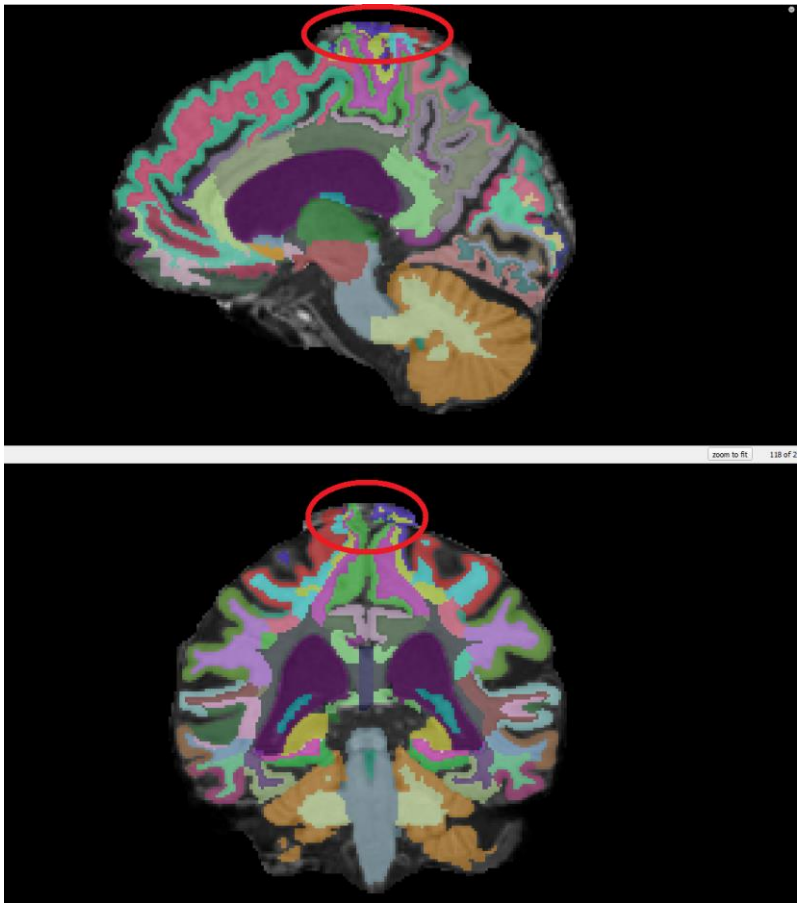


Abbildung 1: Sagittal- und Koronarschnitt der Segmentierung von COKI10030. Rot markiert ist ein fehlerhaftes Areal

Insgesamt liegt hier ein recht gutes Segmentierungsergebnis vor. Zu beanstanden ist, dass die Dura nicht ganz entfernt wurde. Solche Fehler am Hirnrand können durch eine fehlerhafte „brain extraction“ (vgl. Punkt 5.5 der FreeSurfer Pipeline) hervorgerufen werden. Fälschlicherweise wurden einige Labels zu weit nach kranial in die Dura gezogen.

zu Schritt 1

Elf der hier vorliegenden 12 „*leichten Fehler*“ befinden sich im rot markierten Bereich. Die Segmentgrenze wurde im Vergleich zum Gesamtsegment nur wenig nach kranial verletzt.

Daher werden die meisten dieser Fehler in der Tabelle zur Dokumentation mit „cra +“ markiert.

zu Schritt 2

Für die Integrität der Gesamtsegmentierung sind diese Fehler wenig relevant. Eine Pathologie liegt in diesem Bereich offensichtlich nicht vor (daher „*leichte Fehler*“)

zu Schritt 3

Überwiegend leichte Fehler, bei Abwesenheit von schweren Fehlern führt zur Gesamtwertung GUT.

Beispiel 2 (GUT)

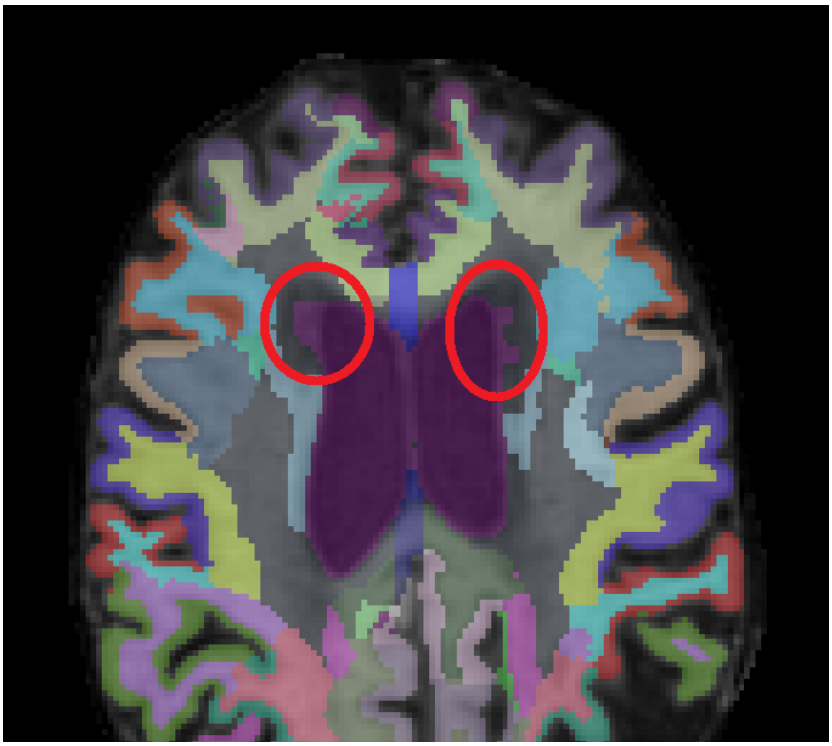


Abbildung 2: Axialschnitt der Segmentierung von COKO10067. Rot markiert sind fehlerhafte Areale

Auch hier liegt eine gelungene Segmentierung vor, allerdings ist die Dura nicht ganz entfernt.

Schritt 1

Im axial betrachteten Bild sind die Label der Ventrikel zu weit nach lateral gezogen (rote Markierungen). In der Dokumentationstabelle notiert man hierfür „lat ++“

Schritt 2

Die vorliegende Volumenabweichung der Seitenventrikel hat auf ihr Gesamtvolumen nur einen kleinen Einfluss. Vor allem ist die Segmentierung hier wahrscheinlich durch eine mikroangiopathische Marklagerschädigung beeinflusst. „Pathologie nicht ausgeschlossen“ führt zur Kategorie „*mittlerer Fehler*“.

Schritt 3

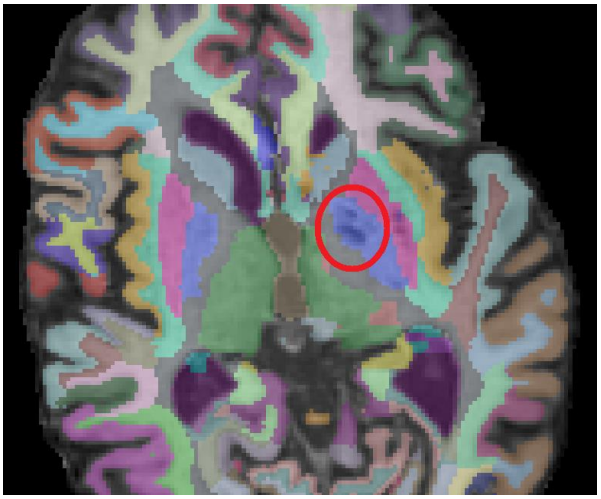
Den 2 *mittleren* Fehlern an den Seitenventrikeln stehen 5 *leichte* gegenüber. Außerdem gibt es keine ***schweren*** Fehler. Daher liegt hier die Gesamtwertung GUT vor.

b) MITTEL

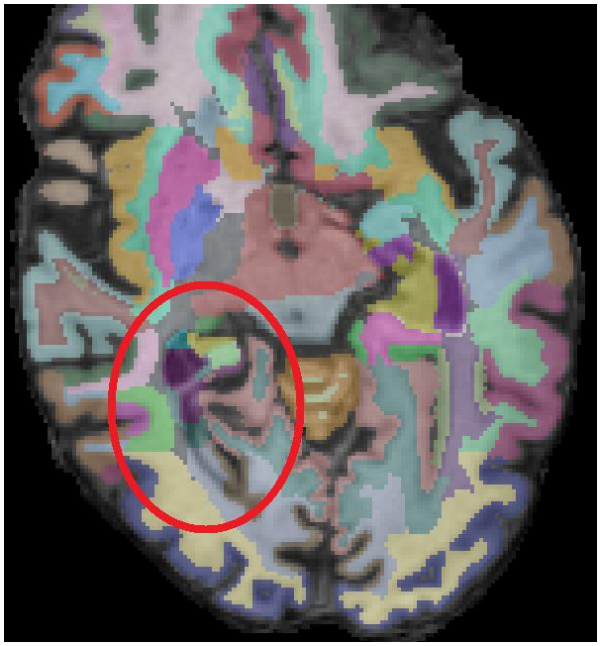
FreeSurfer gestützte Segmentierungen einer multimorbiden, neurologisch oft schwer vorgeschädigten geriatrischen Kohorte ergeben einen breiten Kontrast zwischen den Extremen der möglichen Segmentierungsqualität. So liegt es oft auf der Hand, ein Ergebnis der Kategorie GUT oder SCHLECHT zuzuordnen. Die Kategorie MITTEL soll gerade nicht als Sammelsurium solcher Ergebnisse dienen, die sich den anderen beiden Kategorien nicht zuordnen lassen, sondern stellt eine eigene Rubrik da, die für die weitere Datenverarbeitung ebenfalls akzeptable Segmenteirungen beinhaltet, deren Einschluss allerdings mit einem moderaten Verlust an Qualität einhergeht. Wichtiges Kriterium für die Kategorie MITTEL ist, dass überwiegend ***mittlere*** Fehler vorliegen. ***Mittlere*** Fehler sind durch eine augenscheinlich deutliche Volumenabweichung des Labels charakterisiert, und/oder dadurch, dass im Bereich des Fehlers eventuell eine Pathologie vorliegen könnte, die die Segmentierung irritiert. Falls es darüber hinaus auch zu ***schweren*** Fehlern kommt, dürfen diese die Integrität der Gesamtsegmentierung nur lokal stören. Angeführt sind 3 Beispiele:

Beispiel 1 (MITTEL)

Zunächst fällt auf, dass der Proband etwas schräg im Scanner lag.



Auf diesem Bild ist ein Fehler der Art „*“ zu sehen. Es könnte sich natürlich um ein Gefäß handeln, allerdings gäbe es für Gefäße ein eigenes Label. Auch eine Läsion kommt in Frage. In jedem Fall hätte hier das Label „Left.Pallidum“ ausgespart werden sollen. „*“ Fehler sind entweder als *mittel* oder ***schwer*** zu bewerten. In diesem Fall wurden sie als *mittel* bewertet, da nicht offensichtlich eine Pathologie vorliegt, sich aber auch keine ausschließen lässt.



In der Markierung auf Abbildung 3 lassen sich weitere *mittlere* Fehler erkennen. Der

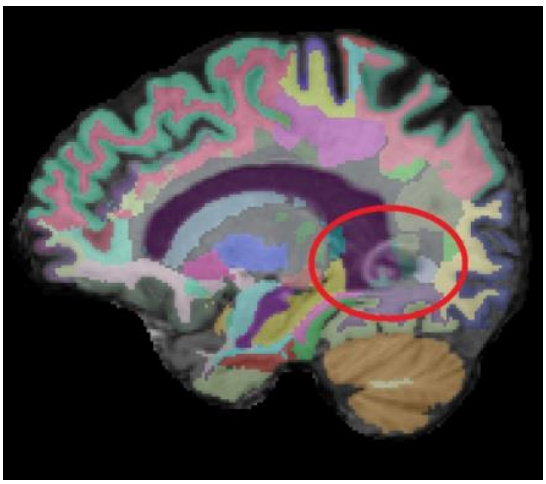


Abbildung 3: Zwei Axialschnitte und ein Sagittalschnitt der Segmentierung von COKI10080. Rot markiert sind fehlerhafte Areale

Ventrikel ragt zu weit nach occipital, sodass andere Labels (ctx.rh.lingual, wm.rh.lingual, wm.rh.pericalcarine) ebenfalls fehlerhaft werden.

Bei der Segmentierung dieses Probanden ergaben sich des Weiteren 4 ***schwere*** Fehler in den Labels „ctx.lh.postcentral“ (occ -), „ctx.lh.supramarginal“ (cra --), „wm.lh.postcentral (occ -), „wm.lh.supramarginal“ (cra --).

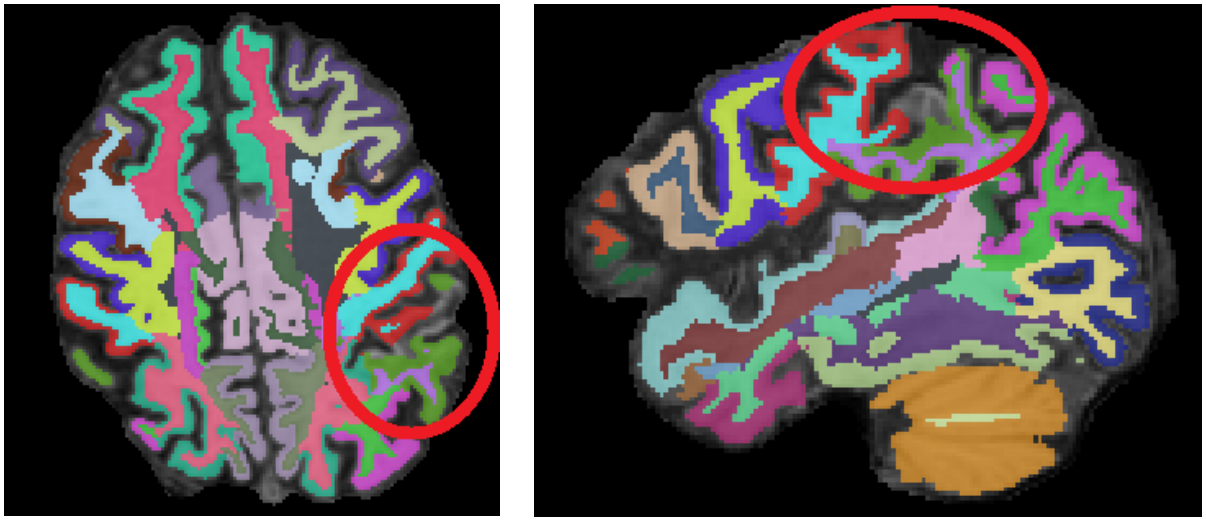


Abbildung 4: Axial- und Sagittalschnitt der Segmentierung von COKI10080. Rot markiert sind fehlerhafte Areale

Betrachtet man hier nur die Grundbilder (Abbildung 5), so fällt eine Infarktnarbe im oben markierten Bereich auf. Da die Pathologie offensichtlich und daher die Segmentierung in diesem Bereich komplex gestört ist, sind die Fehler als ***schwer*** einzustufen.

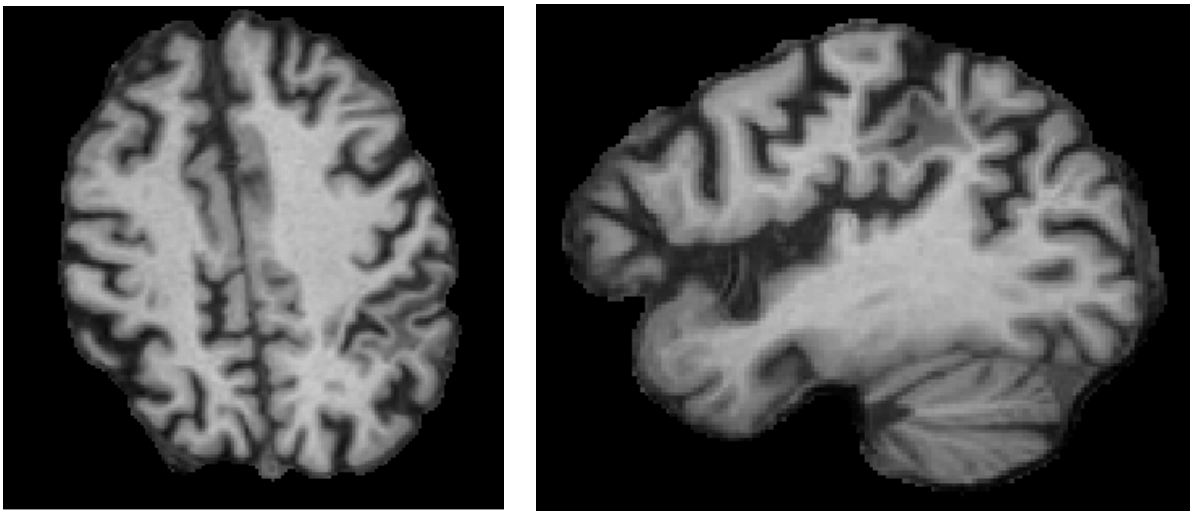


Abbildung 5: Axial- und Sagittalschnitt des T1 Bildes von COKI10080

Insgesamt überwiegen bei COKI10080 die *mittleren* Fehler und die 4 ***schweren*** Fehler stören das Gesamtbild der Segmentierung nur lokal. Damit sind beide Kriterien für MITTEL erfüllt.

SubjectID	Anzahl Fehler "leicht"	Anzahl Fehler "mittel"	Anzahl Fehler "schwer"	Bewertung der Segmentierung
Beispiel 1	1	6	4	MITTEL
Beispiel 2	6	1	2	MITTEL
Beispiel 3	7	7	0	MITTEL

Abbildung 6: Ausschnitt der Bewertungstabelle des Tutorials zur Erklärung der Segmentierungsqualitätsstufe MITTEL

Beispiel 2 (MITTEL)

Beachtlich ist, dass trotz massiver mikroangiopathischer Schädigung des Marklagers immer noch eine passable Segmentierung möglich ist. Neben sechs *leichten* Fehlern, die sich in Randbereichen zur nicht ganz entfernten Dura und im Bereich des Corpus Callosum verorten lassen, ist die auf Abbildung 7 markierte Region nicht ganz erfasst worden.

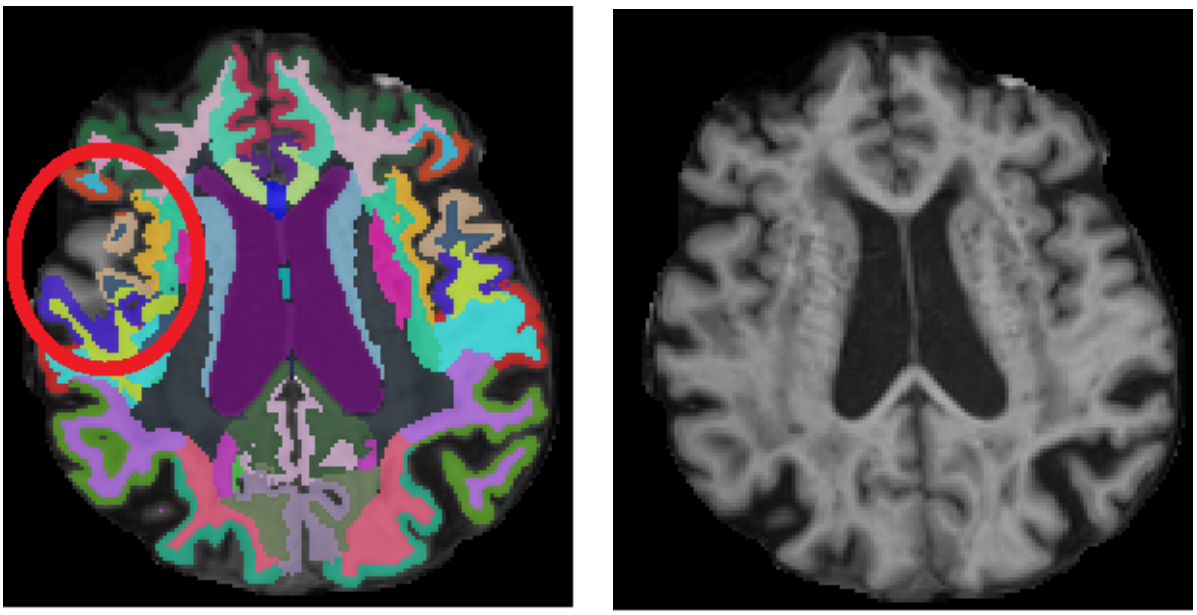


Abbildung 7: Axialschnitte der Segmentierung und des T1 Bildes von COKI10090. Rot markiert ist ein fehlerhaftes Areal

Die Segmentierung im Rindenband unter den Labels „ctx.rh.parsopercularis“ und „ctx.rh.rostralmiddlefrontal“ ist komplex gestört. Hinzu kommt die Marklagerschädigung als Pathologie. Somit liegen ***schwere*** Fehler vor.

Insgesamt überwiegen die *leichten* Fehler. Eine Einordnung als GUT ist allerdings ausgeschlossen, da auch ***schwere*** Fehler vorhanden sind. Die ***schweren*** Fehler stören die Segmentierung nur lokal. Daher wird dieses Ergebnis als MITTEL bewertet.

Beispiel 3 (MITTEL)

Die Bilder der Abbildung 8 sollen zur besseren Orientierung auf mögliche Fehlerregionen hinweisen.

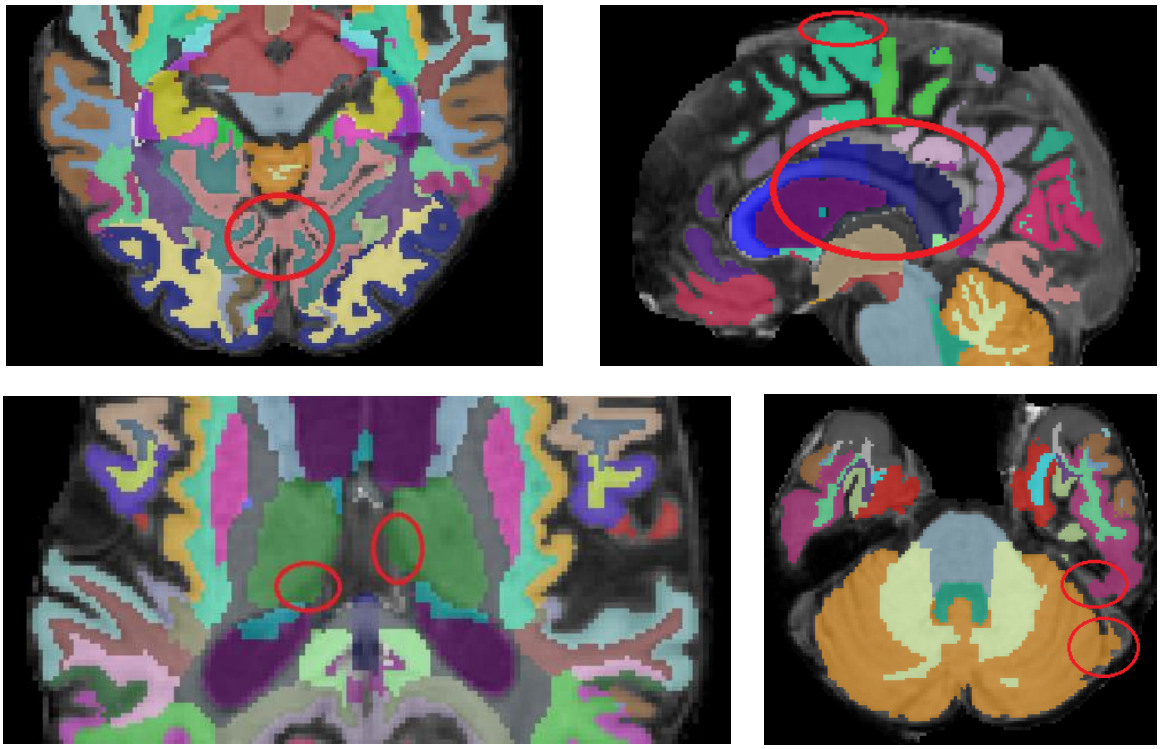


Abbildung 8: Drei Axialschnitte verschiedener Höhen und ein Sagittalschnitt (oben rechts) der Segmentierung von COKI10096. Rot markiert sind fehlerhafte Areale

In der Summe stehen sich 7 *leichte* und 7 *mittlere* Fehler gegenüber (siehe Abbildung 9). In diesem Fall greift die Sonderregel, dass bei Gleichstand die jeweils schlechtere Kategorie mehr ins Gewicht fällt. Somit überwiegen *mittlere* Fehler. Damit fällt die Gesamtbewertung in die Kategorie MITTEL.

SubjectID	Anzahl Fehler "leicht"	Anzahl Fehler "mittel"	Anzahl Fehler "schwer"	Bewertung der Segmentierung
Beispiel 1	1	6	4	MITTEL
Beispiel 2	6	1	2	MITTEL
Beispiel 3	7	7	0	MITTEL

Abbildung 9: Ausschnitt der Bewertungstabelle des Tutorials zur Erklärung der Segmentierungsqualitätsstufe MITTEL

c) SCHLECHT

Der Gesamtsegmentierungsbewertung SCHLECHT liegt eine schwere Hirnpathologie zu Grunde, die die anatomischen Strukturen derart verändert, dass die in FreeSurfer hinterlegten Hirnatlanten nicht mehr anwendbar sind. Dennoch versucht FreeSurfer alle Labels zu vergeben, mit der Folge, dass eine große Anzahl von Fehlern und Folgefehlern entsteht. Für SCHLECHT überwiegen Fehler, die offensichtlich mit einer Pathologie assoziiert sind und die Segmentierung komplex und global stören. Hier sollen 2 Beispiele der Verdeutlichung dienen.

Beispiel 1 (SCHLECHT)

Abbildung 10 zeigt eine Übersicht im Koronarschnitt. Neben der nicht ganz entfernten Dura fällt bei der ersten Betrachtung deutlich der Aspekt eines Normaldruckhydrozephalus auf. Ausdruck dessen sind die weiten Seitenventrikel, deren kraniale Begrenzung im spitzen Winkel in der Balkenregion zusammenläuft.

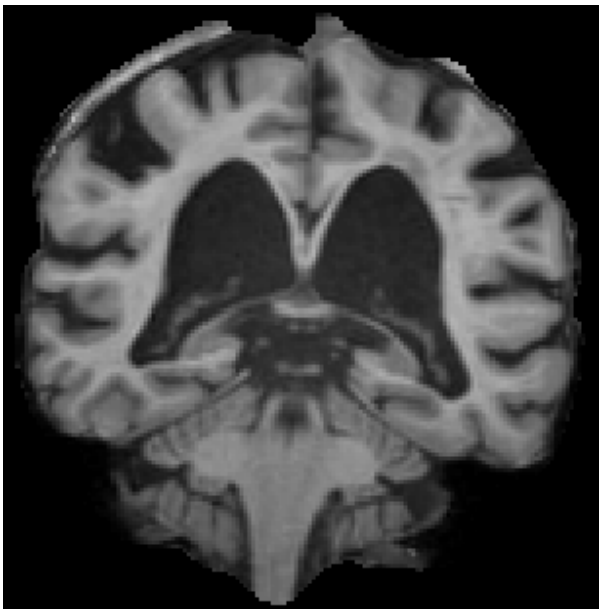


Abbildung 10: Koronarschnitt des T1 Bildes von COKI10068

Auf Abbildung 11 ist die wmparc-Segmentierung über denselben Koronarschnitt des T1-Bildes von Abbildung 10 gelegt. Abbildung 12 zeigt dieselbe Segmentierung in Axial- und Sagittalansicht. Außer die mit Ausrufezeichen markierten fehlenden Ventrikel sind auf diesen Bildern Labels markiert, die in die Ventrikel segmentiert wurden.

Die mit den großen Ventrikeln einhergehenden Fehler sind oft als ***schwer***, mindestens aber als *mittel* einzustufen.

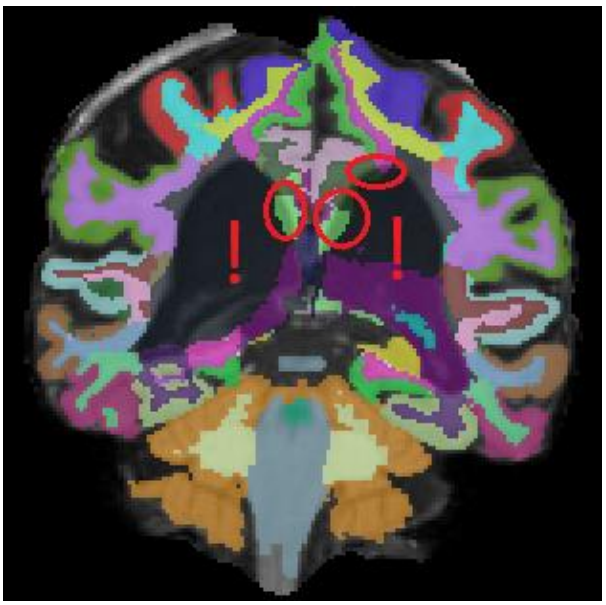


Abbildung 11: Koronarschnitt der Segmentierung von COKI10068

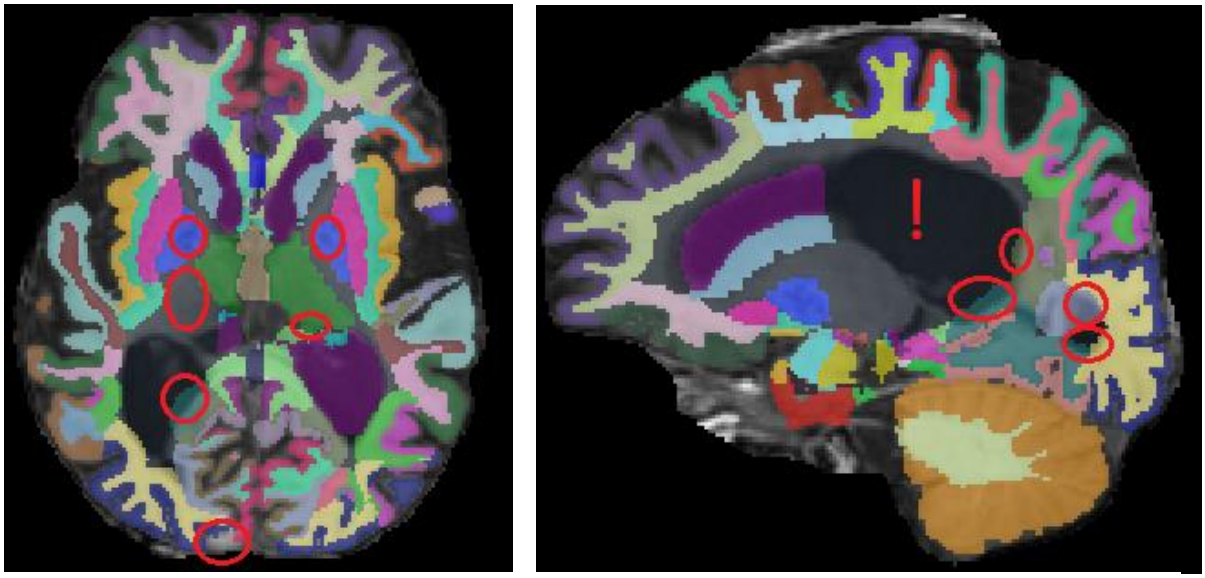


Abbildung 12: Axial- und Sagittalschnitt der Segmentierung von COKI10068. Rot markiert sind fehlerhafte Areale

Betrachtet man zur Gesamtbeurteilung den Ausschnitt der Tabelle (Abbildung 13), so überwiegen hier die *mittleren* Fehler. Die vorliegenden ***schweren*** Fehler beschränken sich allerdings nicht nur auf lokale Störungen, sondern die Integrität der Segmentierung ist global aufgehoben. Daher liegt hier die Kategorie SCHLECHT vor.

SubjectID	Anzahl Fehler "leicht"	Anzahl Fehler "mittel"	Anzahl Fehler "schwer"	Bewertung der Segmentierung
Beispiel 1	2	11	6	SCHLECHT
Beispiel 2	6	0	9	SCHLECHT

Abbildung 13: Ausschnitt der Bewertungstabelle des Tutorials zur Erklärung der Segmentierungsqualitätsstufe SCHLECHT

Beispiel 2 (SCHLECHT)

Bei der Betrachtung des Grundbildes (Abbildung 14) fällt ein erheblicher Substanzverlust auf (rote Markierung), der von einem alten Infarkt verursacht wurde.

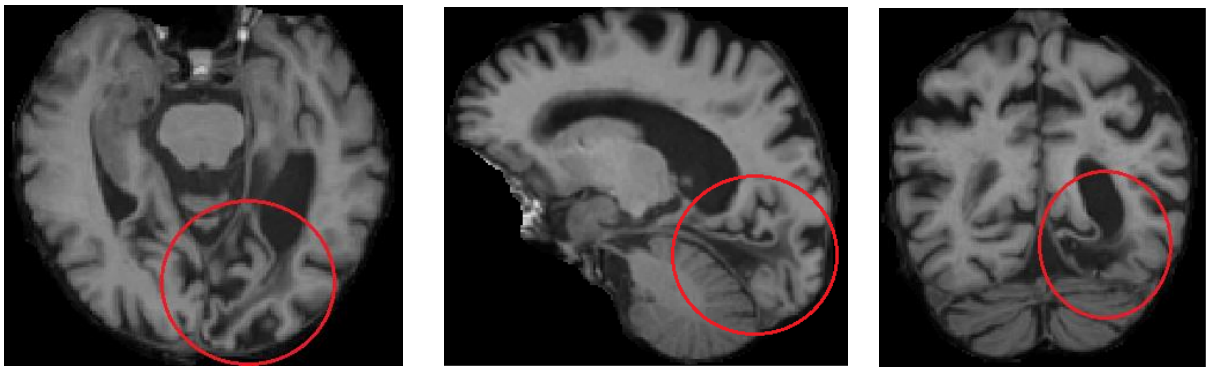


Abbildung 14: Axial-, Sagittal- und Koronarschnitt des T1 Bildes von COKI10086. Rot markiert ist eine ausgeprägte Infarktnarbe mit erheblichem Substanzverlust

In der Segmentierung ergibt sich in diesem Bereich eine Vielzahl von Fehlern. Der Gyrus lingualis scheint anatomisch komplett zu fehlen und ist kompensatorisch quer durch den Ventrikel segmentiert. Dieser wiederum ist nur zum Teil erfasst. Auch der Cuneus und der Gyrus pericalcarinus lassen sich nicht mehr zuordnen. Alle diese Fehler gelten als ***schwer***, da die dazugehörige Pathologie offensichtlich ist. Die Segmentierung ist komplex geschädigt. Außerdem ragt der Kortex des Kleinhirns deutlich über seine kraniale Grenze hinaus.

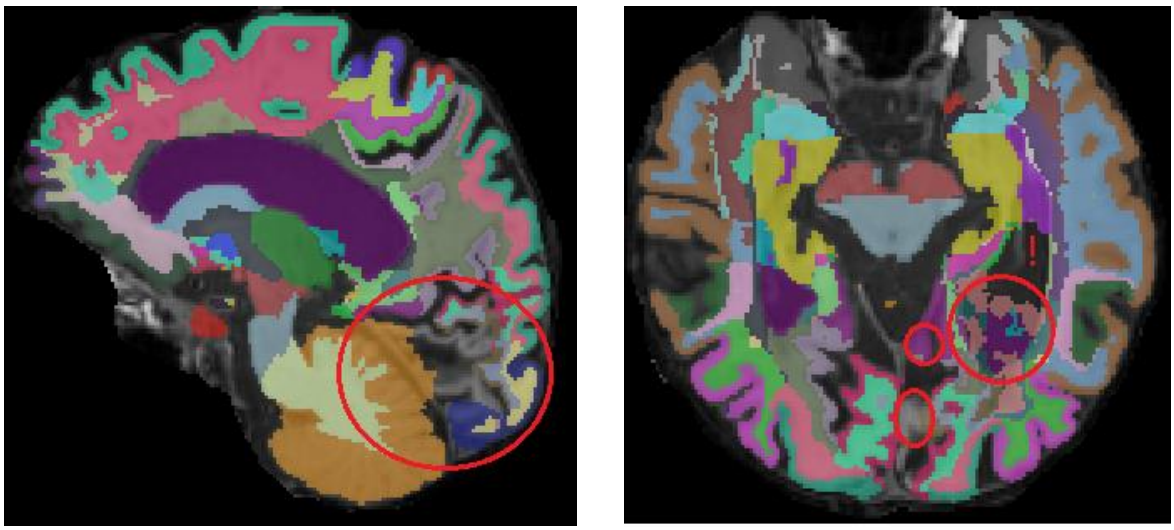


Abbildung 15: Sagittal- und Axialschnitt der Segmentierungen von COKI10086. Rot markiert sind fehlerhafte Areale

In der Summe überwiegen hier die ***schweren*** Fehler (siehe Abbildung 13). Die Segmentierung ist mit Beteiligung von Ventrikeln und Kleinhirn global gestört. Daher erfolgt als Gesamtbewertung SCHLECHT.

2.5 Volumenanalyse

Die für die weiteren volumetrischen Berechnungen benutzten Variablen ergeben sich aus der Summe der von FreeSurfer generierten Volumina einzelner Areale. Die zunächst betrachteten Regionen belaufen sich auf die Frontallappen, die Temporallappen, die Parietallappen, die Okzipitallappen, das Kleinhirn und die Basalganglien. Die Zuordnung der von FreeSurfer bestimmten Regionen zu den oben beschriebenen Hirnpartien erfolgte in Übereinstimmung mit gängiger neuroanatomischer Literatur, vornehmlich wurde das Werk „Prometheus Kopf, Hals und Neuroanatomie“ von Schünke et al., 2018, in der 5. Auflage verwendet. Die folgende Tabelle zeigt die detaillierte Zuordnung.

Tabelle 5: Sechs Hirnregionen und deren Labels in FreeSurfer

Betrachtete Hirnregion	zugrundeliegende FreeSurfer Region	
	linke Hemisphäre	rechte Hemisphäre
Frontallappen	ctx.lh.superiorfrontal wm.lh.superiorfrontal ctx.lh.caudalmiddlefrontal wm.lh.caudalmiddlefrontal ctx.lh.rostralmiddlefrontal wm.lh.caudalmiddlefrontal ctx.lh.parsopercularis wm.lh.parsopercularis ctx.lh.parstriangularis wm.lh.parstriangularis ctx.lh.parsorbitalis wm.lh.parsorbitalis ctx.lh.medialorbitofrontal wm.lh.medialorbitofrontal ctx.lh.lateralorbitofrontal wm.lh.lateralorbitofrontal ctx.lh.frontalpole wm.lh.frontalpole ctx.lh.precentral wm.lh.precentral 0.5 x ctx.lh.paracentral 0.5 x wm.lh.paracentral	ctx.rh.superiorfrontal wm.rh.superiorfrontal ctx.rh.caudalmiddlefrontal wm.rh.caudalmiddlefrontal ctx.rh.rostralmiddlefrontal wm.rh.caudalmiddlefrontal ctx.rh.parsopercularis wm.rh.parsopercularis ctx.rh.parstriangularis wm.rh.parstriangularis ctx.rh.parsorbitalis wm.rh.parsorbitalis ctx.rh.medialorbitofrontal wm.rh.medialorbitofrontal ctx.rh.lateralorbitofrontal wm.rh.lateralorbitofrontal ctx.rh.frontalpole wm.rh.frontalpole ctx.rh.precentral wm.rh.precentral 0.5 x ctx.rh.paracentral 0.5 x wm.rh.paracentral
Temporallappen	ctx.lh.middletemporal wm.lh.middletemporal ctx.lh.superiortemporal wm.lh.superiortemporal ctx.lh.inferiortemporal wm.lh.inferiortemporal	ctx.rh.superiortemporal wm.rh.superiortemporal ctx.rh.middletemporal wm.rh.middletemporal ctx.rh.inferiortemporal wm.rh.inferiortemporal

	ctx.lh.transversetemporal wm.lh.transversetemporal ctx.lh.temporalpole wm.lh.temporalpole ctx.lh.fusiform wm.lh.fusiform	ctx.rh.transversetemporal wm.rh.transversetemporal ctx.rh.temporalpole wm.rh.temporalpole ctx.rh.fusiform wm.rh.fusiform
Parietallappen	ctx.lh.postcentral wm.lh.postcentral ctx.lh.precuneus wm.lh.precuneus ctx.lh.superiorparietal wm.lh.superiorparietal ctx.lh.supramarginal wm.lh.supramarginal ctx.lh.inferiorparietal wm.lh.inferiorparietal 0.5 x ctx.lh.paracentral 0.5 x wm.lh.paracentral	ctx.rh.postcentral wm.rh.postcentral ctx.rh.precuneus wm.rh.precuneus ctx.rh.superiorparietal wm.rh.superiorparietal ctx.rh.supramarginal wm.rh.supramarginal ctx.rh.inferiorparietal wm.rh.inferiorparietal 0.5 x ctx.rh.paracentral 0.5 x wm.rh.paracentral
Okzipitallappen	ctx.lh.cuneus wm.lh.cuneus ctx.lh.lateraloccipital wm.lh.lateraloccipital ctx.lh.lingual wm.lh.lingual ctx.lh.pericalcarine wm.lh.pericalcarine	ctx.rh.cuneus wm.rh.cuneus ctx.rh.lateraloccipital wm.rh.lateraloccipital ctx.rh.lingual wm.rh.lingual ctx.rh.pericalcarine wm.rh.pericalcarine
Kleinhirn	Left.Cerebellum.Cortex Left.Cerebellum.White.Matter	Right.Cerebellum.Cortex Right.Cerebellum.White.Matter
Basalganglien	Left.Caudate Left.Pallidum Left.Putamen	Right.Caudate Right.Pallidum Right.Putamen

Das absolute Volumen der oben beschriebenen Lappen wurde zur Normalisierung für etwaige Unterschiede in Kopfgröße, Geschlecht und Alter durch das totale intrakranielle Volumen (Abkürzung: TIV; in FreeSurfer: eTIV) geteilt, wie es bei volumetrischen Betrachtungen dieser Art üblich ist (Barnes et al., 2010; Malone et al., 2015).

Das oben vorgestellte System zur Bewertung der Segmentierungsqualität wurde von 2 verschiedenen Ratern an 45 Segmentierungsdatensätzen angewandt. Neben dem Rating des Doktoranden wurde das Zweite, nach Studium des im Anhang einsehbaren Tutorials, bestehend aus 8 Beispielsegmentierungen, von einem erfahrenen Neuroradiologen (PD Dr. Christian Riedel) durchgeführt. Die sich jeweils ergebende finale Bewertung der

Segmentierung in die Kategorien „GUT“, „MITTEL“ und „SCHLECHT“ wurden als weitere Variablen hinzugefügt. Um beide Bewertungen in weiteren Analyseschritten zu berücksichtigen, wurden sie so zusammengeführt, dass wenn beide Rater sich für GUT entschieden haben, die Segmentierung auch als GUT kategorisiert wurde. Sollte sich einer für die Kategorie SCHLECHT entschieden haben, so wurde die Segmentierung als SCHLECHT bewertet. Alle übrigen Segmentierungen wurden der Kategorie MITTEL zugeordnet.

2.6 Statistische Methoden

Zur statistischen Auswertung wurde das frei verfügbare Programm RStudios (https://www.rstudio.com/products/rstudio/) in der Version 1.2.1335 verwendet. In Anlehnung an übliche Konventionen medizinischer Forschungsarbeiten wurde für die statistischen Berechnungen das Signifikanzniveau von $\alpha = 0.05$ festgelegt.

2.6.1 Zur ersten Hypothese

Um die Anwendbarkeit und Übertragbarkeit der vorgestellten Methode zu untersuchen, wurde die Interrater-Reliabilität überprüft. Zur Berechnung wurde das frei verfügbare Programm WinPepi genutzt (Abramson, 2004). Zur Quantifizierung der Übereinstimmung der zwei Rater wurde ein für Prävalenz und Bias korrigiertes quadratisches und ein linear gewichtetes Cohens Kappa berechnet (Cohen, 1968). Die Interpretation des Kappa orientiert sich an den von Landis und Koch (1977) vorgestellten Abstufungen. Zur Verifizierung der ersten Hypothese wird mindestens ein $\kappa > 0.4$ verlangt. Ab einem Wert von > 0,4 gilt die Übereinstimmung als moderat, wobei bei ab einem Wert von $\kappa = 0.61$ bis $\kappa = 0.8$ eine substantielle und ab $\kappa > 0.8$ eine exzellente Übereinstimmung angenommen wird (Landis & Koch, 1977).

Kappa	Agreement
<0.20	Poor
0.21 – 0.40	Fair
0.41 - 0.60	Moderate
0.61 – 0.80	Good
0.81 – 1.00	Very good

Abbildung 16: Kappa Werte und deren Interpretation nach Landis und Koch (1977)

2.6.2 Zur zweiten und dritten Hypothese

Um zu überprüfen, dass die Qualität der Segmentierung, in den Abstufungen GUT, MITTEL und SCHLECHT, das Volumen der umschriebenen Hirnareale nicht systematisch beeinflusst, wurde eine Varianzanalyse durchgeführt. Voraussetzung für eine Varianzanalyse ist die Normalverteilung der Variablen.

Diese wurde mit dem Shapiro-Wilk Test und der visuellen Kontrolle der Graphen überprüft. Wenn Normalverteilung angenommen werden konnte, wurden mögliche Unterschiede in der Varianz der Volumina der sechs Hirnregionen (Frontallappen, Temporallappen, Parietallappen, Okzipitallappen, Kleinhirn und Basalganglien) in Bezug auf die drei Gruppen „GUT“, „MITTEL“ und „SCHLECHT“ mit einer einfaktoriellen ANOVA untersucht. Konnte eine Normalverteilung nicht angenommen werden, wurde alternativ der Kruskal Wallis Test angewandt. Anknüpfend wurde in einem weiten Schritt die Verteilung der Varianz untersucht, um der in der dritten Hypothese formulierten möglichen Unterschiede der Volumina der sechs Hirnregionen, bei unterschiedlichen Qualitätsstufen der Segmentierung, nachzugehen. Für normalverteilte Variablen wurde die Homogenität der Varianzverteilung mit dem Bartlett-Test überprüft. Der Bartlett-Test überprüft die Nullhypothese, dass alle Gruppenvarianzen gleich sind. Der Bartlett-Test ist im Vergleich zu anderen statistischen Verfahren (z.B. Levene-Test), ein sensitives Verfahren und reagiert daher empfindlich auf die Verletzung der Varianzhomogentiät (Box, 1953).

Liegt keine Normalverteilung der Variablen vor, wurde der Fligner-Killeen Test verwendet.

Im Rahmen der post-hoc Analyse bei signifikanten Ergebnissen der ANOVA wurde die Korrektur für multiples Testen mit der Bonferroni Methode durchgeführt (Armstrong, 2014), um für Fehler der 1. Art zu korrigieren. Ein solcher Fehler liegt vor, wenn auf Grund von signifikanten Testergebnissen fälschlicherweise eine wahre Null-Hypothese abgelehnt wird.

2.6.3 Zur explorativen Fehleranalyse

Zur genaueren Untersuchung der Fehler wurde zunächst ihre Verteilung auf die einzelnen FreeSurfer Segmente betrachtet werden. Dazu wurde die Summe aller leichten, mittleren und schweren Fehler pro Segment über die 53 Datensätze gebildet und eine Rangfolge erstellt. Im Weiteren wurde mit den 10 FreeSurfer Labels mit der höchsten Fehlersumme weitergearbeitet. Für ebendiese wurde auch hier eine Varianzanalyse der Volumina in Bezug auf die Gesamtsegmentierungsqualität in den Stufen GUT, MITTEL und SCHLECHT durchgeführt. Hierbei handelt es sich nicht um die unter 2.5 beschriebene zusammengeführte Bewertung der zwei Rater, sondern nur um die des Doktoranden, da für das Ranking der Top 10 auch keine zusammengeführte Fehlersumme benutzt wird. Für die Varianzanalyse gelten die gleichen Bedingungen zur Normalverteilung, wie oben schon

beschrieben. Die weiteren Berechnungen zur Varianzanalyse und der Verteilung der Varianz folgten dem gleichen Schema, wie es unter 2.6.2 bereits beschrieben wurde.

3.0 Ergebnisse

Von den MRT Aufnahmen der 45 Probanden wurden 21 insgesamt in die Gruppe „GUT“, 10 in die Gruppe „MITTEL“ und 14 in die Gruppe „SCHLECHT“ eingeordnet. Der folgenden Tabelle können detailliertere demographische Informationen zu den Probanden im Kontext mit der jeweiligen Bewertungsstufe ihrer Segmentierung entnommen werden (Tabelle 6).

Tabelle 6: Übersicht über die demographischen Daten und Diagnosen der Gesamtkohorte (2. Spalte) , sowie der drei nach Segmentierungsqualität aufgeteilten Gruppen (3.-5. Spalte)

Parameter	Gesamt	Gruppe GUT	Gruppe MITTEL	Gruppe SCHLECHT
N für zwischen beiden Ratern harmonisierte Bewertung	45	21	10	14
Alter: Mittelwert (Standardabweichung)	78,7 (6,5)	77,8 (5,8)	79,0 (9,5)	79,8 (4,9)
N weiblich (%)	21 (47%)	13 (62%)	3 (30%)	5 (36%)
Jahre Bildung: Mittelwert (Standardabweichung)	9,6 (1,8) N=41	10,2 (1,9) N=19	8,1 (0,4) N=8	9,6 (1,7) N=14
MOCA: Mittelwert (Standardabweichung)	20,3 (5,5) N=36	20,8 (4,4) N=18	20,8 (5,9) N=6	19,2 (6,9) N=12)
N Diagnose PD (% als Anteil an der Gesamtzahl in der jeweiligen Qualitätsstufe)	19 (42%)	9 (43%)	3 (30%)	7 (50%)
N Diagnose Progressive Supranukleäre Blickparese (% als Anteil an der Gesamtzahl in der jeweiligen Qualitätsstufe)	6 (13%)	3 (14%)	3 (30%)	0
N Diagnose Schlaganfall (% als Anteil an der Gesamtzahl in der jeweiligen Qualitätsstufe)	3 (7%)	1 (5%)	0	2 (14%)
N Diagnose Epilepsie (% als Anteil an der Gesamtzahl in der jeweiligen Qualitätsstufe)	3 (7%)	1 (5%)	0	2 (14%)

In Hinsicht auf Alter, Geschlecht, MOCA und die häufigsten Diagnosen (siehe Tabelle 6), unterscheiden sich die drei Gruppen nicht signifikant. Lediglich ein Test der Jahre an Bildung war signifikant (Chi-Quadrat (2) = 10.2397, p = 0.006). Die Posthoc Tests ergaben, dass

dieser Effekt von der MITTEL Gruppe getrieben ist, welche signifikant weniger Jahre Bildung als die GUT (p=0,004) und die SCHLECHT (p=0,014) Gruppe aufwiesen.

3.1 Vergleichbarkeit der Daten zwischen zwei Ratern (Hypothese 1)

Zur Berechnung der Interrater-Reliabilität wurden von jedem Rater 45 Datensätze beurteilt. Tabelle 7 zeigt die Bewertungen der beiden Rater und Tabelle 8 zeigt das Vorgehen, nach dem die beiden Ratings zusammengeführt wurden. Die Bewertung von „Rater 1“ geht auf den Doktoranden zurück, der n = 32 Segmentierungen als GUT, n = 10 als MITTEL und n = 3 als SCHLECHT bewertete. Dem gegenüber steht die Bewertung von PD Dr. Christian Riedel (Rater 2) mit n = 23 für GUT, n = 9 für MITTEL und n = 13 für SCHLECHT.

Tabelle 7: Bewertung der zwei Rater über 45 Segmentierungen

	Rater 2			
Rater 1	Kategorien	GUT	MITTEL	SCHLECHT
	GUT	21	6	5
	MITTEL	1	3	6
	SCHLECHT	1	0	2

Tabelle 8: Schema zur Zusammenführung der Bewertungen der beiden Rater

Kategorien	Rater 2: GUT	Rater 2: MITTEL	Rater 2: SCHLECHT
Rater 1:GUT	GUT	MITTEL	SCHLECHT
Rater 1: MITTEL	MITTEL	MITTEL	SCHLECHT
Rater 1: SCHLECHT	SCHLECHT	SCHLECHT	SCHLECHT

Das Prävalenz- und Bias- korrigierte Cohen‘s Kappa mit quadratischer Wichtung der Kategorien betrug **κ = 0.59**.

Das Prävalenz- und Bias- korrigierte Cohen’s Kappa mit linearer Wichtung der Kategorien betrug **κ = 0.44**

Maximum attainable Kappa: **κ = 0.61**

3.2 Vergleich des Volumens und der Variabilität des Volumens zwischen den drei Qualitätsstufen

Um einen systematischen Einfluss der Segmentierungsqualität mit den Abstufungen GUT, MITTEL und SCHLECHT, auf das Volumen auszuschließen, wurden einfaktorielle Varianzanalysen für die sechs oben beschriebenen Hirnlappen durchgeführt.

Das Volumen des Frontallappens war gemäß des Shapiro-Wilk-Tests normalverteilt, $p > 0.05$. Mithilfe des Bartlett-Tests wurde das Vorliegen von Varianzhomogenität überprüft. Da $p > 0.05$, konnte die Nullhypothese, nach der die Gruppenvarianzen sich nicht unterscheiden sollen, nicht verworfen werden. Die Voraussetzungen für eine einfaktorielle ANOVA waren somit erfüllt. Diese ergab, dass sich das Volumen des Frontallappens statistisch nicht signifikant für die einzelnen Qualitätskategorien unterschied, $F(2,42) = 1,112$, $p = 0.338$.

Die Anwendung des Shapiro-Wilk-Tests und des Bartlett-Tests konnte das Vorliegen der Voraussetzungen für die Anwendung der ANOVA für den Temporallappen, den Parietallappen, den Okzipitallappen und das Kleinhirn jeweils bestätigen (p jeweils > 0.05).

Die Varianzanalysen für den Parietallappen, den Okzipitallappen und das Kleinhirn zeigten keinen signifikanten Einfluss der Qualitätsstufen auf das von FreeSurfer segmentierte Volumen (p-Werte siehe Tabelle 9).

Die Berechnung für den Temporallappen ergab einen gerade signifikanten Einfluss der Segmentierungsqualität auf das Volumen, $F(2,42) = 3,223$, $p = 0.0499$ auf unkorrigiertem Signifikanzniveau. Nach Bonferronikorrektur war dieses Ergebnis jedoch nicht signifikant. Trotzdem wurde hier Post-hoc T-Tests zwischen allen Gruppen durchgeführt, um zu schauen, welche Gruppen diesen Effekt lenken. Hier war nur der T-Test zwischen der Gruppe GUT und SCHLECHT signifikant ($p=0,045$ nach Bonferronikorrektur).

Die Anwendung des Shapiro-Wilk Tests ergab keine Normalverteilung des Volumens der Basalganglien, sodass anstelle des Bartlett-Tests auf einen Fligner-Killeen Test für die Überprüfung der Varianzhomogenität zurückgegriffen wurde. Dieser ergab mit einem p-Wert von $p > 0.05$ eine homogene Verteilung der Varianzen, in dessen Folge ein Kruskal-Willis Test für die Varianzanalyse angewendet wurde. Dieser ergab mit $p = 0.2774$ kein signifikantes Ergebnis, welches gegen einen Einfluss der Qualitätsstufe auf das FreeSurfer Volumen der Basalganglien spricht.

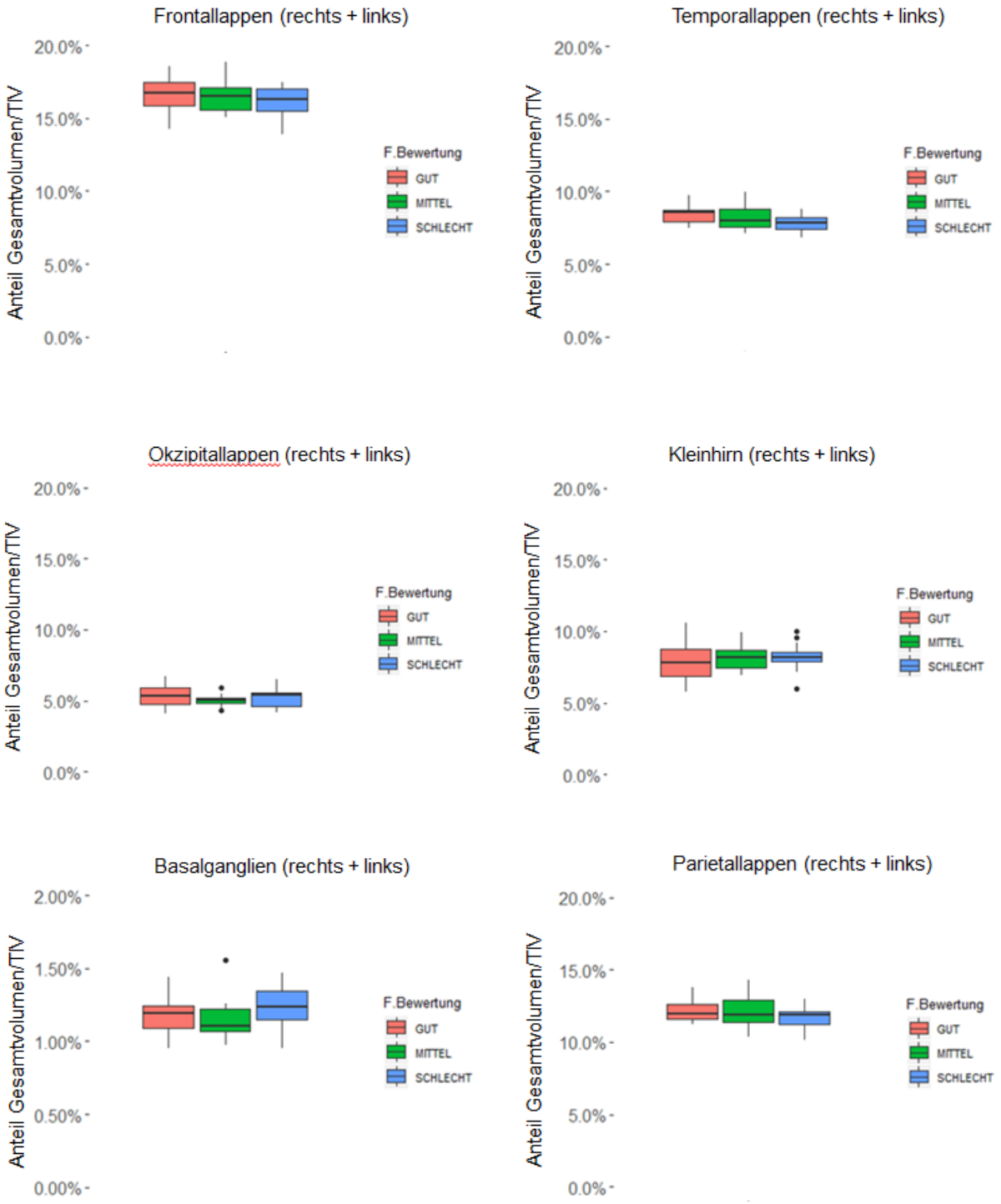


Abbildung 17: Boxplots der relativen Volumina in Prozent (y-Achse) der sechs untersuchten Hirnregionen in den Stufen GUT (rot), MITTEL (grün) und SCHLECHT (blau) der Gesamtqualität der Segmentierung

Tabelle 9: p-Werte und F-Statistik der Varianzanalysen für die sechs untersuchten Hirnregionen

Region	Test	F-Statistik	p-Wert Volumenunterschiede	Korrektur nach Bonferroni
Frontallappen	einfaktorielle ANOVA	1.112	0.338	1.0000
Temporallappen	einfaktorielle ANOVA	3.223	**0.0499**	0.4491
Parietallappen	einfaktorielle ANOVA	1.336	0.2660	1.0000
Okzipitallappen	einfaktorielle ANOVA	0.832	0.442	1.0000
Basalganglien	Kruskal Wallis	-	0.2774	1.0000
Kleinhirn	einfaktorielle ANOVA	0.447	0.6430	1.0000

Tabelle 10: Verwendete Testverfahren zur Überprüfung der Varianzhomogenität und zugehörige p-Werte für die sechs untersuchten Hirnregionen

Region	Test	p-Wert Varianztest
Frontallappen	Bartlett	0.7842
Temporallappen	Bartlett	0.2885
Parietallappen	Bartlett	0.1257
Okzipitallappen	Bartlett	0.3237
Basalganglien	Fligner-Killeen	0.6396
Kleinhirn	Bartlett	0.3710

3.3 Explorative Analyse der Regionen mit den meisten Fehlern

Tabelle 11 zeigt die Fehlerverteilung und deren Summe für die Regionen in FreeSurfer, die am meisten von Fehlern betroffen waren. Anhand der absoluten Fehlersumme wurde eine Reihenfolge bestimmt und die ersten zehn Regionen in die fortführenden Analysen eingeschlossen.

Tabelle 11: FreeSurfer Regionen sortiert nach den meisten Fehlern und ihrer Verteilung

Region	Name der Region in FreeSurfer	Fehlersumme	leichte Fehler	mittlere Fehler	schwere Fehler
mittlerer posteriorer Balken	CC_Mid_Posterior	21	6	12	3
posteriorer Balken	CC_Posterior	20	5	12	3
zentraler Balken	CC_Central	20	5	12	3
linker Seitenventrikel	Left.Lateral.Ventricle	19	2	8	9
rechter Seitenventrikel	Right.Lateral.Ventricle	19	6	10	3
linker Kleinhirnkortex	Left.Cerebellum.Cortex	16	9	4	3
linker Lobulus parietalis superior (Graue Substanz)	ctx.lh.superiorparietal	15	10	2	3
rechter Cuneus (Graue Substanz)	ctx.rh.cuneus	15	15	0	0
rechter Thalamus	Right.Thalamus.Proper	15	3	4	6
rechter Globus pallidus	Right.Pallidum	11	0	8	3

Als nächstes schloss sich die Überprüfung der Normalverteilung der Variablen mit dem Shapiro-Wilk Test an. Tabelle 12 zeigt für welche der zehn Regionen eine Normalverteilung vorlag. In 7/10 Fällen waren die Variablen nicht normalverteilt, sodass zur Untersuchung der Varianzverteilung der Volumina in Bezug auf die drei Qualitätsstufen der Fligner-Killeen Test benutzt wurde. In den 3/10 normalverteilten Fällen wurde dieselbe Untersuchung mit dem Bartlett-Test durchgeführt. In allen Fällen konnte Homogenität in der Verteilung der Varianz angenommen werden.

Tabelle 12: Testverfahren und deren p-Werte zur Überprüfung der Normalverteilung der Variablen und der Verteilung der Varianz für die zehn fehleranfälligsten Regionen

Region	Normalverteilung mit Shapiro-Wilk Test	Test auf Homogenität der Varianz	Test auf Homogenität der Varianz p-Wert	Homogenität der Varianz
mittlerer posteriorer Balken	nicht angenommen	Fligner-Killeen	0.1377	angenommen
posteriorer Balken	nicht angenommen	Fligner-Killeen	0.5861	angenommen
zentraler Balken	nicht angenommen	Fligner-Killeen	0.9888	angenommen
linker Seitenventrikel	nicht angenommen	Fligner-Killeen	0.6268	angenommen
rechter Seitenventrikel	nicht angenommen	Fligner-Killeen	0.9587	angenommen
linker Kleinhirnkortex	angenommen	Bartlett	0.0996	angenommen
linker Lobulus parietalis superior (Graue Substanz)	angenommen	Bartlett	0.7076	angenommen
rechter Cuneus (Graue Substanz)	nicht angenommen	Fligner-Killeen	0.1043	angenommen
rechter Thalamus	angenommen	Bartlett	0.1541	angenommen
rechter Globus pallidus	nicht angenommen	Fligner-Killeen	0.7367	angenommen

Wie unter 3.2 bereits für die sechs Hirnregionen beschrieben, wurde auch für die zehn explorativ untersuchen FreeSurfer Regionen, nach gleichem Schema, der Zusammenhang von Segmentierungsqualität und Volumenunterschieden überprüft. Bei erfüllten Voraussetzungen ergab die einfaktorielle ANOVA für den linken Kleinhirnkortex und den rechten Thalamus, dass sich die Volumina dieser Regionen nicht statistisch signifikant für die Qualitätsabstufungen unterschieden, wohingegen sich für die graue Substanz des linken Lobulus parietalis superior eine Signifikanz auf unkorrigiertem Niveau zeigt zeigte, F(2, 50) =

4.223, p = 0.0202. Posthoc durchgeführte T-Tests ergaben keine signifikanten Ergebnisse, jedoch einen Trend zu einem Unterschied zwischen den Gruppe GUT und MITTEL (p(korrigiert) = 0,069).

Für die sieben Regionen, für welche die Voraussetzungen für eine ANOVA nicht gegeben waren, wurde ein Kruskal-Wallis Test zur Varianzanalyse benutzt. Hier ergaben sich keine signifikanten Ergebnisse (siehe Tabelle 13).

Tabelle 13: p-Werte und F-Statistik der Varianzanalysen für die zehn fehleranfälligsten Regionen

Region	Varianzanalyse Test	F-Statistik	Varianzanalyse p-Wert	Korrektur nach Bonferroni
mittlerer posteriorer Balken	Kruskal-Wallis	-	0.0606	0.4850
posteriorer Balken	Kruskal-Wallis	-	0.6862	1.0000
zentraler Balken	Kruskal-Wallis	-	0.2506	1.0000
linker Seitenventrikel	Kruskal-Wallis	-	0.1769	1.0000
rechter Seitenventrikel	Kruskal-Wallis	-	0.3862	1.0000
linker Kleinhirnkortex	einfaktorielle ANOVA	0.237	0.7900	1.0000
linker Lobulus parietalis superior (Graue Substanz)	einfaktorielle ANOVA	4.223	**0.0202**	0.1818
rechter Cuneus (Graue Substanz)	Kruskal-Wallis	-	0.9318	1.0000
rechter Thalamus	einfaktorielle ANOVA	0.6730	0.515	1.0000
rechter Globus pallidus	Kruskal-Wallis	-	0.8856	1.0000

Da 3/10 Labels in der Auflistung dem Balken (Corpus Callosum) zuzuordnen sind, nahmen wir das zum Anlass die explorative Untersuchung auf den gesamten Balken auszuweiten.

Dazu erstellten wir, so wie bereits bei den sechs Hirnlappen, eine Summe aller Areale des Balkens („CC_Anterior“, „CC_Mid_Anterior“, „CC_Central“, „CC_Mid_Posterior” und “CC_Posterior”). Der Shapiro-Wilk Test ergab eine Normalverteilung der Variablen für den Balken und der Bartlett-Test eine homogene Verteilung der Varianz ($p = 0.0570$). Eine anschließende einfaktorielle ANOVA ergab keinen statistisch signifikanten Unterschied des Volumens des Balkens in den verschiedenen Qualitätsstufen, $F(2, 50) = 1.253$, $p = 0.228$, p (korrigiert) = 1.000.

4. Diskussion

Diese Arbeit untersuchte ein System zur Bewertung von FreeSurfer Segmentierungen, anhand von T1-gewichteten cMRT Bildern einer multimorbiden Kohorte, hinsichtlich dessen Anwendbarkeit und der Übertragbarkeit der Ergebnisse zweier unabhängiger Benutzer, die diese Methode anwandten. Des Weiteren wurde eine mögliche systematische Beeinflussung der von FreeSurfer generierten Volumina für sechs Hirnlappen durch die Segmentierungsqualität untersucht. Schließlich erfolgte eine explorative Ausweitung der Betrachtungen auf die zehn am häufigsten von Fehlern betroffenen FreeSurfer Labels und des Balkens. Diese Aspekte werden nun im Detail diskutiert.

4.1 Interrater-Reliabilität (Hypothese 1)

Das errechnete Cohens Kappa für die Einteilung der Probanden anhand der Segmentierungsqualität in drei Gruppen lag mit 0.59 (quadratische Wichtung) bzw. 0.44 (lineare Wichtung), in der Kategorie „moderat“, wenn man die Interpretationsabstufungen von Landis und Koch (1977) zu Grunde legt. Somit lässt sich die erste Hypothese verifizieren und es kann angenommen werden, dass die hier beschriebenen Gütekriterien für ein System zur Bewertung von FreeSurfer Segmentierungen von zwei Ratern vergleichbar, und damit grundsätzlich anwendbar sind. Der Vollständigkeit halber sind beide Kappa Wichtungen hier aufgeführt, da beide Verfahren im Fall einer 3x3 Matrix ihre Berechtigung haben (Hallgren, 2012; Warrens, 2013). Im vorliegenden Fall soll mit der quadratischen Wichtung der Kategorien bewirkt werden, dass die Uneinigkeit der zwei Rater, die nur eine Kategorie voneinander entfernt sind (GUT-MITTEL, MITTEL-SCHLECHT) nicht so schwer wiegt wie die komplette Uneinigkeit, die zwei Kategorien voneinander entfernt ist (GUT-SCHLECHT), da sich aus der Annäherung der Rater auf den Abstand einer Kategorie bereits eine Tendenz zur Übereinstimmung in der Bewertung der Qualitätsabstufung erkennen lässt, wohingegen die komplette Diskonkordanz für ein schwerwiegendes Problem in der Bewertung des Einzelfalls spricht. Der erreichte Wert liegt mit 0.59 nicht weit entfernt von der zweitbesten Stufe „gut“, für welche Werte von 0,61 bis 0,8 gelten (Landis & Koch, 1977). Somit liegt ein akzeptables, wenn auch keineswegs perfektes, Ergebnis vor. In den Überlegungen, wie man eine bessere Übereinstimmung erzielen kann, sollte man die Anzahl der Kategorien berücksichtigen. Rosen et al. (2018) konnten an einer Pilotstudie zur Untersuchung von cMRT Scanqualität feststellen, dass mit einem System, welches 5 Stufen beinhaltet, nur sehr geringe Interrater-Übereinstimmung erzielt werden konnte. Gerade weil eine steigende Anzahl von Kategorien, bei quadratischer Wichtung, zu höheren Kappa Werten führt (Brenner & Kliebsch, 1996) und daher mit diesem komplexeren Bewertungssystem an sich „bessere“ Kappa-Werte zu erwarten wären, kann dies für besondere Bedingungen in der Neuroradiologie sprechen. Rosens Ratingabstufungen mussten nach erster Evaluation in ein

3 stufiges System verändert werden, in dem 3 erfahrene Rater Kappa Werte von κ=0.64 bis κ=0.81 erzielten. In dieser Konsequenz ein 2 stufiges System vorzustellen, könnte allerdings nur bedingt eine weitere Verbesserung der Interrater Reliabilität bringen. Bei einem zweistufigen System, welches zwischen „akzeptabler“ und „inakzeptabler“ Segmentierungsqualität unterscheidet, würden viele Ergebnisse, die der Kategorie MITTEL zugeordnet wurden in die Ausschlusskategorie fallen, obwohl die Segmentierungsqualität nur minimal an lokal umschriebenen Stellen vom Optimum abweicht. Bei einer Studiengröße von N=45 (nach Abzug des Trainingsdatensatzes) würde das zu einem zu großen Verlust an volumetrisch auswertbaren Daten führen. Darüber hinaus könnte es auch passieren, dass der Kategorie GUT solche Bilder zugeordnet werden würden, die zu gut sind, um sie zu verwerfen, aber dennoch nicht ganz perfekt sind. Somit könnte auch die Stufe GUT einen generellen Qualitätsverlust erleiden. Das hier vorgestellte System möchte mit der Kategorie GUT die Segmentierungen höchstmöglicher Qualität zusammenfassen, und mit der Kategorie MITTEL eine mögliche Erweiterung des akzeptablen Datensatzes, für fortführende volumetrische Analysen, bereitstellen. Eine mögliche Erklärung für die noch zu verbessernde Übereinstimmung in der Interrater Untersuchung könnte das deutlich unterschiedliche Niveau der Erfahrung der beiden Betrachter sein. Der Verfasser der vorliegenden Dissertation hat sich im Rahmen des Entwurfs des Systems intensiv mit der systematischen Betrachtung und Begutachtung von Hirn MRT Bildern beschäftigt. Der zweite Betrachter jedoch war ein neuroradiologischer Oberarzt, der auf langjährige Erfahrung auf dem Sektor der Bildbetrachtung und -beurteilung, intensives Training und fachspezifische Weiterbildungen zurückblicken kann. Diese Differenz kann zu einer unterschiedlich kritischen Beurteilung der Daten führen, die für den zweiten Rater tendenziell zu Gunsten der Kategorien MITTEL und SCHLECHT ausfällt. Dieses Phänomen könnte sich mit den Regeln zur Fehlerkategorisierung erklären lassen. Wie in Tabelle 2 zusammengefasst, ist zum einen die geschätzte Volumenabweichung, die Relevanz und Komplexität des Fehlers für die Gesamtsegmentierung, zum anderen aber auch das Vorhandensein von Pathologien ausschlaggebend, ob ein Fehler als „leicht“, „mittel“ oder „schwer“ eingestuft wird. Bei leichten Fehlern liegt offensichtlich keine Pathologie vor. Bei schweren Fehlern liegt definitiv eine Pathologie vor und bei mittleren Fehlern ist eine Pathologie nicht ausgeschlossen. Diese Einteilung ist im Besonderen von der Erfahrung des Bildbetrachters abhängig, sodass der erfahrene Neuroradiologe wahrscheinlich ein Vielfaches mehr an Pathologien im zugrundeliegenden T1 Bild sieht, die Ursachen einer gestörten Segmentierung sein könnten, als ein Bildbetrachter mit wenig Vorerfahrung. Mehr Fehler der Stufen „mittel“ und „schwer“ führen in der Summe zu einer schlechteren Gesamtbewertung der Segmentierung bei Rater 2. Der hier diskutierte Aspekt sollte Gegenstand weiterführender Untersuchungen sein, bei denen beispielsweise die Interrater-Reliabilität zwischen zwei erfahrenen Ratern und zwei

unerfahrenen Ratern betrachtet wird. Eine solche Untersuchung ist insbesondere von Bedeutung, da im Rahmen von nicht-klinischen Fragestellungen an einen derartigen Datensatz vermutlich eher von Laien der Bildbetrachtung visuell kontrolliert wird. Vor dem Hintergrund, dass das Regelwerk dieses Systems den Anspruch erhebt simpel genug zu sein, um von einem Laien gelernt und angewandt zu werden, aber gleichzeitig differenziert genug ist, um von einem kritischeren Experten verwertbare Ergebnisse zu liefern, ist das erreichte Kappa von $\kappa=0.59$ ein akzeptabler Wert.

4.2 Vergleich des Volumens (Hypothese 2) und der Variabilität (Hypothese 3) des Volumens zwischen den drei Qualitätsstufen

Eingebettet in die Arbeiten der explorativen ComOn Studie zur Erforschung des Zusammenhangs von kognitiven und motorischen Funktionseinschränkungen an einer neurogeriatrischen Kohorte stellen die mit dem Kieler Geriatrie-Protokoll erhobenen cMRT Daten einen wertvollen und möglicherweise einzigartigen Datensatz dar. Die Möglichkeit, algorithmusbasierend quantitative volumetrische Daten zu gewinnen und so klinische Fragestellungen zu erforschen, stellt eine nützliche Erweiterung der ComOn-Datenbank dar. Gerade weil die untersuchte Kohorte wahrscheinlich starke neurostrukturelle Vorschädigungen aufweist und die visuelle Kontrolle der mit FreeSurfer segmentierten MRT Datensätze von sehr variabler Qualität war, war die Notwendigkeit zu einer grundsätzlichen Untersuchung der Methode gegeben, bevor sich weiterführende (inhaltliche, z.B. Vergleich zwischen verschiedenen Diagnosen, etc.) Forschung anschließen kann. Die Untersuchung zur Beantwortung der zweiten Hypothese, ob sich bei der verwendeten Segementierungsmethode systematische Unterschiede hinsichtlich der Volumendurchschnitte in Bezug auf die Qualitätsstufen GUT, MITTEL und SCHLECHT ergeben, orientiert sich an einer ähnlichen Vorgehensweise, wie sie von McCarthy et al. (2015) angewandt wurde. Im Unterschied zu der benannten Studie (McCarthy et al., 2015), handelt es sich hier nicht um eine Untersuchung inwieweit manuelle Korrektur (mutmaßlich verschiedene Qualitätsstufen) der FreeSurfer Segmentierung einen Unterschied auf das Volumen hat. Auch war die Kohorte bei McCharthy deutlich jünger (Altersduchschnitte der Gruppen zwischen 17-20 Jahre vs. 78-80 Jahre). Die hier gezeigten Ergebnisse bestätigen die bereits angeführte Vermutung und verifizieren die zweite Hypothese, dass die Volumina der sechs Regionen Frontallappen, Temporallappen, Parietallappen, Okzipitallappen, Basalganglien und Kleinhirn nicht systematisch durch die drei Qualitätsstufen beeinflusst werden, wenn man die nach Bonferroni korrigierten p-Werte der Varianzanalysen zugrunde leget (siehe letzte Spalte Tabelle 9). Daraus lässt sich schließen, dass auf Ebene der Hirnlappen, FreeSurfer zur volumetrischen Untersuchung von multimorbiden neurologisch kranken Patienten grundsätzlich angewendet werden kann, ohne dass eine Verfälschung der

Daten aufgrund mangelnder Segmentierungsqualität befürchtet werden muss. Eine mögliche Erklärung für dieses –durchaus vielversprechende- Ergebnis könnte in der Definition eines Segmentierungsfehlers liegen. Fehler sind, nach der vom Verfasser der vorliegenden Arbeit entworfenen Methode, eine unzulässige Verschiebung der Segmentgrenze, in dessen Folge einer anatomischen Region entweder zu viel oder zu wenig Volumen zugeordnet wird. Dies kann zu Lasten oder zu Gunsten eines benachbarten Labels gehen. Zu Hirnlappen aufsummierte Labels von FreeSurfer beinhalten bei schlechterer Qualität mehr Fehler, aber absolut gesehen das gleiche Volumen, da sich dieses bei Verschiebung der Grenzen der Sublabels innerhalb eines Lappens, nicht verändert. Auf Ebene der Hirnlappen scheint das Verfahren zuverlässig volumetrische Daten zu liefern und unterstützt somit auch McCarthys Schlussfolgerung, dass eine manuelle Korrektur von Fehlern in FreeSurfer an fest definierten Stellen innerhalb des Segmentierungsablaufs, zu keinem Unterschied in der Hirnvolumetrie führt (McCarthy et al., 2015). Betrachtet man die unkorrigierten p-Werte, ergibt sich ein ähnliches Bild, mit der Ausnahme des Temporallappens, der mit $p = 0.0499$ knapp unterhalb des festgelegten Signifikanzniveaus liegt. Die post-hoc Analyse bestätigt, dass es einen Unterschied zwischen den Gruppen GUT und SCHLECHT hinsichtlich des Volumens gibt. Die Grafik in Abbildung 17 zeigt, dass dieser Volumenunterschied als eine Volumenminderung bei schlechter werdender Qualität ausfällt. Dieser Trend könnte entweder durch technische oder klinische Unterschiede erklärt werden. Für ein technisches Problem bei FreeSurfer sprechen Beobachtungen, dass die Region des Temporallappens von vielen Strukturen, wie z.B. Dura mater, Knochen oder Kleinhirn, umgeben ist, was zu Schwierigkeiten und Unschärfe in der Segmentierung führen kann (Desikan et al., 2010; McCarthy et al., 2015). Desikan et al. beschreibt 2010, dass hier besonders auf den medialen Temporallappen geachtet werden soll. Betrachtet man die FreeSurfer Labels, die den mittleren Temporallappen ausmachen (ctx.lh.middletemporal, wm.lh.middletemporal, ctx.rh.middletemporal, wm.rh.middletemporal) in der Bewertungstabelle, so kann man feststellen, dass sich für die 53 bewerteten Probanden in diesen Regionen keine große Anzahl von Fehlern finden lässt. Um der anfangs erwähnten Möglichkeit einer pathologischen Beeinflussung des Volumens des Temporallappens nachzugehen, bedarf es weiterer Auswertungen, die Gegenstand von Arbeiten sein sollen, die sich an diese Dissertation anschließen können. So könnte es sein, dass in dem hier vorgestellten, von neurologischen Krankheiten betroffenen Patientengut vermehrt zu Atrophien im Temporallappen kommen könnte, die sich in einer erhöhten Anzahl von Fehlern in der Segmentierung niederschlagen, die wiederum zu einer schlechteren Bewertung der Gesamtqualität führen. Mögliche erste Hinweise auf einen Zusammenhang zwischen Pathologie, Atrophie und Segmentierungsqualität liefert die Tabelle zu den demographischen Daten (siehe Tabelle 6), die beispielsweise zeigt, dass der höchste Anteil von PD Diagnosen

in der Gruppe der Segmentierungsqualität SCHLECHT (50%) liegt. Auch der Anteil an Probanden mit Schlaganfall und Epilepsie war in der Gruppe SCHLECHT (je 14%) höher als in der Gruppe GUT (je 7%). Diese Daten sind aber nicht signifikant, und bedürfen daher für eine adäquate Beantwortung weiterer und größerer Kohorten. Um das Ergebnis der ANOVA für den Temporallappen genauer zu untersuchen, sollte in einem nächsten Schritt die Differenzierung in rechte und linke Hemisphäre erfolgen. Eine rein einseitige Volumenminderung könnte in diesem Zusammenhang eher für einen „echten" klinischen Befund, eine beidseitige eher für eine technische Unschärfe der FreeSurfer Software sprechen, da die an den Temporallappen angrenzenden problematischen Strukturen schließlich beidseits bestehen (Desikan et al., 2010). Bis zur genaueren Ergründung des Phänomens sollten FreeSurfer-basierte volumetrische Analysen insbesondere des Temporallappens innerhalb entsprechender Kohorten kritisch beurteilt werden. Das soll nicht in Widerspruch zu der oben bereits statuierten Aussage, dass FreeSurfer auf Lappenebene zuverlässige volumetrische Informationen liefert, stehen, da zu beachten ist, dass nach wie vor die Möglichkeit eines Fehlers 1. Art, bei dem eine wahre Null-Hypothese auf Grund eines signifikanten Ergebnisses fälschlicherweise abgelehnt wird, besteht. Daher wurde die Korrektur nach Bonferroni durchgeführt, die als robustes Mittel gegen ebendiese Fehler bekannt ist (Armstrong, 2014) und in diesem Fall keine signifikanten Ergebnisse für die einfaktoriellen ANOVAs mehr liefert.

Bezüglich der dritten Hypothese, ob eine schlechtere Segmentierungsqualität mit einer erhöhten Variabilität der Varianz einhergeht, legen die oben präsentierten Ergebnisse nahe, dass dies nicht der Fall ist. Weder die Bartlett-Tests für die normalverteilten Daten (alle analysierten Regionen außer die Basalganglien) noch der Fligner-Killeen Test für die nicht normalverteilten Daten (Basalganglien) ergaben signifikant unterschiedliche Werte, was für eine homogene Verteilung der Varianz spricht. Entgegen der Vermutung, dass bei schlechterer Qualität die Werte von mehr Labels durch eine erhöhte Zuordnung von Volumen oder durch einen erhöhten Verlust an Volumen verfälscht sind, scheint es auch bei der visuellen Kontrolle der Boxplots in Abbildung 17 keine Tendenz dieser Art zu geben. Insgesamt lässt sich aus den Ergebnissen zu Verteilung der Varianz schließen, dass auf Ebene der Hirnlappen (rechte und linke Hemisphäre aussummiert) volumetrische Analysen mit den von FreeSurfer generierten Daten für die untersuchte Kohorte durchgeführt werden können, auch wenn die mit dem vorgestellten System bewertete Qualität schlechter ist, da kein großer Effekt von Varianzstreuung zu erwarten ist. In zukünftigen Arbeiten könnten Ausreißer genauer überprüft werden, da hier am ehesten grobe Volumenunterschiede vorliegen könnten. Zu beachten ist, dass die vorgelegten Untersuchungen zunächst nur für die Ebene der Hirnlappen, aufsummiert für beide Hemisphären, durchgeführt wurden. Weiterführende Untersuchungen könnten die Hemisphären getrennt voneinander

untersuchen, um eventuelle Lateralisierungseffekte von FreeSurfer zu identifizieren. Ein anderes mögliches Verfahren, um Volumenausreißer in den Daten zu identifizieren, könnte darin bestehen, zunächst alle Volumenmittelwerte zu erfassen, die von fehlerfreien Segmentierungen generiert worden sind. Diese müssten als Normwerte in eine Datenbank eingelesen werden, wo sie zu Gruppen für Geschlecht und Alter zusammengefasst werden. Für neue Segmentierungen müssten die Volumenmittelwerte mit den Werten dieser Datenbank verglichen werden und ab einer Grenze (z.B. eine Standardabweichung) als Ausreißer identifiziert werden. Ließe sich ein solches Verfahren vollautomatisch direkt an die ohnehin schon vollautomatische Segmentierung anschließen, könnten bereits vor der visuellen Kontrolle/Bewertung der Segmentierung auf mögliche Probleme in einzelnen Regionen aufmerksam gemacht werden. Allerdings bräuchte es für ein solches Verfahren eine weitaus größere Anzahl an Datensätzen, damit für alle Gruppen (z.B. verschiedene Altersgruppen und Geschlecht-spezifisch) genügend Daten für die Volumennormwerte vorliegen. Um eine solche Menge an quantitativ auswertbaren Daten zu gewinnen, besteht die eingangs schon erwähnte Notwendigkeit zur standardisierten Erhebung von Daten. Gerade in der Erforschung und der Behandlung von Atrophie bei neurologischen Erkrankungen wird die automatisierte quantitative Bildverarbeitung in naher Zukunft an Bedeutung gewinnen, sodass zum schnelleren Erheben von Daten, am besten auf Multicenterebene, einheitliche MRT Protokolle einen großen Fortschritt darstellen würden (Marciniewicz et al., 2019).

Insgesamt lässt sich aus den Ergebnissen für die Varianzanalysen und Varianzstreuung ableiten, dass Regionen wie Hirnlappen in Gruppenanalysen mit einbezogen werden können, auch wenn die individuelle Segmentierung in die Kategorie SCHLECHT fällt.

4.3 Untersuchung besonders fehleranfälliger Gehirnregionen (explorative Analyse)

Gerade für detaillierte, klinisch orientierte wissenschaftliche Fragestellungen ist die automatisierte quantitative Segmentierung einzelner Areale von cMRT Bildern mit einer Software wie z.B. FreeSurfer hochinteressant und –relevant. Dies beruht auch darauf, dass klinische und Mobilitäts-Daten vermehrt quantitativ erfasst werden, z.B. mittels tragbarer Sensoren (Maetzler, Klucken, & Horne, 2016). So könnte die Hirnvolumetrie von einzelnen Arealen, die beispielsweise den Motorkortex oder andere funktionell relevante Regionen repräsentieren, quantitative Informationen über ein mögliches strukturelles anatomisches Korrelat von Defiziten im klinischen Assessment oder in der häuslichen Mobilität liefern. Wie auch bei den Untersuchungen auf Lappenebene stellt sich die Frage, ob bei multimorbiden Patienten mit neurologischen Erkrankungen die variable Qualität der Segmentierung einen

systematischen Einfluss auf die volumetrischen Parameter hat und ob sich bei schlechterer Qualität eine erhöhte Streuung der Varianz feststellen lässt (McCarthy et al., 2015). In dieser Arbeit wurden daher die zehn am häufigsten von Fehlern betroffenen FreeSurfer Areale bestimmt. In den Varianzanalysen zeigte sich, dass nur in der grauen Substanz des linken Lobulus parietalis superior ein statistisch signifikanter Volumenunterschied zwischen den Qualitätsstufen bestand. Dies steht in gewisser Übereinstimmung mit den Ergebnissen von McCarthy et al. (2015), die für die weiße Substanz derselben Region Volumenunterschiede bei unterschiedlicher Qualität feststellen konnten. Allerdings liegen bei McCarthy Bilder eines 1.5 Tesla Magnetresonanztomographen zu Grunde. Nach Bonferroni Korrektur zeigte keines der betrachteten Areale signifikant unterschiedliche Werte, sodass davon ausgegangen werden kann, dass auch Areale mit schlechter Segmentierungsqualität in weitere Analysen einbezogen werden können. Allerdings scheint es Regionen zu geben, bei denen eine Tendenz zu systematischen Volumenbeeinflussung besteht, sodass dennoch zu Vorsicht geraten werden sollte. Da nun auf Ebene der einzelnen Areale kein „Ausgleichseffekt“ (siehe 4.2) die homogene Verteilung von Volumenunterschieden und Varianzstreuung erklären kann, muss hier beachtet werden, dass sich die betrachteten Daten von denen der unter 4.2 untersuchten unterscheiden. In die explorative Untersuchung wurden nur Datensätze einbezogen, die vom Doktoranden bewertet wurden. Dies hat den rein pragmatischen Hintergrund, da nur hier die detaillierte Zuordnung der Fehler zu den Labels vollständig vorliegt. Da sich das Level der Bildbetrachtungserfahrung, wie unter 4.1 beschrieben, deutlich von der des zweiten Raters, der ein erfahrener Neuroradiologe ist, unterscheidet, könnte sich darin eine mögliche Beeinflussung der Ergebnisse verbergen. Wie Tabelle 7 zeigt, sind bereits in den Gesamtbewertungskategorien des erfahrenen Bildbetrachters die Kategorien MITTEL und SCHLECHT häufiger vertreten, welches für eine kritischere Bewertung spricht. Wie bereits unter 4.1 beschrieben, ist die Verschlüsselung der gefundenen Pathologien in der Bewertung der Fehlerschwere eine Erklärung. Da der erfahrene Neuroradiologe mehr Pathologien sieht und sich diese in einer anderen Fehlerverteilung niederschlägt könnten hier bereits andere Labels die zehn häufigsten von Fehlern betroffenen Areale darstellen, da hierfür lediglich die Summe der Fehler „leicht“, „mittel“ und „schwer“ für jedes Label herangezogen wird. Um die Daten der explorativen Untersuchung in einen Kontext mit denen der zweiten und dritten Hypothese zu bringen, welche ihrerseits eine zusammengeführte Gesamtbewertung der beiden Rater verwendet (siehe Tabelle 8), müsste man auch hier eine Rangliste der 10 meistbetroffenen Areale generieren. Eine weitere methodische Beeinflussung der Ergebnisse könnte auch in der Aufstellung der 10 Areale liegen, die mittels einfachem Aufsummieren aller Fehler eines Labels erstellt wird. Um die Fehlerlast eines Labels zu bestimmen, geht beim simplen Aufsummieren der Fehler der einzelnen Kategorien die Information über die Qualität des

Fehlers verloren, da ein leichter Fehler und ein schwerer Fehler in der Summe gleich ins Gewicht fallen. Hier besteht die Möglichkeit mit Faktoren die Fehlerschwere in Form eines Scores zu ermitteln und dann zur Fehlerlast aufzusummieren. So könnte die Anzahl der Fehler der Kategorie „schwer“ mit dem Faktor x3, die Fehler der Kategorie „mittel“ mit dem Faktor x2 und die Fehler der Kategorie „leicht“ mit dem Faktor x1 verrechnet werden. Auch eine quadratische Wichtung ist an dieser Stelle möglich („leicht“ x1, „mittel“ x2, „schwer“ x4). Bildet man nun die Summe, sollten solche Areale die von vielen schweren Fehlern betroffen sind mehr Berücksichtigung finden. Diese Unterscheidung ist auch insofern relevant, da bei schweren Fehlern laut Tabelle 4 sicher eine Pathologie zu Grunde liegt, wohingegen bei leichten Fehlern eine Pathologie sicher ausgeschlossen ist, es sich in letzterem Fall also vielmehr um ein technisches Problem von FreeSurfer handelt. Wie bereits dargelegt, erkennt der Neuroradiologe weit mehr Pathologien, sodass die bei einer faktorengewichteten Aufstellung seiner 10 am meisten von Fehlern betroffenen Labels mehr pathologisch veränderte Areale gelistet werden würden. Ebendiese Areale beinhalten vermutlich auch mehr Volumenabweichungen, sodass gerade hier weitere Untersuchungen sich mit systematischen Fehlern im Zusammenhang von Volumen und Segmentierungsqualität beschäftigen sollten, um sicherzugehen, dass von FreeSurfer generierte volumetrische Daten einer Kohorte wie hier vorgestellt für weitere Analysen benutzt werden können, ohne methodisch bedingte Fehler zu riskieren.

Da 3/10 der am häufigsten von Fehlern betroffenen Arealen dem Balken (Corpus Callosum) zuzuordnen sind, erweiterten wir die explorative Untersuchung auf den gesamten Balken, wie er sich aus der Summe der 5 Subregionen ergibt. Anfangs war der Balken im *Desikan-Killiany-Atlas* nicht als volumetrisch belastbare Region angelegt, sondern vielmehr als nötige Struktur zu Begrenzung anliegender Areale wie zum Beispiel der Gyri cinguli (Desikan et al., 2006). Die nun vorhandene Aufteilung der Balkenregion in die vorliegenden 5 Sublabels (Vogt et al., 2003, 2006) ist eine der 2009 vorgenommenen Anpassungen des *Destrieux-Atlasses*. Die Relevanz und das Interesse an Balkenvolumetrie besteht insbesondere für psychiatrische Forschung, da ein Zusammenhang zwischen reduziertem Volumen des Balkens mit Bipolarität und Suizidalität besteht (Gifuni et al., 2017). Bei der visuellen Inspektion der Balkenregion war eine Vielzahl von Fehlern zu beobachten. Dennoch konnten die berechneten Tests keine systematische Beeinflussung der Volumina in den verschiedenen Qualitätsstufen feststellen, sodass davon auszugehen ist, dass auch die fehleranfällige Region des Balkens in volumetrische Analysen eingeschlossen werden kann.

4.4 Limitationen, Methodenkritik und Ausblick

Ziel der vorgelegten Arbeit war neben der oben beschriebenen Überprüfung der vollautomatischen Segmentierung mit FreeSurfer an einer Kohorte multimorbider Patienten mit neurologischen Erkrankungen, die Vorstellung einer neu entworfenen Methode zur Bewertung der Qualität ebendieser Segmentierungen. Da es sich bei der vorgestellten, erstmals in dieser Kohorte angewandten Methode um ein vielschrittiges Verfahren handelt, ist diese einer ganzen Reihe von Limitationen unterworfen. Die hier vorgestellte Kohorte stammt von einer Akutstation und kann so möglicherweise nicht das gesamte Spektrum multimorbider Patienten mit neurologischen Erkrankungen abbilden. Interessant wäre in diesem Zusammenhang sicherlich auch eine Auswertung von Bilddatensätzen aus einer Kohorte mit multimorbiden Patienten, die aber nicht neurologisch erkrankt sind.

Ziel dieser Methode ist es, anhand klarer Regeln akzeptable Segmentierungsergebnisse von hoher Qualität (Kategorie „GUT“), weiterführenden (beispielsweise volumetrischen) Analysen zuführen zu können. Auch wenn mit den hier diskutierten Kriterien eine möglichst objektive Bewertung möglich werden soll, verbleibt, wie bei allen Studien mit menschlichen Betrachtern, eine Rest-Subjektivität. So sind die Angaben zur Raumausdehnung eines Fehlers kritisch zu sehen, da Angaben wie zum Beispiel „über Segmentgrenze deutlich (mehr als eine Schicht, aber max. +1/3 des Gesamtvolumens der betroffenen Struktur) hinaus, fälschlicherweise ist anderen Strukturen dieses Label leicht zugeordnet“ allein vom Augenmaß des Betrachters abhängt, der wiederum die betroffene Struktur in alle Raumrichtungen untersucht haben muss.

Eine weitere Unschärfe entsteht, wenn sich in einem Segmentierungsareal zwei Fehler befinden. In diesem Fall ist es an dem Rater, den schwerwiegenderen Fehler für das Areal zu listen. An dieser Stelle geht eine Information bezüglich der Segmentierungsqualität verloren.

Des Weiteren berücksichtigt die zugrundeliegende Kategorisierung der Fehler einer Segmentierung nur eingeschränkt, ob es sich bei dem Fehler um einen solchen handelt, der auf schlechte Scanqualität oder auf ein hirnpathologisches Korrelat zurückzuführen ist. Im Hinblick auf die Scanqualität unterscheidet Mortamet et al. (2009) Scanner-assoziierte Artefakte von Artefakten, die nicht technisch bedingt sind. Zu letzteren gehören Kopfbewegungen während des Scans, was zu einem systematischen Bias in automatisierter Hirnsegmentierung führen kann (Reuter et al., 2015). Mortamet et al. stellten 2009 ein automatisches System zur Quantifizierung von Scanqualität vor (Mortamet et al., 2009). Eine vorgeschobene Scanqualitätsüberprüfung in dieser oder ähnlicher Form könnte die vorgestellte Methode verbessern. Die Fehlerlast der Segmentierung, die auf scanner- oder patientenassoziierte Artefakte zurückgeht, würde reduziert werden und so zu einer Stärkung der Qualitätsstufe „GUT“ führen. Übrig bleiben mehr Fehler, die ihren Ursprung in einer

Hirnpathologie haben könnten. Eine deutliche Erhöhung der Fallzahl könnte eine quantitative Analyse ebensolcher Fehler in ihrer Verteilung auf die ca. 180 Labels in Assoziation mit zugrundeliegenden Hirnpathologien möglich machen. Grundsätzlich ist es von hohem klinischem Interesse, einen Weg zu finden der von einem Fehlermuster auf eine Pathologie schließen lässt. Um das Bewertungssystem in diese klinisch nutzbare Form weiterzuentwickeln, bedarf es noch einer Reihe von Anpassungen. Nach dem aktuellen Stand würden über die Interpretation von Fehlermuster nur solche Pathologien entdeckt werden, die so schwer sind, dass FreeSurfer mit seiner Segmentierung Probleme in einem großen Ausmaß hat. Dennoch gibt es auch Pathologien, die nicht notwendigerweise zu einem Segmentierungsfehler führen. Eine Erweiterung des Klassifikationssystems, um korrekt segmentierte Pathologien zu detektieren, ist dringend notwendig. Auch hierfür bedarf es einerseits einer größeren Fallzahl und andererseits einer adäquaten neuroradiologischen Expertise für die Evaluation der FreeSurfer-Ergebnisse und deren Einordnung in einen neuropathologischen Kontext. Da mit der oben erwähnten erhöhten Scanqualität sicher nicht alle scanner- oder patientenassoziierten Artefakte eliminiert werden können, sollte als weitere Anpassung der Methode die Trennung von technisch- und pathologiebedingten Fehlern klarer vollzogen werden können.

Auch sollten die Entwicklungen im Bereich maschinelles Lernen insbesondere für die Auswerteschritte mit aufgegriffen werden, da sich die Radiologie hin zum vollautomatisierten Arbeiten entwickelt (Langlotz et al., 2019). Eine Weiterentwicklung des Bewertungssystems für die Qualität von FreeSurfer Segmentierungen hin zu einer algorithmusgestützten vollautomatischen Durchführbarkeit ist zwingende Notwendigkeit, um in der zunehmend vollautomatischen Bildverarbeitung der Neuroradiologie zu bestehen und einen nützlichen Beitrag zu leisten. Außerdem ließe sich so die oben erwähnte Limitation der menschlichen Subjektivität minimieren.

Am 29.4.2020 erschien eine neue Version von FreeSurfer: Stable v7.0.0. Möglicherweise beinhaltet diese Version Verbesserungen der Segmentierung. Die aktualisierte Version konnte nicht mehr getestet werden. In weiteren Studien könnte zudem untersucht werden, in wieweit eine Anpassung der verwendeten FreeSurfer Parameter für die Segmentierung an eine multimorbide Kohorte möglich und sinnvoll wäre, um die Anzahl der Segmentierungsfehler zu senken.

Zusammenfassend liefern bisherige Forschungsergebnisse und die vorliegende Arbeit ein Fundament für die Entwicklung objektiver und effizienter Bildverarbeitung, die den Besonderheiten einer multimorbiden Kohorte mit neurologischen Erkrankungen gerecht wird. Insofern zukünftige Forschungsarbeiten die oben genannten Modifikationen, insbesondere eine Steigerung der Fallzahlen, berücksichtigen, ist davon auszugehen, dass

weiterentwickelte algorithmusbasierte Verfahren einen nützlichen Beitrag für die radiologische Arbeit in Klinik und Forschung liefern werden.

5. Zusammenfassung

Strukturelle Bildgebung und deren vollautomatische Verarbeitung haben in der Neuroradiologie ihren festen Platz und unterliegen einer permanenten Weiterentwicklung. Eingebettet in eine explorative, multizentrische Studie zur Untersuchung des Zusammenhangs von Kognition und Motorik (ComOn) an einer Kohorte von multimorbiden Patienten mit neurologischen Erkrankungen, die einer frührehabilitativen geriatrischen Komplexbehandlung zugeführt wurde, wurden an einer Subgruppe von n=53 standardisierte magnetresonanztomographische T1 Aufnahmen des Gehirns gefertigt. Diese wurden mit dem frei verfügbaren Programm FreeSurfer segmentiert, sodass den T1-Bildpunkten (Voxel) anatomische Regionen zugeordnet werden, und Volumenwerte generiert werden konnten. Da die visuelle Kontrolle der Segmentierungsergebnisse sehr variable Qualität ergeben hat, war es Gegenstand dieser Arbeit, einerseits ein neu entwickeltes Qualitätsbewertungssystem der FreeSurfer Segmentierung vorzustellen und andererseits die Anwendbarkeit von FreeSurfer-Volumetrie an der Kohorte zu überprüfen. Im vorgestellten Verfahren werden Fehler in der Segmentierung zunächst in ihrer Raumausdehnung beschrieben und danach ihre Schwere bewertet. Anhand der Verteilung und der Schwere der Fehler wird das gesamte Segmentierungsergebnisses des Datensatzes eines Probanden einer der drei Kategorien „GUT“, „MITTEL“ oder „SCHLECHT“ zuführt. Angewandt wurde es von zwei Ratern an n = 45 Segmentierungen, wobei einer der beiden der Doktorand selbst und der andere ein erfahrener neuroradiologischer Oberarzt war. Ein für Prävalenz und Bias adjustiertes Cohens Kappa mit quadratischer Wichtung der Kategorien ergab einen Wert von $\kappa = 0.59$, welches nach gängiger Interpretation für eine moderate Übereinstimmung der beiden Rater spricht und die Übertragbarkeit und Anwendbarkeit der vorgestellten Methode bestätigt.

Die Überprüfung der Anwendbarkeit von FreeSurfer als Methode zur Volumetrie an der beschriebenen Kohorte erfolgte an den Volumina von sechs Hirnarealen (Summe aus linker und rechter Hemisphäre für Frontallappen, Temporallappen, Parietallappen, Okzipitallappen, Basalganglien und Kleinhirn). Um auszuschließen, dass ein systematischer Fehler in Form einer Beeinflussung der Volumenwerte durch schlechtere Segmentierungsqualität vorliegt, wurde, je nach Status der Normalverteilung der Variablen, eine einfaktorielle ANOVA bzw. ein Kruskal-Wallis Test zur Varianzanalyse berechnet und die Varianzhomogenität mittels Bartlett- oder Fligner-Killeen Test untersucht. Nach Korrektur für Fehler erster Art mit der Bonferroni-Methode ergaben sich keine statistisch signifikanten Unterschiede der Volumendurchschnitte der sechs Hirnareale hinsichtlich der Qualitätsstufen GUT, MITTEL oder SCHLECHT, sodass auch Segmentierungen von schlechter Qualität in weiterführende volumetrische Untersuchungen eingeschlossen werden können, falls man volumetrischen Fragestellungen auf Ebene der Hirnlappen nachgeht. Eine explorative Analyse der 10 am

häufigsten von Fehlern betroffenen Regionen, sowie des mutmaßlich fehleranfälligen Balkens, ergab ein ähnliches Ergebnis. Hier sollten die Untersuchungen insbesondere auf Areale, die am meisten von schweren Fehlern betroffen sind, ausgeweitet werden, da sich bei der Betrachtung der nicht für multiples Testen korrigierten Werte Hinweise auf eine systematische Beeinflussung der Volumenwerte durch schlechtere Segmentierungsqualität für einzelne Areale (Lobulus parietalis superior, Trend für mittleren posterioren Balken) ergeben hat.

Um das vorgestellte System zur Qualitätsbewertung von FreeSurfer Segmentierungen klinisch nutzbar zu machen, bedarf es noch einiger Weiterentwicklungen. In Zukunft müsste es vollautomatisch durchführbar sein, eine schärfere Differenzierung von technisch- versus pathologiebedingten Fehlern in der Segmentierung gewährleisten und um eine Erfassung von korrekt segmentierten Hirnpathologien erweitert werden. Für diese Entwicklungsschritte sind weitere cMRT Datensätze mit größeren Probandenzahlen notwendig.

6. Literaturverzeichnis

Abramson, J. H. (2004). WINPEPI (PEPI-for-Windows): computer programs for epidemiologists. *Epidemiologic Perspectives & Innovations : EP+I*, *1*(1), 6. https://doi.org/10.1186/1742-5573-1-6

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics*, *34*(5), 502–508. https://doi.org/10.1111/opo.12131

Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry - The methods. *NeuroImage*, *11*(6 I), 805–821. https://doi.org/10.1006/nimg.2000.0582

Barnes, J., Ridgway, G. R., Bartlett, J., Henley, S. M. D., Lehmann, M., Hobbs, N., … Fox, N. C. (2010). Head size, age and gender adjustment in MRI studies: A necessary nuisance? *NeuroImage*, *53*(4), 1244–1255. https://doi.org/10.1016/j.neuroimage.2010.06.025

Barnett, K., Mercer, S. W., Norbury, M., Watt, G., Wyke, S., & Guthrie, B. (2012). Epidemiology of multimorbidity and implications for health care, research, and medical education: A cross-sectional study. *The Lancet*, *380*(9836), 37–43. https://doi.org/10.1016/S0140-6736(12)60240-2

Bojorquez, J. Z., Bricq, S., Acquitter, C., Brunotte, F., Walker, P. M., & Lalande, A. (2017). What are normal relaxation times of tissues at 3 T? *Magnetic Resonance Imaging*, *35*, 69–80. https://doi.org/10.1016/j.mri.2016.08.021

Box, G. E. P. (1953). Non-Normality and Tests on Variances. *Biometrika*, *40*(3/4), 318. https://doi.org/10.2307/2333350

Brenner, H., & Kliebsch, U. (1996). Dependence of Weighted Kappa Coefficients on the Number of Categories. *Epidemiology*, *7*(2), 199–202. Retrieved from http://www.jstor.org/stable/3703036

Brodtmann, A., Pardoe, H., Li, Q., Lichter, R., Ostergaard, L., & Cumming, T. (2012). Changes in regional brain volume three months after stroke. *Journal of the Neurological Sciences*, *322*(1–2), 122–128. https://doi.org/10.1016/j.jns.2012.07.019

Busch, M. A., Schienkiewitz, A., Nowossadeck, E., & Gößwald, A. (2013). Prävalenz des Schlaganfalls bei Erwachsenen im Alter von 40 bis 79 Jahren in Deutschland: Ergebnisse der Studie zur Gesundheit Erwachsener in Deutschland (DEGS1). *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, *56*(5–6), 656–660. https://doi.org/10.1007/s00103-012-1659-0

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. https://doi.org/10.1037/h0026256

Desikan, R. S., Cabral, H. J., Settecase, F., Hess, C. P., Dillon, W. P., Glastonbury, C. M., … Fischl, B. (2010). Automated MRI measures predict progression to Alzheimer's disease. *Neurobiology of Aging*, *31*(8), 1364–1374. https://doi.org/10.1016/j.neurobiolaging.2010.04.023

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., … Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/DOI: 10.1016/j.neuroimage.2006.01.021

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human

cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, *53*(1), 1–15. https://doi.org/10.1016/j.neuroimage.2010.06.010

Ewert, T., & Stucki, G. (2007). Die Internationale Klassifikation der Funktionsfähigkeit, Behinderung und Gesundheit (ICF) : Einsatzmöglichkeiten in Deutschland. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, *50*(7), 953–961. https://doi.org/10.1007/s00103-007-0285-8

Fabbri, E., Zoli, M., Gonzalez-Freire, M., Salive, M. E., Studenski, S. A., & Ferrucci, L. (2015). Aging and Multimorbidity: New Tasks, Priorities, and Frontiers for Integrated Gerontological and Clinical Research. https://doi.org/10.1016/j.jamda.2015.03.013

Fischl, B. (2012). FreeSurfer Authos Manuscript. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021.FreeSurfer

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., … Dale, A. M. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355. https://doi.org/10.1016/S0896-6273(02)00569-X

Fischl, B., Van Der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., … Dale, A. M. (2004). Automatically Parcellating the Human Cerebral Cortex. *Cerebral Cortex*, *14*(1), 11–22. https://doi.org/10.1093/cercor/bhg087

FreeSurferWiki. (2009). Suggested morphometry protocols for optimal FreeSurfer reconstruction. In *FreeSurferWiki* (p. 2). Retrieved from https://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferWiki?action=AttachFile&do=get&target=FreeSurfer_Suggested_Morphometry_Protocols.pdf

Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P., & Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, *6*(2), 67–77. https://doi.org/10.1038/nrneurol.2009.215

Gajawelli, N., Tsao, S., Hwang, D., Wilkins, B., Kriger, S., & Singh, M. (2011). FreeSurfer Parcellation of Brains Containing Large Infarcts.

Geritz, J., Maetzold, S., Steffen, M., Pilotto, A., Corrà, M. F., Moscovich, M., … Maetzler, W. (2020). Motor, cognitive and mobility deficits in 1000 geriatric patients: protocol of a quantitative observational study before and after routine clinical geriatric treatment – the ComOn-study. *BMC Geriatrics*, *20*(1), 45. https://doi.org/10.1186/s12877-020-1445-z

Gifuni, A. J., Olié, E., Ding, Y., Cyprien, F., le Bars, E., Bonafé, A., … Jollant, F. (2017). Corpus callosum volumes in bipolar disorders and suicidal vulnerability. *Psychiatry Research - Neuroimaging*, *262*, 47–54. https://doi.org/10.1016/j.pscychresns.2017.02.002

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Helms, G. (2016). Segmentation of human brain using structural MRI. *Magnetic Resonance Materials in Physics, Biology and Medicine*, *29*(2), 111–124. https://doi.org/10.1007/s10334-015-0518-z

Jenkinson, M., Pechaud, M., & Smith, S. (2005). *BET2 : MR-Based Estimation of Brain, Skull and Scalp Surfaces*. *Eleventh Annual Meeting of the Organization for Human Brain Mapping*. Retrieved from http://ci.nii.ac.jp/naid/10030066593/en/

Klein, A., & Tourville, J. (2012). 101 Labeled Brain Images and a Consistent Human Cortical Labeling Protocol. *Frontiers in Neuroscience*, *6*. https://doi.org/10.3389/fnins.2012.00171

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Langlotz, C. P., Allen, B., Erickson, B. J., Kalpathy-Cramer, J., Bigelow, K., Cook, T. S., … Kandarpa, K. (2019). A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop. *Radiology*, *291*(3), 781–791. https://doi.org/10.1148/radiol.2019190613

Lehmann, S. (2013). Qualitätsvergleich in der funktionellen MRT zwischen 1,5 T und 3 T. Master Arbeit, Martin-Luther-Universität, Halle-Wittenberg

Maetzler, W., Grond, M., & Jacobs, A. H. (2016). Neurogeriatrie (pp. 959–970). https://doi.org/10.1007/978-3-662-46892-0_40

Maetzler, W., Klucken, J., & Horne, M. (2016). A clinical view on the development of technology-based tools in managing Parkinson's disease. *Movement Disorders*, *31*(9), 1263–1271. https://doi.org/10.1002/mds.26673

Malone, I. B., Leung, K. K., Clegg, S., Barnes, J., Whitwell, J. L., Ashburner, J., … Ridgway, G. R. (2015). Accurate automatic estimation of total intracranial volume: A nuisance variable with less nuisance. *NeuroImage*, *104*, 366–372. https://doi.org/10.1016/j.neuroimage.2014.09.034

Marciniewicz, E., Bladowska, J., Podgórski, P., & Sąsiadek, M. (2019). The role of MR volumetry in brain atrophy assessment in multiple sclerosis: A review of the literature. *Advances in Clinical and Experimental Medicine*, *28*(7), 0–0. https://doi.org/10.17219/acem/94137

Masters, C. L., Bateman, R., Blennow, K., Rowe, C. C., Sperling, R. A., & Cummings, J. L. (2015). *Alzheimer's disease*. *Nature Reviews Disease Primers* (Vol. 1). https://doi.org/10.1038/nrdp.2015.56

McCarthy, C. S., Ramprashad, A., Thompson, C., Botti, J. A., Coman, I. L., & Kates, W. R. (2015). A comparison of FreeSurfer-generated data with and without manual intervention. *Frontiers in Neuroscience*, *9*(OCT), 1–18. https://doi.org/10.3389/fnins.2015.00379

Mortamet, B., Bernstein, M. A., Jack, C. R., Gunter, J. L., Ward, C., Britson, P. J., … Krueger, G. (2009). Automatic quality assessment in structural brain magnetic resonance imaging. *Magnetic Resonance in Medicine*, *62*(2), 365–372. https://doi.org/10.1002/mrm.21992

Mugler, J. P., & Brookeman, J. R. (1990). Three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP RAGE). *Magnetic Resonance in Medicine*, *15*(1), 152–157. https://doi.org/10.1002/mrm.1910150117

Pabst, C. (2013). Grundlagen der Magnetresonanz-Tomographie. *Universitätsklinikum Giessen Und Marburg - Lernskript Für Mediziner*. Retrieved from http://www.ukgm.de/ugm_2/deu/umr_rdi/Teaser/Grundlagen_der_Magnetresonanztomographie_MRT_2013.pdf

Reuter, M., Tisdall, M. D., Qureshi, A., Buckner, R. L., van der Kouwe, A. J. W., & Fischl, B. (2015). Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *NeuroImage*, *107*, 107–115.

https://doi.org/10.1016/j.neuroimage.2014.12.006

Rosen, A. F. G., Roalf, D. R., Ruparel, K., Blake, J., Seelaus, K., Villa, L. P., … Satterthwaite, T. D. (2018). Quantitative assessment of structural image quality. *NeuroImage*, *169*(2018), 407–418. https://doi.org/10.1016/j.neuroimage.2017.12.059

Schepkin, V. D., Bejarano, F. C., Morgan, T., Gower-Winter, S., Ozambela, M., & Levenson, C. W. (2012). In vivo magnetic resonance imaging of sodium and diffusion in rat glioma at 21.1 T. *Magnetic Resonance in Medicine*, *67*(4), 1159–1166. https://doi.org/10.1002/mrm.23077

Schick, F. (2006). Sequenzen in der MRT: Teil I. *Radiologe*, *46*(7), 615–630. https://doi.org/10.1007/s00117-006-1364-9

Schünke, M., Schulte, E., Schumacher, U., Voll, M., & Wesker, K. (2018). *PROMETHEUS Kopf, Hals und Neuroanatomie*. Stuttgart: Georg Thieme Verlag. https://doi.org/10.1055/b-006-149644

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143–155. https://doi.org/10.1002/hbm.10062

Thomalla, G., Audebert, H. J., Berger, K., Fiebach, J. B., Fiehler, J., Kaps, M., … Röther, J. (2009). Bildgebung beim Schlaganfall ± eine Übersicht und Empfehlungen des Kompetenznetzes Schlaganfall. *Akt Neurol*, (36), 354–367. https://doi.org/10.1055/s−0029−1220430

Tinetti, M. (2016). Mainstream or Extinction: Can Defining Who We Are Save Geriatrics? *Journal of the American Geriatrics Society*, *64*(7), 1400–1404. https://doi.org/10.1111/jgs.14181

Tinetti, M. E., Bogardus, S. T., & Agostini, J. V. (2004). Potential Pitfalls of Disease-Specific Guidelines for Patients with Multiple Conditions, 2870–2874.

Vogt, B. A., Berger, G. R., & Derbyshire, S. W. G. (2003). Structural and functional dichotomy of human midcingulate cortex. *European Journal of Neuroscience*, *18*(11), 3134–3144. https://doi.org/10.1111/j.1460-9568.2003.03034.x

Vogt, B. A., Vogt, L., & Laureys, S. (2006). Cytology and functionally correlated circuits of human posterior cingulate areas. *NeuroImage*, *29*(2), 452–466. https://doi.org/10.1016/j.neuroimage.2005.07.048

Warrens, M. J. (2013). Weighted kappas for 3 × 3 tables. *Journal of Probability and Statistics*, *2013*. https://doi.org/10.1155/2013/325831

Weishaupt, D., Köchli, V. D., & Marincek, B. (2014). *Wie funktioniert MRI?* Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41616-3

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., & Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage*, *31*(3), 1116–1128. https://doi.org/10.1016/j.neuroimage.2006.01.015

7. Erklärung zum Eigenanteil

Beitrag des Doktoranden zur ComOn Studie:

- Erhebung klinischer Daten (neurogeriatrische Station A2 und neurologische Allgemeinstationen, Klinik für Neurologie, UKSH Campus Kiel) in Form von
    - standardisierte Anamnese
    - Gangparameter mittels tragbarer Sensoren
    - neuropsychologische Testung
    - Restharnmessung
    - Messung der Bioimpedanz
- Einspeisen erhobener Daten in die Forschungsdatenbank „RedCap“
- Initiierung und Organisation der Erhebung standardisierter Bildgebungsdaten in Form des „Kieler Geriatrie Protokolls“ unter Anleitung von Dr. med. Michael Weiß, PD Dr. med. Christian Riedel und Stephan Wolff
- Datenpflege der Bildgebungsdaten unter Anleitung von Oliver Granert

Eigenanteil der vorliegenden Arbeit:

- Eigenständige Entwicklung der Arbeitshypothesen unter enger Betreuung von Prof. Dr. med. Walter Maetzler, Dr. Kirsten Emmert und Oliver Granert
- Betrachtung und Interpretation der FreeSurfer Segmentierungen unter Anleitung von PD Dr. med. Christian Riedel und Oliver Granert
- Eigenständiger Entwurf des Systems zur Bewertung von FreeSurfer Segmentierungen unter Rücksprache mit PD Dr. med. Christian Riedel und Oliver Granert
- Statistische Auswertung und Interpretation der Daten unter enger Betreuung von Dr. Kirsten Emmert und Edjola Naka
- Eigenständige Literatursuche und Implementation in die Arbeit
- Eigenständige Diskussion

8. Danksagungen

Einer Vielzahl von Personen, die mich im Rahmen dieser Doktorarbeit unterstützt haben, gilt mein Dank. Nicht alle können hier namentlich erwähnt werden.

Zunächst möchte ich mich bei meinem Doktorvater Prof. Dr. Walter Maetzler für die enthusiastische und umfassende Betreuung dieser Arbeit bedanken. Zu jederzeit nahm er sich meinen Problemen an und leitete mich mit kreativen Denkanstößen durch die Entstehung dieser Arbeit. Er ermöglichte mir wegweisende Einblicke in Klinik und Forschung. Besonders in Erinnerung bleiben wird mir sein äußerst empathischer, stets zuvorkommender und respektvoller Umgang mit Patienten, Kollegen, Mitarbeitern der Forschungsgruppe, bis hin zu Studenten u.v.m.. Nach einem kurzen Unterricht am Krankenbett mit ihm stand für mich fest, ihn nach einer Möglichkeit zur Promotion zu fragen.

Herzlich möchte ich mich bei meiner Betreuerin Dr. Kirsten Emmert für die hochmotivierte und weitreichende Unterstützung mit dieser Arbeit bedanken. Nicht nur ihr unzähliges Korrekturlesen, sondern vor allem auch der richtige Einfall zur richtigen Zeit und ihr besonderes Maß an Geduld bei sich wiederholenden Fragen zur Statistik, waren eine große Hilfe. Ebenso gilt mein Dank Edjola Naka für die Unterstützung rund um die Statistik.

Besonders bedanke ich mich bei Oliver Granert, der mich herzlich in das Bildgebungslabor aufgenommen hat. Von der Erstellung der FreeSurfer Datensätze, über die vielen konstruktiven Diskussionen und kreativen Denkanstöße, bis hin zu der Einführung in ITK Snap und R Studios, war er eine große Unterstützung und hat in mir so die Begeisterung für die Bildgebung geweckt. Besonders beeindruckt haben mich seine Hilfsbereitschaft, Ruhe und stets präzisen Erklärungen zu komplizierten Technikfragen.

Ein großes Dankeschön auch an Christian Riedel für die neuroradiologische Betreuung. Ebenso an Stephan Wolff für die geduldigen Erklärungen und interessanten Demonstrationen rund um die MRT Physik.

Des Weiteren möchte ich mich bei der ComOn Forschungsgruppe für die herzliche Aufnahme bedanken. Besonders Johanna Gerlitz, Sara Mätzold und Dr. Clint Hansen haben mich auf meinen verschiedenen Einsatzstellen für die ComOn Studie gewissenhaft geleitet.

Außerdem möchte ich mich auch besonders bei meinen Eltern Hubert und Heidi für die vielfältige Unterstützung während der gesamten Studienzeit bedanken.

Meiner Freundin Maren danke ich neben viel Korrekturlesen, ganz besonders für die große Stütze, die sie mir für diese Doktorarbeit, das Studium und unseren gemeinsamen Weg jederzeit war und ist.

Abschließend möchte ich mich bei allen Probanden der ComOn Studie bedanken. Meine Hochachtung gilt diesen betagten Damen und Herren, die trotz einer Vielzahl von medizinischen Problemen und trotz einer ohnehin schon anstrengenden neurogeriatrischen Komplexbehandlung immer noch bereit waren, in den Dienst für die Allgemeinheit zu treten und sich den beengenden und lärmenden Bedingungen im Inneren eines MRT Gerätes aussetzten. Diese Generation, der ich viel mehr als nur Datensätze verdanke, näher kennenlernen zu dürfen, festigte ein weiteres Mal und im besonderen Maß meine Motivation den Beruf des Arztes zu erlernen.

9. Veröffentlichungen

Weiss, M. M., Kress, M., Hobert, M. A., Maetzold, S., Geritz, J., Jansen, O., Wolff, S., Riedel, C., Granert, O., & Maetzler, W. (2018, August). Standardized and quantitative cerebral magnetic resonance imaging in geriatric patients: presentation of study protocol. In ZEITSCHRIFT FUR GERONTOLOGIE UND GERIATRIE (Vol. 51, pp. 124-125). TIERGARTENSTRASSE 17, D-69121 HEIDELBERG, GERMANY: SPRINGER HEIDELBERG.

Kress, M., Oliver, O., Welzel, J., Gerlitz, J., Emmert, K., Naka, E., Riedel, C., & Maetzler, W. (2020, January 10-11*). Fehleranalyse zur Beurteilung der Qualität automatisierter Segmentierungen von cMRT Aufnahmen in einer neurogeriatrischen Kohorte* [Conference poster presentation]. Jahrestagung des Wissenschaftsforums Geriatrie, Berlin, Germany

Geritz, J., Maetzold, S., Steffen, M., Pilotto, A., Corrà, M. F., Moscovich, M., […] , Kress, M., […] , Maetzler, W. (2020). Motor, cognitive and mobility deficits in 1000 geriatric patients: protocol of a quantitative observational study before and after routine clinical geriatric treatment – the ComOn-study. BMC Geriatrics, 20(1), 45. https://doi.org/10.1186/s12877-020-1445-z