Dissertation

Aus dem Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE)

der Ludwig-Maximilians-Universität München

Direktor: Prof. Dr. Ulrich Mansmann

Evaluationsverfahren für eine komplexe Intervention der beruflichen Gesundheitsförderung

Dissertation zum Erwerb des Doktorgrades der Humanbiologie

an der Medizinischen Fakultät der

Ludwig-Maximilians-Universität zu München

vorgelegt von

Dipl.-Stat. Michael Simang

aus München

2015

Mit Genehmigung der Medizinischen Fakultät

der Universität München

Berichterstatter:	Prof. Dr. Ulrich Mansmann
Mitberichterstatter:	Prof. Dr. Klaus Kuhn
	Prof. Dr. Konstantin Strauch
	Prof. Dr. Christian Heumann
Dekan:	Prof. Dr. Dr. h.c. M. Reiser, FACR, FRCR
Tag der mündlichen Prüfung:	14.07.2015

Meiner verstorbenen Mutter gewidmet

Hindernisse sind jene entsetzlichen Dinge, die wir sehen,
wenn wir unsere Augen von unserem Ziel abwenden.

Henry Ford (1863-1947)

Danksagung

Zunächst gilt meine Verbundenheit Herrn Prof. Dr. rer. nat. Ulrich Mansmann für die Möglichkeit, diese Arbeit anfertigen zu können. Darüber hinaus danke ich ihm für die wegweisende Betreuung und fachliche Weiterbildung während der letzten Jahre.

Des Weiteren möchte ich ein Dankeschön an die Siemens Betriebskrankenkasse für die Unterstützung meiner Dissertation richten.

Mein besonderer Dank gilt ferner Frau Dipl.-Stat. Veronika Neppl, die mir gerade in schwierigen Phasen motivierend zur Seite gestanden hat. Dabei leisteten unsere fachlichen Diskussionen ihren Beitrag zur Qualität dieser Arbeit.

Abschließend geht mein Dank an meine Familie, die sich verständnisvoll gezeigt hat und mir mit Unterstützung und Geduld entgegengekommen ist.

Inhaltsverzeichnis

Kapitel 1

Einleitung

In unserer modernen Gesellschaft werden Belastungen durch Rückenschmerzen für immer mehr Menschen zu einem schwerwiegenden Problem. Häufig leiden die Betroffenen nicht lediglich an vorübergehenden, sondern an wiederkehrenden oder gar chronischen Beschwerden. Aufgrund der Komplexität des Symptoms Rückenschmerz mit seiner Vielzahl an möglichen Ursachen gewinnt die Entwicklung neuartiger Therapieverfahren zunehmend an Bedeutung. Im Fokus innovativer Interventionen steht über die Linderung akuter Symptome hinaus eine nachhaltige Verbesserung des Gesundheitszustandes und der Lebensqualität der Patienten.

1.1 Volkskrankheit Rückenschmerz

Als Motivation für die Ausrichtung einer Schmerzintervention auf Rückenschmerzpatienten genügt bereits die Betrachtung der Auftretenshäufigkeit des Krankheitsbildes. Mit einer Lebenszeitprävalenz von ca. 80 % gehören Rückenschmerzen zu den am häufigsten diagnostizierten Erkrankungen. Dabei nimmt die Wahrscheinlichkeit einer Erkrankung mit fortschreitendem Alter zu, wobei Frauen in sämtlichen Altersklassen häufiger betroffen sind als Männer. Eine detaillierte Darstellung der krankheitsspezifischen Aspekte dieses Teilkapitels findet sich in der *Gesundheitsberichterstattung des Bundes* (Robert Koch-Institut (Hrsg.), 2012).

Die wachsende Bedeutung und Präsenz des Krankheitsbildes Rückenschmerz im deutschen Gesundheitswesen ist allerdings nicht allein auf die hohe Prävalenz zurückzuführen. Vielmehr spiegelt sich die Komplexität der Erkrankung hinsichtlich der schwierigen Diagnostik und Behandlung auch im gesundheitsökonomischen Hintergrund wider. Neben direkten Kosten für diverse Untersuchungen und Behandlungen stehen dabei insbesondere die indirekten Kosten im Fokus des Interesses. So findet sich beispielsweise bei den Gesundheits-

kosten aufgrund von Arbeitsunfähigkeiten die Ursache Rückenschmerz an erster Stelle.
Insgesamt steht jede neu entwickelte Rückenschmerzintervention zahlreichen Erwartungen medizinischer und ökonomischer Natur gegenüber. Dementsprechend sollen neben einer Präventionsmöglichkeit und einer akuten Verbesserung des Krankheitsbildes auch langfristige therapeutische Effekte erzielt werden können.

Ausschlaggebend für die Komplexität einer Rückenschmerzbehandlung sind die Schwierigkeiten hinsichtlich der Ursachenermittlung. Bereits der anatomische Aufbau des Rückens bietet vielfältige Möglichkeiten für das Schmerzsymptom. Neben der Wirbelsäule, den Muskeln, den Bändern und den Nervensträngen kann der Schmerz auch durch innere Organe bedingt sein und in den Rücken ausstrahlen. Selbst bei der Identifizierung einer konkreten Ursache (spezifischer Rückenschmerz) existiert keine Garantie für einen Behandlungs- oder gar Operationserfolg. Beispielsweise können die Bandscheiben älterer Menschen eine Deformation aufweisen, ohne für Beschwerden verantwortlich zu sein. Ein operativer Eingriff hätte demnach keinen relevanten Einfluss auf die Symptomatik.
Die überwiegende Anzahl der Rückenschmerzen (ca. 80 %) sind ferner nicht auf eine isolierte organische Ursache zurückzuführen (unspezifischer Rückenschmerz). Entsprechend gestaltet sich eine Behandlung ungleich komplizierter und meist langwieriger, da der Erfolg von der individuellen Konstellation eines Patienten abhängt. Hierbei ist das Risiko für eine Chronifizierung des Schmerzleidens infolge zunehmender Verunsicherung des Patienten in besonderem Maße gegeben. Der Auslöser besteht aus dem Zusammenwirken biologischer, psychologischer und sozialer Faktoren, was sich aus der Perspektive eines Patienten leicht nachvollziehen lässt: Zunächst führt das Auftreten von intensiven Rückenschmerzen zu Funktionseinschränkungen und dem Mobilitätsverlust des Patienten (biologisch). Mit anhaltenden oder häufig wiederkehrenden Beschwerden stellt sich eine erhöhte Sensibilität auf Schmerzsignale ein, welche die Empfindung zusätzlich steigern kann. Infolgedessen entwickeln sich aus einer längerfristigen Problematik und Behandlungsdauer heraus Gefühle der Angst und Hoffnungslosigkeit auf baldige Heilung (psychologisch). Letztlich resultieren aus dem Schmerzleiden der soziale Rückzug und Verhaltensänderungen infolge körperlicher Schonung (sozial).
Um einer derartigen Entwicklung und ihren langfristigen Folgen (Chronifizierung) entgegenzuwirken, finden sich in der Zielsetzung innovativer Schmerztherapien Veränderungen

im gesamten Lebensalltag eines Patienten. So steht neben der Schmerzreduktion insbesondere die Förderung der Eigenverantwortung im Fokus. Von Bedeutung sind ferner die Beibehaltung alltäglicher Aktivitäten und der Erhalt der Arbeitsfähigkeit, wodurch sich ein positiver Einfluss auf die Lebensqualität einstellt.

Durch das Inkrafttreten der *Nationalen Versorgungsleitlinie Kreuzschmerz* (Bundesärztekammer et al., 2010) wurde eine wichtige Grundlage geschaffen, um ein standardisiertes Vorgehen der ärztlichen Versorgung bei Rückenschmerzen zu gewährleisten. Zu einem systematischen Vorgehen gehört unter anderem die Einteilung sämtlicher Vorerkrankungen, Symptome und Risikofaktoren eines Patienten in ein Flaggenmodell gemäß deren Intensität und Tragweite. Gefährlichere Faktoren wie konkrete, behandlungsbedürftige Ursachen (z. B. Tumorerkrankungen, Lähmungserscheinungen) werden durch rote Flaggen gekennzeichnet, wodurch meist das Bild eines spezifischen Rückenschmerzes wiedergegeben wird. Psychologische Auffälligkeiten (z. B. Depressionen, Unzufriedenheit, negative Krankheitseinstellung) werden hingegen als gelbe Flaggen bezeichnet und lassen üblicherweise auf einen nichtspezifischen Rückenschmerz schließen. Anhand der Flaggeneinteilung kann für jeden Patienten eine individuelle und zielgerichtete Therapie initiiert werden.

1.2 Das Projekt Münchner Naturheilkundliches Schmerzintensivprogramm – Rücken

Ein moderner Ansatz einer Rückenschmerzintervention konnte im Rahmen eines *Integrierten Versorgungsprojektes* realisiert werden. Im Allgemeinen basieren derartige Behandlungskonzepte auf der Kooperation verschiedener Institutionen des Gesundheitssystems, wodurch eine qualitätsgesicherte Versorgung der Patienten gewährleistet werden soll. Die Interaktion der beteiligten Kooperationspartner und deren Vernetzung führen dabei zu einem Informationsaustausch, welcher insbesondere bei komplexen Behandlungsprozessen eine organisierte Abfolge der einzelnen Therapieelemente sicherstellt. Zu den wichtigsten Vorteilen solcher koordinierter Behandlungsketten gehört die Vermeidung unnötiger Belastungen der Patienten durch Facharztsuche und Mehrfachbehandlungen, was insbesondere bei chronischen Erkrankungen eine wesentliche Erleichterung darstellt. Ferner wird anhand standardisierter Nachuntersuchungen die Wahrscheinlichkeit eines nachhaltigen Be-

handlungserfolges gesteigert. Für Patienten mit chronischen Rückenschmerzleiden ist dieser Aspekt von besonderer Bedeutung, da die Gefahr eines Zurückfallens in alte Verhaltensweisen minimiert werden kann.

Die Behandlung des Münchner Naturheilkundlichen Schmerzintensivprogramms – Rücken (MNS-R) stellt eine *komplexe Intervention* dar (Definition und Beschreibung siehe Kapitel 2). Die Kombination derartiger Konzepte mit dem Konstrukt der Integrierten Versorgung haben sich als besonders sinnvoll für die Behandlung von Volkskrankheiten wie Diabetes mellitus oder Bandscheibenerkrankungen herausgestellt. Demnach existieren mehrere vergleichbare Projekte wie zum Beispiel die Programme des Berliner Rückenschmerzzentrums und der FPZ in Köln.
Neben einer behandelnden medizinischen Institution haben sich Krankenkassen als geeignete Kooperationspartner erwiesen. Infolgedessen wurde bei der Einrichtung des MNS-R eine Zusammenarbeit zwischen der Klinik für Anästhesiologie der Ludwig-Maximilians-Universität München (LMU) und der Siemens Betriebskrankenkasse (SBK) beschlossen.

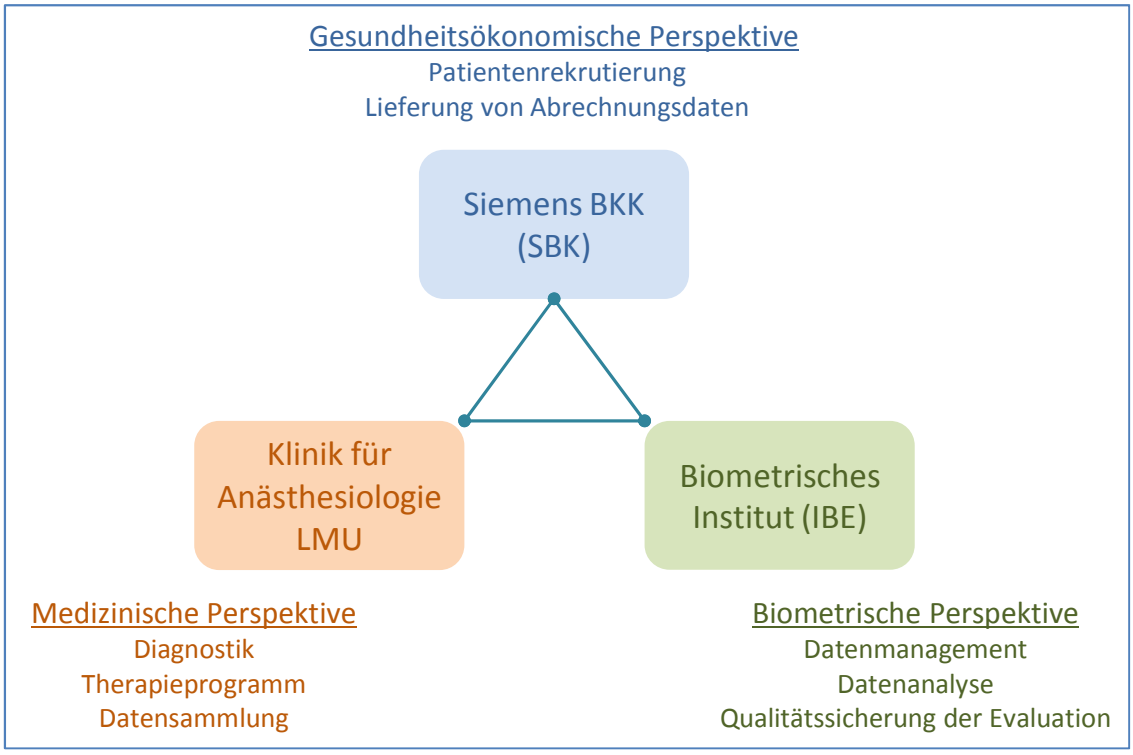

Abbildung 1.2.1: Übersicht der Kooperationsvereinbarungen im Projekt MNS-R

Darüber hinaus erfolgte für die wissenschaftliche Evaluation des Therapieprogramms die Einbindung des Instituts für medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE) der LMU in die Partnerschaft (Abbildung 1.2.1).
Den Fachkompetenzen entsprechend gliedern sich die Anforderungen auf die einzelnen Partner wie folgt auf: Während die Rekrutierung der Patienten und die Sammlung gesundheitsökonomischer Daten seitens der SBK durchgeführt werden, erfolgt die Zusammenstellung einer individuellen Diagnostik und Therapieempfehlung wie auch die medizinische Behandlung der Patienten durch die Klinik für Anästhesiologie (Kapitel 2.1). Die Verarbeitung und Validierung des gesammelten Datenmaterials sowie die Anfertigung einer qualitätsgesicherten Evaluation des Therapieprogramms wird hingegen seitens des IBE gewährleistet (Kapitel 1.3, 2, 3, 5 und 6). Insgesamt ermöglicht dieses Konzept eine unkomplizierte Implementierung und Organisation der Therapie inklusive einer ganzheitlichen Beurteilung des individuellen Gesundheitszustandes jedes einzelnen Patienten. Dabei werden wissenschaftliche Erkenntnisse stichhaltig erarbeitet und bieten eine nachhaltige Grundlage für weitere Untersuchungen.

1.3 Wissenschaftliche Evaluation – gesundheitliche und ökonomische Effektivität

Neben der Durchführung des Schmerztherapieprogramms werden im Rahmen einer Evaluationsstudie die Therapieeffekte genauer quantifiziert. Aus den unterschiedlichen Perspektiven der Kooperationspartner ergeben sich verschiedenartige Fragestellungen, sodass sich die Evaluation des MNS-R in zwei Hauptbestandteile aufgliedert:

Der gesundheitliche Evaluationsteil beinhaltet die Wirksamkeitsprüfung der implementierten Therapie im MNS-R. Zwar konnte eine grundsätzliche Effektivität multimodaler Schmerztherapieprogramme zur Behandlung chronischer Schmerzleiden in zahlreichen wissenschaftlichen Untersuchungen nachgewiesen werden (Guzmán et al., 2002 / Jensen et al., 2005). Allerdings wird der unmittelbare Behandlungseffekt meist anhand von vorher-nachher-Analysen untersucht. Nachdem die Wahrscheinlichkeit einer Reaktivierung bereits überwundener Verhaltensweisen bei chronischen Schmerzpatienten in besonderem Maße gegeben ist, besteht ferner die Notwendigkeit einer langfristigen Beobachtung. Dabei weist

das MNS-R eine intensive Ausrichtung auf die Nachhaltigkeit möglicher Therapieeffekte auf (Kapitel 2.1). Somit wird in der Evaluationsstudie der individuellen Krankheitsentwicklung der Patienten im Langzeitverlauf besondere Bedeutung beigemessen (Kapitel 2.2, 3). Vom biometrischen Standpunkt aus stellt die Überprüfung der Wirksamkeit innerhalb der zugrunde liegenden Patientenstichprobe einen wesentlichen Teil der Untersuchung dar.

Neben dem Interesse hinsichtlich einer positiven gesundheitlichen Entwicklung der Patienten folgt aus der Perspektive der Krankenkasse eine gesundheitsökonomische Fragestellung. Dementsprechend beinhaltet die Evaluation des MNS-R zusätzlich eine kostenanalytische Komponente zur Bewertung des Mehrkostenaufwandes derartiger Therapieformen im Vergleich zu konventionellen Therapiemaßnahmen (Kapitel 6). Im Gegensatz zu anderen Untersuchungen eröffnet sich durch die Kooperation mit der SBK die Möglichkeit, auf objektives Datenmaterial hinsichtlich der Behandlungsgeschichte zurückzugreifen (Kapitel 5). Neben stichhaltigen Informationen zu chirurgischen Eingriffen, Arztbesuchen, Rehabilitationsmaßnahmen und medikamentösen Behandlungsformen erlaubt der Datenbestand die Einbeziehung indirekter Kostenfaktoren wie etwa die Dauer der Arbeitsunfähigkeit. Darauf basierend können die Krankheitsverläufe der Programmteilnehmer jenen konventionell behandelter Patienten gegenübergestellt werden. Infolgedessen bildet die Entwicklung eines Matching-Verfahrens zum Auffinden geeigneter Kontrollpatienten einen wesentlichen Bestandteil der ökonomischen Analyse (Kapitel 5).

Bei der Analyse des ersten Evaluationsteils bedarf es der Anwendung statistischer Instrumente aus dem Bereich der Analyse longitudinaler Daten (Kapitel 3). Nachdem neben parametrischen auch semi-parametrische Techniken eingesetzt werden, ergibt sich die Frage nach einer adäquaten Modellselektion. Insbesondere bei der Verwendung generalisierter Schätzgleichungen bedarf es komplexer Vorgehensweisen für die Beurteilung verschiedener Modellierungen. Eine mögliche Vereinfachung dieser Problematik wird anhand einer Simulationsstudie untersucht (Kapitel 4).

Kapitel 2

Studienbeschreibung des Evaluationsteils I

Im Gegensatz zu medikamentösen Behandlungen stellen komplexe Interventionen eine therapeutische Maßnahme dar, deren Wirkung sich nicht auf eine einzelne Behandlungsmethode stützt. Vielmehr zeichnet sich die Intervention durch mehrere unterschiedliche Komponenten aus, deren Interaktion erst eine therapeutische Wirkung hervorbringt. Dementsprechend ist eine gemeinsame und aufeinander abgestimmte Anwendung der Maßnahmen von zentraler Bedeutung. Die Evaluation einer derart konstruierten Therapieform ist aufgrund der interaktiven Abhängigkeiten zwischen den einzelnen Elementen grundsätzlich mit Schwierigkeiten verbunden (Campbell et al., 2000). So kann der Einfluss einer einzelnen Methode kaum isoliert bewertet werden, da die Modifikation eines Therapiebestandteils gegebenenfalls ausreicht, um die Behandlungseffektivität insgesamt zu verändern.

2.1 Die komplexe Intervention des MNS-R

Basierend auf der Auffassung, dass chronische Schmerzerkrankungen das Resultat eines vielschichtigen Zusammenspiels aus somatischen, kognitiv-emotionalen, behavioralen und sozialen Komponenten sind, stellen komplexe Interventionen eine wirksame Therapiemöglichkeit dar. Die Effektivität dieser Art der Schmerztherapie wird durch zahlreiche wissenschaftliche Studien untermauert (Guzmán et al., 2002 / Jensen et al., 2005). Ein spezielles Konzept auf Basis einer multimodalen Gruppentherapie stellt das Münchner Naturheilkundliche Schmerzintensivprogramm (MNS) dar, welches unabhängig der Schmerzlokalisation und Diagnostik eines Patienten angeboten wird. Das Alleinstellungsmerkmal des MNS bildet die Kombination von Anwendungen und Techniken der klassischen Naturheilkunde, der Traditionellen Chinesischen Medizin (TCM), dem Konzept der Salutogenese, der Psychosomatik und Psychologie sowie der modernen evidenzbasierten Schmerztherapie.

Eine Weiterentwicklung des MNS speziell für Rückenschmerzpatienten (MNS-R) wird in Kooperation mit der Siemens Betriebskrankenkasse durchgeführt (Kapitel 1.2). Je nach Ausprägung der Chronifizierung eines Patienten können drei unterschiedlich intensive Therapievarianten angewendet werden. Ein wesentlicher Aspekt des Erfolgs derartiger Therapieformen besteht aus der Wirkung der Gruppendynamik. Diese entfaltet ihr maximales Moment bei Gruppengrößen von acht Personen (Yalom & Leszcz, 2005). Demzufolge umfassen die einzelnen Therapiegruppen mindestens sechs und maximal zehn Patienten.

Zunächst erfolgt eine Untersuchung der Patienten im Rahmen eines interdisziplinären Assessments durch vier verschiedene Ärzte und Therapeuten (Anästhesisten, Ärzte für Physikalische Medizin, Physiotherapeuten, Psychologen). In der folgenden Schmerzkonferenz wird ein ganzheitliches Bild auf den Patienten und seine Schmerzsymptomatik entwickelt. Im Vordergrund steht auch hier die Erfassung des Patienten im bio-psycho-sozialen Kontext.
Gemäß dem Grad der Chronifizierung kommen drei unterschiedlich intensive Gruppenprogramme zum Einsatz. Das Intensivprogramm (MNS-R 120) läuft über einen Zeitraum von vier Wochen und beinhaltet 120 Stunden. Es ist für Patienten mit einem hohen Grad der Chronifizierung und sehr langer Krankheitsgeschichte konzipiert. Eine reduzierte Version in Form eines zweiwöchigen Programms mit 60 Therapiestunden (MNS-R 60) kann auch als Präventionsmaßnahme eingesetzt werden, sofern ein Patient ein hohes Chronifizierungsrisiko aufweist. Für arbeitsfähige Patienten mit einem niedrigen oder mäßigen Risikoprofil besteht außerdem die Möglichkeit eines berufsbegleitenden Programms (MNS-R 30) in den Abendstunden und am Wochenende. Der Vorteil liegt hierbei in der Vereinbarkeit der Therapie mit der Berufstätigkeit. Das Programm umfasst 30 Therapiestunden und dient vor allem der Senkung des Chronifizierungsrisikos. Neben dem multimodalen Schmerzprogramm erhält der Patient eine individuelle medikamentöse Versorgung (Abbildung 2.1.1).

Das Ziel dieses Konzeptes besteht darin, über eine Linderung der Schmerzen hinaus langfristige Verbesserung der gesamten Lebenssituation insbesondere im Sinne der Lebensqualität zu bewirken. Die Patienten sollen an verschiedene Möglichkeiten, selbst auf ihr Schmerzgeschehen Einfluss nehmen zu können, herangeführt werden. Demnach ist die Vermittlung von Informationen in Seminaren über physiologische und psychologische Aspekte des Schmerzempfindens und über Schmerztherapie im Allgemeinen von Bedeutung.

Eingefahrene Denk- und Verhaltensmuster können dadurch bei den Patienten bereits aufgelöst werden.

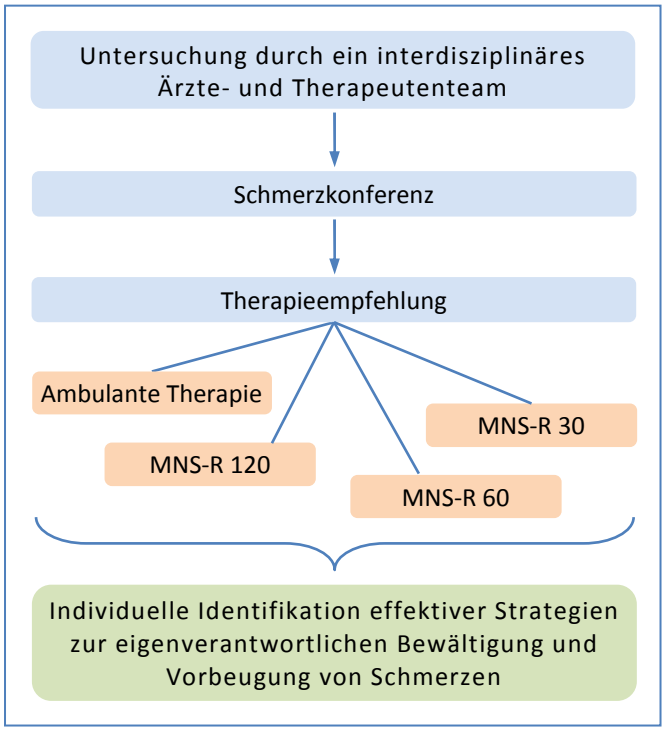

Abbildung 2.1.1: Überblick über den Behandlungsverlauf im MNS-R

Eine der zentralen Anforderungen an das MNS-R besteht in der Steigerung der Körperwahrnehmung. Insbesondere bei chronischen Schmerzerkrankungen wird die körperliche Empfindsamkeit häufig durch das Schmerzgeschehen überlagert. Für die Betroffenen ergeben sich daraus kontinuierliche körperliche Einschränkungen mit nur wenigen Möglichkeiten der aktiven Schmerzreduzierung. Körper- und bewegungsorientierte Verfahren wie Psychotonik, Qigong und Atemtherapie helfen, die Wahrnehmung des eigenen Körpers und somit auch des „Selbst“ im sozialen Kontext zu steigern. Für eine nachhaltige Reaktivierung bewegungsfördernder Aktivitäten ist eine Optimierung des Umgangs mit dem Schmerz unerlässlich. Entspannende und kreative Methoden wie Meditation oder Kunsttherapie unterstützen die Wiedererlangung der differenzierten Wahrnehmung und bieten eine Möglichkeit, der Schmerzempfindung Ausdruck zu verleihen.

Insgesamt führt die eigenständige Ausführung sämtlicher im Therapieprogramm enthaltenen verhaltenstherapeutischen und Körper-Geist-Seele orientierten Maßnahmen zu einer

eigenen Erfahrung des Patienten, welche Methode für ihn persönlich die beste Wirkung erzielt. Nachdem der langfristige Erfolg an die Fortführung und Bewahrung der erlernten Erkenntnisse und Übungen geknüpft ist, wird die aktive Teilnahme an der Gruppentherapie lediglich als eine Stufe (Stufe 1) zum Ziel betrachtet. Diese wird durch zwei weitere Stufen ergänzt, welche die Möglichkeit einer regelmäßigen Wiederholung sicherstellen sollen. Zunächst können Patienten im Rahmen von offenen Gruppen ausgewählte Verfahren systematisch weiterführen (Stufe 2). Als weiteres Angebot stehen Selbsthilfegruppen, Seminare und Vorträge zur Verfügung, um die erlernten Informationen rund um das Thema Schmerz erneut ins Gedächtnis zu rufen (Stufe 3). Ferner ergibt sich durch die Veranstaltung von Nachtreffen der jeweiligen Therapiegruppen in regelmäßigen Abständen eine weitere Begleitung der Programmteilnehmer über einen Zeitraum von zwei Jahren nach Absolvierung des MNS-R. Hierbei bietet sich die Möglichkeit einer ausführlichen Berichterstattung jedes Patienten – verbunden mit einem intensiven Austausch zwischen den Beteiligten untereinander. Durch die ärztliche Betreuung der Gruppentreffen kann eine fachliche Einordnung der Erfahrungen mit weiterführender Beratung sichergestellt werden.

Durch das dreistufige Konzept und eine weiterführende Betreuung ergeben sich optimale Bedingungen, damit die individuell effektiven Therapieansätze über den Behandlungszeitraum hinaus eigenständig fortgeführt werden und so die Nachhaltigkeit der positiven Veränderungen erreicht wird.

2.2 Fragestellung und Studiendesign

Der erste Teil der Evaluation des Rückenschmerzprogramms MNS-R befasst sich ausschließlich mit den gesundheitlichen Veränderungen der Patienten. Dabei stehen die beiden hauptsächlichen Zielsetzungen einer unmittelbaren Effektivität und einer nachhaltigen therapeutischen Wirkung im Fokus des Interesses. Beide Komponenten werden zur Bearbeitung im Rahmen einer klinischen Studie durch folgende Fragestellungen repräsentiert:

i. Lassen sich im vorher-nachher-Vergleich relevante Veränderungen hinsichtlich des Gesundheitszustandes der Teilnehmer feststellen?

ii. Zeichnen sich die individuellen Entwicklungen im zeitlichen Verlauf durch langfristige Veränderungen im Sinne nachhaltiger Therapieeffekte aus?

Insbesondere der sekundäre Endpunkt (ii) ist ausschlaggebend für die Wahl des Studiendesigns. Hierbei wird näher untersucht, ob sich durch das mehrstufige Therapiekonzept zusätzliche Erfolge hinsichtlich der Nachhaltigkeit von Verbesserungen erzielen lassen. Wie bereits erläutert, sind die Therapiestufen 2 und 3 darauf ausgerichtet, die Veränderungen und gewonnenen Erkenntnisse beim Teilnehmer präsent zu halten und zu vermeiden, dass krankheitsbedingte Empfindungs- und Verhaltensmuster reaktiviert werden. Die Erkenntnisse aus einer Vergleichsanalyse zwischen verschiedenen Zeitpunkten hinsichtlich der Zustände von Patienten sind für eine adäquate Beurteilung dieser Fragestellung jedoch unzureichend. Vielmehr sollen die Entwicklungen im zeitlichen Verlauf in die Berechnungen eingeschlossen werden, was die statistischen Methoden der Analyse longitudinaler Daten bewerkstelligen können.
Darüber hinaus liegt der Untersuchung der Vergleich der Gesundheitszustände zum Zeitpunkt vor mit jenem nach der Therapieteilnahme als primärer Endpunkt (i) zugrunde (vorher-nachher-Vergleich). Im Hinblick auf die oben erwähnte Methodik des sekundären Endpunktes besteht dabei keine Notwendigkeit einer separaten Bearbeitung dieser beiden Fragestellungen. Das anzuwendende Analysevorgehen bietet ausreichend Möglichkeiten, diese prä-post-Analyse unkompliziert in die longitudinalen Berechnungen zu integrieren. Hierfür wird der relevante langfristige Beobachtungszeitraum durch einen zusätzlichen Messzeitpunkt vor der Therapieteilnahme erweitert. Eine detaillierte Erläuterung des Analysevorgehens findet sich in Kapitel 3.

Da aus organisatorischen Gründen auf die Bildung einer Kontrollgruppe im ersten Evaluationsteil verzichtet werden muss, erfolgt die Durchführung einer prospektiven Beobachtungsstudie. Für die Beurteilung der Veränderungen im zeitlichen Verlauf wird die gesundheitliche Entwicklung jedes Patienten einheitlich dokumentiert. Hierzu sind Informationen hinsichtlich des individuellen Zustandes an verschiedenen Zeitpunkten zu einem gesamten Datenbestand zusammenzutragen. Neben krankheitsspezifischen Merkmalen wie der Schmerzstärke oder der körperlichen Einschränkung liegt der Schwerpunkt auf Informationen bezüglich der Lebensqualität oder der psychischen Verfassung der Teilnehmer.

2.3 Datenerhebung

Die Sammlung der relevanten Informationen erfolgt anhand verschiedener Fragebögen, welche von den Patienten bearbeitet werden. Zur Erfassung der direkten Therapiewirkung findet eine Erhebung vor der Erstvorstellung im interdisziplinären Assessment sowie direkt nach Absolvierung des MNS-R-Programms statt. Ferner gilt es zu erfahren, wie sich die Programmteilnahme auf den weiteren Krankheitsverlauf auswirkt. Hierfür erfolgen weitere Befragungen im Abstand von jeweils drei, sechs, zwölf und 24 Monaten nach Programmteilnahme (Abbildung 2.3.1).

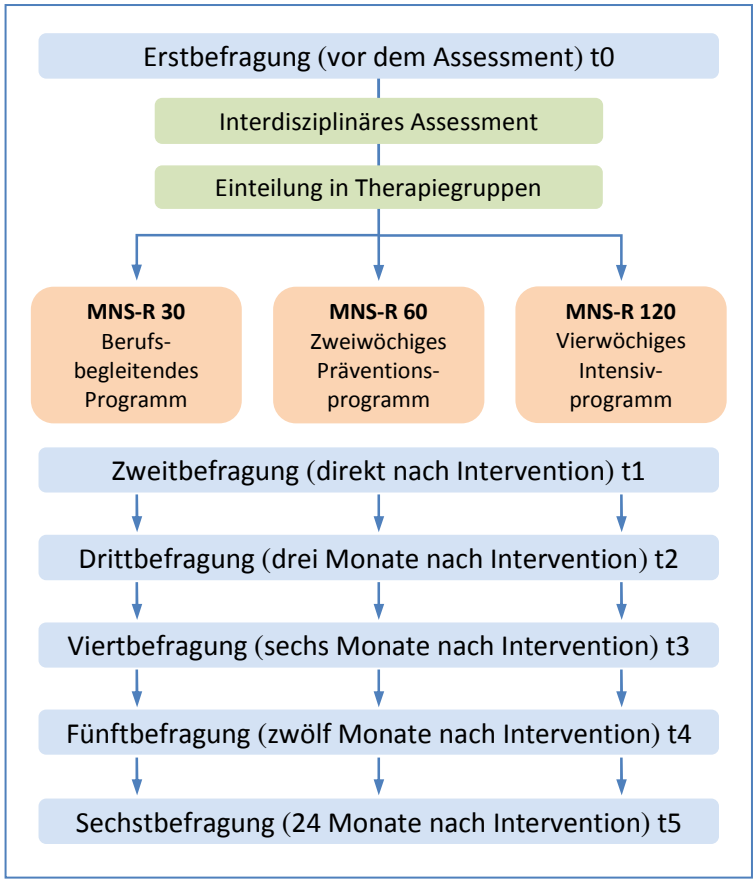

Abbildung 2.3.1: Überblick über die Befragungszeitpunkte der Patienten des MNS-R

Zur Gewährleistung einer umfangreichen Erfassung und Vergleichbarkeit der Patientensituation während des gesamten Studienzeitraums werden sämtliche Fragebögen zu allen Beobachtungszeitpunkten erhoben.

Für die Durchführung der Evaluationsstudie ist es primär nicht von Bedeutung, welcher Therapiegruppe ein Teilnehmer zugeteilt wird. Ebenso ist das Ergebnis der Voruntersuchungen oder Unterschiede hinsichtlich der medikamentösen Versorgung nicht relevant. Vielmehr stellt aus biometrischer Sicht die Gesamtheit sämtlicher therapeutischer Aktivitäten der komplexen Intervention bis hin zu individuellen Patientengesprächen eine komplette, in sich abgeschlossene Intervention dar.

Wie bereits erläutert, basiert der Ansatz der interdisziplinären Diagnostik und multimodalen Therapie auf dem Verständnis, dass chronische Schmerzen nicht auf eine isolierte Ursache zurückzuführen sind. Vielmehr ist der Patient als Ganzes zu verstehen, wodurch chronische Leiden als Ergebnis des Zusammenwirkens mehrerer verschiedener Aspekte angesehen werden können. Diese lassen sich in vier Hauptkategorien eingliedern: die soziale, die somatische, die kognitiv-emotionale und die behaviorale Komponente (Abbildung 2.3.2).

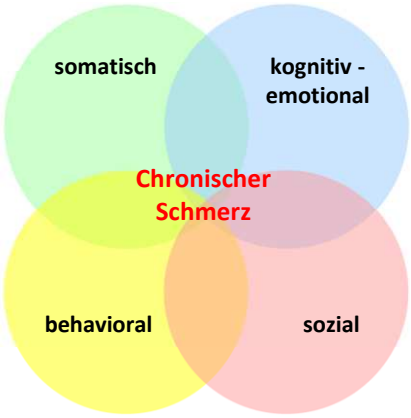

Abbildung 2.3.2: Überblick über Komponenten des chronischen Schmerzes

Für die Beurteilung der Entwicklung eines Patienten in jedem der vier Teilbereiche werden mehrere standardisierte und voneinander unabhängige Fragebogenpakete herangezogen. Einerseits führt dieses Vorgehen zu einem individuellen Datenbestand jedes Programmteilnehmers, auf dessen Grundlage sich Veränderungen in jedem Bereich separat untersuchen lassen. Andererseits bringt das Ausfüllen von Fragebögen einen therapeutischen Vorteil mit sich: Die Patienten werden sich mitunter anhand der Fragestellungen selbst über die Auswirkungen und verschiedenen Facetten ihrer Schmerzempfindung bewusst. Darüber hinaus können viele Erlebnisse der Krankheitsgeschichte, Vorerkrankungen und Begleiterschei-

nungen erneut vor Augen geführt werden, was die Anamnese seitens des Arztes erleichtert und insbesondere einer exakten Einschätzung der Chronifizierung dienlich ist.

2.3.1 Deutscher Schmerzfragebogen (DSF)

Die Grundlage der Befragung und Datenerhebung bildet der standardisierte „Deutsche Schmerzfragebogen (DSF)“ der Deutschen Schmerzgesellschaft. Dieser umfasst verschiedene Fragen rund um die gesundheitliche, soziale und berufliche Situation, wodurch eine umfangreiche Beleuchtung sämtlicher Lebensbereiche eines Patienten erfolgt. Ferner sind einzelne standardisierte und validierte Fragebögen Bestandteil des DSF, welche anhand einer vorgegebenen Auswertungsroutine einen jeweiligen Scorewert hervorbringen. Aufgrund des festen Wertebereichs jedes Scores dient der Individualwert eines Befragten auch als Maßzahl für den zugrunde liegenden Bereich.

Darüber hinaus werden weitere, ebenfalls standardisierte Fragebögen herangezogen, um die somatische, die kognitiv-emotionale, die behaviorale und die soziale Komponente separat und detailliert zu beurteilen (Tabelle 2.3.1). Einerseits helfen die zusätzlichen Angaben in Bezug auf den Umgang mit den Schmerzempfindungen, die aktuelle Situation des Patienten noch genauer zu verstehen. Andererseits vervollständigen die Angaben hinsichtlich der psychischen Beeinträchtigung durch die chronische Erkrankung und der damit verbundenen zukunftsbezogenen Lebensperspektiven das Bild des Programmteilnehmers.

Im DSF enthalten	Zusätzliche Fragebögen
Schmerzempfindungsskala affektiv	Unsicherheitsfragebogen
Schmerzempfindungsskala sensorisch	Lebenszufriedenheit
Allgemeine Depressionsskala (ADS)	FFB-H-R
Pain Disability Index (PDI)	FABQ-D
SF36 – Körperliche Summenskala	FESV
SF36 – Psychische Summenskala	CPAQ-D

Tabelle 2.3.1: Überblick über die verwendeten Fragebögen zur Evaluation des MNS-R

Anhand des Datenbestandes werden nach der Patientenbefragung sämtliche in Tabelle 2.3.1 aufgeführten Scores berechnet, deren individuelle Auswertung im Folgenden kurz erläutert wird. Das meist zur Anwendung kommende Procedere verlangt zunächst die Kodierung sämtlicher im Fragebogen enthaltenen Antworten. Die Kodierungswerte werden durch verschiedenartige Berechnungen zu einem Scorewert für jeden Befragten zusammengefasst. Meist besteht diese Berechnung aus dem Aufsummieren sämtlicher Antworten. Sofern eine oder mehrere Fragen nicht beantwortet wurden, kann kein Scorewert für den Patienten ermittelt werden.

2.3.2 Pain Disability Index (PDI)

Der PDI erfasst die schmerzbedingte, subjektive Beeinträchtigung in sieben Lebensbereichen. Zu jedem der Bereiche wird eine Frage gestellt, welche auf einer ganzzahligen Skala von 0 bis 10 beantwortet werden kann. Diese werden zu einem Gesamtscore mit insgesamt maximal erreichbaren 70 Punkten aufsummiert, was die maximale Beeinträchtigung wiedergeben würde. Keine Beeinträchtigung wird mit einem Wert von 0 deutlich. Die Fragen des PDI beziehen sich auf folgende Lebensbereiche:

- Familiär-häuslicher Bereich
- Erholung
- Soziale Aktivitäten
- Beruf
- Sexualleben
- Selbstversorgung
- Lebensnotwendige Tätigkeiten

2.3.3 Schmerzempfindungsskala (SES)

Die SES enthält insgesamt 28 Fragen mit einer Antwortskala von 1 bis 4 und erlaubt die Messung und differenzierte Beschreibung der subjektiv wahrgenommenen Schmerzen. Die Auswertung der Fragen erfolgt durch Summierung der einzelnen Responsewerte, wobei aus dem Fragebogen zwei unterschiedliche Merkmale gewonnen werden.
Das erste Merkmal umfasst die Fragen 1 bis 14 und beschreibt die affektive Schmerzemp-

findung (14-56 Punkte). Das zweite Merkmal liefert eine Aussage bezüglich der sensorischen Schmerzempfindung, welche mittels der Fragen 15 bis 25 ermittelt wird (10-40 Punkte). Dabei äußert sich eine positive Veränderung durch eine Reduzierung des Wertes des jeweiligen Subscores. Die übrigen vier Fragen gehen nicht in die Auswertung ein.

2.3.4 Allgemeine Depressionsskala (ADS)

Die ADS umfasst 20 Fragen und misst das Vorhandensein und die Dauer von Beeinträchtigungen durch depressive Gemütszustände und negative Denkmuster. Die Auswertung erfolgt durch Summieren der einzelnen Antwort-Scores, welche von 0 bis 3 kodiert werden. Der Wertebereich der Gesamtpunktzahl umfasst folglich eine Spanne von 0 bis 60 Punkten. Die Annahme einer depressiven Erkrankung wird bei einem Scorewert über 23 Punkten getroffen. Demnach kann eine Reduzierung des Wertes als Verbesserung angesehen werden.

2.3.5 Short Form 36 Health Survey Questionnaire (SF-36)

Der SF-36 ist ein standardisierter Fragebogen zur Beurteilung des allgemeinen Gesundheitszustandes und der Lebensqualität. Er besteht aus 36 Fragen mit unterschiedlichen Antwortskalen zum körperlichen Gesundheitszustand, zum seelischen Befinden und zur Schmerz- und Gefühlssituation. Die Auswertung erfolgt durch komplexe Berechnungen, aus denen acht Größen gewonnen werden: Die körperliche Funktionsfähigkeit, die körperliche Rollenfunktion, die körperlichen Schmerzen, die allgemeine Gesundheitswahrnehmung, die Vitalität, die soziale Funktionsfähigkeit, die emotionale Rollenfunktion und das psychische Wohlbefinden. Jede Größe ist normiert und verfügt über einen Wertebereich zwischen 0 und 100 Punkten. Darüber hinaus werden alle Ergebnisse zusätzlich in zwei Subscores zusammengefasst, der *Standardisierte Körperliche Summenskala*, und die *Standardisierte Psychische Summenskala*. Diese bewegen sich ebenfalls zwischen 0 und 100 Punkten. Zur Analyse der Patientenpopulation des MNS-R werden diese beiden Summenskalen herangezogen. Hier bewirkt eine Verbesserung des Patientenzustandes eine Erhöhung der Werte.

2.3.6 Lebenszufriedenheitsfragebogen (FLZ)

Der FLZ liefert eine Aussage zur Zufriedenheit im Hinblick auf die aktuelle Lebenssituation und besteht aus einem allgemeinen und einem gesundheitsspezifischen Teil. Der allgemeine Bereich bezieht sich auf das soziale Umfeld und die Lebenssituation eines Patienten, wogegen der gesundheitsspezifische Teil die subjektiven Fähigkeiten und Empfindungen misst. Ersterer besteht aus 17, zweiterer aus 16 Fragen unter der Verwendung einer siebenstufigen Antwortskala (von „sehr unzufrieden“ bis „sehr zufrieden“), deren Auswertung entsprechend zwei Subscores als Ergebnis hervorbringt. Diese weisen einen Wertebereich zwischen -96 und 160 auf. Je höher der Wert des jeweiligen Subscores ist, desto positiver bewertet der Patient seine Situation im Verhältnis zur subjektiven Priorität des entsprechenden Lebensbereiches.

2.3.7 Unsicherheitsfragebogen (UFB)

Die Erhebung des UFB dient der Beurteilung des Selbstbewusstseins im sozialen Umgang und wird dazu eingesetzt, die behandlungsbedürftige Selbstunsicherheit eines Patienten zu diagnostizieren. Hierzu werden 65 Fragen mit einer Antwortskala von 0 bis 5 gestellt, welche in sechs Subskalen aufgeteilt werden. Diese sind je nach Thema positiv oder negativ gepoolt und geben Einschätzungen über folgende Bereiche:

- Fehlschlag- und Kritikangst (0 bis 75 Punkte)
- Kontaktangst (0 bis 75 Punkte)
- Fähigkeit, fordern zu können (0 bis 65 Punkte)
- Nicht-Nein-Sagen können (0 bis 50 Punkte)
- Schuldgefühle (0 bis 25 Punkte)
- Anständigkeit (0 bis 25 Punkte)

2.3.8 Fear-Avoidance-Belief-Questionnaire (FABQ-D)

Anhand des FABQ-D kann die angstbedingte Vermeidungshaltung bezüglich physischer Aktivität und Arbeit erfasst werden. Dies erfolgt durch die Beantwortung von 16 Fragen unter der Verwendung einer siebenstufigen Antwortskala, welche die Aussagen zwischen „stimmt gar nicht“ und „stimmt genau“ abstuft. Die Auswertung gliedert sich in folgende drei Bereiche und erfolgt über die Summierung der Responsewerte aus jeweils fünf Fragen.

1. Beruf als Ursache von Rückenschmerzen (0 bis 30 Punkte)
2. Prognose über die Wiederaufnahme der Berufstätigkeit (0 bis 30 Punkte)
3. Angstvermeidung bzgl. genereller Aktivität (0 bis 30 Punkte)

Eine Frage bleibt dabei generell ohne Beachtung. Eine Reduzierung der jeweiligen Werte zeigt eine Verbesserung in dem entsprechenden Bereich an.

2.3.9 FESV

Der FESV deckt mit 38 Fragen folgende neun Bereiche der Schmerzbewältigungsfähigkeit eines Patienten ab.

- Handlungs-Planungs-Kompetenzen (4 bis 24 Punkte)
- Kognitive Umstrukturierung (4 bis 24 Punkte)
- Kompetenzerleben (4 bis 24 Punkte)
- Mentale Ablenkung (4 bis 24 Punkte)
- Gegensteuernde Aktivitäten (4 bis 24 Punkte)
- Ruhe- und Entspannungstechniken (4 bis 24 Punkte)
- Schmerzbedingte Hilflosigkeit und Depression (5 bis 30 Punkte)
- Schmerzbedingte Angst (4 bis 24 Punkte)
- Schmerzbedingter Ärger (5 bis 30 Punkte)

Die Fragen können mittels einer sechsstufigen Skala von „stimmt vollkommen“ bis „stimmt überhaupt nicht“ beantwortet werden. Für jeden Bereich wird eine Outcome-Variable durch Summierung von vier bzw. fünf Antwort-Scores berechnet. Je höher der Wert einer Variable, desto stärker ist der entsprechende Mechanismus der Schmerzkontrolle bei dem Patienten ausgeprägt.

2.3.10 Funktionsfragebogen Hannover (FFB-H-R)

Der Funktionsfragebogen gibt eine Einschätzung über die schmerzbedingte Funktionskapazität des Patienten bei der alltäglichen Verrichtung seiner Aufgaben. Er ermöglicht eine Wiedergabe bereits leichter bis mäßiger Funktionseinschränkungen. Hierbei wird aus zwölf Fragen mit dreistufiger Antwortskala ein Merkmal durch Summation der Responsewerte gebildet. Diese Größe bewegt sich zwischen 0 und maximal 24 Punkten, wobei 0 Punkte einer Behinderung von 100 % entspricht – die maximale Punktzahl kann bei einer vollen Funktionsfähigkeit erreicht werden.

2.3.11 CPAQ-D

Der CPAQ-D ermöglicht eine Einschätzung der Schmerz- und Aktivitätsbereitschaft des Patienten. Dazu werden ihm 20 Fragen mit einer Antwortskala von 0 bis 6 gestellt und den beiden Bereichen zugeordnet. Bei der Schmerzbereitschaft gehen die Antworten von acht Fragen ein, wodurch die errechneten Werte zwischen 0 und 48 Punkten liegen. Zur Einschätzung der Aktivitätsbereitschaft sind zehn Fragen maßgeblich, was zu einem Wertebereich zwischen 0 und 60 Punkten führt. Die Auswertung erfolgt wiederum mittels Summierung. Je höher die entsprechende Summe, desto höher ist die Bereitschaft zur Aktivität oder zur Schmerzempfindung. Zwei der gestellten Fragen finden keine Verwendung.

2.3.12 Chronifizierungsgrad

Der Chronifizierungsgrad nach Gerbershagen wird anhand verschiedener Informationen aus dem Deutschen Schmerzfragebogen (DSF) ermittelt. Im Wesentlichen handelt es sich dabei um Eigenschaften des Schmerzleidens, das Auftreten und die Lokalisation betreffend. Des Weiteren sind Angaben zur Medikamenteneinnahme und Behandlungsgeschichte relevant (Abbildung 2.3.3).

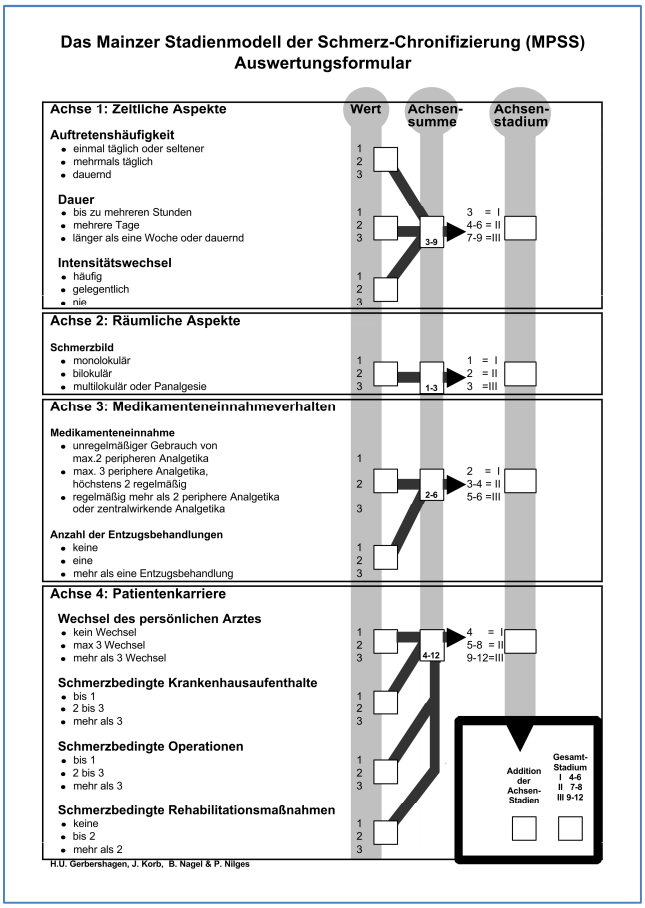

Abbildung 2.3.3: Vorgehen bei der Bestimmung des Chronifizierungsgrades

Die Werte der Antworten werden für jeden Bereich summiert und gemäß dem Resultat einer von drei Kategorien (Grad I, Grad II, Grad III) zugeteilt. Dabei weisen Patienten, welche dem Grad I angehören, lediglich eine geringe Chronifizierung auf, wobei relevante Risikofaktoren erfüllt werden. Befragte mit Chronifizierungsgrad III blicken in der Regel bereits auf eine langjährige Krankheitsgeschichte zurück.

2.3.13 Datenaufbereitung

Sämtliche vorgestellte Fragebögen werden zu jedem Erhebungszeitpunkt in Papierform von den Programmteilnehmern, wie eingangs beschrieben, bearbeitet. Um eine ausreichende Anonymisierung sicherzustellen, erfolgt die Vergabe einer Identitätsnummer für jeden Patienten. Sämtliche vorliegende Fragebögen werden damit versehen und digitalisiert, wodurch eine langfristige Archivierung sichergestellt werden kann.

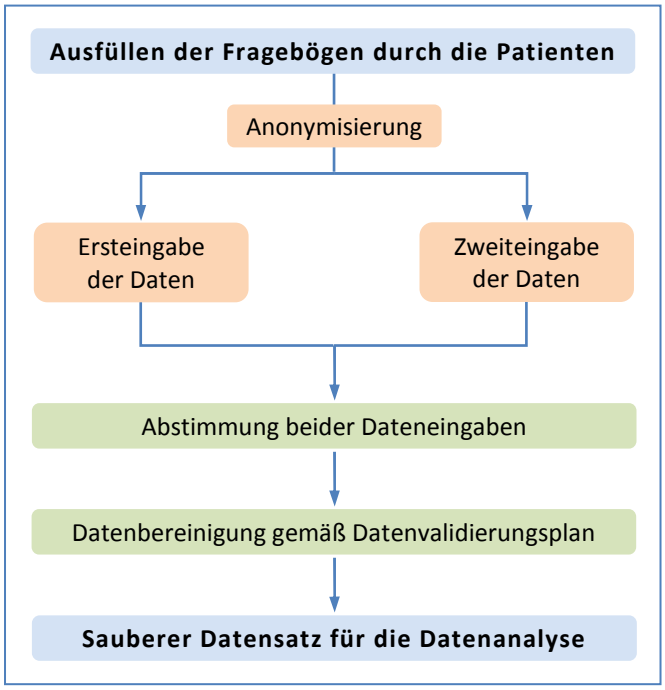

Abbildung 2.3.4: Digitalisierung der Fragebögen und Datenvalidierung

Des Weiteren erlaubt dieses Vorgehen die Datenerfassung anhand eines automatischen Systems. Zwei voneinander unabhängige Dateneingaben, bei denen die Erfassungsprozedur von unterschiedlichen Personen kontrolliert wird, gewährleisten eine Minimierung der Anzahl möglicher Fehleingaben. Durch den Vergleich der resultierenden Datenbanken aus Erst- und Zweiteingabe können anschließend Unstimmigkeiten identifiziert und aufgeklärt werden. Ein weiteres Instrument der qualitätssichernden Datenbereinigung stellt der eigens entwickelte Datenvalidierungsplan dar, anhand dessen unplausible oder auffällige Daten ermittelt und korrigiert werden können. In einigen Fällen erweist sich der Vergleich der Antworten eines Patienten mit jenen aus älteren und neueren Fragebögen als hilfreich, um die Plausibilität besser einschätzen zu können. Insgesamt hat sich insbesondere während der Datenbereinigung eine digitale Archivierung der Fragebögen bewährt. Jederzeit können so Unklarheiten oder Auffälligkeiten kurzfristig und unkompliziert anhand der Originalbögen kontrolliert werden. Zur Gewährleistung eines qualitätsgesicherten Vorgehens bei der Datenaufbereitung ist besondere Sorgfalt auch hinsichtlich der Validierung von Prüfprogrammen und Eingabemasken erforderlich.

Analyse longitudinaler Daten

Das in Kapitel 2.2 vorgestellte Studiendesign des MNS-R zeichnet sich neben der Untersuchung von unmittelbaren Therapieeffekten (vorher-nachher-Vergleich) insbesondere durch einen längerfristigen Beobachtungszeitraum nach der Intervention aus. Aufgrund der mehrfachen Datenerhebung während dieses Zeitraums lässt sich von jedem einzelnen Patienten ein individuelles Entwicklungsprofil im zeitlichen Verlauf zusammenstellen. Diese als longitudinal bezeichnete Datenstruktur enthält spezifische Anforderungen an die Analyse, wodurch die Anwendung speziell entwickelter Verfahren notwendig wird.

3.1 Problemstellung der Datenstruktur

Eines der grundlegenden Unterscheidungskriterien für statistische Standardverfahren richtet sich nach der Abhängigkeit der zu analysierenden Daten. Unabhängige Daten zeichnen sich meist durch vollständig voneinander getrennte Fälle aus und liegen beispielsweise einem Vergleich von Subpopulationen innerhalb einer Erhebung zugrunde. Im Gegensatz dazu stehen Daten eines einzelnen Individuums in Abhängigkeit zueinander, was meist in Form einer wiederholten Messung auftritt. Dies lässt sich bereits an einem trivialen Beispiel leicht nachvollziehen: Ein Patient weist bei einer Untersuchung einen erhöhten Blutdruckwert auf. Die Wahrscheinlichkeit für einen ebenfalls erhöhten Wert bei einer zweiten Messung kurze Zeit später ist größer als für einen Blutdruckwert auf niedrigem Niveau. Infolgedessen existiert in einer Stichprobe im Falle einer Messwiederholung eine positive Korrelation zwischen den erhobenen Daten. Ferner sollte sich vergegenwärtigt werden, dass die Abhängigkeit zwischen zwei Messungen umso stärker ist, je weniger Zeit zwischen den Messzeitpunkten vergangen ist (Liang & Zeger, 1993). Der Verzicht auf eine Integration dieser Korrelationen in statistische Analysen führt zu einer Überschätzung der Varianzen und folglich zu falschen Ergebnissen, was leicht für den *i*-ten Fall $(i = 1, \ldots, n)$ formal darzustellen ist:

$$Var(Y_{i2} - Y_{i1}) = Var(Y_{i1}) + Var(Y_{i2}) \underbrace{- 2 \cdot Cov(Y_{i2}; Y_{i1})}_{\text{Term entfällt bei Ignoranz der Korrelation}} \quad (1.1)$$

Eine Messung mit einer einzelnen Wiederholung wird als gepaarte Stichprobe bezeichnet. Nachdem die Standardmethoden für unabhängige Daten die Korrelation zwischen den Messungen nicht berücksichtigen, werden zur Datenanalyse entsprechend angepasste Verfahren verwendet. Im Falle gepaarter Stichproben existieren beispielsweise eigene Teststatistiken, welche den Abhängigkeiten zwischen den Daten Rechnung tragen.

Die Verallgemeinerung dieser Problemstellung auf mehrere Wiederholungen innerhalb einer Datenerhebung wird als longitudinale Datenstruktur bezeichnet. Demnach kann diese als Erweiterung einer gepaarten Stichprobe auf mindestens drei Beobachtungszeitpunkte verstanden werden, wobei deren Realisationen jeweils untereinander positiv korreliert sind (Pan et al., 2000). Zudem lässt sich durch die chronologische Abfolge der Messzeitpunkte j $(j = 1, \ldots, m)$ für jeden Patienten i eine individuelle Entwicklung im zeitlichen Verlauf nachvollziehen. Entsprechend ergibt sich die Frage nach einem adäquaten Analysevorgehen unter Berücksichtigung der Korrelationen zwischen sämtlichen Erhebungen.

Die paarweisen Vergleiche der Messzeitpunkte stellen eine naheliegende, aber sehr nachteilige Möglichkeit dar, die Korrelationen in die Berechnungen einzubeziehen. Einerseits müssen hierbei sämtliche Kombinationen der Erhebungen miteinander verglichen werden, wodurch bereits bei wenigen Zeitpunkten eine Vielzahl an Berechnungen notwendig wird. Andererseits ist ein derartiges Vorgehen mit einem sehr hohen Anteil an nicht genutztem Informationsgehalt verbunden, da der Verlauf über die Zeit jedes einzelnen Patienten nicht erfasst werden kann. Eine adäquate Vorgehensweise sollte folglich die zusammenfassende Untersuchung sämtlicher Messzeitpunkte in einem Analyseschritt ermöglichen. Bedingt durch die zugrunde liegenden Fragestellungen ist es häufig auch notwendig, neben den Abhängigkeiten in den Daten auch die Unabhängigkeit zwischen den Individuen in die Analyse einzubeziehen. Dadurch lässt sich beispielsweise untersuchen, ob einzelne Patientengruppen eine unterschiedliche Entwicklung aufweisen und auf welche Einflussgrößen dies ggf. zurückgeführt werden kann.

Insgesamt lässt sich bereits durch diese wenigen Überlegungen festhalten, dass die Analyse longitudinaler Daten nicht durch die Bereitstellung eines angepassten Instruments bewerkstelligt werden kann, sondern dass es einer besonderen Vorgehensweise bedarf. Diese wird im Folgenden anhand des Beispiels des Pain Disability Index (PDI) erläutert. Zur Beschreibung der Datenstruktur und der erhobenen Merkmale sei auf Kapitel 2.3 verwiesen.

3.2 Deskriptive Methoden

Die Anforderungen an die Analyse longitudinaler Daten lassen sich im Wesentlichen in zwei Ziele zusammenfassen (Fitzmaurice et al., 2004 - Kapitel 2.2):

i. Eine adäquate Charakterisierung des individuellen Verlaufes eines Patienten über die Zeit (within-individual changes).

ii. Die Untersuchung von intra-individuellen Unterschieden inklusive möglicher Einflüsse durch Kovariablen (between-subject).

Entsprechend der üblichen Vorgehensweise bei statistischen Analysen besteht der erste Schritt in der deskriptiven Beurteilung der vorliegenden Daten. Im Falle stetiger Merkmale wie dem PDI umfasst diese die Berechnung von Mittelwerten, Standardabweichungen, Medianen sowie Minima und Maxima für jeden Erhebungszeitpunkt (Tabelle 3.2.1).

		Pain Disability Index (PDI)					
		Gültige N	Mittelwert	Median	Standard-abweichung	Minimum	Maximum
Time	t0	95	33.20	34.00	14.75	5.00	68.00
	t1	76	23.99	23.00	14.96	3.00	58.00
	t2	64	20.84	16.00	15.03	2.00	57.00
	t3	51	21.33	17.00	14.80	0.00	62.00
	t4	24	23.58	20.00	17.40	0.00	59.00

Tabelle 3.2.1: Numerische Entwicklung des PDI im zeitlichen Verlauf t0 bis t4

Anhand der Resultate lässt sich eine vage Vorstellung der Entwicklung der Patientenpopulation über die Zeit gewinnen. Nach der Intervention (t1) und im weiteren Verlauf haben

sich die Werte deutlich reduziert, wobei sich gegen Ende des Beobachtungszeitraumes (t4) ein leichter Anstieg des PDI zeigt.

Für eine präzise Analyseplanung bei longitudinalen Daten ist neben der deskriptiven Untersuchung die Betrachtung sämtlicher individueller Verläufe in einem Zeitdiagramm unerlässlich. Dabei wird die Entwicklung einer Variable über die Zeit für jeden einzelnen Patienten dargestellt (Abbildung 3.2.1).

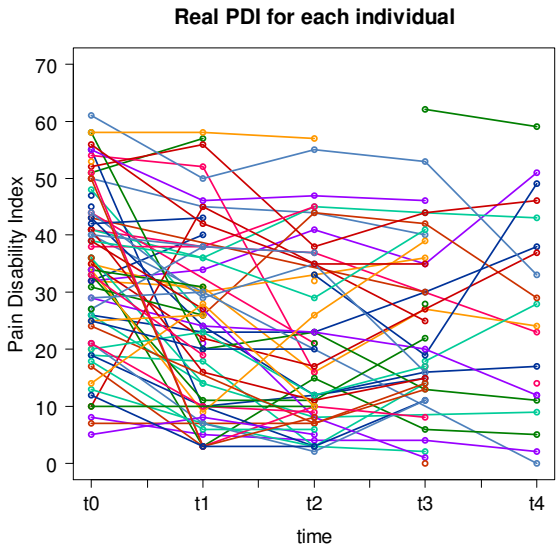

Abbildung 3.2.1: Reale PDI-Werte im zeitlichen Verlauf t0 bis t4

Aus der resultierenden Formation aller Verläufe können zahlreiche Schlussfolgerungen gezogen werden. Neben markanten Auffälligkeiten, welche gegebenenfalls im Laufe der Analyse als Ausreißer behandelt werden müssen, lässt sich eine Vorstellung hinsichtlich der Homogenität einer Population in ihrer Entwicklung im Zeitverlauf ableiten. Damit erfolgt zeitgleich sowohl eine Überprüfung der deskriptiv gewonnenen Erkenntnisse als auch die Formulierung möglicher sekundärer Hypothesen, beispielsweise für die Untersuchung von Subgruppenunterschieden. In Bezug auf den hier untersuchten PDI bestätigt sich die Einschätzung aus der deskriptiven Auswertung – einem zunächst fallenden und gegen Ende des Beobachtungszeitraums leicht ansteigenden Trend. Dabei scheint die vorliegende Population eine recht homogene Entwicklung nach der Intervention zu vollziehen.

Der hierbei gewonnene Eindruck bildet zusammen mit den deskriptiven Ergebnissen einen ersten entscheidenden Schritt für die Wahl eines geeigneten Regressionsmodells. Allerdings kann die Auswertung des Zeitdiagramms schnell Schwierigkeiten bereiten (Fitzmaurice et al., 2004 - Kapitel 3.3). Häufig weisen Studien einen hinreichend großen Stichprobenumfang auf, wodurch die Übersichtlichkeit der Diagramme bis hin zur Unlesbarkeit reduziert wird. Die Entwicklung der Patienten im zeitlichen Verlauf kann hierbei kaum erfasst werden. Ähnliche Erschwernisse bereiten kategoriale Daten, da sich die Schwankungen lediglich zwischen den vorgegebenen Ausprägungen der interessierenden Variable bewegen können. Die alternative Betrachtung von einzelnen Subgruppen oder die Darstellung von Mittelwerten im zeitlichen Verlauf anstatt der Rohdaten stellen Möglichkeiten zur Erleichterung von komplexen Situationen dar.

3.3 Konditionale und marginale Modelle

Gängige Methoden der induktiven Datenanalyse – beispielsweise aus der Test- und Schätztheorie – besitzen keine ausreichende Flexibilität, um die Komplexität longitudinaler Datenstrukturen ausreichend zu berücksichtigen. Für eine adäquate Analyse eignen sich insbesondere Instrumente, welche die verschiedenen Arten der Variabilität erfassen können. Folglich ist es für die Nutzung des maximalen Informationsgehaltes aus den verfügbaren Daten nicht ausreichend, Standardverfahren im Hinblick auf die Korrelation zwischen den Beobachtungszeitpunkten anzupassen. Vielmehr wurden basierend auf der Theorie der Regressionsanalyse Modifikationen statistischer Modelle entwickelt, welche die separate Betrachtung der unterschiedlichen Arten an Variabilität und ihrer Ursprünge ermöglichen (Fitzmaurice et al., 2004 - Kapitel 2.5):

i. Die Heterogenität zwischen den Individuen (random effects): Die einfachste Form dieser Variabilität besteht in einem individuellen Achsenabschnitt (random intercept), welcher zwischen Patienten mit generell hohen Werten von jenen mit niedrigen Werten unterscheidet. Weitaus realitätsnäher ist hingegen die zusätzliche Berücksichtigung eines individuellen Verlaufes (random slope), da Interventionen meist unterschiedliche Wirkungen auf Patienten verursachen und sich daraus unterschiedliche Verläufe ergeben.

ii. Die inhärente biologische Variation: Eine individuelle, gleichmäßige Schwankung um den eigenen Trend im zeitlichen Verlauf, welche möglicherweise sogar abhängig von der Tagesform stärker oder schwächer ausfallen kann.

iii. Der Messfehler: Diese Schwankungen entsprechen der Abweichung zwischen gemessenem und wahrem Wert und bilden die geringste Form der Variabilität. In der Statistik wird der Messfehler daher als „Rauschen“ bezeichnet.

Die zweite und dritte Form der Variabilität kann bereits durch die Anwendung gängiger Regressionsmodelle in die Berechnungen einbezogen werden. Die Heterogenität zwischen den Individuen ist hingegen lediglich durch spezifisch erweiterte Modelle aus der Theorie der Analyse longitudinaler Daten greifbar.

Modellierung mittels GLM

Eine einfache Herangehensweise an die Modellierung von Messwiederholungen bedient sich der Theorie *generalisierter linearer Regressionsmodelle (GLM)* (Definition siehe Kapitel 4.1). Anknüpfend an die deskriptive Analyse des PDI (Kapitel 3.2) richten sich die kommenden Ausführungen an stetig und quasistetig skalierte Merkmale. Hierbei steht die Annahme im Vordergrund, dass die zu untersuchenden Endpunkte zumindest approximativ einer multivariaten Normalverteilung folgen, wobei dies keine Voraussetzung für adäquate Parameterschätzungen darstellt. Grundsätzlich existieren vergleichbare Modellanpassungen an die Struktur longitudinaler Daten auch für diskrete Skalierungen und Zähldaten. Derartige Modifizierungen sind allerdings nicht Gegenstand dieser Arbeit.

Erfahrungen mit longitudinalen Datenstrukturen haben gezeigt, dass die Trends der Mittelwerte über die Zeit im Allgemeinen eher gleichmäßigen Verläufen ohne spontane Sprünge im Graphen folgen. Demnach ermöglicht ein lineares Regressionsmodell mit polynomialen Termen

$$E\left(Y_{ij} \mid time_{ij} \right) = \beta_0 + \beta_1 time_{ij} + \beta_2 time_{ij}^2 + \beta_3 time_{ij}^3 + \ldots + \beta_k time_{ij}^k$$

bereits eine gute Anpassung an die meisten Entwicklungen, wobei auch log(time) als Einflussterm häufig Anwendung findet. Gewöhnlich lässt ein Modell mit Polynomen maximal

dritten Grades eine ausreichende Erklärung der Daten zu. Darüber hinaus ist eine Vorgehensweise empfehlenswert, welche zunächst eine Prüfung Polynome höherer Grade vorsieht. Anhand einer schrittweisen Modellselektion können diese gegebenenfalls ausgeschlossen werden, wodurch sich eine verbesserte Datenanpassung realisieren lässt (Fitzmaurice et al., 2004).

Wie bereits erläutert, liegt der Nachteil gängiger linearer Modelle allerdings in der Ignoranz der Korrelation zwischen den Messzeitpunkten (Kapitel 3.1 / Liang & Zeger, 1993). Dies führt aufgrund der damit verbundenen Überschätzung der Varianz zu verzerrten Schätzungen der Regressionskoeffizienten und zu großen Standardfehlern (siehe (1.1)). Für die Berücksichtigung der Korrelationen zwischen den Zeitpunkten existieren verschiedene Ansätze zur Korrektur des Maximum-Likelihood-Schätzers (MLE) für die Regressionskoeffizienten. Unter der Annahme einer Zielgröße $Y_i = (Y_{i1}, \ldots, Y_{im})$, welche über die Zeitpunkte einer multivariaten Normalverteilung folgt, besitzt der ML-Schätzer folgende Struktur

$$\hat{\beta} = \left[\sum_{i=1}^{n} \left(X_i' \, \Sigma_i^{-1} \, X_i \right) \right]^{-1} \sum_{i=1}^{n} \left(X_i' \, \Sigma_i^{-1} \, y_i \right). \tag{1.2}$$

Nachdem die Kovarianz $Cov(Y_i) = \Sigma_i$ zwischen den Beobachtungszeitpunkten im Allgemeinen unbekannt ist, bedarf es hierfür ebenfalls einer Schätzung auf Grundlage der empirischen Daten. Aus der Bestimmung mittels ML-Methode resultiert eine äußerst flexible Form der Schätzung, da im Vorfeld der Berechnung keinerlei Einschränkungen getroffen werden. Die Kovarianzmatrix weist demnach eine unspezifizierte und damit realitätsnahe Struktur

$$Cov(Y_i) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_m^2 \end{pmatrix}$$

auf. Einerseits eröffnet sich durch dieses Vorgehen der Vorteil einer guten Anpassungsfähigkeit an die Daten, da sämtliche Matrixelemente separat voneinander bestimmt werden. Andererseits erfordert die fehlende Spezifikation eine Schätzung für jede Varianz und Kovarianz, was bereits bei wenigen Zeitpunkten eine Vielzahl zu schätzender Parameter mit sich bringt (Laird & Ware, 1982). Die Anzahl der Schätzparameter potenziert sich entspre-

chend mit der Anzahl der Zeitpunkte. Sofern keine besonders umfangreichen Patientenpopulationen untersucht werden, besteht demnach die Gefahr der Überparametrisierung.

Eine Alternative bietet die Vorgabe von Strukturmodellen für die Kovarianzmatrizen, welche sich durch einen deutlich einfacheren Aufbau auszeichnen. Die beiden bekanntesten Modelle werden als *zusammengesetzte Symmetrie* (compound symmetry) und als *autoregressiv* bezeichnet:

$$Cov_{CS}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} \qquad Cov_{AR}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{m-1} \\ \rho & 1 & \rho & \cdots & \rho^{m-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{m-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{m-1} & \rho^{m-2} & \rho^{m-3} & \cdots & 1 \end{pmatrix}$$

Die sich bei der unspezifizierten Kovarianzannahme ergebende Problematik der zahlreichen zu schätzenden Parameter wird durch die Vorgabe der Strukturmodelle vermieden. Beide hier präsentierte Formen verlangen eine Berechnung von lediglich zwei Parametern – einen für die Varianzen und einen für die Korrelation. Als nachteilig erweist sich die schlechtere Anpassung an die Daten aufgrund der geringeren Flexibilität. Die autoregressive Form der Kovarianzmatrix berücksichtigt durch die Exponenten in den Nebendiagonalelementen zumindest die bereits erwähnte Reduzierung der Korrelation bei weiter auseinanderliegenden Zeitpunkten. Beide Modelle setzen allerdings Varianzhomogenität über den zeitlichen Verlauf voraus, was in der Praxis nur selten in den Daten vorzufinden ist.

Modellierung mittels konditionaler Modelle

Das optimale Gleichgewicht zwischen der Anzahl der Schätzparameter und der Anpassungsfähigkeit der Kovarianzen liefert die *random-effects*-Kovarianz. Wie der Name vermuten lässt, folgt sie nicht notwendig einer vorgegebenen Struktur, sondern setzt sich aus der Variation der Gesamtpopulation und der jedes einzelnen Individuums zusammen. Dabei bildet sie eine Funktion über die Zeit und verlangt somit keine Varianzhomogenität der Messzeitpunkte. Weitere detaillierte Ausführungen finden sich bei Fitzmaurice et al., 2004 - Kapitel 8 und bei Cnaan et al., 1997.

Dieses hohe Maß an Flexibilität zeigt sich bereits in der Namensgebung derartiger Modellierungen. So beschreibt der Begriff *konditional* die Bedingung auf die individuellen Vorgaben eines Falles, was besondere Vorteile vorzugsweise für medizinische Untersuchungen mit sich bringt. Nachdem jede Behandlung auf die Verbesserung des Gesundheitszustandes jedes einzelnen Patienten abzielt, ermöglicht auch eine adäquate Analyse eine individuelle Interpretationsmöglichkeit der Ergebnisse. Für den ersten Evaluationsteil des MNS-R bieten demnach konditionale Methoden mit ihren sensiblen Anpassungsmöglichkeiten ideale Voraussetzungen für eine Untersuchung aus der medizinischen Perspektive.

Das in der Praxis zur Anwendung gebrachte Verfahren wird als *linear mixed effects model (LME)* bezeichnet. Diese Modelle bilden eine Sonderform der linearen Modelle (LM) und wurden primär für die Berücksichtigung der Besonderheiten longitudinaler Daten entwickelt. Technisch gesehen besteht der Unterschied zu den klassischen LM aus einer Modellerweiterung um weitere Regressionsparameter b_{0i} und b_{1i}

$$Y_i = \underbrace{\beta_0 + \beta_1 x_{1i}}_{\text{Prädiktor des LM}} + \underbrace{b_{0i} + b_{1i} z_{1i}}_{\substack{\text{random intercept} \\ \text{\& random slope}}} + \varepsilon_i \tag{1.3}$$

für einen individuellen Achsenabschnitt (random intercept) sowie für einen individuellen Einfluss der Prädiktoren, woraus sich ein für jeden Patienten eigener Kurvenverlauf (random slope) bestimmen lässt. Die zusätzlichen Koeffizienten b_{0i} und b_{1i} beschreiben dabei lediglich die Abweichung des Individuums *i* von dem Modell der Gesamtpopulation, wodurch die Bestimmung eines individuellen Modells für jeden Patienten bewerkstelligt werden kann. Insgesamt wird diese Modellstruktur meist in einer Matrix-Schreibweise

$$Y_i = \underbrace{X_i \beta}_{\substack{\text{fixed} \\ \text{effects}}} + \underbrace{Z_i b_i}_{\substack{\text{random} \\ \text{effects}}} + \varepsilon_i$$

zusammengefasst, wobei die Elemente des klassischen linearen Modells die Populationseffekte (fixed effects) darstellen, und die Modellerweiterungen sämtliche subjekt-spezifischen Effekte (random effects) repräsentieren. Entsprechend bietet diese Separation die Möglichkeit, zwischen der inner-subjektiven Variabilität und jener zwischen den Individuen zu unterscheiden. Die patientenspezifische Interpretation der Ergebnisse dient ferner als Grund-

lage für weitere Analyseschritte, wie beispielsweise der Identifikation möglicher Subgruppen, bei denen eine Intervention weniger erfolgreich verlaufen ist. Entsprechend lässt sich die bereits erwähnte Bedeutung für medizinische Untersuchungen ausmachen.

Aus den gewonnenen Informationen der deskriptiven Auswertung des PDI lässt sich lediglich die klassische Fragestellung nach relevanten Veränderungen der Zielgröße im zeitlichen Verlauf ableiten. Aufgrund der als homogen einzustufenden Entwicklung der Patientenpopulation ergibt sich primär keine Forderung nach Subgruppenanalysen oder nach der Untersuchung des Einflusses von Kovariablen. Technisch beschränkt sich folglich die Auswahl der Einflussgrößen auf die Zeitvariable **`time`**. Basierend auf den Erkenntnissen der grafischen Betrachtung der Rohdaten erscheint ein Modell mit linearem und logarithmischem Einfluss der Zeit hinsichtlich der festen sowie der individuellen Effekte

$$E\left(Y_{ij} \mid b_i, time_{ij}\right) = \beta_0 + \beta_1 \cdot time_{ij} + \beta_2 \cdot \log\left(time_{ij}\right) + b_{0i} + b_{1i} \cdot time_{ij} + b_{2i} \cdot \log\left(time_{ij}\right)$$

als adäquate Schätzung für die Entwicklung des PDI. Alternative Modellierungen mit einer quadratischen Form der Zeitvariable lieferten eine schlechtere Modellanpassung und wurden infolgedessen verworfen. Grundsätzlich liegt den Berechnungen ein Signifikanzniveau von 5 % zugrunde.

Die Beurteilung der Ergebnisse erfolgt gemäß der verschiedenen Bestandteile der LMEs. Die populationsspezifischen Resultate (fixed effects) werden analog zu den Ergebnissen linearer Modelle interpretiert (Tabelle 3.3.1).

	Value	**Std.Error**	**DF**	**t-value**	**p-value**
(Intercept)	34.53214	1.601224	165	21.566087	0.000
time	6.63436	1.974985	165	3.359196	0.001
I(log(time+1))	-22.67404	4.676639	165	-4.848362	0.000

Tabelle 3.3.1: Resultat der fixed effects aus der LME-Modellierung des PDI mit Einflussgröße Zeit

Im Falle des PDI findet sich demnach ein gleichgerichteter linearer sowie ein gegensinniger logarithmischer Einfluss der Zeit. Die Wirkung beider Kovariablen ist von signifikantem Ausmaß.

Schlussfolgerungen aus den individuellen Effekten (random effects) bedürfen einer ausführlicheren Betrachtung der Details, da für jeden Patienten eine individuelle Schätzkurve berechnet wurde. Die grafische Gegenüberstellung der Rohwerte mit den Modellschätzungen des PDI zeigt das durch die Regression erhöhte Maß an Ordnung und Übersichtlichkeit (Abbildung 3.3.1). Infolgedessen können individuelle Tendenzen innerhalb der zeitlichen Entwicklung weitaus einfacher identifiziert werden.

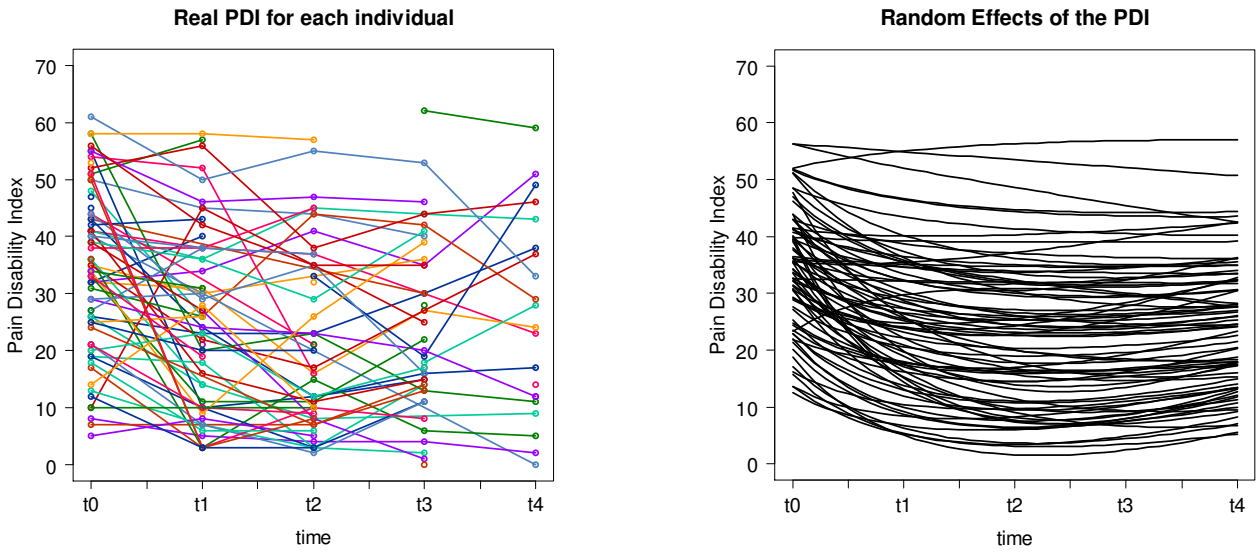

Abbildung 3.3.1: Reale PDI-Werte (links) und individuelle Modellierungen (rechts) im zeitlichen Verlauf t0 bis t4

Zur Vollständigkeit einer seriösen Datenanalyse ist es von Bedeutung, die Anpassung der Modelle an die realen Patientendaten zu beurteilen. Eine Modellierung, welche die realen Daten nicht in angemessenem Umfang erklärt, verfügt über keine belastbare Aussagekraft. Für die Beurteilung der Modellanpassung existieren Maßzahlen, welche sämtliche Patienten und Zeitpunkte einbeziehen. Nachdem jedoch bei der Analyse longitudinaler Daten mittels LME insbesondere die individuellen Verläufe im Vordergrund stehen, bedarf es eines Vergleiches der Graphen der Modellschätzungen mit den Rohdaten. Hierzu werden exemplarisch einige Fälle ausgewählt, deren Entwicklungen über die Zeit einen im Vergleich zur Gesamtpopulation speziellen Verlauf aufweisen.

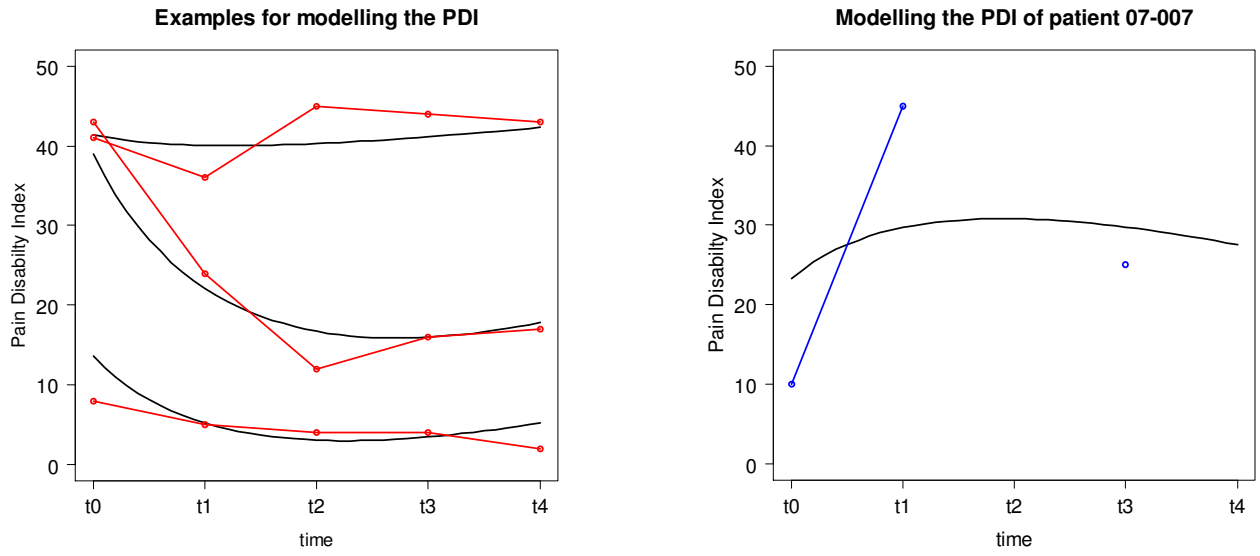

Abbildung 3.3.2: Exemplarische Darstellung von LME-Modell und realen Daten für drei Patienten (— Modell; — Patient)

Anhand der ausgewählten Patienten (Abbildung 3.3.2 links) lässt sich die hohe Anpassungsfähigkeit der LME demonstrieren. Es zeigt sich sowohl bei starken wie auch bei geringen Veränderungen im zeitlichen Verlauf eine Nähe der Modellierungen zu den realen Daten. Ferner weisen die gewählten Patienten deutliche Unterschiede in den Wertebereichen auf, was keinen Einfluss auf die Modellanpassung mit sich bringt. Darüber hinaus findet sich bei der Betrachtung sämtlicher individuellen Modelle ein einzelner Patient mit einem außergewöhnlichen Verlauf gegen den allgemeinen Trend der Population. Die grafische Überprüfung seiner Modellschätzung (Abbildung 3.3.2 rechts) bescheinigt dem LME eine optimale Leistung. Demnach zeichnet sich der Verlauf des Patienten durch erhebliche Schwankungen insbesondere zwischen den Zeitpunkten t0 und t1 aus. Zusätzliche Erschwernisse für die Modellschätzung ergeben sich durch die fehlenden Daten zu den Zeitpunkten t2 und t4.

Als Verallgemeinerung der grafischen Einzelbetrachtungen (Abbildung 3.3.2) können die individuellen Residuen der Modelle jedes einzelnen Patienten

$$r_{ij} = y_{ij} - \left(x'_{ij}\,\hat{\beta} + z'_{ij}\,\hat{b}_i \right) \qquad r_{ij} \sim N\left(0, \sigma_i\right)$$

zusammengefasst werden. Die Darstellung der resultierenden Abstände liefert eine Vorstellung bezüglich der Datenanpassung unter Berücksichtigung fester und zufälliger Effekte

(Abbildung 3.3.3). Abgesehen von den Daten zum Zeitpunkt t4 bewegt sich der überwiegende Teil der Residuen in einem Wertebereich bis maximal vier Punkte. Unter Berücksichtigung des Wertebereiches des PDI zwischen 0 und 70 Indexpunkten zeigt sich demnach eine sehr zufriedenstellende Datenanpassung.

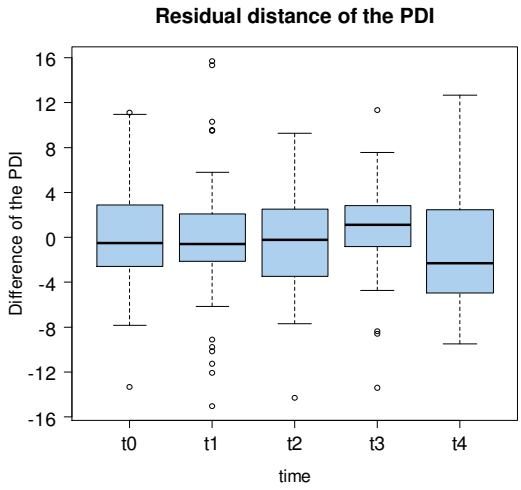

Abbildung 3.3.3: Individuelle Residuen im zeitlichen Verlauf (t0 bis t4)

Überdurchschnittliche Abstände zwischen Schätzwerten und empirischen Daten bilden Ausreißer unter den Residuen. Diese können meist auf fehlende Werte bei anderen Zeitpunkten zurückgeführt werden. Ferner ist in dieser Darstellung das Vorliegen einer Normalverteilung mit Mittelwert 0 leicht nachvollziehbar, welcher die Residuen im Allgemeinen folgen.

Abgesehen von den Einzelbetrachtungen individueller Modelle können anhand der Berechnung der random effects Aussagen hinsichtlich der Heterogenität in der Entwicklung der Patientenpopulation über die Zeit getroffen werden. Hierzu wird die Standardabweichung der random effects für jede Einflussgröße betrachtet (Tabelle 3.3.2).

	Standard deviation
(Intercept)	11.936646
time	7.867018
log(time)	22.196440

Tabelle 3.3.2: Standardabweichung der random effects aus der Modellierung des PDI

Im Fall des PDI bewegen sich die Variationen der individuellen Effekte in einem unauffälligen Bereich, was auf eine homogene Veränderung der Patientenpopulation im zeitlichen Verlauf schließen lässt. Demnach bestätigt sich der erste Eindruck zu Beginn der Datenanalyse.

Insgesamt ist die hier vorgestellte Vorgehensweise auf stetige Merkmale wie den PDI ausgerichtet. Analog zu der Generalisierung linearer Modelle für nicht normalverteilte Zielgrößen existieren auch bei gemischten Modellen modifizierte Erweiterungen. Demnach kann beispielsweise beim Vorliegen einer Poisson- oder Binomialverteilung auf entsprechend spezifisch angepasste Formen der *generalized linear mixed effects models (GLME)* zurückgegriffen werden. Weiterführende Beschreibungen finden sich etwa bei Fitzmaurice et al., 2004 - Part III.

Modellierung mittels marginaler Modelle

Aufgrund der unterschiedlichen Perspektiven innerhalb der Kooperation zwischen klinischer Einrichtung und Betriebskrankenkasse erfordert die zugrundeliegende Fragestellung der Evaluation sowohl eine medizinische als auch eine gesundheitsökonomische Interpretation der Ergebnisse. Somit werden die Daten neben der Berechnung von LME auch mittels marginaler Regressionsmodelle analysiert. Der Begriff *marginal* ist hierbei als *populationsspezifisch* zu verstehen, was direkt mit einer gesundheitsökonomischen Deutung der Ergebnisse einhergeht. Marginale Modelle zeichnen sich durch ein besonders hohes Maß an Anpassungsfähigkeit aus, wodurch sich diese Modellierungstechnik gut zur Analyse longitudinaler Daten eignet. Grundsätzlich wird diese Flexibilität anhand des im Vergleich zu klassischen GLM sehr geringen Umfangs an Vorgaben erzeugt. Somit bedarf es zur Anwendung marginaler Modelle keinerlei Verteilungsannahme, sondern lediglich eine voneinander unabhängige Spezifikation der ersten und zweiten Momente (Erwartungswert und Varianzfunktion). Aufgrund dieser Konstellation eröffnet sich die Möglichkeit, die Besonderheiten longitudinaler Daten direkt in die Modellschätzung einzubeziehen. Eine sehr übersichtliche und detaillierte Zusammenstellung der technischen Hintergründe findet sich bei Pan et al., 2000 und deren Quellen. Ein praktischer Leitfaden zum optimalen Analysevorgehen, welcher auch als Grundlage zur Evaluation des MNS-R gedient hat, ist in Fitzmaurice et al., 2004 - Kapitel 11 enthalten.

Das Konzept der statistischen Inferenz basiert auf der theoretischen Grundlage generalisierter Regressionsmodelle (GLM). Dabei wird die allgemeine, verteilungsunabhängige Form der Scorefunktion (sog. Quasi-Scorefunktion) zur Modellschätzung herangezogen. Zusammen mit den spezifizierten ersten beiden Momenten liegt eine sogenannte *generalisierte Schätzgleichung (GEE)* vor, anhand derer nach der gängigen Maximum-Likelihood-Methode die Schätzung der Regressionskoeffizienten bewerkstelligt werden kann. Daher werden die GEE auch als semi-parametrische Schätzer bezeichnet. Für eine weitere Detaillierung der theoretischen Hintergründe statistischer Inferenz und Modellselektion bei GEE sei auf Kapitel 4 verwiesen.

Grundsätzlich wird für die Analyse longitudinaler Daten die Maximum-Likelihood-Schätzung generalisiert und erweitert, um die Korrelationen zwischen den Zeitpunkten in die Berechnung zu integrieren. Somit kommt eine angepasste Varianzfunktion

$$V_i = A_i^{1/2} \underbrace{Corr(Y_i)}_{\substack{Erweiterung\ für \\ longitudinale\ Daten}} A_i^{1/2}$$

zum Einsatz, in der die Korrelationsmatrix $Corr(Y_i)$ des Responsevektors Y_i eingebettet wird. Folglich beschreibt A_i die übliche Varianzfunktion $\mathrm{Var}(Y_{ij})$, welche bei GEEs Anwendung findet (demnach ist $A_i^{1/2}$ eine Diagonalmatrix und enthält die Standardabweichungen). Diese modifizierte Form wird als Arbeitskovarianz (working covariance) bezeichnet. Eine Besonderheit der GEE-Schätzung sorgt hierbei für einen zusätzlichen Vorteil im Hinblick auf die Problematik longitudinaler Daten: Für eine konsistente Schätzung der Regressionsparameter besteht keine Notwenigkeit einer korrekten Spezifikation der Varianzfunktion. Verständlicherweise wäre bei Verwendung einer falschen Varianzfunktion die Effizienz und Präzision der Schätzung eingeschränkt, was einen zu großen Standardfehler zur Folge hätte. Dennoch konvergiert die Mittlere Quadratische Abweichung (MSE) des Schätzers gegen Null für $n \to \infty$ (Pan et al., 2000).

Wie bereits erläutert, lässt sich die populationsspezifische Analyse longitudinaler Daten durch die Verwendung von GEE umsetzen. Analog zur Berechnung des LME wird das Analysevorgehen exemplarisch am PDI demonstriert. Vorausgesetzt wird die deskriptive Beurtei-

lung der Daten inklusive der grafischen Erkenntnisse (Kapitel 3.2). Für die Regression wird eine zu den festen Effekten des LME äquivalente Modellstruktur

$$E\left(Y_{ij} \mid time_{ij}\right) = \beta_0 + \beta_1 \cdot time_{ij} + \beta_2 \cdot \log\left(time_{ij}\right)$$

mit linearem und logarithmischem Einfluss der Zeitvariable **`time`** angewandt. Dabei liegt ein Signifikanzniveau von 5 % zugrunde. Bei dem numerischen Resultat der Modellberechnung sind insbesondere die Regressionsparameter von Bedeutung. Deren populationsspezifische Interpretation gestaltet sich analog zur klassischen linearen Regression unter Vorliegen einer Normalverteilung. Ferner fokussiert die Beurteilung der Ergebnisse die verschiedenen Schätzungen der Varianzen und Kovarianzen. Erhebliche Abweichungen zwischen *naiven* und *robusten* Standardfehlern lassen auf falsche Modellannahmen hinsichtlich der Varianzfunktion schließen (Tabelle 3.3.3).

	Estimate	**Naive S.E.**	**Naive z**	**Robust S.E.**	**Robust z**	**p.value**
(Intercept)	34.705971	1.750607	19.825109	1.628004	21.317662	0.000
time	8.605371	3.275224	2.627415	3.417857	2.517768	0.012
I(log(time+1))	-26.850015	7.486139	-3.586630	7.621629	-3.522871	0.000

Tabelle 3.3.3: Resultat der GEE-Modellierung des PDI mit Einflussgröße Zeit

Im Falle des PDI kann ein gleichsinniger linearer Einfluss der Variable **`time`** mit signifikantem Ausmaß festgestellt werden. Noch deutlich ausgeprägter ist der gegengerichtete Einfluss der logarithmierten Form der Zeitvariable. Mit dem zeitlichen Verlauf stellt sich somit eine Reduzierung des PDI ein.

Diese Erkenntnis wird anhand einer grafischen Darstellung der Rohwerte und des GEE-Modells kontrolliert (Abbildung 3.3.4). Da sich die Berechnung marginaler Modelle generell auf den Mittelwert einer Population bezieht, erfolgt zudem eine Überprüfung der Datenanpassung durch die Gegenüberstellung von Schätz- und Mittelwerten des PDI zu jedem Zeitpunkt. Es zeigt sich eine präzise Modellschätzung, welche eine gute Erklärung der empirischen Daten liefert. Als Ergebnis des grafischen Vergleichs mit den realen Daten kann die Schätzung ferner als plausibel eingestuft werden.

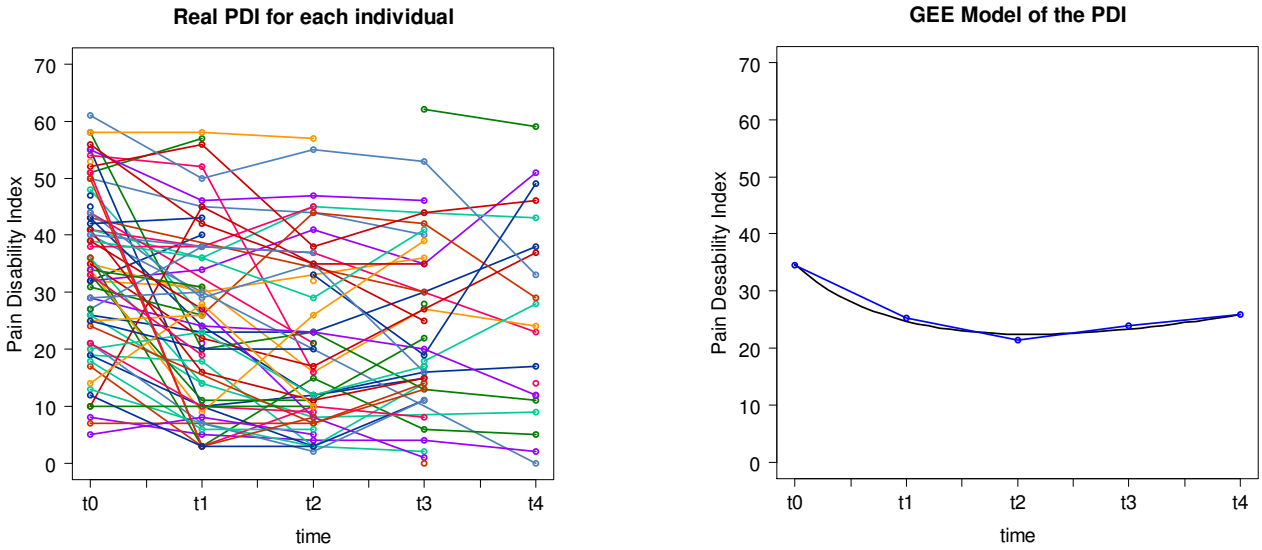

Abbildung 3.3.4: Reale Werte des PDI (links) und GEE-Modellierung (rechts) t0 bis t4

Kapitel 4

Modellselektion bei marginalen Modellen

Der erste Evaluationsteil des MNS-R umfasst die Verwendung zweier unterschiedlicher Regressionsmodelle. Dabei orientiert sich die Wahl der jeweiligen Methode insbesondere an der zu untersuchenden Fragestellung und somit an der Interpretation der Ergebnisse. Die Anwendung von *linear mixed effects models (LME)* zeichnet sich durch die Bestimmung von random effects aus, wodurch eine Modellierung der individuellen Patientenentwicklung über die Zeit ermöglicht wird. Durch diese individuelle Betrachtungsweise ergibt sich ein medizinischer Interpretationsansatz. *Generalized estimating equations (GEE)* hingegen zeichnen sich im Vergleich zu LMEs durch eine erhöhte Flexibilität aus, erlauben allerdings keine Ermittlung von random effects. Demnach stellen die GEEs ein optimales Instrument bei der Untersuchung populationsspezifischer Effekte dar. Insbesondere bei der Bearbeitung gesundheitsökonomischer Fragestellungen stellt diese Modellform Vorteile hinsichtlich der Ergebnisinterpretation bereit.

4.1 Modellselektion bei LMEs und GEEs

Modellselektion bei GLM

LMEs bzw. GLMEs basieren auf der Theorie *generalisierter linearer Regressionsmodelle (GLM)*. Zur Erläuterung der formalen Hintergründe der Modellselektion bedarf es daher eines Überblicks über die Definition und Schätzmethode der GLM (Nelder & Wedderburn, 1972 sowie Fahrmeir et al., 2007):

1. **Verteilungsannahme**

   Die Verteilung einer abhängigen Variablen Y lässt sich in Form einer einparametrischen Exponentialfamilie schreiben. Deren Verteilungsdichte ist somit durch

$$f(y_i \mid \theta_i) = \exp\left(\frac{y_i \theta_i - b(\theta_i)}{\phi}\,\omega_i + c(y_i, \phi, \omega_i)\right)$$

gegeben. Dabei wird θ_i als natürlicher Parameter bezeichnet. Der Parameter ϕ ist hingegen ein Dispersionsparameter und von *i* unabhängig. Grundsätzlich gehören neben der Normalverteilung unter anderem auch Binomial-, Poisson- und Gammaverteilungen zu den Exponentialfamilien.

2. Strukturannahme

Der bedingte Erwartungswert der Zielvariablen $\mu_i = E(y_i \mid x_i)$ für gegebene Kovariablen $x_i = (1, x_{i1}, \ldots, x_{ip})$ wird mit dem linearen Prädiktor

$$\eta_i = x_i' \beta = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

durch eine Responsefunktion *h*, bzw. durch eine Linkfunktion $g = h^{-1}$ verknüpft:

$$\mu_i = h(\eta_i) = h(x_i' \beta) \quad bzw. \quad \eta_i = g(\mu_i)$$

Die Responsefunktion *h* ist ferner eine eineindeutige und zweimal differenzierbare Funktion.

Durch das Vorliegen einer konkreten Verteilungsannahme mit entsprechender Wahrscheinlichkeitsdichte (Likelihood-Funktion) besteht die Möglichkeit einer parametrischen Schätzung der Modellparameter $\beta_1, \ldots, \beta_p$ (Likelihood-Inferenz). Zusammenfassend lässt sich das Vorgehen für die Schätzung wie folgt beschreiben:
Gesucht wird für jeden Modellparameter ein Schätzwert, welcher die Wahrscheinlichkeitsdichte maximiert (Maximum-Likelihood-Prinzip). Technisch entspricht die durchzuführende Prozedur meist der Bestimmung der lokalen Maxima durch Nullsetzen der Scorefunktion

$$s(\beta) = \frac{\partial \log\left(f(y \mid \beta)\right)}{\partial \beta} = \sum_{i=1}^{n} x_i \frac{d_i}{\sigma_i^2}\,(y_i - \mu_i) \stackrel{!}{=} 0, \qquad (1.4)$$

wobei für eine vereinfachte Rechnung die logarithmierte Form der Wahrscheinlichkeitsdichte verwendet wird.

Des Weiteren bringt das Vorliegen einer Likelihood verschiedene weitere Vorteile mit sich: Beispielsweise kann diese zum Testen linearer Hypothesen herangezogen werden, etwa bei der Anwendung von Likelihood-Quotienten- oder Score-Statistiken. Eine besondere Erleichterung bietet sich ferner in der Bewertung von Regressionsmodellen. Hierfür existieren verschiedene Maßzahlen, die die Anpassungsgüte eines Modells an die realen Daten quantifizieren können. Dabei bildet die logarithmierte Likelihood-Funktion *l* bei der Berechnung der gängigen Maße wie beispielsweise der Devianz

$$D(y,\hat{\mu}) = -2\phi\{l(y;\hat{\mu},\phi) - l(y;y,\phi)\}$$

den zentralen Bestandteil. Das wohl gebräuchlichste und in der Evaluation des MNS-R zum Einsatz gebrachte Anpassungsmaß nennt sich Akaikes Informationskriterium (AIC) und ist wie folgt definiert:

$$AIC = -2\,l(\hat{\beta}) + 2p$$

In der Praxis liegen meist mehrere Regressionsmodelle mit einer unterschiedlichen Kombination von Kovariablen vor, aus denen die Modellschätzung mit der besten Anpassungsgüte gewählt werden soll. Durch den Vergleich der Resultate der AIC-Berechnung werden die anpassungsstärksten LMEs für jedes Merkmal der Evaluation identifiziert. Dabei repräsentiert ein kleinerer AIC-Wert einen geringeren Abstand zwischen reellen Daten und dem Schätzwert, wodurch auf eine bessere Erklärung der Daten durch die Modellschätzung geschlossen werden kann. Entsprechend geht die Anzahl der geschätzten Parameter *p* bestrafend in die Berechnung ein, was durch Addition zu dem Wert der Likelihood erreicht wird. Infolgedessen wird das Modell mit dem kleinsten AIC-Wert zur Präsentation und Interpretation des Ergebnisses herangezogen.

Modellselektion bei GEE

Im Gegensatz zu den LMEs findet sich in der Theorie der marginalen Modelle keinerlei Verteilungsvoraussetzung an die Daten. Die Schätzung der Modellparameter basiert auf einer Quasi-Likelihood-Funktion $Q(\beta,\phi)$, deren Struktur aus der Theorie der GLM stammt. Analog zum Vorgehen im Bereich der GLM ergibt sich aus der Ableitung der Quasi-Likelihood eine Quasi-Scorefunktion, der *generalisierten Schätzfunktion (GEE)*

$$Q(\beta) = \sum_{i=1}^{n} x_i \, \frac{d_i}{\sigma_i^2} \, (y_i - \mu_i). \tag{1.5}$$

Die Vorgabe der ersten beiden Momente (Erwartungswert und Varianz) impliziert direkt die Schätzung der Modellparameter durch Verwendung der Quasi-Scorefunktion. Die Besonderheit dieser Theorie liegt in der Flexibilität einer getrennten Festlegung von Erwartungswert- und Varianzfunktion. Wie bei der echten Scorefunktion wird dennoch eine konsistente Parameterschätzung ermöglicht (Fahrmeir et al., 2007). Sogar eine unkorrekte Spezifizierung der Varianz hebt die Konsistenz der Schätzung

$$MSE\left(\hat{\beta}_{GEE}\right) \xrightarrow{n\to\infty} 0$$

nicht auf. Allerdings könnte eine derartige Konstellation die Effizienz der Schätzung beeinträchtigen.

Der Vorteil dieses semi-parametrischen Schätzverfahrens verursacht durch das Fehlen einer konkreten Wahrscheinlichkeitsdichte gleichzeitig den schwerwiegendsten Nachteil im Umgang mit GEEs. Während der Einsatz konditionaler Regressionsmethoden wie dem LME eine Berechnung des AIC ermöglicht, kann dieser nicht zur Modellbewertung zwischen mehreren GEEs herangezogen werden. Auch weitere Likelihood-basierende Maße zur Modellwahl stehen hier nicht zur Verfügung. Entsprechend verlangt die Modellselektion bei GEEs stets schwierige und aufwendige Verfahren, wie beispielsweise das Bootstrapping.

Pan, 2001 stellt für eine Erleichterung dieser Problemstellung eine Modifikation des AIC vor, welche basierend auf der Verwendung einer Quasi-Likelihood eine Modellselektion bei GEEs ermöglichen soll:

$$QIC(R) \equiv -2\, Q\left(\hat{\beta}(R); I, D\right) + 2\, trace\left(\hat{\Omega}_I \, \hat{V}_r\right) \tag{1.6}$$

Im Gegensatz zur generellen Verwendung des AIC sind kaum Studien zu finden, welche sich kritisch mit der Qualität des QIC auseinandersetzen. Derartige Untersuchungen könnten eine Prüfung des Potentials für die Modellselektion unter verschiedenen Datensituationen anregen. Nachdem die Verwendung des QIC lediglich auf vereinzelte statistische Analysen beschränkt ist, wird dessen Leistungsfähigkeit im Folgenden näher untersucht. Hierfür liegt die longitudinale Datenstruktur aus der Evaluation des MNS-R zugrunde.

4.2 Aufbau der Simulationsstudie

Die wissenschaftliche Grundlage der Statistik zielt auf die Beantwortung komplexer Fragestellungen unter einem minimalen Aufwand an Datenerhebung ab. Die Problemstellungen werden hierzu in mathematische Sachverhalte übersetzt, welche sich anhand von empirischen Daten bearbeiten lassen. Nachdem in der Regel die Eigenschaften der Grundgesamtheit unbekannt sind, ist der hohe Aufwand einer gewissenhaften Stichprobenplanung und Datenerhebung unerlässlich. Basierend auf den Daten sollen schließlich stichhaltige Ergebnisse erarbeitet werden, welche möglichst repräsentative Aussagen über die Grundgesamtheit bewerkstelligen.
Anhand dieser Überlegungen werden die zahlreichen Vorteile einer Simulationsstudie für die Beurteilung eines neuen statistischen Verfahrens deutlich. Im Gegensatz zu empirischen Daten kann einer Simulation eine Grundgesamtheit vorgegeben werden, aus welcher beliebig viele Stichproben unterschiedlicher Repräsentativität generiert werden können. Anhand der Kenntnisse über die Grundgesamtheit kann die zu prüfende Analysemethode nachhaltig beurteilt werden.

4.2.1 Zielsetzung und Fragestellung

Der Überprüfung des QIC liegen im Wesentlichen drei Fragestellungen zugrunde:

i. Welches Potential weist das QIC generell auf, um bessere von schlechteren Regressionsmodellen zu unterscheiden?

ii. Wie nachhaltig ist die Leistung des QIC unter verschiedenen Korrelationsstrukturen der random effects?

iii. Beeinflussen kleine Stichprobenumfänge die Leistungsfähigkeit des QIC?

Die Untersuchung der ersten Fragestellung soll klären, ob das QIC ein geeignetes Instrument zur Modellselektion darstellt, wofür es einer hinreichenden Trennschärfe zur Unterscheidung zwischen guten und schlechteren Modellanpassungen bedarf. Darüber hinaus soll überprüft werden, ob eine ähnliche Trennschärfe unter erschwerten Bedingungen beibehalten werden kann. Beispielsweise könnten kleine Stichprobenumfänge oder die ver-

schiedenen Korrelationsstrukturen longitudinaler Daten möglicherweise zu häufigeren Fehlentscheidungen führen.

4.2.2 Berechnung der abhängigen Variable Y

Da die Motivation für eine detaillierte Untersuchung des QIC aus der Evaluation des MNS-R entstand, soll die vorliegende Daten- und Modellstruktur aus dem ersten Evaluationsteil die Grundlage der Simulationsstudie bilden. Die der Berechnung der GEEs zugrunde liegende abhängige Variable Y unterliegt der Vorgabe dreier unterschiedlicher Verteilungstypen: Die Annahme einer Normalverteilung stammt aus den Scoreberechnungen standardisierter Fragebögen wie beispielsweise der PDI, aus denen stetige Merkmale resultieren. Darüber hinaus existieren Poisson-verteilte Variablen, welche sich z. B. aus der Frage nach der *Anzahl an Arztbesuchen innerhalb der letzten drei Monate* ergeben. Die dritte Vorgabe besteht aus einer Binomialverteilung, welche aus Fragen mit dichotomen Antwortmöglichkeiten entstehen, wie beispielsweise *Beziehen Sie eine Rente – ja/nein.* Dem Studiendesign der Evaluation folgend werden die Daten für fünf Beobachtungszeitpunkte simuliert, wobei der erste und zweite Zeitpunkt den vorher-nachher-Vergleich für die Therapie darstellen.

Die Simulationsberechnung der fixed effects erfolgt unter formaler Verwendung eines linearen Regressionsmodells, wobei lediglich die Zeit t und die logarithmierte Zeit log(t) als Einflussgrößen dienen

$$y = a + bt + c\log(t) + \varepsilon$$

Dabei werden die Koeffizienten a, b und c aus den Modellberechnungen des MNS-R verwendet. Durch Einbeziehung der normalverteilten Fehlervariablen ε einer linearen Regression wird ein allgemeines Rauschen in den Daten erzeugt, was generellen Abweichungen durch Messfehler Rechnung tragen soll.
Zusätzlich zu den festen Effekten lassen sich anhand geeigneter Variationen der Regressionskoeffizienten a, b und c random effects erzeugen. Hierzu werden Hilfskoeffizienten α_i, β_i und γ_i aus einer multivariaten Normalverteilung generiert und zu den Regressionskoeffizienten addiert

$$\tau_i = \left(\mathrm{a}+\alpha_i\right) + \left(\mathrm{b}+\beta_i\right)\mathrm{t} + \left(\mathrm{c}+\gamma_i\right)\log(\mathrm{t}) + \varepsilon_i \,.$$

$$\sim MN\left(0,\Sigma_{ranef}\right) \qquad \sim N(0,2)$$

Die Erzeugung der simulierten Werte erfolgt durch Ziehen aus der jeweiligen Verteilung nach dem Zufallsprinzip. Statistische Programmpakete beinhalten hierzu geeignete Funktionen (Tabelle 4.2.1). Während die Bestimmung einer normalverteilten abhängigen Variable durch den direkten Einfluss des Terms τ_i erfolgt, beschreibt dieser im Falle der Poissonverteilten Variable lediglich den Erwartungswert $\lambda_i = \tau_i$. Für die Simulation eines binomialverteilten Merkmales wird logit(τ_i) als Wahrscheinlichkeitsparameter π_i verwendet.

Verteilung von Y_i	**Bestimmung der Werte von Y_i**
Normal	$Y_i = \tau_i$
Poisson	$Y_i = \mathrm{rpois}(\tau_i)$
Binomial	$Y_i = \mathrm{rbin}\left(n, \mathrm{logit}(\tau_i)\right)$

Tabelle 4.2.1: Berechnung der abhängigen Variable Y_i entsprechend der jeweiligen Verteilungsannahme

4.2.3 Simulationsumgebung und Berechnung der GEEs

Aufgrund des beschriebenen Vorgehens zur Erzeugung der simulierten Daten existiert ein *wahres* Regressionsmodell mit den Einflussgrößen t und log(t). Es bildet exakt die zur Datensimulation vorgegebene Struktur nach. Infolgedessen lässt sich kein anderes Modell finden, welches eine bessere Datenanpassung erreichen könnte – selbst unter der Verwendung weiterer Einflussgrößen. Die Anwendung eines optimalen Instruments zur Modellwahl sollte also stets dieses wahre Modell bevorzugen.

$$E(Y_i) = \overbrace{\underbrace{\beta_0 + \beta_1 t_i + \beta_2 \log(t_i)}_{\text{wahres Modell}} + \beta_3 x_i}^{\text{alternatives Modell}} \tag{1.7}$$

Die Modellalternative besteht lediglich aus der zusätzlichen Integration einer unabhängigen Kovariable *X* in die Regressionsanalyse, woraus sich eine entsprechend schlechtere Modellanpassung ergibt. Für die unabhängige Einflussgröße *X* wird ebenfalls eine Normal-, eine

Poisson- und eine Binomialverteilung verwendet. Die Berechnung des QIC erfolgt sowohl für die wahre als auch für die drei alternativen Modellierungen. Als Ergebnis der Simulationsstudie werden die QIC-Werte der jeweiligen Modelle verglichen. Die direkte Gegenüberstellung zeigt unmittelbar, ob es möglich ist, mittels des QIC das korrekte Modell zu identifizieren.

4.2.4 Variationsmöglichkeiten im Studienverlauf

Nachdem sich die Schwächen derartiger Instrumente erst unter erschwerten Bedingungen offenbaren, werden stets verschiedene Stichprobenumfänge (n = 10, 20, 50, 100, 500) in den Simulationen getestet. Darüber hinaus erfolgt hinsichtlich der random effects eine Prüfung unterschiedlicher Kovarianzstrukturen und deren Wirkung auf die Trennschärfe des QIC. Dabei werden für die Kovarianzmatrix der random effects sowohl eine unabhängige, eine symmetrische als auch eine frei definierte Struktur vorgegeben.

4.3 Simulationsergebnisse

Insgesamt fallen die Durchläufe der Simulationen hinsichtlich der Erfolgserwartungen zufriedenstellend aus, sodass neue Erkenntnisse aus der Untersuchung gewonnen werden können. Lediglich vereinzelte Simulationen mit sehr geringer Fallzahl bringen keine verwertbaren Resultate. Bei der folgenden Präsentation der Ergebnisse werden diese gesondert erwähnt.

4.3.1 Allgemeine Erkenntnisse

Die Auswahl eines Modells unter gleichzeitiger Verwerfung eines anderen richtet sich nach den Werten der QIC-Berechnung. Dabei soll das Modell mit der besseren Datenanpassung den kleineren QIC-Wert liefern. Diesem Gedanken folgend kann die Modellselektion unter Einsatz eines Anpassungskriteriums auch als Testsituation begriffen werden. Für eine korrekte Bewertung der Diskriminanzeigenschaft des QIC bieten sich somit *Receiver-Operating-Characteristic (ROC)*-Kurven an. Diese spiegeln direkt das Verhältnis zwischen korrekten und fehlerhaften Entscheidungen wider. Dabei repräsentiert die richtig-positiv-Rate den Anteil korrekt getroffener positiver Ergebnisse eines Testverfahrens und wird zur

grafischen Darstellung auf der Ordinate eines Koordinatensystems abgetragen. Die Werte der Abszisse entsprechen der falsch-positiv-Rate, welche den Anteil der positiven Entscheidungen beschreibt, deren Einschätzung falsch gewesen ist.
Somit stellt sich eine optimale Unterscheidungseigenschaft eines Testverfahrens durch eine ROC-Kurve dar, welche sich zunächst durch einen nahezu parallel zur Ordinate verlaufenden starken Anstieg bis nahe an den Wert 1 auszeichnet und weiter parallel zur Abszisse verläuft. Dieser charakteristische Verlauf leuchtet insofern ein, als sich bei geringen Testwerten nahezu ausschließlich korrekte Testresultate einstellen. Gleichzeitig sind kaum Falschentscheidungen vorhanden. Im Bereich höherer Testwerte ändert sich das Verhältnis zu wenigen korrekten und überwiegend falschen Entscheidungen, was einem parallelen Verlauf zur Abszisse entspricht. Neben der grafischen Begutachtung kann die Trennschärfe eines Tests auch numerisch anhand der Fläche unter der ROC-Kurve dargestellt werden. Hierzu wird die *Area under the curve (AUC)* berechnet und fungiert folglich als Diskriminanzmaß. Die AUC verfügt über einen Wertebereich zwischen 0 und 1, wobei ein Wert von 0.5 der Winkelhalbierenden entspricht und eine quasi nicht vorhandene Diskriminanzleistung bescheinigt.

Als erste Erkenntnis aus den Simulationsläufen lässt sich festhalten, dass die Leistung des QIC zur Unterscheidung zwischen Regressionsmodellen mit besserer und schlechterer Datenanpassung kaum an die Kovarianzstruktur der random effects gebunden scheint. Sowohl die Verläufe der ROC-Kurven als auch die Werte der AUC weisen bei sämtlichen Simulationen lediglich geringe Abweichungen auf (Tabelle 4.3.1).

	independent	**compound symmetry**	**free structured**
AUC	0.9945	0.9951	0.9873

Tabelle 4.3.1: AUC-Werte unter verschiedenen Kovarianzstrukturen simulierter random effects; Y~Norm, n=20

Aus diesem Ergebnis kann geschlossen werden, dass die Trennschärfe des QIC nicht durch eine inhomogene Entwicklung der Patientenpopulation über die Zeit beeinflusst werden kann.

Eine weitere Beobachtung hinsichtlich des Wertebereichs des QIC lässt sich durch den Vergleich der Simulationen mit unterschiedlich verteilten Einflussgrößen verzeichnen. Während aus dem wahren Modell lediglich ein kleiner Wertebereich resultiert, ergibt sich unter Einbeziehung normal- und Poisson-verteilter Kovariablen eine ähnliche und deutlich vergrößerte Spannweite. Die größte Variabilität unter den QIC-Resultaten erzeugt hierbei der Einfluss binomialverteilter Merkmale, was mit einem entsprechend vergrößerten Wertebereich im Vergleich zu anderen Verteilungen einhergeht (Abbildung 4.3.1).

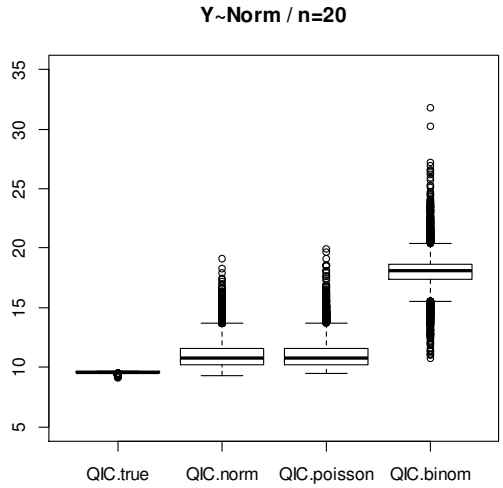

Abbildung 4.3.1: Wertebereiche des QIC im wahren Modell und unter verschieden verteilten Kovariablen

Je geringer die Überschneidung der Wertebereiche der verschiedenen Boxplots mit dem des wahren Modells ist, desto höher kann die Erwartung an den QIC gestellt werden, eine zufriedenstellende Trennschärfe zwischen wahrem und falschem Modell zu erbringen.

4.3.2 Ergebnisse unter normalverteilter abhängiger Variable

Nachdem sich keine relevanten Unterschiede hinsichtlich der verschiedenen Kovarianzstrukturen der random effects gezeigt haben, wird die weitere Ergebnispräsentation auf den Fall der Unabhängigkeit beschränkt. Bei Modellierungen von normalverteilten abhängigen Variablen zeichnet sich das QIC durch eine gute Leistung zur Modellselektion aus (Tabelle 4.3.2).

	AUC (true vs. incorrect model)		
n	normal	poisson	binomial
10	0.8069	0.8031	1.0000
20	0.9945	0.9957	1.0000
50	1.0000	1.0000	1.0000

Tabelle 4.3.2: AUC-Ergebnisse unter verschieden verteilten Kovariablen bei n=10, 20, 50; Y~Norm

Insbesondere bei Stichprobenumfängen von 20 oder mehr kann eine optimale Trennschärfe nachgewiesen werden. Bei kleineren Datensätzen zeigt sich mit einem AUC-Wert von ca. 0.8 eine gute Diskriminanzeigenschaft (Abbildung 4.3.2).

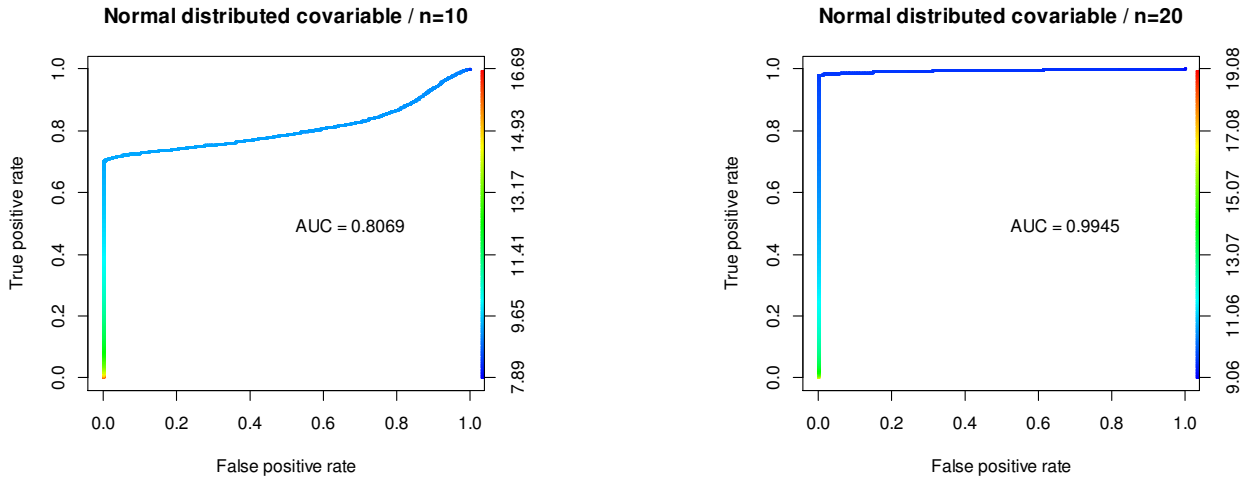

Abbildung 4.3.2: ROC-Kurven unter normal verteilter abhängiger Variable Y_i

4.3.3 Ergebnisse unter Poisson-verteilter abhängiger Variable

Die Berechnungen unter Verwendung Poisson-verteilter Zielgrößen bescheinigen dem QIC eine ähnlich gute Leistung, wie unter der Vorgabe einer Normalverteilung (Tabelle 4.3.3).

	AUC (true vs. incorrect model)		
n	normal	poisson	binomial
10	0.8513	0.8548	0.9923
20	0.9786	0.9796	1.0000
50	0.9999	1.0000	1.0000

Tabelle 4.3.3: AUC-Ergebnisse unter verschieden verteilten Kovariablen bei n=10, 20, 50; Y∼Pois

Dabei zeigt sich eine optimale Trennschärfe bei Datensätzen mit 20 oder mehr Fällen. Aus den Resultaten bei geringen Stichprobenumfängen von n = 10 lässt sich ferner eine solide Leistung ableiten, die sich durch eine höhere Effizienz im Vergleich zu normalverteilten Y auszeichnet (Abbildung 4.3.3).

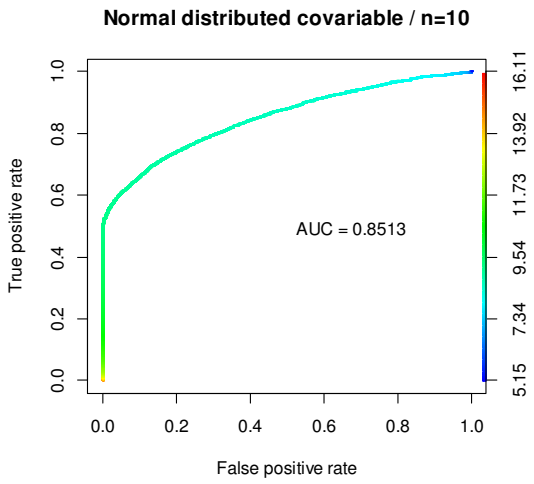

Abbildung 4.3.3: ROC-Kurven unter Poisson-verteilter abhängiger Variable Y_i

4.3.4 Ergebnisse unter binomialverteilter abhängiger Variable

Die Simulationen unter binomialen abhängigen Variablen bringen unter bestimmten Vorgaben keine adäquate GEE-Berechnung hervor. Betroffen ist davon die Kombination aus kleinen Datensätzen (n = 10) und Poisson- bzw. binomialverteilten Kovariaten. Bei den übrigen Simulationen insbesondere im Bereich größerer Datensätze (n = 20, 50) vermittelt das QIC den Eindruck einer beständigen Treffsicherheit hinsichtlich der korrekten Selektion von Modellen mit höherer Anpassungsgüte (Tabelle 4.3.4).

	AUC (true vs. incorrect model)		
n	normal	poisson	binomial
10	0.8541	--	--
20	0.9588	0.9577	0.9949
50	0.9983	0.9983	1.0000

Tabelle 4.3.4: AUC-Ergebnisse unter verschieden verteilten Kovariablen bei n=10, 20, 50; Y∼Binom

Vergleichbar zu den vorher beschriebenen Untersuchungen zeigt sich erneut eine reduzierte Diskriminanzleistung bei einer kleinen Anzahl von Fällen (n = 10) von ähnlichem Ausmaß (Abbildung 4.3.4).

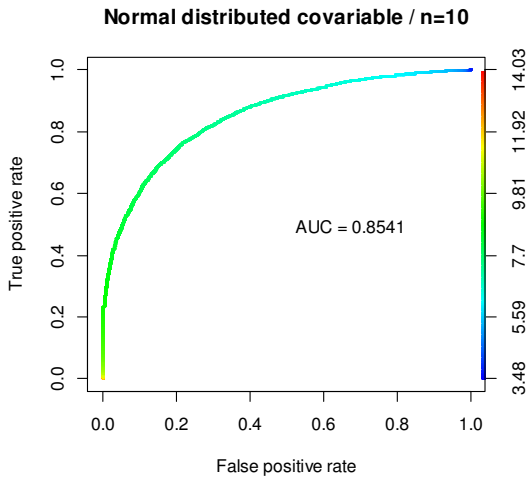

Abbildung 4.3.4: ROC-Kurven unter binomialverteilter abhängiger Variable Y_i

4.4 Diskussion und Zusammenfassung

Die durchgeführte Simulationsstudie attestiert dem QIC insgesamt eine effiziente Eigenschaft zur Modellselektion. Durch die Unabhängigkeit der Trennschärfe von der Korrelationsstruktur der random effects erfüllt sich eine Grundvoraussetzung für eine positive Einschätzung. Die weiteren Untersuchungen mit normal-, Poisson- und binomialverteilten Zielgrößen zeichnen stets ein ähnliches Bild von der Leistungsfähigkeit des QIC. Demnach kann für mittlere und große Datensätze (n = 20, 50) eine sehr gute Diskriminanzeigenschaft nachgewiesen werden, was anhand der sehr hohen AUC-Werte teils nahe dem Maximum (AUC = 1) belegt wird. Im Falle wenig vorteilhafter Vorgaben, wie sehr geringen Stichprobenumfängen (n = 10), kann eine deutlich reduzierte Trennschärfe zwischen Regressionsmodellen mit guter und schlechterer Datenanpassung verzeichnet werden. Die hier vorgegebenen Erschwernisse bereiten teilweise Schwierigkeiten bei der GEE-Berechnung. Demnach erweist sich der angelegte Schwierigkeitsgrad bereits als grenzwertig. Ferner rechtfertigt das Ausmaß der Einschränkung bei der Trennschärfe noch keinen Verzicht auf das QIC als Instrument zur Modellselektion.

Unter Berücksichtigung sämtlicher gewonnener Erkenntnisse hat sich das QIC als adäquates Mittel zur Modellselektion herausgestellt. Von zentraler Bedeutung ist die Abhängigkeit dieser positiven Bewertung von der zugrunde liegenden Datenstruktur des MNS-R. Demnach können sich außerhalb des longitudinalen Datenumfelds anderweitige Ergebnisse einstellen. Hierzu sollte ferner die simulierte Modellstruktur mit lediglich einer Kovariablen bedacht werden. Komplexere Modellierungen mit mehreren Einflussgrößen könnten sich gravierend auf die Diskriminanzeigenschaft auswirken. Weitere Untersuchungen unter anderen Voraussetzungen sind für eine allgemeine Beurteilung unerlässlich, würden aber den Rahmen dieser Arbeit übersteigen.

Kapitel 5

Matching-Vorgehen des Evaluationsteils II

Der zweite Teil der Evaluation des MNS-R beinhaltet, wie in Kapitel 1.3 erläutert, die gesundheitsökonomische Beurteilung der Intervention. Dabei ist die Fragestellung nicht auf eine mögliche Kostenreduzierung bei den Programmteilnehmern infolge der Behandlung limitiert (vorher-nachher-Vergleich). Vielmehr sollen ausgabenrelevante Erkenntnisse aus dem Vergleich zwischen multimodaler Gruppen- und konventioneller Schmerztherapie gewonnen werden (Fall-Kontroll-Vergleich).

5.1 Matching-Procedere in der Statistik

Die dafür erforderlichen Kontrollen werden anhand eines Matching-Verfahrens aus einer Population konventionell behandelter Schmerzpatienten ausgewählt. Aus diesem Grunde wird zunächst der technische Hintergrund von Matching-Verfahren in der Statistik erläutert.

Das Ziel eines Matchings besteht in der Identifikation und Zuordnung zweier zueinander ähnlicher Fälle. Idealerweise zeichnen sich diese durch nahezu gleiche Ausprägungen innerhalb definierter Variablen aus, wodurch sich optimale Voraussetzungen für eine Vergleichsanalyse ergeben. Meist werden Matching-Procedere im Bereich der Soziologie z. B. zur Datenfusion eingesetzt (Rässler, 2002), wo häufig auch der Begriff *Statistische Zwillinge* für zwei Matching-Partner zu finden ist. Allerdings stellen sich in den letzten Jahren zunehmende Anforderungen im öffentlichen Sektor ein, da auf Grundlage wachsender Umfänge von Registerdaten eine Extraktion spezifischer Datensätze für Analysezwecke notwendig wird (u. a. Lechner, 1999).

In der Statistik dienen Matching-Verfahren beispielsweise der Imputation fehlender Daten (Little & Rubin, 2002). Eine verbreitete Methode verwendet zur Berechnung eines Ersatzwertes die Ausprägungen der k ähnlichsten Fälle und wird daher *k-nearest-neighbors* ge-

nannt. Darüber hinaus kann anhand eines Matchings eine Kontrollgruppe für eine Subgruppenanalyse zusammengestellt werden. Hierbei fällt in Abhängigkeit zur Fragestellung dem Matching-Procedere eine zusätzliche Funktion zu, da sich durch die Anwendung Verzerrungen aufgrund von Selektionseffekten reduzieren lassen. Nachdem dieser Aspekt grundsätzlich von Bedeutung ist und ferner die Motivationsgrundlage in der MNS-R-Evaluation darstellt, erfolgt eine detailliertere Erläuterung anhand eines Beispiels:

Zugrunde liegt eine zu untersuchende Variable *Y*, etwa die Schmerzstärke bei Rückenschmerzleiden. Von Interesse ist nun der Gruppenunterschied zwischen MNS-R-Teilnehmern *i* und konventionell behandelten Patienten *k* hinsichtlich der Schmerzstärke. Die Gruppenzugehörigkeit eines jeden Patienten wird durch die binäre Variable *Therapieteilnahme (T)* beschrieben. Zur Klärung dieser Fragestellung anhand eines statistischen Modells könnte der Einfluss von *T* auf die Schmerzstärke *Y* geprüft werden

$$E\left(y_i \mid x_i\right) = \beta_0 + \beta_1 T_i + \sum_{j=2}^{p} \beta_j x_{ij} \,. \tag{1.8}$$

Dabei muss Beachtung finden, dass zweierlei Quellen für mögliche Schwierigkeiten existieren: Zum einen herrscht hinsichtlich der Größen der beiden Subpopulationen ein ausgeprägtes Ungleichgewicht. Im Vergleich zu konventionell behandelten Rückenschmerzpatienten sind die Programmteilnehmer zahlenmäßig stark unterlegen. Unter den Referenzpatienten finden sich demnach zahlreiche Beispiele, die sich von der Teilnehmergruppe deutlich unterscheiden, wodurch Heterogenität innerhalb dieser Population erzeugt wird. Die Wahrscheinlichkeit für Selektionseffekte, durch deren verzerrenden Einfluss auf die Analyse die Therapieeffekte nicht korrekt identifiziert werden können, erreicht folglich ein erhöhtes Ausmaß.
Zum anderen lassen sich leicht weitere Einflussgrößen $x_2, \ldots, x_p$ ausmachen, welche sowohl mit der abhängigen Variable *Y* als auch mit der Gruppenvariable *T* nicht unerheblich korreliert sind. Als Beispiele hierfür sind die Anzahl der Arbeitsunfähigkeitstage oder die Häufigkeit der Arztbesuche zu nennen. Diese auch als *Störgrößen* bezeichneten Merkmale weisen aufgrund der starken Ungleichgewichte ebenfalls deutlich unterschiedliche Verteilungen zwischen den Gruppen auf.

In dieser nicht unüblichen Konstellation zeichnet sich die Gesamtheit konventionell behandelter Patienten durch strukturelle Defizite aus, wodurch keine adäquate Referenzfunktion erfüllt werden kann. Allerdings besteht durch ein abgestimmtes Matching-Verfahren die Möglichkeit, Referenzpatienten derartig zu selektieren, dass ähnliche Verteilungen in beiden Gruppen resultieren. Entsprechend kann eine ausgewogenere Balance hinsichtlich der Gruppengröße sowie in Bezug auf Störgrößen erreicht werden (Breslow et al., 1978 / Schlesselman, 1982). Infolgedessen lässt sich die Effizienz einer Studie maßgeblich erhöhen, sofern die Matching-Variablen sowohl eine Korrelation mit der abhängigen Variable *Y* als auch mit der Einflussgröße *T* aufweisen (Clayton & Hills, 1993). Es versteht sich, dass der durch das Matching gewonnene Informationsgehalt in der Konstruktion einer späteren Analyse Beachtung finden muss (z. B. Fall-Kontroll-Studien). Genauer betrachtet betrifft dies die Zuordnungen zwischen Interventions- und Kontrollpatienten (Stratum), welche es stets in die Berechnungen zu integrieren gilt. Meist erfolgt dies durch Einbeziehung der Stratumsvariable in eine statistische Modellierung.

Das Vorgehen in der Entwicklung des Matching-Procederes besteht vorwiegend aus zwei unabhängigen Schritten. Zunächst gilt es, diejenigen Merkmale (meist Störgrößen) auszuwählen, welche als Matching-Variablen fungieren sollen. Für jedes dieser Merkmale wird individuell ein Kriterium definiert, das es von den späteren Kontrollen der Analyse zu erfüllen gilt. Hinsichtlich der Anzahl der Kriterien existieren weitreichende Empfehlungen. Während manche Ansätze so wenig wie möglich Kriterien befürworten, raten andere dazu, sämtliche mit *Y* korrelierte Variablen einzubeziehen (Smith, 1997).
Anschließend wird anhand der definierten Kriterien die Ähnlichkeit zwischen den Fällen *i* (hier: MNS-R-Patient) und den Kontrollen *k* bestimmt. Die Wahl der Methodik zur Bestimmung der Ähnlichkeiten erfolgt in Abhängigkeit der Fragestellung der späteren Analyse und der Datenstruktur. Hierzu existieren verschiedene Ansätze, wobei meist die folgenden beiden Methoden Anwendung finden:

Berechnung des *Propensity Score*

Der Berechnung des Propensity Score liegt die Anwendung eines logistischen Regressionsmodells

$$P\left(T_i = 1 \middle| x_i\right) = \frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)}$$

zugrunde. Dabei findet die Indikatorvariable für die Therapieteilnahme *T* (Gruppenvariable in (1.8)) ihre Funktion als abhängige Größe, während sämtliche Matching-Variablen durch Einbindung in den linearen Prädiktor $x_i'\beta$ in die Modellierung integriert werden. Die Durchführung der Regressionsberechnung liefert die entsprechenden Parameterschätzungen $\hat{\beta}_1, \ldots, \hat{\beta}_p$, mit deren Hilfe für jeden Patienten sowohl aus der Interventions- als auch aus der Kontrollgruppe ein bedingter Wahrscheinlichkeitswert (*Propensity Score*) bestimmt werden kann. Durch die Distanzen zwischen den Scorewerten eines Falls *i* und jeder Kontrolle *k*

$$d\left(\ P\left(T_i = 1 \middle| x_i\right), P\left(T_k = 1 \middle| x_k\right)\ \right) = d\left(P_i, P_k\right) = \left|P_i - P_k\right|$$

werden sämtliche Elemente beurteilt und die Matching-Partner für die spätere Analyse ausgewählt. Demnach gilt derjenige Kontrollpatient *k* dem Programmteilnehmer *i* am ähnlichsten, wenn er eine minimale Distanz *d* aufweist, vermöge

$$k^* = \arg\min_k d\left(P_i, P_k\right).$$

Im Falle mehrerer gesuchter Matching-Partner bietet sich die Möglichkeit, entweder die Anzahl der auszuwählenden ähnlichsten Kontrollen festzulegen oder eine Konstante c vorzugeben, die von der Distanz *d* nicht überschritten werden darf. Für die Bestimmung einer geeigneten Konstanten empfiehlt es sich, verschiedene Werte für c zu testen – beginnend mit dem kleinsten (z. B. c = {0.0001, 0.001, 0.01}). Somit lässt sich ein für die zugrunde liegende Datensituation angemessener Wert finden, ohne einer zu großen oder zu kleinen Anzahl an Matching-Partnern zu unterliegen.

Bewertung mittels statistischer Distanzmaße

Ein alternatives Vorgehen bietet sich durch die Berechnung statistischer Distanzmaße. Grundsätzlich liegt der Interpretation dieser Maßzahlen die Vorstellung zugrunde, dass sich zwei Individuen *i* und *k* umso ähnlicher sind, je kleiner das Ergebnis ihrer Distanz

$$d(x_i, x_k) = \|x_i - x_k\|_q = \left\{ \sum_{j=1}^{p} \left| x_{ij} - x_{kj} \right|^q \right\}^{\frac{1}{q}} \tag{1.9}$$

ausfällt. Dabei werden zwei aus (1.9) resultierende Abstandsmaße am häufigsten herangezogen: Die Cityblock-Metrik für $q = 1$ und der euklidische Abstand für $q = 2$.

Im Vergleich zur Methode mittels Propesity Score bietet sich hierbei die Möglichkeit einer äußerst flexiblen Anpassung an die spätere Analysesituation. Als nachteilig erweist sich allerdings der Umstand, dass unterschiedlichen Skalenniveaus Rechnung getragen werden muss. Demnach ist es notwendig, durch Gewichtungen derart auf die Distanzberechnung einzuwirken, dass keine Bevorzugung oder Benachteiligung einer Variable aufgrund ihres Wertebereiches erfolgt. Ohne Korrektur könnte der Einfluss einer Variable auf das Ergebnis unerwünscht groß oder klein ausfallen. Eine geeignete Verdeutlichung bietet beispielsweise der Vergleich einer binären Variable wie Arbeitslosigkeit gegenüber einem stetigen Merkmal wie dem Alter. Bei Nichtübereinstimmung hinsichtlich der Arbeitslosigkeit entspricht hierbei der maximale Strafterm in der Distanzberechnung einem Altersunterschied von einem Lebensjahr zwischen Fall und potentiellem Matching-Partner. Je nach Datenlage hätte dies lediglich einen geringen Einfluss auf das Ergebnis des Abstandsmaßes. Ein häufig verwendeter Ansatz für geeignete Ausgleichsgewichtungen besteht aus einer Multiplikation mit dem Inversen der Standardabweichung der jeweiligen Variable. Neben ihrer Korrekturfunktion bieten die Gewichtungen ferner die Möglichkeit zur Justierung des gesamten Matchings je nach Relevanz und gewünschter Einflussstärke des jeweiligen Kriteriums.

Insgesamt ist es bei der Entwicklung eines optimalen Matchings erforderlich, verschiedene Faktoren zu berücksichtigen. Von besonderer Bedeutung ist demnach die Planung und Evaluation der Prozedur sowie die Validierung der individuellen Programmteile. Dabei müssen verschiedene Fragen im Vorfeld geklärt werden, beispielsweise wie mit einer mehrfachen

Verwendung einer Kontrolle umgegangen werden soll. Denkbar ist hierbei eine limitierte Anzahl an Verwendungen eines Kontrollpatienten, sodass sich die Varianzen nicht zu stark reduzieren. Ferner erweist sich auch eine Vorauswahl der Kontrollpopulation als nützlich, um einige Kriterien mit einer Ausschlussmöglichkeit zu versehen.
Nach der Umsetzung kann anhand von Testläufen die generelle Wirkungsweise der gesetzten Kriterien bewertet werden. Dabei ist auch die exemplarische Prüfung einzelner Patienten von Bedeutung, um die Ähnlichkeit von Kontrollen und Fällen und somit die Leistungsfähigkeit des Matchings zu beurteilen.

5.2 Ausgangssituation im MNS-R

Die ökonomische Analyse erfordert für optimale Vergleichsvoraussetzungen von MNS-R- mit Kontrollpatienten eine jeweils zeitlich parallele Gegenüberstellung. Nachdem im Studiendesign grundsätzlich auf eine randomisierte Kontrollgruppe verzichtet wurde, müssen für den zweiten Evaluationsteil entsprechende Kontrollpatienten ermittelt werden. Genauer betrachtet ist es nicht ausreichend, eine gemäß den Eigenschaften der Patientenpopulation vergleichbare Kontrollgruppe zu generieren. Vielmehr werden für einen adäquaten Vergleich jedem Programmteilnehmer mehrere konventionell behandelte Rückenschmerzpatienten zugeordnet. Von Bedeutung ist hierbei eine möglichst umfassende Ähnlichkeit der Ausprägung ausgesuchter Merkmale zwischen MNS-R-Teilnehmern und Kontrollpatienten.

Der Bestand longitudinaler Daten aus dem ersten Evaluationsteil basiert auf einer exklusiven Befragung der MNS-R-Teilnehmer. Folglich bietet sich hier keine ausreichende Grundlage für die Bewertung weiterer externer Patienten. Darüber hinaus weisen die erhobenen Daten aus ökonomischer Sicht verschiedene Schwachpunkte auf, sodass eine Verwendungseignung für die zugrunde liegende Untersuchung ausgeschlossen werden muss. Beispielsweise resultieren aus den Befragungen lediglich subjektive Einschätzungen der Betroffenen. Zwar bringt dies für die medizinische Bewertung im ersten Evaluationsteil erhebliche Vorteile hinsichtlich der Charakterisierung des individuellen Schmerzleidens. Allerdings stützen sich auch die Antworten ökonomischer Fragestellungen auf die eigene Wahrnehmung und Erinnerung der Befragten. Insbesondere in Bezug auf länger zurückliegende

Behandlungsereignisse wie Operationen und Rehabilitationsbehandlungen stellen daher abrechnungsseitige Datenquellen eine stichhaltigere Alternative dar.

Nachdem die Durchführung des Projekts MNS-R auf der Kooperation zwischen der SBK und der Klinik für Anästhesiologie basiert, ergibt sich die vorteilhafte Situation einer Verfügbarkeit von Krankenkassendaten. Diese Bestände liefern stichhaltige und umfangreiche Informationen zu Behandlungen und deren Abrechnungen nicht nur über die Programmteilnehmer, sondern über sämtliche Versicherte. Somit eröffnet sich die Möglichkeit, weitere Personen mit vergleichbaren Krankheitsverläufen im Datenbestand der SBK zu identifizieren. Ferner kann anschließend ein stichhaltiger Vergleich der Kostenentwicklung zwischen Teilnehmern und ihren konventionell behandelten Kontrollen im zeitlichen Verlauf angestellt werden.

5.3 Matching-Konzept des MNS-R

Im Hinblick auf eine stichhaltige Analyse ist die Identifizierung von Kontrollpatienten und deren Zuordnung zu dem jeweiligen MNS-R-Teilnehmer (Matching) von zentraler Bedeutung. Falls die Kontrollen sich in wesentlichen soziodemografischen oder krankheitsspezifischen Merkmalen zu stark von ihren Partnern unterscheiden, würde eine Analyse inadäquate Ergebnisse hervorbringen. Daher besteht die Notwendigkeit einer umfangreichen Auseinandersetzung mit der Konzeption des Matching-Vorgehens und deren technischen Umsetzung.

Grundsätzlich kann jeder konventionell behandelte SBK-Versicherte als potentieller Matching-Partner für einen MNS-R-Patienten herangezogen werden. Eine optimale Kontrolle gleicht dem Programmteilnehmer nicht nur in wesentlichen soziodemografischen Merkmalen, sondern vor allem in seiner Krankheitsgeschichte. Diesbezügliche Ähnlichkeiten lassen sich anhand der Behandlungsereignisse im zeitlichen Verlauf feststellen. Hierbei findet sich bei jedem Patienten eine unregelmäßige Abfolge gesundheitsökonomischer Ereignisse wie stationäre und ambulante Versorgungen, Rehabilitations- und physiotherapeutische Behandlungen sowie medikamentöse Verordnungen.

Der Krankheitsverlauf eines MNS-R-Patienten beinhaltet den Zeitpunkt der Programmteilnahme, welcher den zeitlichen Verlauf in eine Periode vor und eine nach der Intervention gliedert (Abbildung 5.3.1). Exakt diese Unterteilung wird auf die Verläufe der Kontrollen übertragen, um der Vergleichsanalyse stets dieselben Zeiträume zugrunde zu legen. Generell dienen die Ereignisse vor der Intervention der Beurteilung der Ähnlichkeit zwischen Programmteilnehmer und anderen SBK-Versicherten. Zeigen sich hier entsprechende Übereinstimmungen, folgt die Auswahl der Versicherten als Matching-Partner. Die Beurteilung der Ähnlichkeiten und damit die Identifizierung der Partner erfolgt anhand vorher festgelegter Kriterien. Deren Prüfung wird mithilfe eines eigens entwickelten Algorithmus durchgeführt. Für die Auswahl des relevanten Beurteilungszeitraums eignen sich beispielsweise die letzten zwölf Monate vor Beginn der Intervention.

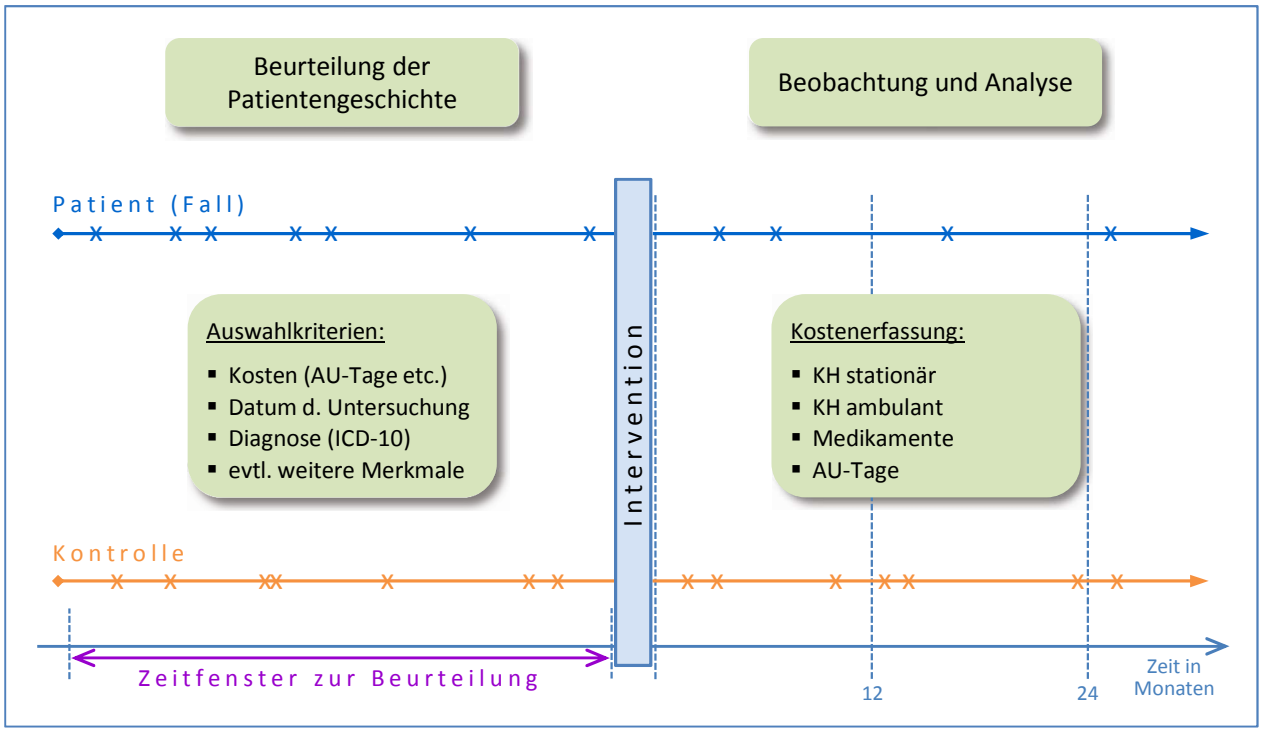

Abbildung 5.3.1: Grafische Darstellung des Matching-Konzepts

Die durch das Matching ermittelten Kontrollpatienten stellen das Vergleichskollektiv für die kostenanalytische Fall-Kontroll-Studie dar. Im Vordergrund steht hierbei die Frage nach der Entwicklung sämtlicher Behandlungskosten im Anschluss an die MNS-R-Teilnahme. Mögliche Reduzierungen bei der Inanspruchnahme der gesundheitlichen Versorgung im Vergleich zu den konventionell behandelten Kontrollen lassen einen Schluss auf eine Kostenef-

fektivität der Programmteilnahme zu. Für eine detaillierte Erläuterung des Analysevorgehens sei auf Kapitel 6 verwiesen.

5.4 Aufbau des Datenbestands

Im Zuge der Realisierung des Matching-Konzepts bei der MNS-R-Evaluation stand zunächst der Aufbau einer geeigneten Datenbank im Vordergrund. Wie bereits erläutert, werden die Daten von der SBK zur Verfügung gestellt. Zur Wahrung des Datenschutzes sind sämtliche Daten durch Patienten-Identitätsnummern pseudonymisiert. Grundsätzlich umfasst die Datenlieferung neun Einzelbereiche (Tabelle 5.4.1) aus den Jahren 2008, 2009 und 2010 und ist in 16 Datenmatrizen aufgeteilt.

Datenbereiche
Stammdaten
Rentendaten
Daten stationärer Behandlungen
Daten ambulanter Behandlungen
Arzneimitteldaten
Daten bzgl. Arbeitsunfähigkeit
Übersicht über Heilmittelverordnungen
Übersicht über Rehabilitationsmaßnahmen
Daten bzgl. Hierarchischer Morbiditätsgruppen (HMG)

Tabelle 5.4.1: Aufgliederung der SBK-Datenlieferung

Aufgrund der klaren Ordnung in der Datenstruktur konnte bei der Planung des zweiten Evaluationsteils der Aufwand des Datenmanagements klar abgeschätzt werden. Ein wichtiger Teil umfasste die Kontrollen des umfangreichen Datenmaterials auf Plausibilität. Dabei müssen sowohl Abstimmungen innerhalb einer jeden Datenmatrix als auch zwischen den einzelnen Datenbankfragmenten eingeschlossen werden. Beispielsweise soll jeder Patient bei einem Arztbesuch einerseits eine Diagnose erhalten haben und andererseits in der Stammdatenmatrix enthalten sein. Dieser Überlegung folgend besteht grundsätzlich die Möglichkeit, den Arbeitsumfang für die Datenvalidierung abzuschätzen. Unabhängig von der Sorgfalt und Datenbankpflege ist die Bearbeitung großer Datenmengen stets mit einem erhöhten Aufwand verbunden und bedarf meist der Korrektur einzelner Datenbankfrag-

mente. Dennoch kann selbst in diesem Fall ein durchschnittlicher Mehraufwand explizit kalkuliert werden, sodass derartige Vorgänge die Arbeitsplanung nicht überstrapazieren.

Die Entwicklung des Matching-Konzepts erforderte die Durchführung mehrerer Datenabrufe, wodurch sich unerwartet beträchtliche Abstimmungsschwierigkeiten einstellten. Als Ursache erwies sich letztlich die Komplexität und Organisation des deutschen Gesundheitswesens. Obwohl die durchgeführte Evaluation im Vergleich zu großen, staatlich gesponserten Studien lediglich einen überschaubaren Einblick ermöglicht, konnte dennoch ein Eindruck über den Umfang und die Zusammensetzung der Datenbestände deutscher Krankenversicherungen gewonnen werden. Grundsätzlich erhalten die Krankenkassen ihre Daten aus unterschiedlichen Bereichen des deutschen Gesundheitswesens. Von außen betrachtet unterliegen die Vorgänge der Datentransfers einer Organisation, deren Effizienz einer erheblichen Steigerung bedarf. Bedingt durch zeitlich verzögerte Meldungen und permanente Datenkorrekturen unterliegt der Datenbestand einer Krankenkasse laufenden Veränderungen. Infolgedessen führen zwei identische, aber zeitlich auseinanderliegende Datenabrufe zu unterschiedlichen Datensätzen. Entsprechend können bereits geringfügige Korrekturen von Studiendaten einen im Vorfeld nicht zu kalkulierenden Aufwand an das Datenmanagement stellen. Erschwerend wirkt die konsequente Einbeziehung entbehrlicher Datenmengen auf die Situation. Beispielsweise enthalten die Datenbestände neben den gesicherten auch sämtliche Verdachtsdiagnosen einer jeden Behandlung.

Insgesamt zeigte sich die bei der Studienplanung veranschlagte durchschnittliche Anforderung an das Datenmanagement um ein Vielfaches unterdimensioniert. Es ist daher festzuhalten, dass eine effiziente Planung auf Krankenkassendaten basierender Studien kaum realisierbar ist. Letztlich verursachen die beschriebenen Bedingungen nicht vorhersehbare Belastungen in Form anhaltender Abstimmungsarbeiten und einem erhöhtem Aufwand an Datenpflege. Die Dimension des zusätzlich notwendigen Arbeitspensums ist demnach bei der Studienplanung im Vorfeld nicht seriös abschätzbar. Infolgedessen wurden im Laufe der letzten Jahre die Studienengagements unter Verwendung eigener Datenbestände seitens der Krankenkassen sukzessive reduziert. Aus biometrischer Sicht bringt diese Entwicklung einen bedeutenden Nachteil mit sich, da stichhaltiges Datenmaterial aus den Abrechnungsvorgängen einer gesundheitsökonomischen Nutzung versagt bleibt.

5.5 Technische Umsetzung des Matchings

Für die methodische Umsetzung des Matching-Konzepts (Kapitel 5.3) stehen grundsätzlich verschiedene technische Ansätze zur Verfügung. Im Folgenden werden zwei Prozeduren näher untersucht:

Eine erste einfache Herangehensweise besteht aus dem sukzessiven Filtern der Versicherungspopulation nach festgelegten Kriterien (Selektionsverfahren). Dieses sehr strikte Vorgehen reduziert die Auswahl potentieller Matching-Partner umso stärker, je größer die Anzahl an Filterkriterien ist. Der Nachteil dieser Methode besteht aus der fehlenden individuellen Selektionsmöglichkeit aus der verbleibenden Auswahl an SBK-Versicherten. Demnach kann das Ziel der Identifikation *möglichst* ähnlicher Kontrollpatienten noch nicht vollständig erreicht werden. Zudem ist zu bedenken, dass der Selektionsmechanismus für die einzelnen Programmteilnehmer sehr unterschiedlich große Kontrollgruppen hervorbringen kann. Im Gegensatz zu jungen MNS-R-Teilnehmern steht beispielsweise älteren Patienten in der Regel eine deutlich größere Anzahl an Kontrollen zur Verfügung. Insbesondere bei einer Vielzahl an potentiellen Matching-Partnern gestaltet sich die Identifikation der vier ähnlichsten allein durch Selektion entsprechend schwierig.

Eine zweite Möglichkeit zur technischen Realisierung des Matchings bietet die Berechnung statistischer Abstandsmaße (Ähnlichkeitsverfahren). Als Resultat liegt zu jedem Paar, bestehend aus Programmteilnehmer und jedem einzelnen konventionell behandelten SBK-Versicherten, eine Ähnlichkeitsaussage innerhalb der zugrunde gelegten Kriterien vor. Entsprechend stellt dies eine Bewertungsmöglichkeit darüber dar, welcher Kontrollpatient dem MNS-R-Teilnehmer insgesamt ähnlicher ist. Die Sortierung sämtlicher potentieller Matching-Partner in eine Rangfolge gemäß ihrer Abstandsergebnisse führt folglich zur Identifikation der ähnlichsten Partner. Zwar kann das Ziel des Matchings mithilfe dieser Methode erreicht werden, dennoch finden sich auch hier erhebliche Nachteile. Zum einen existiert keine Ausschlussmöglichkeit von Matching-Partnern, was allerdings in Bezug auf Alter oder Geschlecht von wesentlicher Bedeutung ist. Zum anderen gestaltet sich die Berechnung eines Abstandsmaßes zwischen jedem Programmteilnehmer und jedem anderweitigen SBK-Versicherten überaus zeit- und ressourcenaufwendig.

Die Umsetzung des Matchings bei der Evaluation des MNS-R besteht folglich aus einer Kombination von Selektions- und Ähnlichkeitsverfahren, da sich die Vorteile beider Methoden optimal ergänzen. In der Praxis werden zunächst allgemeine und teilnehmerspezifische Selektionsvorgänge durchgeführt. Hierbei kommen sämtliche Kriterien zur Anwendung, bei denen ein strikter Ausschluss unabdingbar ist. Als Ergebnis dieses Auswahlvorgangs ist jedem MNS-R-Teilnehmer bereits eine Population konventionell behandelter Rückenschmerzpatienten zugeordnet, welche die individuellen soziodemografischen Eigenschaften aufweisen.

Matching-Kriterien, bei denen Abweichungen keinen Ausschluss zur Folge haben muss, werden innerhalb der anschließenden Ähnlichkeitsberechnung berücksichtigt. Die Kontrollgruppe eines jeden Programmteilnehmers bleibt hierbei bereits unverändert. Im Fokus steht lediglich die Identifizierung von Kontrollen mit möglichst ähnlichen Merkmalsausprägungen. Sofern die Kontrollpopulation eine hinreichende Größe umfasst, wird die Anzahl der Kontrollen auf die vier ähnlichsten reduziert.

5.5.1 Kriterien des Selektionsverfahrens

Während des Selektionsverfahrens wird die Population sämtlicher SBK-Versicherten (exklusive der MNS-R-Patienten) hinsichtlich folgender Kriterien vorselektiert:

Soziodemografische Kriterien

Nachdem eine exakte Übereinstimmung des Alters die Auswahl der Matching-Partner zu stark eingrenzt, wird eine Abweichung von ± 2 Lebensjahren zugelassen. In Bezug auf das Geschlecht erfolgt eine strikte Selektion, da sich die Therapie chronischer Schmerzleiden zwischen Frauen und Männern erheblich unterscheidet (Robert Koch-Institut (Hrsg.), 2012).

Region des Wohnortes

Ein weiterer, für die Kostenanalyse relevanter Sachverhalt besteht in der unterschiedlich ausgeprägten gesundheitlichen Versorgung zwischen Ballungs- und ländlichen Wohngebieten. Entsprechend wird ein Matching-Kriterium konstruiert, welches auf eine vergleichbare Versorgungssituation zwischen den Therapieteilnehmern und ihren Kontrollpatienten abzielt. Als Grundlage wird die Postleitzahl des Wohnortes herangezogen. Unter der Verwen-

dung eines Postleitzahlenclusters mit einem Radius von 30 km um die Ballungsgebiete wird eine Zuordnung von Kontrollen unabhängig der Stadtteile zugelassen. Außerhalb dieser Cluster erfolgt eine Selektion gemäß den ersten beiden Stellen der Postleitzahl.

Rentenstatus

Bei zahlreichen chronischen Schmerzpatienten, welche sich in ihrer Erkrankung eingerichtet haben, zeichnet sich in der Regel eine erschwerte Motivierung für neue Therapieansätze ab. Dies ist insbesondere dann der Fall, wenn die Arbeitsmotivation bereits eingeschränkt und das Bestreben nach einem Rentenstatus wegen Erwerbsminderung vorhanden ist. Dabei ist nicht ausschlaggebend, ob bereits eine Rente bezogen wird oder erst ein Antrag gestellt wurde. Für den Einfluss auf den Therapieerfolg ist lediglich die subjektive Einstellung des Patienten relevant. Diesem Sachverhalt folgend muss der Rentenstatus in das Matching-Procedere eingeschlossen werden. Allerdings ist aufgrund der Komplexität des Deutschen Rentensystems mit seiner Vielzahl an Rentenarten eine detaillierte Berücksichtigung nicht zielführend. Infolgedessen wird lediglich strikt nach der Erwerbsminderung aufgrund von Rückenschmerzen (Antragstellung und Rentenbezug) selektiert.

Komorbiditäten

Eine adäquate Bewertung der Krankheitsgeschichte verlangt grundsätzlich eine ganzheitliche Betrachtung des Patienten. Daher wird ein weiteres Kriterium formuliert, welches neben den Charakteristika des Krankheitsverlaufes chronischer Rückenschmerzen die individuellen Komorbiditäten einschließt. Aufgrund der gesundheitsökonomischen Ausrichtung des zweiten Evaluationsteils soll beim Matching neben den Komorbiditäten selbst auch deren Kostenaufwand erfasst werden. Ein Instrument, welches beide Aspekte zusammenführt, findet sich in der Finanzierungsmethode des Deutschen Gesundheitssystems wieder. Genauer betrachtet werden seit dem Jahr 2009 die unterschiedlichen Risikokonstellationen der Krankenkassen anhand von Morbiditäten ausgeglichen. Dieser sogenannte *Morbiditäts-Risikostrukturausgleich (Morbi-RSA)* wird auf Basis *Hierarchischer Morbiditätsgruppen (HMG)* bestimmt. Der grobe Aufbau dieser Gruppen lässt sich wie folgt darstellen:
Verschiedene Diagnosen, welche einem medizinischen Zusammenhang unterliegen, werden in Diagnosegruppen (Dx-Gruppen) zusammengefasst. Diese sind nach festgelegten ökonomischen Gesichtspunkten zu Morbiditätsgruppen (MG) gebündelt. Mehrere MG werden ih-

rerseits wiederum in insgesamt 25 Morbiditätsklassen zusammengefasst. Nachdem die MG innerhalb einer jeden Klasse gemäß dem Schweregrad der Erkrankung hierarchisiert sind, werden die 25 Klassen auch als *Morbiditätshierarchien* bezeichnet. Die ausschlaggebende Einteilung, der ein Patient nach Häufung einer Diagnose zugeteilt wird, besteht in den MG. Aufgrund deren Organisation innerhalb der Hierarchien werden diese im Allgemeinen HMG genannt.
Die Verarbeitung des Morbi-RSA-Ansatzes als Matching-Kriterium erfolgt in zwei separaten Schritten. Eine strikte Konformität wird lediglich für die Morbiditätshierarchien gefordert (Selektionsmethode), was der übergeordneten Art der Erkrankung Rechnung tragen soll. Eine Übereinstimmung der einzelnen HMG zweier Patienten wird hingegen in die Ähnlichkeitsberechnung integriert. Je geringer die Überschneidung der individuellen HMG ausfällt, desto größer wird der Term, der zur Ähnlichkeit bestrafend addiert wird. Hinsichtlich möglicher Anpassungen während des Matchings sei auf Kapitel 6.1 verwiesen.

Arbeitsunfähigkeit

Als einer der größten Verursacher indirekter Kosten bilden Arbeitsausfälle aufgrund von Rückenschmerzen einen wesentlichen Bestandteil gesundheitsökonomischer Untersuchungen. Potentielle Matching-Partner müssen daher in etwa die Anzahl an Arbeitsunfähigkeitstagen ihres MNS-R-Pendants aufweisen. Es wird eine maximale Abweichung von ± 14 Tagen zugelassen.

Ein weiteres denkbares Kriterium zur Selektion ergibt sich aus der Verfügbarkeit der Kostendaten. Demnach könnte eine Vorgabe erfolgen, welche eine Übereinstimmung des Kostenrahmens aus Arzneimittelverordnungen, Rehabilitationsmaßnahmen und Heilmitteln wie z. B. Physiotherapie verlangt. Bei der Evaluation des MNS-R wurde dies lediglich vorübergehend zu Testzwecken umgesetzt.
Hinsichtlich der Kriterien für das Selektionsverfahren sei abschließend nochmals darauf hingewiesen, dass lediglich Rückenschmerz relevante Ereignisse verarbeitet werden. Dies spiegelt sich insbesondere in einer Einschränkung der Population potentieller Matching-Partner wider. Entsprechend wurden SBK-Versicherte ohne Rückenschmerzdiagnosen bereits vor Anwendung des Selektionsverfahrens ausgeschlossen. Für weitere Informationen sei auf den Abschlussbericht des zweiten Evaluationsteils verwiesen.

5.5.2 Kriterien des Ähnlichkeitsverfahrens

Die zentrale Aufgabe der Ähnlichkeitsberechnung liegt in der Beurteilung der Krankengeschichten. Dabei bleibt für jeden MNS-R-Patienten der Umfang seiner Kontrollpopulation unverändert. Aus dem selektierten Kollektiv werden lediglich die Kontrollpatienten mit den ähnlichsten Krankheitsverläufen identifiziert. Dabei besteht die Möglichkeit, mittels unterschiedlicher Gewichtungen der Kriterien studienrelevante Akzente zu setzen. Somit kann der Einfluss der einzelnen Charakteristika und Ereignisse entsprechend erhöht oder reduziert werden.

Diagnosen

Das wesentliche Merkmal zur Beurteilung, ob sich die Krankheitsgeschichten zweier Individuen gleichen, besteht in der Betrachtung der Diagnosen. Folglich werden zu sämtlichen Rückenschmerzereignissen innerhalb des Bezugszeitraumes die gesicherten Diagnosen zusammengefasst. Relevant sind hierbei stationäre und ambulante Versorgungen sowie Arbeitsunfähigkeitsepisoden. Als Matching-Kriterium dient der prozentuale Anteil an identischen Diagnosen zwischen denen des MNS-R-Teilnehmers und jenen der Kontrollpatienten. Zur Vorbeugung eines Selektionseffektes in Form einer Bevorzugung des MNS-R-Patienten wird dieser Anteil nicht ohne weiteres aus Sicht des Programmteilnehmers bestimmt. Vielmehr wird eine symmetrische Beurteilung beispielsweise mittels der Verwendung einer Minimumfunktion präferiert.

Anzahl der Behandlungen

Insbesondere aus gesundheitsökonomischer Sicht ist für eine Bewertung der Krankheitsgeschichte die Versorgungsanforderung eines Patienten relevant. Entsprechend wird die Anzahl stationärer und ambulanter Behandlungen in die Ähnlichkeitsberechnung eingebunden.

Überblick über Matching-Kriterien

Zusammenfassend kann festgehalten werden, dass die Kombination aus Selektionsmechanismus und Ähnlichkeitsbewertung die nötige Flexibilität mit sich bringt, um die einzelnen Kriterien mit unterschiedlicher Intensität anzuwenden. Abhängig von der Vorgabe, ob eine Diskrepanz zwischen Programmteilnehmer und Kontrollpatient hinsichtlich eines Merk-

mals zum Ausschluss der Kontrolle führen soll, wird dieses Kriterium eines der beiden Verfahren zugeordnet. Dies ermöglicht eine praxisnahe Realisierung passgenauer Matching-Konzepte, ohne Kompromisse aufgrund der technischen Umsetzbarkeit eingehen zu müssen. Für eine bessere Übersichtlichkeit sind im Folgenden sämtliche Matching-Kriterien der MNS-R-Evaluation inkl. der jeweiligen Anwendung zusammengefasst (Tabelle 5.5.1). Überwiegend werden soziodemografische Eigenschaften einer strikten Auswahl unterzogen, während die Bewertung der Krankheitsgeschichte der Ähnlichkeitsbewertung unterliegt.

Kriterien des Selektionsmechanismus	**Kriterien der Ähnlichkeitsberechnung**
Alter ± 2 Jahre	Überschneidung der Diagnosen
Geschlecht	Überschneidung der Morbiditätsgruppen
Region des Wohnortes	Häufigkeit stationärer Behandlungen
Rentenstatus wg. Erwerbsminderung	Häufigkeit ambulanter Behandlungen
HMG-Klassifikation	
Arbeitsunfähigkeitstage ± 14 Tage	

Tabelle 5.5.1: Übersicht der Matching-Kriterien inkl. methodischer Zuordnung

Kapitel 6

Gesundheitsökonomische Analyse

Durch die vorgegebene Patientenpopulation aus MNS-R-Teilnehmern und den jeweils dazu gematchten Kontrollen ergibt sich eine gänzlich andere Datensituation im Vergleich zum ersten Evaluationsteil. Entsprechend ist hinsichtlich der gesundheitsökonomischen Untersuchung eine andere analytische Vorgehensweise notwendig, welche ab Kapitel 6.2 vorgestellt wird.

Die Durchführung des Matchings ist bei den meisten MNS-R-Patienten erfolgreich verlaufen: In diesen Fällen liegt eine Identifikation von mindestens vier Matching-Partnern vor, welche einen ähnlichen Krankheitsverlauf aufweisen. Nachdem das Matching die Datengrundlage für die gesundheitsökonomische Analyse hervorbringt, ist zunächst eine Überprüfung und Bewertung der angelegten Kriterien von Bedeutung.

6.1 Bewertung der Matching-Kriterien

Insgesamt muss eine Bewertung unter dem Gesichtspunkt geführt werden, dass die vorliegende Teilnehmerpopulation ein sehr spezielles Patientenkollektiv darstellt. Aufgrund der langjährigen Krankheitsgeschichte sind psychische Störungen und gravierende Folgeerkrankungen keine Seltenheit. Diese speziellen Umstände spiegeln sich auch im Matching-Verlauf wider. Demnach resultiert ein relativ hoher Anteil an Patienten mit nicht oder nur eingeschränkt erfolgreichen Matching-Versuchen. Unter Betrachtung der beiden methodischen Ansätze ist dies als Aussortierung sämtlicher potentieller Kontrollpatienten während des Selektionsverfahrens zu verstehen. Dabei zeigen sich ausgeprägte Unterschiede zwischen den verschiedenen Kriterien im Hinblick auf die Reduzierung der Versichertenpopulation.

Selektionsverfahren

Bei der Durchführung des selektiven Prozesses innerhalb des Matchings reduziert sich die Anzahl potentieller Matching-Partner schrittweise mit der Umsetzung der einzelnen Kriterien. Meist bewirken die Prozeduren eine dem Merkmal angemessene Verkleinerung der Kontrollgruppe jedes Programmteilnehmers. Beispielsweise folgt aus der Forderung nach der Übereinstimmung der Geschlechter eine Reduzierung um ca. 50 % (Tabelle 6.1.1).

Kriterium	start	age	sex	pension	morbi-class	au.days	final
Anzahl Kontrollen	129913	10556	5375	4980	1192	695	
Relative Veränderung	100 %	-91.87 %	-49.08 %	-7.36 %	-76.06 %	-41.66 %	0.53 %

Tabelle 6.1.1: Durchschnittliche Reduzierung der Kontrollpopulation innerhalb der Ballungsgebiete durch das Matching

Nachdem das Alter mit einem kleinen Toleranzbereich von ± 2 Lebensjahren als erstes Kriterium Anwendung findet, stellt auch eine durchschnittliche Reduzierung um über 90 % kein außergewöhnliches Ergebnis dar. Auffälligkeiten zeichnen sich hingegen im Bereich der Morbiditätsklassen und der Arbeitsunfähigkeitstage (AU-Tage) ab, da hier umfangreiche Reduzierungen der Kontrollgruppen resultieren.

Bei näherer Betrachtung der Patienten mit erfolglosen Matching-Durchläufen sind hinsichtlich der AU-Tage besonders lang andauernde oder häufig vorkommende Arbeitsausfälle erfasst. Vor dem Hintergrund einer chronischen Schmerzerkrankung mit langjährigen Beschwerden sind ausgeprägte Arbeitsunfähigkeitsperioden zwar als plausibel einzustufen. Dennoch können diese Ausreißereigenschaften der Patienten durch das Matching nicht kompensiert werden. Selbst eine testweise Toleranzerweiterung des Kriteriums auf ± 20 % der AU-Tage erzielt keine wesentlich erfolgreichere Kontrollzuweisung. Insgesamt muss bei der Bewertung einzelner Kriterien auch berücksichtigt werden, dass ein nicht erfolgreicher Matching-Durchlauf eines Patienten das Resultat der Kombination mehrerer Merkmale ist. Gegebenenfalls können zwar einzelne Eigenschaften eines Teilnehmers als Matching-Vorgabe von seinen potentiellen Kontrollen erfüllt werden, nicht jedoch die spezielle individuelle Kombination. Ein junger Patient mit besonders vielen AU-Tagen könnte hierfür als Beispiel dienen, da eine derartige Zusammensetzung selten zu finden ist.

In Bezug auf die Kombination von Kriterien muss ferner der hohe Ausschluss von potentiellen Matching-Partnern aufgrund nicht übereinstimmender Morbiditätsklassen näher erläutert werden. Mit einer durchschnittlichen Reduzierung der Kotrollgruppenstärke um ca. 76 % stellt dieses Kriterium wohl eine der größten Hürden für die Kontrollen dar. Zwar kann auch hier die Kombination einer Komorbidität mit einem weiteren Matching-Kriterium Schwierigkeiten bereiten. Die nähere Untersuchung erfolgloser Matching-Durchläufe zeigt allerdings eine Häufung spezieller Gesundheitszustände der Patienten. Demnach existieren häufig mehrere verschiedenartige Komorbiditäten, deren gemeinsames Auftreten kaum bei potentiellen Matching-Partnern vorzufinden ist.
Bei der Beurteilung der Mobiditätsklassen als Matching-Kriterium muss ferner die Frage nach der grundsätzlichen Authentizität dieser Klassifizierung gestellt werden. Beispielsweise wäre davon auszugehen, dass die chronische Schmerzerkrankung der MNS-R-Patienten sich in der Zuweisung der Morbiditätsgruppen ausreichend widerspiegelt. Dennoch kann bei keinem einzigen Programmteilnehmer eine HMG für chronischen Schmerz (HMG 252 und 253) beobachtet werden, obwohl die Krankheitsgeschichten teilweise einen Zeitraum von 10 Jahren weit übersteigen. Infolgedessen kann dem System der Morbiditätsklassen auf Grundlage der Erfahrungswerte dieser Evaluation keine realitätsnahe Wiedergabe des Gesundheitszustandes bescheinigt werden. Möglicherweise ist die identifizierte Diskrepanz zwischen Erkrankung und Klassifizierung auch dem Umstand geschuldet, dass die Diagnostik chronischer Schmerzerkrankungen nicht unwesentlichen Schwierigkeiten unterliegt (Kapitel 1.1).

Insgesamt resultiert aus der Anwendung des Selektionsverfahrens eine erfolgreiche Auswahl der jeweiligen Subpopulation. Dies wird anhand der Kriterien Alter und Geschlecht exemplarisch veranschaulicht, wobei das herangezogene Kollektiv auf Patienten mit Wohnsitz innerhalb der Ballungsräume beschränkt ist (Tabelle 6.1.2). Diese Subgruppe zeichnet sich dadurch aus, dass die Matching-Partner in Abhängigkeit der Programmteilnehmer disjunkte Teilmengen bilden. Folglich wird keine Kontrolle mehr als einem Fall zugeordnet.

Alter MNS-R-Gruppe		
Median	Mean	SD
51.00	50.56	9.72

Alter Kontrollgruppe		
Median	Mean	SD
52.00	50.81	9.77

Geschlecht MNS-R-Gruppe		
männlich	weiblich	Summe
16	25	41
39.0 %	61.0 %	100 %

Geschlecht Kontrollgruppe		
männlich	weiblich	Summe
64	100	164
39.0 %	61.0 %	100 %

Tabelle 6.1.2: Vergleich zwischen Patienten- und Kontrollgruppe innerhalb der Ballungsgebiete nach der Selektion

Während im Bereich des Geschlechtes eine exakte Übereinstimmung vorliegt, zeigen die Berechnungen des durchschnittlichen Alters leichte Abweichungen. Sowohl anhand des Medians als auch des arithmetischen Mittels lässt sich belegen, dass die als Matching-Kriterium vorgegebene Toleranzgrenze von ± 2 Lebensjahren nicht verletzt wird. Zusammen mit einer näherungsweisen Übereinstimmung der Standardabweichung kann für beide Merkmale eine zuverlässige Selektionswirkung festgehalten werden.

Ähnlichkeitsverfahren

Neben der Anwendung des Selektionsverfahrens soll auch die Wirkung der Ähnlichkeitsberechnungen bewertet werden. Aufgrund der komplexen Verarbeitung der Morbiditätsklassen im Matching-Prozess birgt eine quantitative Zusammenfassung der Matching-Ergebnisse keine zielführende Aussagekraft. Daher werden die Ähnlichkeiten mittels dreier ausgewählter Programmteilnehmer exemplarisch visualisiert. Da meist keine stationären Behandlungen durchgeführt worden sind, orientiert sich die Auswahl der präsentierten Teilnehmer an der Anzahl der ambulanten Versorgungsereignisse. Hierzu wird je ein Patient mit einer geringen, einer mittleren und einer großen Anzahl an Behandlungen gewählt (Tabelle 6.1.3).

Der Vergleich der Merkmalsausprägungen zwischen den Programmteilnehmern und ihren Kontrollen spiegelt eine korrekte Identifizierung der ähnlichsten Vergleichsversicherten wider. Hierbei liegt das Potential insbesondere darin, eine abstufende Reihenfolge der Kontrollen gemäß ihrer Ähnlichkeit zum MNS-R-Patienten festzulegen. Erst diese Möglichkeit führt zu der Effektivität des Auswahlmechanismus.

	Patienten-ID	Status	Anzahl amb. Beh.	Anzahl stat. Beh.	HMG
gering	**1052167**	**MNS-R**	**2**	**0**	**58**
	501368	Kontrolle	2	0	58
	856265	Kontrolle	2	0	58
	915981	Kontrolle	2	0	58
	1060179	Kontrolle	2	0	58, 164

	Patienten-ID	Status	Anzahl amb. Beh.	Anzahl stat. Beh.	HMG
mittel	**255551**	**MNS-R**	**8**	**0**	**keine**
	247612	Kontrolle	8	0	keine
	996131	Kontrolle	8	0	keine
	830601	Kontrolle	8	0	keine
	949697	Kontrolle	8	0	keine

	Patienten-ID	Status	Anzahl amb. Beh.	Anzahl stat. Beh.	HMG
groß	**822158**	**MNS-R**	**14**	**0**	**39**
	889953	Kontrolle	14	0	13, 19, 39, 90, 133
	594071	Kontrolle	14	0	39, 40, 58, 90, 109
	498199	Kontrolle	13	0	19, 39, 57
	612046	Kontrolle	13	0	39, 56, 109

Tabelle 6.1.3: Exemplarische Beurteilung der Ähnlichkeitsberechnungen zwischen den Patienten und ihren Kontrollen

Dokumentation der erfolglosen Matching-Durchläufe

Anknüpfend an die Bewertung der Selektions- und Ähnlichkeitskriterien erfolgt eine detaillierte Betrachtung der Programmteilnehmer ohne erfolgreiches Matching. Wie bereits erläutert, kann ein negatives Ergebnis häufig nicht mit einem einzelnen Kriterium begründet werden. Vielmehr führt die Seltenheit einer Kombination mehrerer Merkmale oder Komorbiditäten eines MNS-R-Patienten zum Ausbleiben der erfolgreichen Identifikation von Kontrollen. Vergleichbar zu der Bewertung der Selektionskriterien fallen bei der individuellen Betrachtung der betroffenen Patienten ebenfalls zwei hauptsächliche Ursachen auf. Demnach führt die Anwendung der Kriterien *Anzahl der Arbeitsunfähigkeitstage* und *Morbiditätsgruppen (HMG)* zur umfangreichsten Eliminierung potentieller Matching-Partner. Diesem Eindruck folgend werden die betroffenen MNS-R-Patienten hinsichtlich dieser Merkmale näher betrachtet (Tabelle 6.1.4). Häufig zeigen sich hierbei vielschichtige Nebener-

krankungen, deren Kombination nur bei wenigen oder keinen weiteren SBK-Versicherten zu finden ist. In Verbindung mit einer zum Teil großen Anzahl an Arbeitsunfähigkeitstagen bleibt folglich eine Zuordnung an Matching-Partnern ergebnislos.

Für eine exemplarische Darstellung der speziellen Matching-Vorgaben kann beispielsweise die Patientensituation des Patienten ‚875904' herangezogen werden: Der Patient im Alter von 35 Jahren leidet neben der chronischen Rückenschmerzerkrankung unter anderem an Hypertonie und Hepatitis. Insbesondere in Hinblick auf die Verbindung mit dem geringen Alter ist ein negativer Matching-Versuch durchaus plausibel. Ein weiteres interessantes Beispiel zeigt sich im Fall des Patienten ‚204488': Neben dem Rückenschmerzleiden weist der Patient u. a. eine epileptische Erkrankung auf. Insbesondere im Zusammenhang mit einer Anzahl von 309 Arbeitsunfähigkeitstagen (ca. 10.4 Monate) binnen eines Jahres wegen Rückenschmerzleiden ergibt sich eine selten auftretende Merkmalskonstellation.

PID	AU-Tage	Hierarchische Morbiditätsgruppen (HMG)				Weitere HMG / sonstiges
100183	104	Spinalkanalstenose	Depression	Hypertonie	--	Alter: 44
260740	366	Erkrankungen der Speiseröhre		Depression	--	Rente wg. Erwerbsminderung
269858	159	Leberzirrhose	Spinalkanalstenose; Osteoarthritis, -porose	Depression	Epilepsie	Hemiplegie; chronisch obstruktive Bronchitis
321703	156	Psychosen, psychotische und dissoziative Störungen		Nekrose von Gelenken, Muskeln, oder Knochen		
875904	96	Sonstige virale Hepatitis	Spinalkanalstenose	Depression	Polyneuropathie	Hypertonie; Alter: 35
999646	297	Depression	Polyneuropathie	--	--	
655128	108	Psychosen, psychotische und dissoziative Störungen			--	Außerstädtischer Wohnort
816941	181	Medizinische Komplikationen - andere iatrogene Schäden			--	Außerstädtischer Wohnort
192065	146	Rheumatoide Arthritis	Osteoarthritis in Hüfte oder Knie	Medizinische Komplikationen - andere iatrogene Schäden		Alter: 46
193666	36	Ernste bösartige Neubildungen		Depression	Hypertonie	
204488	309	Chronisch obstruktive Bronchitis		Epilepsie	--	
205311	0	Ernste bösartige Neubildungen	Alkohol- oder Drogenabhängigkeit	Depression	Neurogene Blase	
233682	332	HIV/AIDS	Depressive Episode	Polyneuropathie		
236080	39	Diabetes	Spinalkanalstenose	Polyneuropathie	Hypertonie	
263154	58	Diabetes	Kostenint. schwerwiegende Stoffwechselerkrankung	Depression	Polyneuropathie	Koronare Herzkrankheit; Hypertonie
302829	33	Chronisch obstruktive Bronchitis		Spinalkanalstenose	Psychosen	Polyneuropathie
566491	38	Osteoporose	Epilepsie	Chronisch obstruktive Bronchitis		
861336	153	Chronisch entzündliche Darmerkrankungen		Spinalkanalstenose	--	
20632	10	Erkrankungen / Verletzungen des Rückenmarks		--	--	Alter=34; Außerstädt. W-Ort
521567	172	Erkrankungen der Speiseröhre	Hypertonie / Arrhythmien	Periphere Gefäßerkrankungen	--	Außerstädtischer Wohnort
797795	38	Spinalkanalstenose	--	--	--	Außerstädtischer Wohnort

Tabelle 6.1.4: Übersicht über nicht erfolgreich verlaufene Matching-Durchläufe

6.2 Deskriptive Analyse

Nachdem das Matching erfolgreich abgeschlossen ist, muss die Datensituation eigens für die Kostenanalyse von Neuem betrachtet werden – unabhängig von sämtlichen bisherigen Datenaufbereitungen. Während für das Matching ausschließlich der Zeitraum vor der Intervention relevant war, liegt der Fokus für die Datenanalyse insbesondere auf verschiedenen post-interventionellen Beobachtungszeiträumen. Dabei liegen der Untersuchung folgende Fragestellungen zugrunde:

i. Existieren Unterschiede zwischen den MNS-R-Teilnehmern und Ihren Kontrollen hinsichtlich der Inanspruchnahme der gesundheitlichen Versorgung (Zeitraum nach der Intervention)?

ii. Weichen diese Unterschiede grundlegend vom Status vor der Therapieteilnahme ab?

Zur Untersuchung der primären Aufgabe genügt es nicht, einen einzelnen längeren Zeitraum nach der Programmteilnahme zu betrachten. Vielmehr können durch die Einbeziehung mehrerer unterschiedlich langer Perioden deutlich differenziertere Aussagen über die Entwicklung im zeitlichen Verlauf getroffen werden. Durch dieses Vorgehen wird beispielsweise ersichtlich, welche Situation sich kurz nach dem MNS-R einstellt, oder ob mögliche Therapieeffekte während der Folgemonate Schwankungen unterliegen. Insgesamt werden hierfür drei Zeiträume für eine kurz-, eine mittel- und eine langfristige Analyse festgelegt (Abbildung 6.2.1).

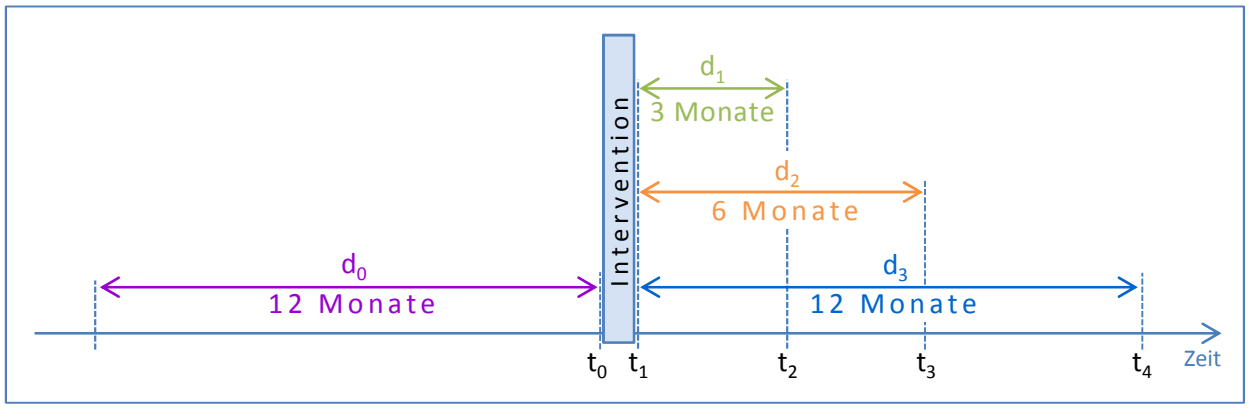

Abbildung 6.2.1: Übersicht der Beobachtungzeiträume

Für eine adäquate Gegenüberstellung der variierenden post-interventionellen Zeiträume mit dem Kalenderjahr vor der Intervention ist es notwendig, den unterschiedlichen Längen der Beobachtungsperioden Rechnung zu tragen. Dies erfolgt anhand einer Normierung der absoluten Werte eines Zeitraums – etwa die Anzahl notwendig gewordener Behandlungsereignisse (Tabelle 6.2.1).

Normierung der Beobachtungszeiträume		
Zeitpunkt	Messzeitraum	Normierungsfaktor
prä d_0	Kalenderjahr vor Intervention	1
post d_1	Drei Monate nach Intervention	4
post d_2	Sechs Monate nach Intervention	2
post d_3	Zwölf Monate nach Intervention	1

Tabelle 6.2.1: Normierung der Beobachtungszeiträume für eine adäquate Vergleichbarkeit

Während in der Evaluation des MNS-R eine Hochrechnung auf 12 Monate zur Anwendung kommt, ist grundsätzlich jede anderweitige Abstimmung mit entsprechend angepassten Normierungsfaktoren denkbar.

Im Folgenden wird anhand des Beispiels der Arzneimittelverschreibungen das methodische Vorgehen der Fall-Kontroll-Studie erläutert. Aufgrund der unterschiedlichen Komorbiditäten der Patienten sind die Ergebnisse lediglich eingeschränkt aussagekräftig. Durch die geringere klinische Relevanz im Vergleich zu den anderen Endpunkten eignet sich dieses Merkmal besonders zur Demonstration. Für detaillierte Informationen hinsichtlich der Studienergebnisse und medizinischen Schlussfolgerungen sei auf den Abschlussbericht des Evaluationsteils II verwiesen.
Grundsätzlich werden sämtliche Analyseschritte sowohl für die Ereignishäufigkeiten, wie beispielsweise die Anzahl verschriebener Medikamente, als auch für die resultierenden Kosten durchgeführt. Analog zur Analyse longitudinaler Daten erfolgt zunächst eine visuelle Darstellung der zu untersuchenden Daten. Dadurch soll ein erster Eindruck über die Entwicklungen im zeitlichen Verlauf vermittelt werden. Allerdings birgt die Darstellung der zur numerischen Analyse herangezogenen Datenstruktur erhebliche Erschwernisse für die Interpretation. Nachdem etwa der dreimonatige Beobachtungszeitraum post Intervention

auch Bestandteil des sechsmonatigen ist, beinhalten beide Perioden teilweise dieselben Ereignisse. Diese Überschneidung der Zeiträume führt in der Visualisierung dazu, dass zum Beispiel eine sinkende Anzahl an Arztbesuchen nicht durch einen fallenden Graphen repräsentiert wird. Interpretiert werden muss demnach nicht der Verlauf selber, sondern der Wertezuwachs im Vergleich zur letzten Periode. Diesem Effekt kann mittels einer separaten Datenaufbereitung entgegengewirkt werden, welche sich durch voneinander getrennte Zeiträume auszeichnet. Demnach ergibt sich eine einfachere Darstellung, bei der die Verläufe direkt den Veränderungen über die Zeit entsprechen (Abbildung 6.2.2).

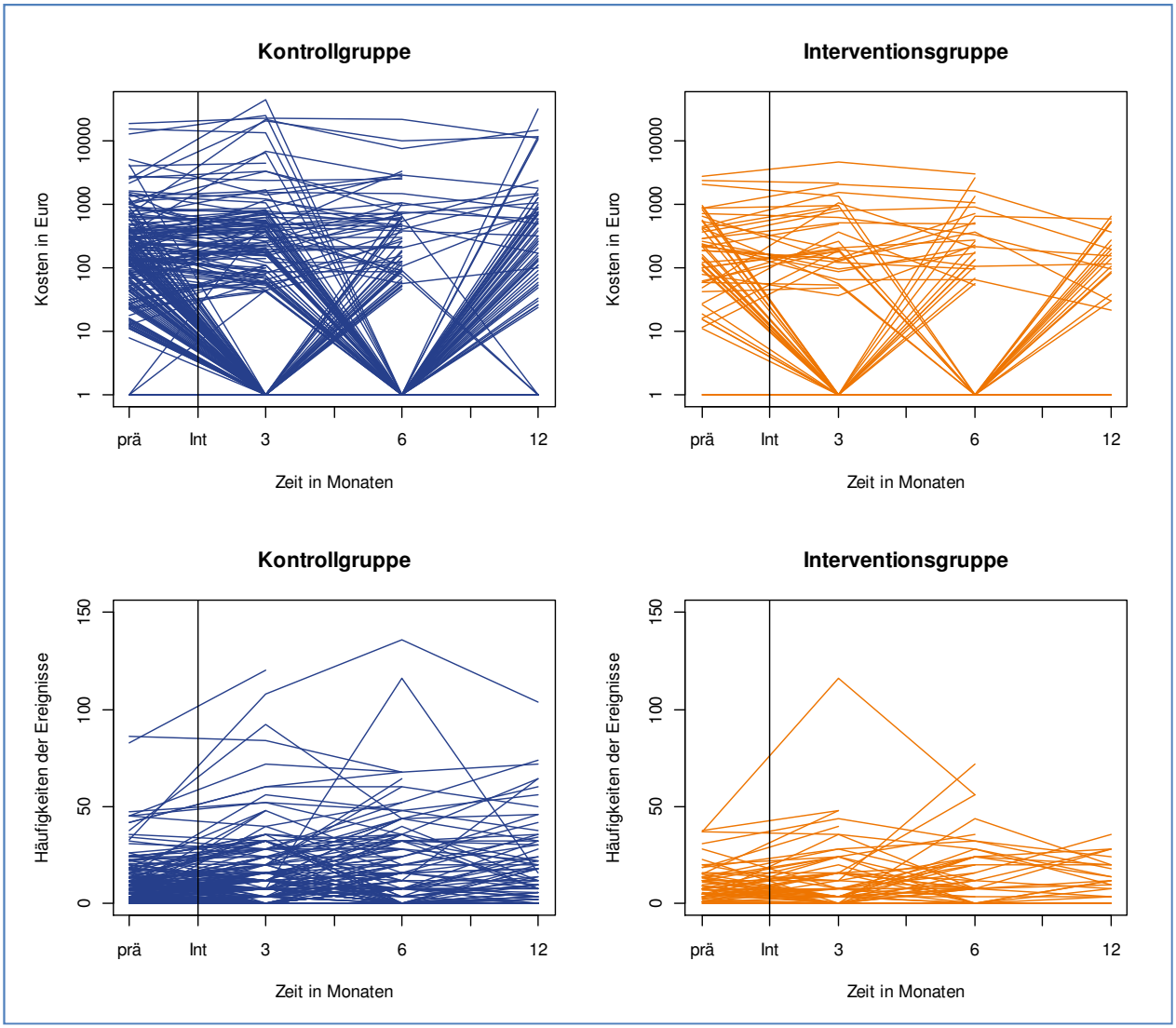

Abbildung 6.2.2: Verlauf der Kosten und Häufigkeiten von Arzneimittelverschreibungen jedes Patienten

Für eine erste grafische Übersicht erscheint diese vom Stratum unabhängige Gegenüberstellung von Kontroll- und Interventionsgruppe als zweckmäßig. Somit können die Entwicklungen der MNS-R-Patienten (Fälle) einerseits und der konventionell behandelter Matching-Partner (Kontrollen) andererseits gebündelt dargestellt und verglichen werden. Neben der grafischen Darstellung ist eine numerische Zusammenfassung essenziell für die Vermittlung eines umfassenden Eindrucks bezüglich der Daten. Hierzu werden gängige Lageparameter wie Mittelwerte und Mediane aus der Stichprobe bestimmt (Tabelle 6.2.2).

Kosten verschriebener Arzneimittel in Euro

	MNS-R-Gruppe				Kontrollgruppe				$\delta_{\text{casecontrol}}$		
	N	mean	median	Quantile (75 % ; 25 %)	N	mean	median	Quantile (75 % ; 25 %)	mean	median	Quantile (75 % ; 25 %)
$\mathbf{d_0}$	58	368.6	177.7	59.3 ; 420.8	208	592.7	147.8	43.4 ; 388.0	-203.4	-35.5	-274.8 ; 33.3
$\mathbf{d_1 - d_0}$	58	-2.9	-45.8	-167.2 ; 103.8	208	445.2	-27.2	-142.1 ; 52.6	580.9	-27.5	-150.7 ; 251.7
$\mathbf{d_2 - d_0}$	51	-27.0	-31.8	-166.3 ; 77.2	180	169.0	-37.4	-138.0 ; 12.2	270.4	-11.7	-171.1 ; 150.2
$\mathbf{d_3 - d_0}$	25	-155.6	-73.6	-270.2 ; 0.6	90	411.8	-14.4	-108.8 ; 25.7	610.0	46.93	-64.1 ; 400.0

Verschreibungshäufigkeiten von Arzneimitteln

	MNS-R-Gruppe				Kontrollgruppe				$\delta_{\text{casecontrol}}$		
	N	mean	median	Quantile (75 % ; 25 %)	N	mean	median	Quantile (75 % ; 25 %)	mean	median	Quantile (75 % ; 25 %)
$\mathbf{d_0}$	58	10.5	8.0	4.0 ; 14.0	208	10.4	7.0	3.0 ; 13.0	0.2	-0.5	-3.9 ; 3.8
$\mathbf{d_1 - d_0}$	58	3.96	0.5	-5.0 ; 10.8	208	2.96	0.0	-4.0 ; 6.0	-1.0	0.0	-9.4 ; 9.2
$\mathbf{d_2 - d_0}$	51	3.55	1.0	-3.0 ; 7.5	180	3.18	1.0	-3.0 ; 7.0	-0.5	-0.5	-7.5 ; 6.4
$\mathbf{d_3 - d_0}$	25	1.52	0.0	-2.0 ; 6.0	90	4.24	1.5	-2.0 ; 8.0	2.2	-1.0	-7.0 ; -7.3

Tabelle 6.2.2: Durchschnittliche Häufigkeiten und Kosten verschriebener Arzneimittel pro Patient

Wie bereits erläutert, sind die Daten für einen adäquaten Vergleich auf einen 12-monatigen Zeitraum normiert. Die hier durchgeführten Berechnungen orientieren sich stets an der Situation vor der Intervention (Zeitraum d_0). Die Durchschnittswerte des Zeitverlaufes nach der Teilnahme am Therapieprogramm werden entsprechend als Differenz zu den prä-Werten dargestellt. Negative Resultate weisen demnach auf eine Reduzierung der Kosten bzw. der Verschreibungshäufigkeiten hin. Positive Werte beschreiben eine gegenläufige Entwicklung.

Der wesentliche Unterschied zwischen der Analyse randomisierter, kontrollierter Studien (RCT) und Fall-Kontroll-Studien besteht in der Verarbeitung der Kontrollgruppen. Aufgrund der Randomisierung bei RCTs gelten Interventions- und Kontrollgruppen als balanciert hinsichtlich der Verteilung der interessierenden Endpunkte. Beide Gruppen lassen sich daher als unabhängig betrachten und gleichwertig gegenüberstellen. Fall-Kontroll-Studien beziehen hingegen stets die Verbindung der Fälle und deren Kontrollen mit in die Untersuchung ein. Durch diese Zuweisungen, genannt *Stratifikationen*, beinhaltet der Datensatz einen zusätzlichen Informationsgehalt, welcher bei der Analyse genutzt werden kann. Im Fall der MNS-R-Evaluation resultiert nach Vollendung des Matchings etwa ein Verhältnis Fall:Kontrolle = 1:m mit m = 4, wodurch ein Stratum meist aus fünf Patienten besteht. Wegen der Bedeutsamkeit der Stratifikation in der weiteren Analyse erfolgt bereits bei den deskriptiven Betrachtungen neben der gruppenspezifischen Präsentation zusätzlich ein Fall-Kontroll-Vergleich. Realisiert wird dies anhand einer über die Teilnehmer gemittelte Differenz zwischen der Merkmalsausprägung des *i*-ten Falls und dem Mittelwert seiner Kontrollen

$$\delta_{ij} = Y_{ij} - \frac{1}{n}\sum_{k=1}^{4} X_{ijk} \qquad \text{für } j \in \{d_0, \ldots, d_3\}.$$

Dabei erfolgt hinsichtlich der postinterventionellen Beobachtungen analog zu den Gruppenwerten stets eine Darstellung als Differenz mit den prä-Resultaten (Tabelle 6.2.2).

6.3 Fall-Kontroll-Studie

Anhand der deskriptiven Analyse konnte ein grober Eindruck hinsichtlich der Entwicklung von Verschreibungshäufigkeiten und der Arzneimittelkosten vor und nach der Therapie gewonnen werden. Unter Verwendung induktiver statistischer Methoden werden im Weiteren mögliche Effekte näher spezifiziert, was einen Rückschluss auf das Verhalten der Grundgesamtheit zulässt. Wie bereits beschrieben, erfolgt bei der Fall-Kontroll-Studie stets eine Berücksichtigung der Stratifikation. Demnach werden die Fälle ausschließlich den ihnen zugeteilten Kontrollen gegenübergestellt.

Insbesondere bei Daten mit umfangreichen Schwankungen empfiehlt sich bei einem Vergleich zweier Gruppen eine Berechnung sowohl auf Grundlage des arithmetischen Mittels als auch zusätzlich des Medians. Somit können sich aufgrund dessen ausreißerresistenter Eigenschaft weitere Erkenntnisse eröffnen. Während die Anwendung einer klassischen statistischen Regressionsanalyse eine Modellierung des Mittelwertes beinhaltet, verlangt die Orientierung auf den Median ein anderweitiges Instrument. Hierfür eignen sich beispielsweise nonparametrische Testverfahren wie der Wilcoxon-Test (Tabelle 6.3.1).

Für eine Integration der zugrunde liegenden Fragestellungen in die Testsituation werden grundsätzlich die Merkmalsausprägungen der maximal vier Kontrollen gemittelt dem Wert des betreffenden Therapieteilnehmers gegenübergestellt. Für Untersuchungen innerhalb eines Zeitraums wie bei baseline-Unterschieden (Testsituation 1) können direkt die Rohwerte herangezogen werden. Im Hinblick auf die vorher-nachher-Effekte ist allerdings ein Vergleich zwischen den Zeiträumen erforderlich (Testsituation 2-4). Hierfür liegen den Testberechnungen die Differenzen aus prä- und post-Werten anstatt der gemittelten Rohwerte zugrunde.

Situationen für Wilcoxon-Test: Angefallene Kosten in Euro pro Patient		**p-Wert**
1	$MNS\text{-}R\text{-}Gr_{prä}$ vs. $KGr_{prä}$	0.019
2	Δ ($MNS\text{-}R\text{-}Gr_{prä}$; $MNS\text{-}R\text{-}Gr_{d1}$) vs. Δ ($KGr_{prä}$; KGr_{d1})	0.536
3	Δ ($MNS\text{-}R\text{-}Gr_{prä}$; $MNS\text{-}R\text{-}Gr_{d2}$) vs. Δ ($KGr_{prä}$; KGr_{d2})	0.951
4	Δ ($MNS\text{-}R\text{-}Gr_{prä}$; $MNS\text{-}R\text{-}Gr_{d3}$) vs. Δ ($KGr_{prä}$; KGr_{d3})	0.114

Situationen für Wilcoxon-Test: Verschreibungshäufigkeiten pro Patient		**p-Wert**
1	$MNS\text{-}R\text{-}Gr_{prä}$ vs. $KGr_{prä}$	0.751
2	Δ ($MNS\text{-}R\text{-}Gr_{prä}$; $MNS\text{-}R\text{-}Gr_{d1}$) vs. Δ ($KGr_{prä}$; KGr_{d1})	0.948
3	Δ ($MNS\text{-}R\text{-}Gr_{prä}$; $MNS\text{-}R\text{-}Gr_{d2}$) vs. Δ ($KGr_{prä}$; KGr_{d2})	0.768
4	Δ ($MNS\text{-}R\text{-}Gr_{prä}$; $MNS\text{-}R\text{-}Gr_{d3}$) vs. Δ ($KGr_{prä}$; KGr_{d3})	0.657

Tabelle 6.3.1: Testergebnisse für Verschreibungshäufigkeiten und Kosten sämtlicher Medikamente pro Patient

Nachdem dieses Vorgehen die Einbeziehung der Stratifizierung vorsieht, lassen die baseline-Untersuchungen einen Schluss auf die Qualität der Matching-Prozedur zu. Wie bereits in Kapitel 5.1 dargestellt, soll aus einem optimalen Matching eine Balance zwischen den Gruppen resultieren.

In Bezug auf die Ergebnisse muss darauf hingewiesen werden, dass sich das zur Demonstration gewählte Merkmal der Arzneimittelverschreibungen wesentlich von den in der Evaluation untersuchten Endpunkten unterscheidet. Aufgrund der Vielzahl unterschiedlicher Nebenerkrankungen zeigt sich hier lediglich eine geringe Aussagekraft, was zum Beispiel die gravierenden Abweichungen von den Daten der Analgetikaverschreibungen untermauern. Somit überraschen die prä-interventionellen Unterschiede signifikanten Ausmaßes nicht, obwohl derartige Abweichungen gerade durch das Matching vermieden werden sollen. Im Widerspruch zu diesem Resultat sind bei keinem der primären Studienendpunkte relevante baseline-Unterschiede zu verzeichnen, sodass sich allein aus der Beobachtung dieses Merkmals keine mangelhafte Leistungsfähigkeit des angewandten Matching-Verfahrens ableiten lässt.

Für eine Schätzung des arithmetischen Mittels im Rahmen der Fall-Kontroll-Studie verfügt die Theorie der Regressionsanalyse über erhebliche Vorteile gegenüber anderweitigen Methoden. Obwohl im Gegensatz zu der Analyse longitudinaler Daten gegenwärtig nicht die Veränderungen im zeitlichen Verlauf im Fokus stehen, kann dennoch die Methode der *linear mixed effects model (LME)* gewinnbringend eingesetzt werden (Modellbeschreibung siehe Kapitel 3). Durch die separate Schätzung der festen und zufälligen Effekte eröffnet sich eine komfortable Möglichkeit, die Stratifikation in die Modellberechnung zu integrieren. Während das interessierende Modell mit der abhängigen Variable *Y* und einem linearen Prädiktor als fixed effects formuliert wird, erfüllt die Einbindung der Stratum-Variable als random effects gänzlich die Differenzierung innerhalb der Kontrollgruppe. Basierend auf der Schätzmethode der Maximum-Likelihood-Theorie ergibt die beschriebene Anwendung der LMEs eine Modellschätzung des Mittelwertes der Differenzen zwischen den Fällen und ihrer Kontrollen. Ein weiterer Vorteil der Regressionsanalyse liegt in der flexiblen Gestaltung des linearen Prädiktors. Hierdurch kann die Bearbeitung beider Hauptfragestellungen präzise umgesetzt werden. Anhand des Vergleiches zweier berechneter Modellierungen

$$\begin{aligned} &[1] \quad \text{Fixed effects:} \quad E(Y_i \mid b_i) = \beta_0 + \beta_1 \cdot MNS_i && \text{random effects} = \texttt{Strat} \\ &[2] \quad \text{Fixed effects:} \quad E(Y_i \mid b_i) = \beta_0 + \beta_1 \cdot MNS_i + \beta_2 \cdot Y_{i,prä} && \text{random effects} = \texttt{Strat} \end{aligned} \tag{1.10}$$

lässt sich exakt der Einfluss der prä-Werte auf die Modellschätzungen spezifizieren. Inner-

halb der Modellformel steht *Y* für die interessierende, also abhängige Variable, wobei *b* die Regressionskoeffizienten der random effects beschreiben. Die Einflussgröße **MNS** besteht aus einer 0-1-kodierten Gruppenvariable, welche zwischen den MNS-R-Teilnehmern (**MNS = 1**) und Kontrollpatienten (**MNS = 0**) differenziert. Die Einbeziehung der prä-Werte in das Modell wird im Allgemeinen auch als Modelladjustierung bezeichnet. Anhand des Vergleichs der Resultate mit und ohne Adjustierung kann eine Quantifizierung des vorher-nachher-Effektes inklusive Signifikanzaussage gewährleistet werden.

Die Interpretation des LME-Ergebnisses erfolgt analog zu den generalisierten linearen Modellen. Im Bereich der Kosten entsprechen die Regressionsparameter demnach direkt dem Wert in Euro – in dem gewählten Beispiel (Zeitraum d1) liegen also die mittleren Arzneimittelkosten der Kontrollpatienten bei 1051 Euro ohne Berücksichtigung der prä-Werte (Abbildung 6.3.1). Im Falle der Programmteilnahme liegen die durchschnittlichen Kosten 686 Euro niedriger, wobei der Unterschied zwischen Fällen und Kontrollen kein signifikantes Ausmaß erreicht (p = 0.25). Ferner kann aus dem adjustierten Modell ((1.10) [2]) eine ausgeprägte Abhängigkeit der Zielgröße von den Werten vor der Intervention abgeleitet werden. Mit jeder zusätzlichen Einheit (Euro) vor der Behandlung (d0) steigen die Kosten im Zeitraum d1 um 1.33 Euro.

```
Fixed effects: medi.d1 ~ MNS
                Value Std.Error  DF   t-value p-value
(Intercept)  1051.2765  301.0267 207  3.492304  0.0006
MNS          -685.6124  589.9288 207 -1.162195  0.2465

Fixed effects: medi.d1 ~ MNS + medi.prä
                 Value Std.Error  DF   t-value p-value
(Intercept)   250.0714  239.2588 206  1.045192  0.2972
MNS          -374.3590  492.1431 206 -0.760671  0.4477
medi.prä        1.3292    0.1141 206 11.647561  0.0000
```

Abbildung 6.3.1: LME-Ergebnisse für die Kosten verschriebener Arzneimittel mit und ohne prä-Adjustierung

Für die Berechnung der Verschreibungshäufigkeiten ergibt sich ein ähnliches Vorgehen, wobei für Zähldaten andere Modellvoraussetzungen getroffen werden müssen: Da die Anzahl der Ereignisse pro Zeitintervall einer Poisson-Verteilung folgt, kommen hier *generalized linear mixed effects models (GLME)* zum Einsatz. Der Unterschied zum LME besteht lediglich in der Anpassung der Modellschätzung an nicht-normalverteilte Zielgrößen. In der

Anwendung der GLME macht sich diese Modifikation insbesondere bei der Interpretation der Ergebnisse bemerkbar, da die festen Effekte des Poisson-GLME

$$[1] \quad \text{Fixed effects:} \quad \log\left(E\left(Y_i \mid b_i\right)\right) = \beta_0 + \beta_1 \cdot MNS$$

$$[2] \quad \text{Fixed effects:} \quad \log\left(E\left(Y_i \mid b_i\right)\right) = \beta_0 + \beta_1 \cdot MNS + \beta_2 \cdot Y_{i,prä}$$

der Familie der loglinearen Modelle angehört. Dabei existieren zwei Möglichkeiten für die Beschreibung der geschätzten Modellparameter:
Die unmittelbare Aussage der Resultate bezieht sich zunächst auf die logarithmierten Häufigkeiten. Der Einfluss der Variable **MNS** wirkt also wie bisher additiv auf das Ergebnis. Demnach beträgt die logarithmierte Häufigkeit der Arzneimittelverschreibungen bei den Kontrollpatienten ca. 2.12 und bei den MNS-R-Teilnehmern ca. 2.19 (Abbildung 6.3.2).
Die zweite und im Allgemeinen anschaulichere Interpretationsvariante eröffnet sich durch das Exponieren der Modellformel. Infolgedessen wirken einerseits die Einflüsse direkt auf die Häufigkeiten, wodurch sich die Aussagen aufgrund der nicht-logarithmischen Form greifbarer und verständlicher darstellen. Andererseits folgt aus dem Exponieren des linearen Prädiktors jetzt ein multiplikativer Einfluss sämtlicher Kovariablen. Entsprechend ergibt sich für die Kontrollpatienten eine mittlere Häufigkeit von exp(2.11) = 8.33 Besuchen im Zeitraum d1. Für die MNS-R-Patienten hingegen erhöht sich die Anzahl der Besuche um den Faktor exp(0.07) = 1.08, was einer Erhöhung um ca. 8 % entspricht.

```
Formula: medi.d1 ~ MNS + (1 | strat)
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.1194     0.1528   13.87   <2e-16 ***
MNS           0.0741     0.0400    1.85   0.0643 .

Formula: medi.d1 ~ MNS + medi.prä + (1 | strat)
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.7372     0.1201  14.461  < 2e-16 ***
MNS           0.1253     0.0410   3.058  0.00223 **
medi.prä      0.0336     0.0013  26.866  < 2e-16 ***
```

Abbildung 6.3.2: Ergebnisse der Poisson-Modelle für die Verschreibungshäufigkeiten aller Medikamente

Erwartungsgemäß erfolgt analog dazu die Interpretation der Modellvariante mit Adjustierung. Demnach ist eine deutliche Abhängigkeit von den Werten vor der Intervention zu re-

gistrieren. Bei einer mittleren Häufigkeit von exp(1.74) = 5.68 Verschreibungen bei den Kontrollen führt jede weitere Verschreibung im vorher-Zeitraum zu einem Anstieg um den Faktor exp(0.03) = 1.03 oder um eine dreiprozentige Erhöhung. Für die MNS-Gruppe führt damit jede zusätzliche Verschreibung vor der Teilnahme zu einer Steigerung um 17.2 % (exp(0.12+0.03) = 1.172) im Vergleich zu den Kontrollen.
Darüber hinaus beinhaltet die Evaluation des MNS-R – wie bei der Berechnung der Wilcoxon-Tests ausgeführt – die Analyse mehrerer Perioden nach der Programmteilnahme. Anhand der Gegenüberstellung der Ergebnisse verschiedener Beobachtungzeiträume lässt sich aus den Modellberechnungen die Entwicklung der Kosten und Häufigkeiten über die Zeit herausarbeiten. Eine Betrachtung der Ergebnisse über einen längeren Zeitraum inklusive der Schlussfolgerungen kann dem Abschlussbericht der Evaluation entnommen werden.

Abweichend von dem geplanten Analysevorgehen kann es vorkommen, dass die Zähldaten zu geringe Häufigkeiten aufweisen und infolgedessen die Modellschätzung auf Basis der Poisson-Verteilung gegebenenfalls keinen sinnvollen Schluss zulassen. Eine Möglichkeit, dennoch eine seriöse Analyse durchzuführen, besteht in der Dichotomisierung der Daten. Aus den Verschreibungshäufigkeiten lässt sich entsprechend eine binäre Variable konstruieren, welche lediglich angibt, ob ein Patient in dem betreffenden Zeitraum ein Ereignis aufweisen kann oder nicht. Die Zielgröße der Modellberechnung folgt demnach einer Binomialverteilung, was die Verwendung eines logistischen Regressionsmodells

$$[1] \quad \text{Fixed effects:} \quad \log\left(\frac{P(Y_i=1\,|\,b_i)}{1-P(Y_i=1\,|\,b_i)}\right) = \beta_0 + \beta_1 \cdot MNS_i$$

$$[2] \quad \text{Fixed effects:} \quad \log\left(\frac{P(Y_i=1\,|\,b_i)}{1-P(Y_i=1\,|\,b_i)}\right) = \beta_0 + \beta_1 \cdot MNS_i + \beta_2 \cdot Y_{i,prä}$$

für die Berechnung der festen Effekte des GLME nach sich zieht. Für die Interpretation der Ergebnisse ist es insbesondere von Bedeutung, dass hierbei kein Erwartungswert, sondern die logarithmierten Chancen (log odds) modelliert werden. Analog zum Umgang mit dem Poisson-Modell führt das Exponieren der Modellformel zu einer anschaulicheren Aussagekraft der Resultate. Anstatt der log odds können demzufolge die Chancen in nicht-

logarithmischer Form gedeutet werden, wobei die Kovariablen einen multiplikativen Einfluss ausüben. Ferner lassen sich für den Vergleich zweier Populationen auch die relativen Chancen (odds ratio) bestimmen, wodurch das Chancenverhältnis der beiden Populationen beschrieben wird. Durch die Aussagekraft der relativen Chancen werden diese auch als Zusammenhangsmaß eingesetzt (Tutz, 2000).

Kapitel 7

Diskussion

Grundsätzlich dienen statistische Untersuchungen dem Zweck, Verhaltensweisen oder Eigenschaften einer Grundgesamtheit von Merkmalsträgern zu quantifizieren. In der Medizin könnte beispielsweise die Frage erörtert werden, ob und in welchem Maße ein neuer analgetischer Wirkstoff die gewünschte Schmerzerleichterung erzielt. Im Gegensatz zur Durchführung einer Vollerhebung genügt es durch die Anwendung statistischer Analysemethoden, lediglich eine Teilpopulation (Stichprobe) zu erfassen und dennoch stichhaltige Aussagen über die Grundgesamtheit treffen zu können. Demzufolge haben sich zahlreiche methodische Vorgehensweisen zur statistischen Beantwortung verschiedenster Fragestellungen etabliert.

Bei der Prüfung neuer Interventionen kommen häufig *Randomisierte Kontrollierte Studien (RCT)* zur Anwendung, da diese Studienform besondere qualitative Vorteile mit sich bringt. Der Kerngedanke zeichnet sich durch den Vergleich zweier unabhängiger Gruppen aus, wobei eine Gruppe mit dem zu untersuchenden Präparat behandelt wird (Verumgruppe). Die zweite Teilpopulation dient dem Vergleich und repräsentiert in der Regel die etablierten Behandlungsmethoden (Kontrollgruppe). Die Kontrollpatienten zeichnen sich je nach Fragestellung entweder durch eine Behandlung mit konventionellen Verfahren oder durch die Verabreichung von Placebo-Präparaten aus. Das Ziel der Kontrollbehandlung besteht aus der Zusammenstellung einer Referenzpopulation, welche eine adäquate Beurteilung der Verumintervention ermöglicht. Die Zuweisung eines Patienten in die jeweilige Gruppe erfolgt dabei nach dem Zufallsprinzip, was anhand standardisierter Randomisierungsprocedere gewährleistet wird. Insbesondere die Unabhängigkeit der Randomisierung ermöglicht es, unerwünschten Selektionseffekten entgegenzuwirken und bildet eines der zentralen Qualitätsmerkmale von randomisierten Studien. Abgesehen von den gewählten Randomisierungsvariablen können dabei die Gruppen auch hinsichtlich unbekannter Störfaktoren ausbalanciert werden (Senn, 1994 / Signorini et al., 1993 / Pocock & Simon, 1975). Infolgedessen hat diese Studienform im wissenschaftlichen Umfeld zunehmend an Bedeutung ge-

wonnen und erfüllt bei entsprechenden Voraussetzungen qualitativ höchste Anforderungen (Abbildung 7.1 / Windeler et al., 2008).

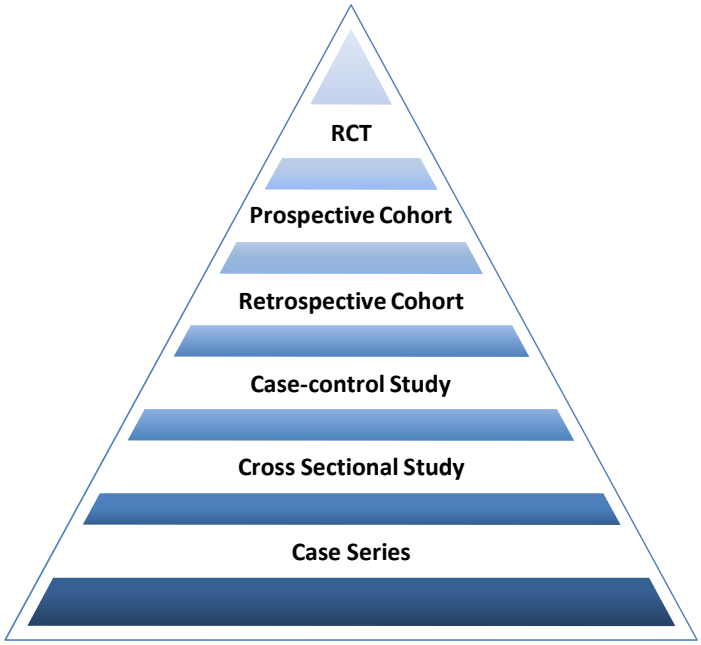

Abbildung 7.1: Pyramide der Evidenz (Ho et al., 2008)

Für die Durchführung einer qualitativ hochwertigen Studie genügt es allerdings nicht, die Anwendung einer RCT zu forcieren. Vielmehr ist die Auswahl eines adäquaten Studiendesigns stets von den Rahmenbedingungen wie Zielsetzung und Fragestellung einer Untersuchung abhängig. Somit können Voraussetzungen existieren, unter denen ein Design mit einem niedrigeren Niveau in der Evidenz-Hierarchie durchaus höhergestellten Studien vorzuziehen ist. Deutlich wird dies anhand von Situationen, bei denen von der Bildung einer Referenzgruppe grundsätzlich Abstand genommen werden muss – meist aufgrund der Verletzung ethischer Grundsätze. Bei der Therapie von Krebspatienten würde beispielsweise eine Behandlung mit Placebo-Präparaten drastische Folgen für die Patienten mit sich bringen und muss daher ausgeschlossen werden.

Ferner kann auch – wie im Fall der MNS-R-Evaluation – die Komplexität der Studienorganisation gegen die Bildung einer Referenzgruppe sprechen. Zunächst erscheint die Anwendung eines kontrollierten Studiendesigns bei Rückenschmerzpatienten unkompliziert umsetzbar zu sein, könnten die Kontrollpatienten doch anhand üblicher schulmedizinischer Methoden behandelt werden. Eine spezifische Betrachtung des gesamten Integrierten Ver-

sorgungsprojekts führt hingegen zu der Erkenntnis, dass sämtliche Elemente des Patientenkontakts auf den besonderen Umgang mit den Patienten ausgerichtet sind. Angefangen bei dem Rekrutierungsvorgang der SBK, welcher die Initiative seitens der Krankenkasse einschließt, bis hin zur Arbeitsorganisation der Schmerzambulanz, bei der sich der Patient ganzheitlich versorgt fühlen soll, unterscheidet sich die Therapieform wesentlich vom Vorgehen der konventionellen Schmerztherapie. Bei einer Behandlung der Kontrollpatienten im Umfeld des Projekts könnte daher die Bildung einer validen Referenzgruppe nicht gewährleistet werden.

Studienvorgaben, welche der Bildung einer Kontrollgruppe widersprechen, können häufig durch die Anwendung eines alternativen Studienansatzes entschärft werden. So lässt sich die Behandlung sämtlicher Studienteilnehmer mit der Notwendigkeit einer Referenzpopulation etwa durch die Durchführung eines cross-over-Designs vereinbaren. Hierbei erhalten sämtliche Teilnehmer zeitlich versetzt sowohl die Referenz- als auch die Verumtherapiemaßnahmen (Abbildung 7.2).

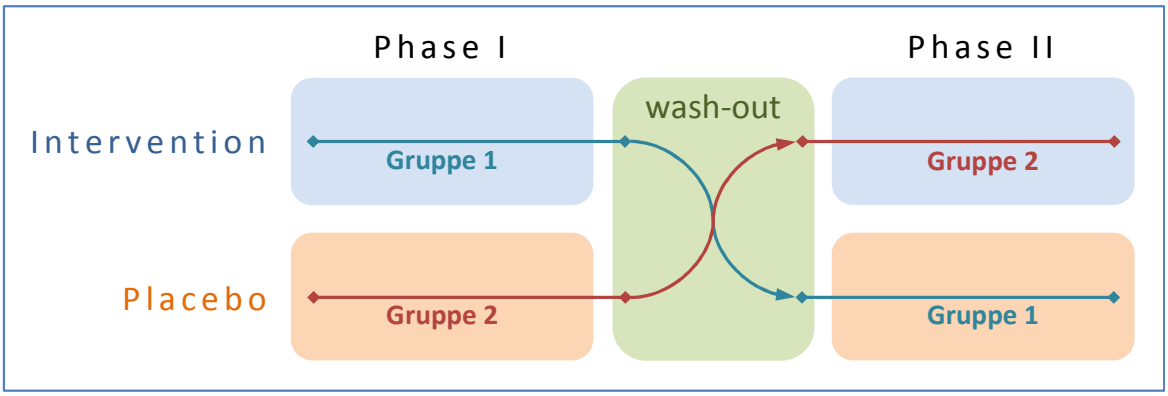

Abbildung 7.2: Design von cross-over-Studien

Besondere Sorgfalt muss bei dieser Studienform allerdings auf die wash-out-Phase gelegt werden. So wird der Zeitverlauf zwischen den beiden aktiven Behandlungsphasen bezeichnet. Insbesondere im Bezug auf die Verumgruppe dient diese Periode zum Abbau der Therapieeffekte, damit eine anschließende Funktion derselben Patienten als Kontrollen nicht verfälscht wird (Senn, 2002). Diese zentrale Forderung kann mit der Therapie des MNS-R grundsätzlich nicht in Einklang gebracht werden, da die während der Therapie erlangten Erfahrungen und Erkenntnisse langfristig erhalten bleiben sollen. Somit ergäbe sich ein di-

rekter Widerspruch zwischen der Studienvoraussetzung und einem nachhaltigen therapeutischen Effekt, welcher einer der primären Therapieziele darstellt.

Nachdem die RCT, wie auch die Alternative einer cross-over-Studie, keine geeigneten Optionen für die Evaluation des MNS-R darstellen, wurde im ersten Teil lediglich eine Beobachtungsstudie durchgeführt. Infolgedessen stützt sich der zweite Evaluationsteil auf ein völlig eigenständiges Studiendesign, dessen qualitativer Hintergrund zusammen mit den implementierten Vorarbeiten einer näheren Betrachtung bedarf:
Grundsätzlich weist das für die Kostenanalyse herangezogene retrospektive Vorgehen einer Fall-Kontroll-Studie naturgemäß qualitative Defizite im Vergleich zu einer RCT auf. Zahlreiche Gegenüberstellungen der beiden Studienformen untermauern die allgemeine Auffassung, nach der Interventionseffekte nur bedingt durch Fall-Kontroll-Studien nachweisbar sind. Exemplarisch kann dies anhand der häufig diskutierten Mammographie-Daten einer Studie aus Malmö (Schweden) (Andersson et al., 1988) dargestellt werden, da hier sowohl ein Fall-Kontroll-Design als auch eine RCT zur Anwendung kamen. So beschreibt Jørgensen, 2011 sehr übersichtlich, wie relevante und signifikante Effekte zwar durch die Untersuchung mittels Fall-Kontroll-Methode identifiziert werden konnten. Dennoch war es nicht möglich, diese ebenfalls durch die ursprünglichen Berechnungen einer RCT nachzuweisen. Entsprechend lässt sich durch diese Vergleichsuntersuchung darstellen, dass Therapieeffekte in Fall-Kontroll-Studien unter geeigneten Voraussetzungen lediglich auf Verzerrungen zurückgeführt werden können (McCartney, 2012 / Demissie et al., 1998).
Aufgrund dieses hohen Maßes an Sensibilität leistet folglich das Auswahlverfahren der Kontrollen einen substanziellen Beitrag zur Studienqualität und darf bei der Studienbewertung nicht außer Acht gelassen werden. Vergleichbar zu anderweitigen Analysen bezüglich der Mammographie-Thematik (Demissie et al., 1998) werden die Kontrollen in der Malmö-Studie per Zufall ausgewählt und wird auf die Entwicklung eines Matching-Verfahrens nach definierten Kriterien gänzlich verzichtet. Ursprünglich sind Selektionseffekte als Resultat dieses Vorgehens als vernachlässigbar eingestuft worden. Nach zahlreichen Diskrepanzen wird mittlerweile seitens der WHO die Empfehlung ausgesprochen, Fragestellungen bezüglich der Effektivität präventiver Mammografie-Screenings nicht mehr durch den Einsatz von Fall-Kontroll-Studien zu untersuchen (Jørgensen, 2011).

Grundsätzlich kann aus den erarbeiteten Vergleichsstudien die Erkenntnis festgehalten werden, dass sich Fall-Kontroll-Studien durch eine erhöhte Sensibilität auf Ungleichgewichte zwischen Fällen und Kontrollen auszeichnen. Optimale Voraussetzungen finden sich somit lediglich bei sehr homogenen Patientenpopulationen, wobei es sich stets als aufwendig erweisen dürfte, Verzerrungen in Form von Selektionseffekten entgegenzuwirken. Nachdem der MNS-R-Evaluation mitunter spezielle Schmerzpatienten mit teils jahrzehntelangen chronischen Beschwerden zugrunde liegen, kann offensichtlich nicht von einer homogenen Patientenpopulation ausgegangen werden. Daher wurde der hohen Sensibilität des Fall-Kontroll-Designs Rechnung getragen und vor der Durchführung der Kostenanalyse ein umfangreiches Matching-Verfahren entwickelt. Als Folge dieser Vorarbeiten sollen lediglich ähnliche Patienten (Fälle und Kontrollen) hinsichtlich ihrer Krankheitsgeschichte verglichen werden. Dabei wird neben den diagnostischen Eigenschaften des Rückenschmerzleidens besondere Aufmerksamkeit auf die Notwendigkeit der medizinischen Versorgung gerichtet. Die Zielsetzung des Matchings besteht demnach in der Reduzierung der Ungleichgewichte zwischen den Programmteilnehmern (Fälle) und ihren Kontrollen. Durch eine ähnliche Verteilung beider Populationen hinsichtlich möglicher Störgrößen kann Verzerrungen entgegengewirkt und deren Einfluss auf die Analyse reduziert werden (Kupper et al., 1981 / Breslow et al., 1978).

Zur Einschätzung einer erfolgreichen Homogenisierung der Populationen durch das Matching-Vorgehen muss das spätere Analysevorgehen der resultierenden Datensituation gegenübergestellt werden:
Nach Durchführung des Matchings sind jedem Patienten maximal vier Kontrollen zugeordnet. Basierend auf den definierten Matching-Kriterien besteht zwischen Fällen und Kontrollen eine Ähnlichkeitsbeziehung, deren Intensität vorgegeben werden kann. Infolgedessen verfügt jeder Patient, welcher im Hinblick auf ein zu untersuchendes Merkmal hohe Werte aufweist (high responder), über mehrere Kontrollen mit ebenfalls hohen Werten. Zweifelsfrei besteht diese Gegebenheit analog dazu bei Patienten mit geringen Werten (low responder) sowie bei Ausreißern im oberen und unteren Wertebereich.
Die Analyseberechnungen beinhalten die Gegenüberstellung von Patienten- und Kontrollwerten, etwa durch die Bildung von Differenzen. Teilweise wird dies mit zeitlichen Vergleichen verbunden (prä-post-Analyse), wodurch wiederum Differenzwerte im Fokus stehen

(Abbildung 7.3). Primär schließen die Analyseberechnungen folglich keinerlei Informationen hinsichtlich des Verhältnisses zum Wertebereich einer zu untersuchenden Variable ein. Die Fragestellung nach der Position der Messwerte eines Patienten im theoretischen Wertebereich des Merkmals ist also für das Analyseergebnis nicht zwingend relevant.

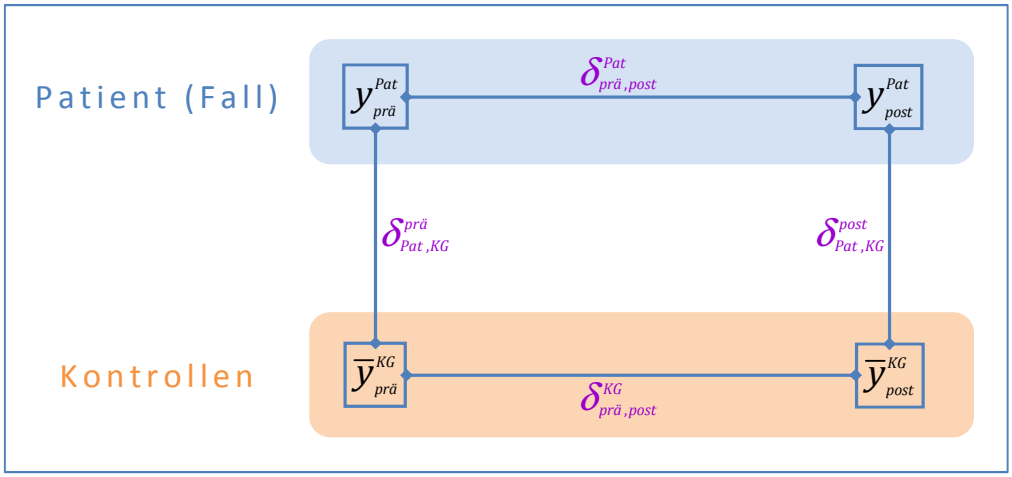

Abbildung 7.3: Theoretisches Analyseschema für Fall-Kontroll-Studien

Eine Auswahl der Kontrollen nach dem Zufallsprinzip, wie es in einigen medizinischen Studien praktiziert worden ist (Demissie et al., 1998), würde einen Vergleich zwischen Fällen mit teilweise sehr unterschiedlichen Kontrollen zulassen. Die fehlende Abstimmung der individuellen Wertebereiche hätte bei der Analyse von Differenzen erhebliche Verzerrungen zur Folge. Vorwiegend bei heterogenen Populationen mit zahlreichen Extremfällen würden besonders hohe Differenzwerte resultieren und die Analyse verfälschen. Die im Vorfeld durchgeführte Matching-Prozedur leistet demnach einen wesentlichen Beitrag dazu, Ungleichgewichte aufgrund heterogener Populationen zu mildern.
Darüber hinaus können komplexe statistische Methoden beispielsweise aus der Theorie longitudinaler Daten dazu beitragen, zusammen mit der im Matching erzeugten Datenbasis Verzerrungen entgegenzuwirken. So bieten linear mixed effects models die Möglichkeit, sowohl die Zuordnung der Kontrollen (Stratum) als auch die Rohwerte vor und nach der Intervention anstelle der Differenzen zu berücksichtigen (Kapitel 6).

Obwohl durch das vorgeschlagene Vorgehen die Untersuchungsvoraussetzungen deutlich verbessert werden können, bietet sich dennoch keine ausreichende Gewährleistung für die

korrekte Identifikation möglicher Therapieeffekte. Letztlich steht der Studienerfolg in starker Abhängigkeit der Matching-Kriterien, weshalb es bei der Auswahl und Justierung besonderer Sorgfalt bedarf.
Des Weiteren können Ausreißer nur bedingt adäquat beurteilt werden. Zwar ist es jederzeit möglich, in einer Population einen extremen Wert zu identifizieren. Für eine Evaluierung dieser Beobachtung fehlt es allerdings an Referenzwerten. Im Gegensatz dazu lässt die Verwendung einer randomisierten Kontrollgruppe eine detaillierte Vergleichsuntersuchung der Basiszustände zwischen Fällen und Kontrollen zu (baseline-Analyse). Hierbei ausgemachte Ausreißer können, wenn nötig, anschließend einer gesonderten Betrachtung unterzogen werden. Durch die Durchführung des Matchings fallen hingegen die Unterschiede zwischen Fällen und Kontrollen vor der Intervention gering aus. Allein durch das Vorliegen von Extremwerten innerhalb der prä-post-Analyse lässt sich folglich nicht eindeutig feststellen, ob ein Patient einen Spezialfall darstellt. Wegen der Unabhängigkeit der Ergebnisse zum Wertebereich ist eine Differenzierung zwischen Ausreißern und Patienten mit schlicht überdurchschnittlicher Reaktion auf die Intervention nicht möglich. Infolgedessen findet sich keine ausreichende Entscheidungsgrundlage für den korrekten Umgang mit einem Patienten, welcher auffällige Resultate aufweist. Eine Fehleinschätzung von Ausreißern könnte letztlich zu Selektionseffekten führen und einen relevanten Einfluss auf das Studienergebnis nach sich ziehen.

Zusammenfassend kann festgehalten werden, dass durch den Einsatz von Matching-Verfahren die Sensibilität von Fall-Kontroll-Studien gehemmt werden kann. Komplexe statistische Methoden optimieren dabei die Verwendung des Informationsgehaltes. Dennoch können trotz verbesserter Rahmenbedingungen die Vorteile einer Referenzgruppe in einer RCT nicht kompensiert werden. Sofern von der Bildung einer Kontrollgruppe Abstand genommen werden muss, haben sich in der Praxis verschiedene Studiendesigns als hilfreich erwiesen, um weniger verzerrungsanfällige Optionen zu erarbeiten (Ho et al., 2008).

Kapitel 8

Zusammenfassung

Aufgrund ihrer hohen Prävalenz und Komplexität gehören Rückenschmerzleiden zu den bedeutungsvollsten Krankheiten des deutschen Gesundheitswesens. Bedingt durch den vielschichtigen anatomischen Aufbau des Rückens gestaltet sich eine zielgerichtete Diagnostik und Behandlung meist problematisch. Eine Optimierung der Behandlungseffizienz insbesondere bei unspezifischen Rückenschmerzen liegt demnach im Interesse eines medizinischen sowie gesundheitsökonomischen Standpunktes.
Innerhalb des Integrierten Versorgungsprojekts *Münchener Naturheilkundliches Schmerzintensivprogramm – Rücken (MNS-R)* unter Beteiligung der Klinik für Anästhesiologie der Universität München, der Siemens Betriebskrankenkasse (SBK) und des Instituts für medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE) wurde ein multimodales Gruppentherapieprogramm zur Behandlung unspezifischer Rückenschmerzen zusammengestellt. Die Zielsetzung dieser komplexen Intervention besteht in der Remobilisierung der Patienten sowie einer nachhaltigen Reduzierung und Vorbeugung des Chronifizierungsrisikos. Methodisch basiert das therapeutische Vorgehen auf Elementen der klassischen Naturheilkunde, der Traditionellen Chinesischen Medizin (TCM) sowie der evidenzbasierten Schmerztherapie.
Hinsichtlich der wissenschaftlichen Evaluation derartiger Behandlungsformen weist das MNS-R einen grundlegenden Unterschied zu bisherigen Untersuchungen auf. Während die klassische Hauptfragestellung primär auf die medizinische Effektivität ausgerichtet ist, ermöglicht die Partnerschaft mit der SBK einen zusätzlichen gesundheitsökonomischen Schwerpunkt auf Grundlage stichhaltiger Abrechnungsdaten.

Demzufolge bildet die vorher-nachher-Analyse des Gesundheitszustandes eines Teilnehmers zusammen mit einer besonderen Beachtung der langfristigen Entwicklung über 24 Monate lediglich einen ersten Evaluationsteil. Den statistischen Berechnungen liegen Daten aus erhobenen Fragebögen zugrunde, welche von den Patienten zu verschiedenen Zeit-

punkten bearbeitet werden. Als statistische Methode werden spezifische Instrumente der longitudinalen Datenanalyse zum Einsatz gebracht, wobei den unterschiedlichen Perspektiven der verschiedenen Projektpartner Rechnung getragen werden muss. Somit eignet sich für eine medizinische Interpretation der Ergebnisse die Verwendung konditionaler Modelle, da deren individuelle Effekte eine Aussage zu jedem Patienten zulassen. Für populationsspezifische Aussagen sind marginale Modelle mit ihrem hohen Maß an Flexibilität vorzuziehen.

Nachdem die Berechnung marginaler Modellschätzungen nicht auf einer konkreten Wahrscheinlichkeitsfunktion (Likelihood) basiert, ist deren Einsatz stets an Schwierigkeiten hinsichtlich der Modellselektion geknüpft. So existiert keine etablierte Maßzahl für die Modellanpassung vergleichbar zum Akaike Informationskriterium (AIC) für likelihood-basierte Modelle. Ein in der Literatur diskutierter Vorschlag eines auf die Quasi-Likelihood-Funktion angepasste Form des AIC namens QIC hat sich im Rahmen einer Simulationsstudie als effektive Maßzahl dargestellt. Demnach kann unter der zugrunde liegenden Daten- und Modellstruktur des ersten Evaluationsteils eine Verwendung empfohlen werden. Für eine generelle Beurteilung des QIC sind weitere Simulationsstudien unter alternativen Datenbeständen und komplexeren Modellen unerlässlich.

Der zweite Teil der MNS-R-Evaluation, die gesundheitsökonomische Bewertung, beinhaltet eine Vergleichsanalyse zwischen Programmteilnehmern und konventionell behandelten Rückenschmerzpatienten. Dabei werden sowohl direkte als auch indirekte Kosten der medizinischen Versorgung in Zeiträumen vor und nach der Intervention gegenübergestellt. Die Zusammenstellung einer konventionell therapierten Referenzgruppe erfolgt unter Verwendung des Versichertenkollektivs der SBK. Zur Vorbeugung von Selektionseffekten werden Referenzpatienten mittels eines eigens entwickelten Matching-Procederes identifiziert. Als Ergebnis verfügt jeder Programmteilnehmer über mehrere Kontrollen, welche eine starke Ähnlichkeit zu ihm aufweisen. Das Ziel des Matchings besteht in der Balance zwischen den Populationen hinsichtlich der Verteilung in Endpunkten und Störgrößen.

Unter Berücksichtigung der Zuordnung zwischen Programmteilnehmern und Kontrollen erfolgt die Umsetzung der Kostenanalyse als Fall-Kontroll-Studie. Dadurch eröffnet sich eine Nutzung des durch das Matching erzeugten zusätzlichen Informationsgehaltes. Neben der Anwendung nonparametrischer Testverfahren bieten *linear mixed effects models*

aus der longitudinalen Datenanalyse einen besonderen Vorteil, da sich die Stratifikation als random effects in die Berechnung einbinden lässt.

Insgesamt muss konstatiert werden, dass Analysen ohne randomisierte Referenzgruppe stets qualitativen Nachteilen unterliegen. Trotz der Korrekturbemühungen durch die Entwicklung und Anwendung eines Matching-Vorgehens lassen sich die Defizite nicht vollständig kompensieren. Demnach werden für eine nachhaltige Klärung der Kosteneffektivität multimodaler Schmerztherapieprogramme weitere Studien als notwendig erachtet.

Literaturverzeichnis

[1] Andersson, I. et al., 1988. *Mammographic screening and mortality from breast cancer: the Malmö mammographic screening trial*. British Medical Journal 297(6654): pp. 943-948.

[2] Breslow, N.E. et al., 1978. *Estimation of multiple relative risk functions in matched case-control studies*. American Journal of Epidemiology 108(4): pp. 299-307.

[3] Bundesärztekammer, Kassenärztl. Bundesvereinigung & Arbeitsgem. d. Med. Fachgesellschaften, 2010. *Nationale VersorgungsLeitlinie Kreuzschmerz - Langfassung*. [Online] Available at: http://www.kreuzschmerz.versorgungsleitlinien.de [Accessed 15 März 2013].

[4] Campbell, M. et al., 2000. *Framework for design and evaluation of complex interventions to improve health*. British Medical Journal 321: pp. 694-696.

[5] Clayton, D. & Hills, M., 1993. *Statistical models in epidemiology*. New York: Oxford University Press.

[6] Cnaan, A., Laird, N.M. & Slasor, P., 1997. *Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data*. Statistics in Medicine 16: pp. 2349-2380.

[7] Demissie, K., Mills, O.F. & Rhoads, G.G., 1998. *Empirical comparison of the results of randomized controlled trials and case-control studies in evaluating the effectiveness of screening mammography*. Journal of Clinical Epidemiology 51(2): pp. 81-91.

[8] Fahrmeir, L., Hamerle, A. & Tutz, G., 1996. *Multivariate statistische Verfahren*. 2nd ed. Berlin: Walter de Gruyter.

[9] Fahrmeir, L., Kneib, T. & Lang, S., 2007. *Regression - Modelle, Methoden und Anwendungen*. Berlin, Heidelberg: Springer.

[10] Fahrmeir, L. & Tutz, G., 1994. *Multivariate statistical modelling based on generalized linear models*. New York: Springer.

[11] Fitzmaurice, G.M., Laird, N.M. & Ware, J.H., 2004. *Applied longitudinal analysis*. New Jersey: John Wiley & Sons.

[12] Guzmán, J. et al., 2002. *Multidisciplinary bio-psycho-social rehabilitation for chronic low-back pain*. The Cochrane Database of Systematic Reviews.

[13] Ho, P.M., Peterson, P.N. & Masoudi, F.A., 2008. *Evaluating the evidence: Is there a rigid hierarchy?* Circulation 118: pp. 1675-1684.

[14] Jensen, I.B., Bergström, G., Ljungquist, T. & Bodin, L., 2005. *A 3-year follow-up of a multidisciplinary rehabilitation programme for back and neck pain*. Pain 115: pp. 273-283.

[15] Jørgensen, K.J., 2011. *Flawed methods explain the effect of mammography screening in Nijmegen*. British Journal of Cancer 105: pp. 592-593.

[16] Kupper, L.L. et al., 1981. *Matching in epidemiologic studies: Validity and efficiency considerations*. Biometrics 37: pp. 271-291.

[17] Laird, N.M. & Ware, J.H., 1982. *Random-effects models for longitudinal data*. Biometrics 38: pp. 963-974.

[18] Lechner, M., 1999. *Identification and estimation of causal effects of multiple treatments under the conditional independence*. Bonn: IZA, Discussion Paper No. 91.

[19] Liang, K.-Y. & Zeger, S.L., 1993. *Regression analysis for correlated data*. Annual Review of Public Health 14: pp. 43-68.

[20] Little, R.J.A. & Rubin, D.B., 2002. *Statistical analysis with missing data*. 2nd ed. New York: John Wiley & Sons.

[21] McCartney, M., 2012. *Breast screening: case-control vs. rct - the problems*. [Online] Available at: http://margaretmccartney.com/2012/09/08/breast-screening-casecontrol-vs-rct-the-problems/ [Accessed 09 Oktober 2013].

[22] Nelder, J.A. & Wedderburn, R.W.M., 1972. *Generalized linear models*. Journal of the Royal Statistical Society, Series A 135(3): pp. 370-384.

[23] Pan, W., 2001. *Akaike's information criterion in generalized estimating equations*. Biometrics 57: pp. 120-125.

[24] Pan, W., Louis, T.A. & Connett, J.E., 2000. *A note on marginal linear regression with correlated response data*. The American Statistician 54(3): pp. 191-195.

[25] Pocock, S.J. & Simon, R., 1975. *Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial*. Biometrics 31: pp. 103-115.

[26] Rässler, S., 2002. *Statistical matching - A frequentist theory, practical applications and alternative bayesian approaches*. Lecture Notes in Statistics, 168. New York: Springer.

[27] Robert Koch-Institut (Hrsg.), 2012. *Rückenschmerzen. Gesundheitsberichterstattung des Bundes. Heft 53*. Berlin.

[28] Schlesselman, J.J., 1982. *Case-Control Studies: Design, Conduct, Analysis*. Oxford: Oxford University Press.

[29] Senn, S., 1994. *Testing for baseline balance in clinical trials*. Statistics in Medicine 13: pp. 1715-1726.

[30] Senn, S., 2002. *Cross-over trials in clinical research*. 2nd ed. Chichester: John Wiley & Sons.

[31] Signorini, D.F. et al., 1993. *Dynamic balanced randomization for clinical trials*. Statistics in Medicine 12: pp. 2343-2350.

[32] Smith, H.L., 1997. *Matching with multiple controls to estimate treatment effects in observational studies*. Sociological Methodology 27: pp. 325-353.

[33] Toutenburg, H., 2003. *Lineare Modelle*. 2nd ed. Heidelberg: Physica.

[34] Tutz, G., 2000. *Die Analyse kategorialer Daten*. München: Oldenbourg.

[35] Windeler, J. et al., 2008. *Randomisierte kontrollierte Studien: Kritische Evaluation ist ein Wesensmerkmal ärztlichen Handelns*. Deutsches Ärzteblatt 105(11): pp. A 565-570.

[36] Yalom, I.D. & Leszcz, M., 2005. *The theory and practice of group psychotherapy*. 5th ed. New York: Basic Books.

Abbildungsverzeichnis

Tabellenverzeichnis

Eidesstattliche Versicherung

Ich, *Dipl.-Stat. Michael Simang*, erkläre hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Thema

„Evaluationsverfahren für eine komplexe Intervention der beruflichen Gesundheitsförderung“

selbständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

München, 20. Mai 2014

_____

Dipl.-Stat. Michael Simang