Qualitätssicherung durch (faire) Einrichtungsvergleiche?

Zum Umgang mit dem Problem fehlender Werte im Kontext der einrichtungsvergleichenden Qualitätssicherung medizinischer Rehabilitation

Dissertation

zur Erlangung der Würde des Doktors der Philosophie

der Universität Hamburg

vorgelegt von

Sven Rabung

aus München

Hamburg 2007

Referent: Prof. Dr. Dr. Uwe Koch
Korreferent: Prof. Dr. Franz Petermann

Datum der letzten mündlichen Prüfung: 04.04.2007

„The only really good solution to the missing data problem is not to have any."

(Paul D. Allison)

Kontaktadresse:

Dipl.-Psych. Sven Rabung

Institut und Poliklinik für Medizinische Psychologie
Universitätsklinikum Hamburg-Eppendorf
Martinistr. 52, Haus S35
20246 Hamburg

srabung@uke.uni-hamburg.de

Danksagung

Ohne die Unterstützung, die mir – vor allem in den letzten Monaten – von verschiedener Seite zuteil wurde, hätte ich die vorliegende Arbeit niemals so schnell vollenden können.

Ich möchte mich daher ganz herzlich bedanken…

…bei *Holger Schulz*, dem Ehrenvorsitzenden meines Buchclubs, für die zahlreichen sokratischen Dialoge, die sachgerechte Umsetzung sozialpsychologischer Befunde zur Förderung optimaler Leistungsfähigkeit unter den Bedingungen maximaler Kontrolle vs. maximaler Freiheit und das eigens entwickelte HDI-System (HDI = „HolgerDissInkasso"), mit dem er unerbittlich Fortschritte einforderte.

…bei *Stephan Kawski*, der meiner Motivation als Mitstreiter im Doktorandenvergleich sehr förderlich war, und mich auch noch, als er bereits konfounderbedingt zurückgefallen war, stets aufzubauen wusste.

…bei *Ane Bleich*, die mich nach zahllosen durchwachten Nächten mit ihrer koffeinhaltigen Unterstützung über so manches Nachmittagstief gerettet hat.

…bei *Uwe Koch*, der es hoffentlich nicht bereut hat, mich im vergangenen Jahr nach Hamburg geholt zu haben, für die Bereitstellung der Rahmenbedingungen und die sehr wohltuende Unterstützung aus dem Hintergrund.

…bei den Menschen, die mein Interesse am wissenschaftlichen Arbeiten geweckt und gefördert haben. Dies waren vor allem (in chronologischer Reihenfolge) *Isa Sammet*, *Kai Sassenberg*, *Henning Schauenburg*, *Eric Leibing* und *Falk Leichsenring*.

…bei *meinen Eltern*, die mich in der Wahl und Verwirklichung meiner Ziele schon immer uneingeschränkt unterstützt haben und damit einen wichtigen Grundstein für meine Entwicklung bis zum heutigen Tage gelegt haben.

…bei meiner Lebensgefährtin *Sylke Andreas*, die nach eigener Promotion und Approbation nun auch einmal „die andere Seite" kennen lernen musste und mich trotz nicht unerheblicher eigener Beanspruchung stets nach Kräften unterstützt hat.

Euch allen vielen Dank!

Hamburg, im Dezember 2006 Sven Rabung

Inhalt

1. **Hintergrund** 9

2. **Das Problem fehlender Daten** 12
   2.1. Systematik fehlender Werte 14
   2.2. Umgang mit fehlenden Werten 16
   2.3. Determinanten der Güte von Fehlwert-Ersetzungen 26

3. **Fragestellung** 30

4. **Material und Methoden** 35
   4.1. Untersuchungsrahmen und -design 36
   4.2. Instrumente und berücksichtigte Variablen 37
   4.3. Stichprobe 43
   4.4. Analyse fehlender Werte 47
   4.5. Simulationsstudie zum Umgang mit fehlenden Werten 48
   4.6. Hypothesen und Hypothesenprüfung 61
   4.7. Exemplarische Übertragung der Befunde zur Ersetzbarkeit fehlender Entlassungswerte auf den Kontext des Einrichtungsvergleichs 67

5. **Ergebnisse** 69
   5.1. Fehlende Werte 69
   5.2. Simulationsstudie zur Ersetzbarkeit fehlender Werte 74
   5.3. Exemplarische Darstellung eines Einrichtungsvergleiches mit Multipler Imputation fehlender Werte 99

6. **Diskussion** 109
   6.1. Notwendigkeit der Berücksichtigung fehlender Werte 110
   6.2. Ersetzbarkeit fehlender Werte 113
   6.3. Konsequenzen der Ersetzung fehlender Werte 118
   6.4. Empfehlungen zum Umgang mit fehlenden Werten 120
   6.5. Validität der ermittelten Befunde 123

7. **Fazit** 125

8. **Zusammenfassung** 126

9. **Literaturverzeichnis** 128

10. **Abbildungsverzeichnis** 134

11. **Tabellenverzeichnis** 135

12. **Anhang** 137
   12.1. Berücksichtigte Variablen 137
   12.2. Verteilungseigenschaften der Ergebnismaße nach Dropoutsimulation 138
   12.3. Intraklassenkorrelationskoeffizienten für die einzelnen Ersetzungsvarianten 143

1. Hintergrund

Einrichtungsvergleiche stellen im Bereich der medizinischen Versorgung einen zentralen Bestandteil von Qualitätssicherungsprogrammen dar, der im Bereich der medizinischen Rehabilitation sogar explizit vom Gesetzgeber gefordert wird (vgl. §20 Abs. 1, SGB IX). Der direkte Vergleich verschiedener Einrichtungen eines Indikationsbereichs soll dabei als Grundlage für die Beurteilung der Qualität der erbrachten medizinischen Leistungen einzelner Kliniken dienen. Gegenüber alternativen Bewertungsverfahren, wie z.B. der Anwendung absoluter oder ipsativer Standards, zeichnet sich die Methode des Einrichtungsvergleichs vor allem durch ihre unmittelbare Praxisrelevanz aus: Der Vergleich konkurrierender Einrichtungen soll es nämlich ermöglichen, die zentrale Frage zu beantworten, in welchen Einrichtungen Patienten faktisch am besten behandelt werden (Farin et al., 2004).

Die wichtigste Aufgabe bei der Konzeption von Qualitätssicherungsprogrammen besteht zunächst in der adäquaten Operationalisierung der zu bewertenden Qualität. Nur, wenn die relevanten Qualitätsindikatoren identifiziert und mittels geeigneter Messverfahren reliabel und valide erfasst werden können, ist eine vergleichende Bewertung der Qualität der in verschiedenen Einrichtungen durchgeführten Behandlungen möglich. Die vorrangige Bedeutung kommt dabei im Bereich der medizinischen Versorgung naheliegenderweise dem Behandlungsergebnis zu („Ergebnisqualität"). Ergänzend sind jedoch auch strukturelle und prozessbezogene Faktoren zu berücksichtigen, die sich für das Erreichen eines optimalen Behandlungsergebnisses als bedeutsam erwiesen haben bzw. als bedeutsam angenommen werden („Struktur- und Prozessqualität", vgl. Kawski & Koch, 1999; Kawski & Koch, 2002; Kawski & Koch, 2004).

Außerdem ist bei solchen Einrichtungsvergleichen auch immer zu berücksichtigen, dass insbesondere die erzielten Behandlungsergebnisse nicht allein in der Verantwortung der jeweiligen leistungserbringenden Einrichtung liegen, sondern zusätzlich durch zahlreiche äußere Faktoren beeinflusst werden können (Shwartz et al., 1997). Neben der tatsächlichen Qualität der erbrachten Leistungen determinieren vor allem die spezifischen Eigenschaften der behandelten Patienten das Therapie-

ergebnis, ohne dass diese durch die leistungserbringenden Einrichtungen zu kontrollieren wären (vgl. Abbildung 1).

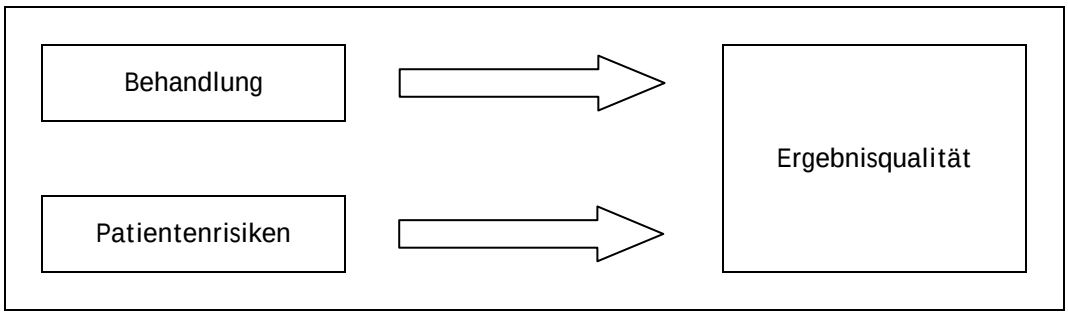

Abbildung 1: Determinanten der Ergebnisqualität[1]

Faktoren, die mit dem Behandlungsergebnis assoziiert sein könnten, aber nicht durch die zu beurteilenden Einrichtungen zu beeinflussen sind (sog. „Konfounder" oder „Risikofaktoren"), sind daher bei der Erstellung von Einrichtungsvergleichen entsprechend zu berücksichtigen, um eine faire Bewertung der verglichenen Einrichtungen zu gewährleisten. Zur Kontrolle derartig mit dem Behandlungsergebnis konfundierter Faktoren haben sich statistische Verfahren der Risikoadjustierung bewährt (Farin et al., 2004; Schulz et al., 2004; Wegscheider, 2004).

Doch selbst wenn Ergebnisqualität und patientenbezogene Risikofaktoren optimal operationalisiert wurden, ist der Rückschluss von den gemessenen Behandlungsergebnissen auf die Qualität der erbrachten Leistungen einer Einrichtung nicht zwingend zulässig. In nahezu allen Untersuchungen ergeben sich nämlich auch umschriebene Anteile von Fällen, die aufgrund von fehlenden Daten nicht in den Qualitätsanalysen berücksichtigt werden können (sog. „Dropouts" oder „Nonresponder"). Mit steigender Dropout-Rate ist die Generalisierbarkeit ermittelter Ergebnisse jedoch zunehmend gefährdet. Um Dropout-behaftete Daten also überhaupt in Einrichtungsvergleichen verwenden zu dürfen, sollte sichergestellt sein, dass sich die Fälle, für die verwertbare Daten vorliegen (sog. „Responder"), nicht systematisch von den Dropouts unterscheiden. Gemäß der gängigen Praxis wird zu diesem Zwecke in Dropoutanalysen überprüft, ob sich die jeweils resultierende Untersu-

[1] Streng genommen müsste hier auch noch der Zufall als weitere Einflussgröße auf die Ergebnisqualität aufgeführt werden. Da zufällige Einflüsse auf das Behandlungsergebnis jedoch durch die Untersuchung entsprechend großer Stichproben kontrolliert werden können, soll dieser Faktor an dieser Stelle vernachlässigt werden.

chungsstichprobe einer Einrichtung zumindest hinsichtlich der bekanntermaßen mit dem interessierenden Qualitätskriterium konfundierten Merkmale als repräsentativ für die interessierende Grundgesamtheit der in dieser Einrichtung behandelten Patienten belegen lässt (vgl. hierzu auch Kap. 2). Bislang ist jedoch nicht gesichert, inwieweit ein solches Vorgehen tatsächlich ausreicht, um faire Einrichtungsvergleiche auch bei sehr unterschiedlichen Teilnahmequoten der einzelnen Einrichtungen zu gewährleisten. Außerdem fehlen Konzepte zum Umgang mit Einrichtungen, für die sich die Repräsentativität der erfassten Daten nicht bestätigen lässt.

Zusammenfassend lässt sich somit festhalten, dass die *gemessene* Ergebnisqualität also nicht nur durch Merkmale der behandelnden Einrichtung und Merkmale der behandelten Patienten, sondern – nicht zuletzt - auch durch verschiedene Merkmale der Untersuchung, wie insbesondere die Vollständigkeit der erfassten Daten und den Umgang mit fehlenden Werten, determiniert wird (vgl. Abbildung 2 sowie das folgende Kapitel).

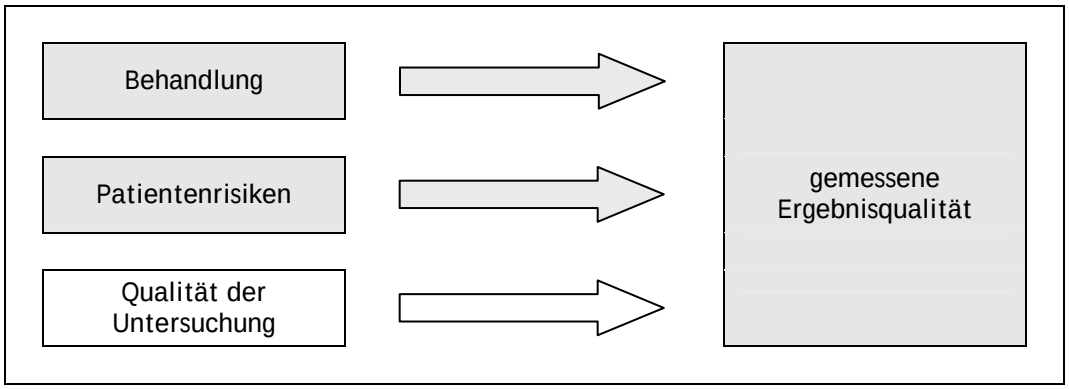

Abbildung 2: Determinanten der gemessenen Ergebnisqualität

In der vorliegenden Arbeit sollen vor diesem Hintergrund alternative Strategien zum Umgang mit fehlenden Werten untersucht und hinsichtlich der Konsequenzen ihres Einsatzes im Kontext der vergleichenden Qualitätssicherung überprüft werden. Nachdem im nächsten Kapitel zunächst die verschiedenen Mechanismen fehlender Werte sowie gängige Verfahren zum Umgang mit fehlenden Werten dargestellt werden, wird die Zweckmäßigkeit der verschiedenen Imputationsverfahren im empirischen Teil der Arbeit im Rahmen einer Simulationsstudie zur Ersetzbarkeit fehlender Werte sowie anhand eines praxisorientierten Anwendungsbeispiels beleuchtet.

2. Das Problem fehlender Daten

In nahezu allen empirischen Untersuchungen ergeben sich, insbesondere im Falle wiederholter Messungen, mehr oder weniger große Raten fehlender Werte. Zum Teil ergeben sich umschriebene Fehlwertquoten aufgrund einzelner fehlender Angaben bei ansonsten vorliegenden Daten (sog. „Item-Nonresponse“), zum Teil fehlen die Angaben einzelner Patienten aus bestimmten Erhebungseinheiten jedoch auch vollständig (sog. „Unit-Nonresponse“). Das Auftreten von fehlenden Werten führt insbesondere bei multivariaten Analysen, die auf einer vollständigen Datenmatrix beruhen, in jedem Falle zu einer Fallzahlreduktion und damit verbundenen Verringerung der statistischen Power. Durch die ausschließliche Berücksichtigung von Fällen mit vollständigen Daten („Complete Case“-Analysen) ergibt sich ein Informationsverlust, die Repräsentativität der in der Auswertung berücksichtigten Stichproben wird fraglich und statistische Ergebnisse können verfälscht werden. Mit steigender Fehlwertquote ist also nicht zuletzt auch die Generalisierbarkeit ermittelter Befunde zunehmend eingeschränkt. Um in solchen Fällen sicherzustellen, dass Befunde, die auf der Grundlage von eingeschränkten Stichproben ermittelt wurden, dennoch Rückschlüsse auf die interessierende Grundgesamtheit erlauben, werden üblicherweise Repräsentativitätsanalysen durchgeführt: In diesen wird die Stichprobe, für die vollständig verwertbare Daten vorliegen, hinsichtlich als zentral angenommener Maße mit der Stichprobe der Dropout-Patienten verglichen. Sollen Dropout-behaftete Daten in Einrichtungsvergleichen verwendet werden, muss zuvor zumindest sichergestellt sein, dass sich die jeweils resultierende Responderstichprobe einer Einrichtung hinsichtlich der bekanntermaßen mit dem Zielkriterium konfundierten Merkmale als repräsentativ für die Grundgesamtheit der in dieser Einrichtung behandelten Patienten belegen lässt. Lässt sich die Repräsentativität jedoch nicht bestätigen, so dürfte die entsprechende Einrichtung konsequenterweise auch nicht im Einrichtungsvergleich berücksichtigt werden.

Um dem Problem eingeschränkter Stichproben zu begegnen, sollten fehlende Werte also nach Möglichkeit ersetzt werden. Während dies bei einzelnen fehlenden Werten unter bestimmten Voraussetzungen verhältnismäßig gut möglich ist, stellen größere Fehlwertquoten oder das vollständige Fehlen von Daten zu einem Messzeit-

punkt höhere methodische Anforderungen dar. In der gängigen Forschungspraxis führen sie daher bislang noch häufig zu einem Ausschluss der entsprechenden Fälle von den weiteren Analysen (sog. „Dropout“). Derartige Dropout-Quoten schwanken je nach Studiendesign und untersuchter Stichprobe erheblich, die Spannweite reicht dabei üblicherweise von wenigen Prozenten bis hin zu Ausfallquoten von über 50 Prozent. Vollständig verwertbare Datensätze bilden die große Ausnahme.

In den letzten Jahren hat sich jedoch mehr und mehr die Forderung durchgesetzt, die Auswertung klinischer Studien in jedem Falle nach dem „Intention-to-treat“-Ansatz (ITT) durchzuführen. In seiner strengen Form verlangt das ITT-Prinzip, dass jeder in eine Studie aufgenommene Patient in den Auswertungen berücksichtigt wird. Da die einfacheren „Complete Case“-Analysen dem ITT-Prinzip widersprechen, sind also Methoden gefragt, die verhindern, dass Patienten mit teilweise fehlenden Werten aus der Auswertung ausgeschlossen werden. Bei der praktischen Umsetzung des ITT-Prinzips stellt sich somit die Frage, wie mit fehlenden Werten bezüglich relevanter Zielgrößen umgegangen werden kann.

Da die Möglichkeiten des Umgangs mit fehlenden Daten maßgeblich von der Systematik der Datenausfälle determiniert werden, soll im Folgenden zunächst ein Überblick über die verschiedenen möglichen Ausfallmechanismen gegeben werden. Im Anschluss werden dann die verschiedenen Optionen zum Umgang mit fehlenden Werten sowie ihr Einfluss auf das Ergebnis der auf ihrer Grundlage durchgeführten statistischen Analysen im Einzelnen dargestellt. Abschließend sollen in diesem Kapitel weitere Faktoren benannt werden, die die Güte der Ersetzung fehlender Werte zusätzlich beeinflussen können.

2.1. Systematik fehlender Werte

In klinischen Studien fehlen Daten in den seltensten Fällen zufällig. Zumeist ist ein Datenausfall systematisch bedingt, das heißt, dass es einen Grund für das Nichtvorliegen bestimmter Daten gibt. Nach den Ursachen, die das Auftreten von fehlenden Werten bedingen, lassen sich grundsätzlich vier verschiedene Ausfallmechanismen unterscheiden (vgl. Tabelle 1), die für den Umgang mit den fehlenden Daten von entscheidender Bedeutung sind (z.B. Allison, 2002; Collins et al., 2001; Rubin, 1976):

(1) Nur, wenn das Auftreten fehlender Werte weder von der Ausprägung der (nicht angegebenen) Werte der Variablen selbst, noch von der Ausprägung anderer erhobener Variablen abhängt, spricht man von „vollständig zufällig fehlenden" Daten (MCAR = *Missing Completely at Random*).

(2) Ein zweiter Ausfallmechanismus wird irreführenderweise als „zufällig fehlend" (MAR = *Missing at Random*) bezeichnet, obwohl die fehlenden Werte in dieser Bedingung alles andere als zufällig fehlen. „Zufälliges Fehlen" liegt nämlich dann vor, wenn das Fehlen von Werten zwar nicht von der Ausprägung der (nicht bekannten) Werte der Variablen selbst abhängt, aber durch die übrigen Variablen im Datensatz erklärt werden kann (Schafer, 1997). Insofern wäre hier die Bezeichnung „bedingt zufällig fehlend" angemessener.

(3a/b) Wenn nun aber die im Datensatz fehlenden Werte direkt von der Ausprägung der (nicht bekannten) Werte der jeweiligen Variablen selbst abhängen, spricht man von „nicht zufällig fehlenden" Daten (MNAR = *Missing not at Random*). In den meisten Fällen „nicht zufällig fehlender" Informationen ist das Auftreten fehlender Werte zusätzlich mit der Ausprägung anderer Variablen assoziiert. Ist dies nicht der Fall, spricht man von als „zufällig fehlend beobachteten" Daten (OAR = *Observed at Random*). Letzterer Spezialfall stellt ein besonderes methodisches Problem dar, da er empirisch nicht von MCAR zu unterscheiden ist. Letztlich ist er aber dem Mechanismus „nicht zufällig fehlender" Werte (MNAR) zuzuordnen.

Tabelle 1: Unterscheidung von Ausfallmechanismen nach den Ursachen für fehlende Daten

Ausfallmechanismus	**Ausfall abhängig von fehlenden Daten?**	**Ausfall abhängig von vorhandenen Daten?**
Missing Completely at Random (MCAR)	nein	nein
Missing at Random (MAR)	nein	ja
Missing not at Random (MNAR)	ja	ja/nein
MNAR-Spezialfall: Observed at Random (OAR)	ja	nein

„Vollständig zufällig fehlende“ Werte (MCAR) bilden in klinischen Studien eine große Ausnahme, da es zumeist eine systematische Ursache dafür gibt, dass Probanden ihre Angaben verweigern oder nicht erreicht werden konnten (Rubin, 1976). MCAR könnte aber beispielsweise auftreten, wenn fehlende Werte durch Eingabefehler entstanden oder einzelne Fragebögen verloren gegangen sind. „Nicht zufällig fehlende“ Werte (MNAR) könnten zum Beispiel vorliegen, wenn im Rahmen einer Untersuchung zur Depression ein Item zur aktuellen Suizidalität besonders häufig von denjenigen Probanden nicht beantwortet wird, die aktuell suizidal sind. Würden im gleichen Fragebogen jedoch noch einige weitere Variablen zur Depressivität erhoben, die in engem Zusammenhang mit dem Merkmal Suizidalität und somit auch mit dem Fehlen der entsprechenden Informationen stehen, könnte bereits lediglich „zufälliges Fehlen“ (MAR) vorliegen.

Anhand einer Missing-Data-Diagnose kann festgestellt werden, ob die Bedingung MCAR verletzt ist. Dies wäre zum Beispiel der Fall, sobald bedeutsame Zusammenhänge zwischen dem Fehlen von Werten und anderen Informationen im Datensatz bestehen. Prüfen lässt sich die MCAR-Annahme unter anderem mittels des MCAR-Tests nach Little (1987). Liegt dem Ausfall der Daten jedoch eine Systematik zugrunde, so lässt sich nicht empirisch ermitteln, ob die Daten „zufällig“ (MAR) oder „nicht zufällig“ (MNAR) fehlen. Um diese Unterscheidung treffen zu können, müsste nämlich die Ausprägung der fehlenden Daten bekannt sein, da nur so überprüft werden könnte, ob die nicht vorliegenden Informationen notwendig sind, um den Datenausfall zu erklären. Zur Entscheidung dieser Frage muss daher auf inhaltliches Wissen über den Forschungsgegenstand zurückgegriffen werden. In der Regel

ist jedoch keine eindeutige Beantwortung dieser Fragestellung möglich (Wirtz, 2004).

Wenn fehlende Daten das MCAR- oder MAR-Kriterium erfüllen, kann der Ausfallmechanismus als „ignorable“ bezeichnet werden. Dies bedeutet, dass die Systematik fehlender Werte im Ersetzungsprozess nicht eigens modelliert werden muss. Sind die Daten nicht MAR, wird der Ausfallmechanismus als „nonignorable“ bezeichnet. Die Ausfallsystematik muss in diesem Falle in den entsprechenden Ersetzungsmodellen berücksichtigt werden, um gute Schätzungen der interessierenden Parameter zu erhalten. Dies setzt jedoch eine sehr gute Kenntnis der zugrunde liegenden Ausfallmechanismen voraus, die nicht aus den vorliegenden Daten erschlossen werden können. Aus diesem Grund und weil Modelle für „nonignorable“ fehlende Daten jeweils auf den individuellen Anwendungskontext zugeschnitten werden müssen, sollen sich die weiteren Ausführungen maßgeblich auf solche Fälle beschränken, in denen fehlende Daten mindestens das MAR-Kriterium erfüllen.

2.2. Umgang mit fehlenden Werten

Seit den späten 50er Jahren des vergangenen Jahrhunderts wurde eine Vielzahl von Methoden zum Umgang mit fehlenden Werten entwickelt (vgl. Afifi & Elashoff, 1966; Little & Rubin, 1987; Schafer, 1999; Scheffer, 2002). Die verschiedenen Strategien zum Umgang mit fehlenden Werten können sich unterschiedlich auf die Bewertung der Qualität untersuchter Einrichtungen (z.B. durch Verzerrungen) sowie auf die Power der durchgeführten Tests (z.B. durch Stichprobenreduktion) auswirken. Obwohl sich in der Literatur zunehmend Hinweise auf mögliche Gefahren der Anwendung (zu) einfacher Strategien des Umgangs mit fehlenden Daten finden, ist deren unkritische Anwendung bislang in der Praxis weit verbreitet.

Grundsätzlich lassen sich zwei Hauptgruppen von Verfahren zum Umgang mit fehlenden Werten unterscheiden: Methoden, bei denen fehlende Werte von den weiterführende Analysen ausgeschlossen werden (sog. „Eliminierungsverfahren“), und Verfahren, bei denen fehlende Werte vor der Durchführung weitergehender Analy-

sen auf der Basis gültiger Werte in anderen Variablen bzw. Fällen der Stichprobe ersetzt werden (sog. „Imputationsverfahren"). Eine weitere Unterscheidung der Verfahren lässt sich nach den notwendigen Voraussetzungen in Bezug auf die dem Datenausfall zugrunde liegenden Mechanismen anstellen: Hier sind im Wesentlichen Verfahren, die „vollständig zufällig fehlende" Werte (MCAR) voraussetzen, von solchen abzugrenzen, die lediglich an die Bedingung „zufällig fehlender" Werte (MAR) gebunden sind. Im Folgenden sollen die gängigsten Verfahren der beiden genannten Hauptgruppen (Eliminierungs- und Imputationsverfahren) im Einzelnen dargestellt werden. Da die Eliminierungsverfahren sowie das Imputationsverfahren der Mittelwertersetzung an die MCAR-Bedingung geknüpft sind, die in der Realität so gut wie nie vorliegt, soll der Schwerpunkt dabei auf die komplexeren Imputationsverfahren gelegt werden (vgl.Tabelle 2).

Tabelle 2: Gängige Verfahren zum Umgang mit fehlenden Werten

Verfahren	**Beschreibung**	**Bedingung**
Fallweiser Ausschluss („Complete Case Analysis")	Alle Fälle mit mindestens einem fehlenden Wert werden ausgeschlossen.	MCAR
Paarweiser Ausschluss („Available Case Analysis")	Variablen werden paarweise analysiert, wobei alle Fälle mit fehlenden Werten ausgeschlossen werden.	MCAR
Ersetzung durch Mittelwerte („Mean Imputation")	Alle fehlenden Werte werden durch Variablenmittelwerte ersetzt.	MCAR
Regressionsschätzungen („Regression Imputation")	Alle fehlenden Werte werden durch eine regressionsbasierte Schätzung ersetzt.	MAR
Maximum-Likelihood-Schätzungen (z.B. EM-Algorithmus)	Alle fehlenden Werte werden durch Maximum-Likelihood-Schätzungen ersetzt.	MAR
Multiple Imputationen	Jeder fehlende Wert wird durch mehrere zufällige oder geschätzte Werte ersetzt.	MAR (MNAR)

Fallweiser Ausschluss („Complete Case Analysis“)

Die auch als „Complete Case Analysis“ bezeichnete Methode des fall- bzw. listenweisen Ausschlusses besteht in der Eliminierung aller unvollständigen Datensätze und der anschließenden statistischen Analyse der vollständig beobachteten Untersuchungseinheiten. Für die statistische Auswertung sind dadurch auch die Voraussetzungen zur Anwendung von multivariaten Standardverfahren, die eine vollständige Datenmatrix verlangen, gegeben. Durch die „Complete Case Analysis“ wird außerdem die Vergleichbarkeit verschiedener univariater Statistiken gewährleistet, da die einzelnen Analysen jeweils auf der gleichen Stichprobengröße basieren (vgl. Little & Rubin, 2002).

Ein wesentlicher Nachteil dieses Verfahrens besteht jedoch im teilweise immensen Informationsverlust, der sich durch die Nichtberücksichtigung von vorhandenen Beobachtungen in multivariaten Datensätzen ergibt, in denen gemäß des Ausschlusskriteriums zum Teil nur ein einzelner Wert fehlen muss. Aus der Verringerung der Stichprobe ergeben sich ein vergrößerter Standardfehler und damit eine verringerte statistische Power. Im Extremfall ist dabei eine Reduktion der Stichprobe bis hin zur völligen Unbrauchbarkeit des resultierenden Datensatzes möglich. Zusätzlich einschränkend ist, dass die Anwendung des fallweisen Ausschlusses nur dann zu unverzerrten Ergebnissen führt, wenn der Datenausfall vollständig zufällig (MCAR) ist und die vollständigen Datensätze eine repräsentative Stichprobe der Grundgesamtheit bilden. Diese Bedingung ist jedoch in der Forschungspraxis nur selten gegeben. Ist die MCAR-Bedingung verletzt, führt die „Complete Case Analysis“ zu verzerrten Ergebnissen, die nicht generalisiert werden dürfen.

Das Verfahren des fallweisen Ausschlusses sollte also selbst beim Vorliegen vollständig zufällig fehlender Werte im Sinne des MCAR-Kriteriums nur dann angewandt werden, wenn die Daten geringe Fehlwertquoten aufweisen, große Stichproben verfügbar sind und/oder von großen aufzudeckenden Effekten auszugehen ist.

Paarweiser Ausschluss („Available Case Analysis“)

Bei der Methode des paarweisen Ausschlusses handelt es sich um ein weiteres, auch als „Available Case Analysis“ bezeichnetes Eliminierungsverfahren, das für die Datenauswertung die jeweils beobachteten Merkmale berücksichtigt. Hierdurch tritt zwar ein geringerer Informationsverlust als bei Anwendung der „Complete Case Analysis“ auf, es können sich jedoch aufgrund der unterschiedlichen zugrunde gelegten Stichprobenumfänge für verschiedene Statistiken inkonsistente Ergebnisse ergeben (Schafer & Graham, 2002). Wie die Methode des listenweisen Fallausschlusses ist auch die Durchführung der „Available Case Analysis“ an die Voraussetzung vollständig zufällig fehlender Werte (MCAR) gebunden. Aufgrund der genannten Einschränkungen sollte auf die Anwendung dieses Verfahrens möglichst gänzlich verzichtet werden.

Ersetzung durch Mittelwerte („Mean Imputation“)

Das einfachste und dadurch in der Praxis am weitesten verbreitete Verfahren zur Ersetzung fehlender Werte wird als „Mean Imputation“ bezeichnet. Um vervollständigte Datensätze zu erhalten, werden dabei fehlende Werte durch das arithmetische Mittel der beobachteten Werte der entsprechenden Variablen ersetzt (Cochran, 1977). Diese Vorgehensweise führt jedoch oft zu verzerrten Schätzern und in vielen Fällen sogar zu schlechteren Ergebnissen als die zuvor beschriebenen Eliminierungsverfahren. Selbst unter MCAR-Bedingungen ist die Mittelwertersetzung nämlich nur bei der Berechnung von Mittelwerten oder Summen verzerrungsfrei. Die Ersetzung fehlender Werte durch das arithmetische Mittel der jeweiligen Variable verzerrt die wahren Verteilungseigenschaften des entsprechenden Merkmals; die wahre Varianz und damit einhergehend die wahren Zusammenhänge zu anderen Variablen werden unterschätzt. Insofern ist auch dieses Verfahren nicht zur Anwendung zu empfehlen.

Regressionsschätzungen („Regression Imputation“)

Bei der Methode der „Regression Imputation“ werden fehlende Werte durch (multiple) Regression auf der Basis bekannter Beziehungen zwischen den vorhandenen Variablen geschätzt und ersetzt. Voraussetzung für diese Methode ist, neben dem MAR-Kriterium, das Vorliegen substantieller Zusammenhänge zwischen den fehlwertbehafteten und anderen erhobenen Variablen sowie nach Möglichkeit eine geringe Quote von weit verstreuten Fehlwerten.

Durch die Regressionsschätzung wird die wahre Varianz der zu ersetzenden Werte allerdings zumeist unterschätzt, bestehende Zusammenhänge zu anderen Variablen werden verstärkt bzw. verzerrt. Zur Berichtigung der verzerrten Korrelationsmatrix existiert jedoch ein von Buck entwickeltes Korrekturverfahren (Buck, 1960). Im Gegensatz zur Mittelwertsimputation können beim Verfahren der Regressionsschätzung zudem Schätzwerte außerhalb des zulässigen Wertebereichs auftreten. Da sich die Schätzungen konkret auf die Zusammenhänge innerhalb der vorliegenden Daten stützen, kann außerdem die Generalisierbarkeit der resultierenden Datenmatrix beeinträchtigt sein.

Likelihood-basierte Schätzungen und EM-Algorithmus

In der Literatur zur Behandlung von fehlenden Werten finden sich zunehmend Lösungsansätze, die auf der Bestimmung von Maximum-Likelihood-Schätzern beruhen. Diese likelihood-basierten Verfahren sind insbesondere in Fällen anwendbar, in denen der zugrunde liegende Ausfallmechanismus ignorierbar ist (MCAR/MAR). Ein Vorteil dieser Methoden besteht im Gegensatz zu den meisten traditionellen Verfahren in der Einbeziehung aller verfügbaren Informationen des Datenbestandes. Außerdem sind likelihood-basierte Verfahren unabhängig von der Erscheinungsform der fehlenden Werte in der Datenmatrix einsetzbar. Traditionelle Methoden sind demgegenüber häufig auf ein monotones Ausfallmuster beschränkt oder auf die Erfüllung der restriktiven MCAR-Annahme angewiesen.

Die Likelihood-Funktion ist im Allgemeinen kompliziert und nicht in geschlossener Form darstellbar (Schafer, 1997). Das Maximierungsproblem kann jedoch durch die Anwendung von Näherungsverfahren, wie z.B. dem Expectation-Maximization-

Algorithmus (EM-Algorithmus), gelöst werden. Der EM-Algorithmus stellt somit eine allgemeine Methode zur Berechnung von Maximum-Likelihood-Schätzungen dar, die sich prinzipiell auf jedes Problem fehlender Daten anwenden lässt (Little & Rubin, 1987; 2002).

Der EM-Algorithmus ist ein iteratives Verfahren. Jede Iteration besteht aus einem E-Schritt („Expectation-Step“), in dem die "averaged log-likelihood" berechnet wird, und einem M-Schritt („Maximization-Step“), in dem diejenigen Parameter gesucht werden, die die "averaged log-likelihood" maximieren (vgl. Little, 1983). Unter der Annahme einer multivariaten Normalverteilung und einer einfachen Zufallsstichprobe stellen der Vektor der Mittelwerte und die Kovarianzmatrix die suffizienten Statistiken dar. Dementsprechend besteht der E-Schritt des EM-Algorithmus darin, die Erwartungswerte der suffizienten Statistiken unter den aktuellen Schätzungen des Mittelwertvektors und der Kovarianzmatrix zu finden. Der M-Schritt besteht dann aus der Berechnung der neuen Schätzungen. Ein fehlender Wert wird jeweils durch eine lineare Regression mit allen vorhandenen Variablenwerten dieses Falles geschätzt (Schafer & Graham, 2002).

Der vollständige EM-Algorithmus läuft folgendermaßen ab:

1. Zunächst wird eine ursprüngliche Schätzung des Mittelwertvektors und der Kovarianzmatrix aus den vollständigen Fällen berechnet.
2. Für jeden Fall mit fehlenden Werten werden dann die Matrizen in einen vollständigen und einen unvollständigen Teil aufgeteilt.
3. Anschließend werden die fehlenden Werte durch multiple Regression mit allen vorhandenen Variablen unter Benutzung der geschätzten Mittelwerte und der Kovarianzmatrix geschätzt.
4. Danach werden ein neuer geschätzter Mittelwertvektor und eine neue geschätzte Kovarianzmatrix berechnet.
5. Die resultierende Kovarianzmatrix wird korrigiert, indem für jeden Fall mit fehlenden Werten die residuale Kovarianz der Variablen zu dem entsprechenden Element der Kovarianzmatrix addiert wird.

6. Abschließend wird ein Konvergenzkriterium bestimmt, indem die Parameterschätzungen von zwei aufeinander folgenden Iterationen auf Unterschiede überprüft werden. Solange keine Konvergenz erreicht wurde, werden die Schritte 2-6 wiederholt. Erst, wenn zwischen zwei aufeinander folgenden Iterationen keine wesentlichen Unterschiede mehr bestehen, wird der EM-Algorithmus beendet.

Da die Schritte des EM-Algorithmus mit Ausnahme von Schritt 5 lediglich Anwendungen üblicher „complete data“ Methoden darstellen, soll im Folgenden nur noch etwas näher auf die Korrektur der Kovarianzmatrix eingegangen werden: Da bei der Berechnung eines Elementes der jeweils neuen geschätzten Kovarianzmatrix auch Elemente der Datenmatrix benutzt werden, die über multiple Regressionen geschätzt wurden, würden die Diagonalelemente der Kovarianzmatrix, also die Varianzen, ohne Korrektur unterschätzt, die Kovarianzen würden unter- oder überschätzt. Die entsprechend notwendige Korrektur erfolgt dadurch, dass bei der Berechnung der Elemente der Kovarianzmatrix zu jedem Summanden, also für jeden Fall, die residuale Kovarianz addiert wird. Die residuale Kovarianz zweier Variablen entspricht der Kovarianz der Residuen beider Variablen bezüglich der zur Vorhersage benutzten Variablen. Damit hängt die residuale Kovarianz jeweils vom Muster der fehlenden Werte ab, d.h. sie variiert zwischen den Fällen mit unterschiedlichen Fehlwert-Mustern, weil für die Vorhersagen jeweils eine andere Menge von Prädiktoren benutzt wird, nämlich diejenigen Variablen, die in dem jeweiligen Fehlwert-Muster vorhanden sind. Die residuale Kovarianz ist nur dann ungleich Null, wenn beide Variablen für den entsprechenden Fall fehlen, da sonst mindestens eine der beiden Variablen zur Prädiktorenmenge der anderen Variablen gehört. In letzterem Fall würde die Kovarianz zwischen Residuen berechnet, aus denen der durch die jeweils andere Variable linear erklärbare Anteil bereits eliminiert ist, die Kovarianz beträgt also Null. Bei der Berechnung der Varianzen fehlen bei unvollständigen Fällen immer "beide" Variablen oder keine der "beiden" Variablen, daher ist für jeden Fall mit fehlenden Werten auf einer bestimmten Variablen eine Korrektur der Varianz dieser Variablen erforderlich.

Nach Anwendung des EM-Algorithmus können die in der letzten Iteration generierten, vervollständigten Daten für die weitere statistische Analyse verwendet werden. Wie Beale u. Little zeigen konnten (Beale & Little, 1975), ist die Ersetzung fehlender Werte mittels des EM-Algorithmus identisch mit der multiplen Regressionsersetzung nach Buck (1960; s.o.), sofern diese iteriert angewandt wird.

Der EM-Algorithmus wird, zumindest unter der MAR-Annahme, in der neueren Literatur allgemein als das unter allen Umständen zu bevorzugende Verfahren zur einfachen Imputation fehlender Werter empfohlen („State of the Art"). Obgleich die Performanz des EM-Algorithmus in den meisten der bisher vorliegenden Simulationsstudien als den alternativen (einfachen) Verfahren zum Umgang mit fehlenden Werten belegt werden konnte (z.B. Musil et al., 2002; Scheffer, 2002), liegen bislang nur begrenzte Erkenntnisse über seine Robustheit bei Verletzung der MAR-Annahme vor.

Ein verbleibendes Manko der EM-Ersetzung besteht jedoch darin, dass der Algorithmus lediglich die bedingten Erwartungswerte für die Ersetzung verwendet und somit die Unsicherheit bezüglich der Ergänzungen vernachlässigt wird. Insofern sind die Parameterschätzungen, die auf den vervollständigten Daten basieren, bei Gültigkeit der MAR-Annahme zwar unverzerrt, die resultierenden Standardfehler und Teststatistiken sind hingegen nicht unbedingt verlässlich. Vor dem Hintergrund dieser Einschränkung wurden weitere likelihood-basierte Ersetzungsverfahren, wie beispielsweise die verschiedenen Methoden der „Multiplen Imputation" entwickelt.

Multiple Imputationen

Bei jeder Ersetzung fehlender Werte ergibt sich das Problem, dass die ergänzten Werte auch bei Bekanntheit oder Ignorierbarkeit des Ausfallmechanismus von den wahren Ausprägungen der unbeobachteten Daten abweichen können. Werden fehlende Werte jedoch durch jeweils nur eine Ausprägung ersetzt (Einfache Imputation), so bleibt diese Unsicherheit der Schätzung unberücksichtigt. Bei den anschließenden statistischen Analysen wird dann irrtümlicherweise davon ausgegangen, dass die ergänzten Daten genau wie die tatsächlichen Beobachtungen erhoben wurden und somit bekannt wären.

Ein nahe liegender Ansatz zur Lösung dieser Problematik besteht in der mehrfachen Ergänzung von fehlenden Werten (Multiple Imputation, MI), wobei mehrere, d.h. m>1, plausibel vervollständigte Datenbestände erzeugt werden, die wiederum durch statistische Standardmethoden einzeln auswertbar sind. Durch die geeignete Zusammenfassung der m ermittelten Einzelergebnisse lassen sich dann unverzerrte Parameterschätzer generieren.

Ein wesentlicher Vorteil des MI-Verfahrens besteht zusätzlich darin, dass die Unterschiede zwischen den multiplen Analyseergebnissen die Unsicherheit bezüglich der Ersetzung der fehlenden Werte widerspiegeln und somit auch Aussagen über die Genauigkeit der Schätzer getroffen werden können (vgl. Rubin, 1987; Rubin & Schenker, 1986). Die Unsicherheit der Schätzung lässt sich dabei über Konfidenzintervalle für das jeweils zusammengefasste Ergebnis darstellen. Dabei ist allerdings zu berücksichtigen, dass die Schätzer und ihre Varianz bei einer begrenzten Anzahl von m Ersetzungen gegebenenfalls nicht effizient bestimmt werden können. Möglicherweise ließe sich das jeweilige Konfidenzintervall nämlich jeweils noch minimieren, wenn m wesentlich größer gewählt würde. Es konnte jedoch gezeigt werden, dass sich selbst bei hohen Anteilen fehlender Werte in den meisten Fällen eine verhältnismäßig geringe Anzahl an Imputationen als ausreichend erweist, da der Schätzfehler dort nur unwesentlich größer als bei einer hoch gewählten Anzahl von Schätzungen ausfällt. Üblicherweise werden in der Literatur zwischen drei und zehn Ersetzungen pro fehlendem Wert als ausreichend angegeben (Rubin, 1987; Schafer, 1997).

Ein genereller Vorteil der Imputation von fehlenden Werten besteht in der strikten Trennung zwischen Ersetzungsprozess und anschließender Datenauswertung. Dabei ist die Ersetzung als ein vorbereitender Schritt zu betrachten, durch den sichergestellt werden sollte, dass die Berechnung jeglicher Statistiken aus den vervollständigten Datenbeständen zu validen Aussagen über die Grundgesamtheit führt. So ist es z.B. notwendig, dass neben dem Schätzer für die mittlere Ausprägung eines fehlwertbehafteten Merkmals auch die aus den Daten ermittelte Varianz des Schätzers unverzerrt ist. Während diese Bedingungen von einfachen Imputationsverfahren nicht erfüllt werden, soll das Verfahren der Multiplen Imputation diesem Anspruch genügen.

Zusammenfassende Bewertung

Ein Überblick über die Vor- und Nachteile der einzelnen vorgestellten Verfahren zum Umgang mit fehlenden Werten findet sich in Tabelle 3. Dieser Übersicht zufolge wären die likelihood-basierten Verfahren (EM-Algorithmus und Multiple Imputationen) den Eliminierungs-Verfahren sowie den einfacheren Imputationsverfahren in den meisten Fällen vorzuziehen. Multiple Imputationsverfahren zeichnen sich zusätzlich durch die Berücksichtigung der Unsicherheit bei der Fehlwert-Schätzung aus.

Tabelle 3: Vor- und Nachteile gängiger Verfahren zum Umgang mit fehlenden Werten

Verfahren	**Vor-/Nachteile**	**Fazit**
Fallweiser Ausschluss	MCAR-Voraussetzung; erhebliche Stichprobenreduktion	Nur bei Daten mit geringer Fehlwertquote (< 5%), großen Stichproben, großen Effekten
Paarweiser Ausschluss	Effizienter als fallweiser Ausschluss; MCAR-Voraussetzung; unterschiedliche Datengrundlage für jedes Variablenpaar führt zu inkonsistenten Parameterschätzungen	Nie anwenden!
Ersetzung durch Mittelwerte	Einfachheit und augenscheinliche Plausibilität; MCAR-Voraussetzung; liefert auch bei MCAR verzerrte Parameterschätzungen	Nie anwenden!
Regressionsschätzungen	Ausnutzung der vorhandenen Information; nur MAR-Bedingung; überhöhte Korrelationsschätzung; Unterschätzung des Standardfehlers	Nur in seltenen Fällen nützlich
Maximum-Likelihood-Schätzungen (EM-Algorithmus)	Ausnutzung der gegebenen Information; effiziente und unverzerrte Parameterschätzung; nur MAR-Bedingung; Unterschätzung des Standardfehlers	Zur Parameterschätzung anwenden, zufallskritische Absicherung ergänzen
Multiple Imputationen	Ermöglicht valide statistische Entscheidungen; nur MAR-Bedingung; (zeitlich) äußerst aufwändig; kann aufgrund der gleichen Daten und der gleichen Methode zu unterschiedlichen Ergebnissen führen (Zufallskomponente)	Am besten in Verbindung mit ML-Verfahren anwenden

Übertragen auf die Perspektive der Ausfallmethodik lässt sich festhalten, dass „vollständig zufällig fehlende“ Daten (MCAR) am einfachsten zu handhaben sind, sich aber auch lediglich „zufällig fehlende“ Werte (MAR) mit entsprechenden statistischen Verfahren noch verhältnismäßig gut verarbeiten lassen. Fehlen Werte jedoch „nicht zufällig“ (MNAR), so stellt ihre Handhabung erheblich höhere methodische Anforderungen an entsprechende Verfahren (vgl. Tabelle 4). Für den MNAR-Fall zeichnen sich in jüngerer Zeit jedoch so genannte „Pattern-Mixture Models“ als aussichtsreiche Methode ab (vgl. Allison, 2002; Hedeker & Gibbons, 1997).

Tabelle 4: Handhabbarkeit der verschiedenen Mechanismen fehlender Werte

Ausfallmechanismus	**Handhabung**
MCAR	leichte statistische Handhabung
MAR	mit komplexen statistischen Verfahren handhabbar
OAR	sehr schwere Handhabung
MNAR	sehr schwere Handhabung

2.3. Determinanten der Güte von Fehlwert-Ersetzungen

Wie oben ausgeführt wurde, hängt die Güte der Ersetzung fehlender Werte maßgeblich von der zugrunde liegenden Systematik des Datenausfalls und der gewählten Ersetzungsmethodik ab. Neben diesen beiden Aspekten gibt es jedoch noch eine Vielzahl weiterer Faktoren, die die Güte der Fehlwert-Ersetzung ebenfalls nicht unerheblich beeinflussen können. Eine Auswahl der bedeutsamsten Einflussfaktoren soll daher im Folgenden dargestellt werden.

Ausfallmechanismus und Ersetzungsmethode

In Bezug auf den Ausfallmechanismus kann gemäß der Ausführungen in den vorangegangenen Abschnitten im Allgemeinen davon ausgegangen werden, dass Werte, die „vollständig zufällig fehlen“ (MCAR), besser geschätzt werden können als Wer-

te, die lediglich „zufällig fehlen“ (MAR), und diese wiederum besser als Werte, die gar „nicht zufällig fehlen“ (MNAR). Bezogen auf die verfügbaren Methoden zur Ersetzung fehlender Werte, verspricht das likelihood-basierte Verfahren der Multiplen Imputation bei „zufällig fehlenden“ Werten (MAR) gegenüber den einfachen Imputationsverfahren, wie dem EM-Algorithmus und der Regressionsschätzung, die besseren Schätzergebnisse.

Ausmaß fehlender Werte

An erster Stelle ist neben den den fehlenden Daten zugrunde liegenden Ausfallmechanismen und der gewählten Ersetzungsmethodik das *Ausmaß des Datenausfalls* zu berücksichtigen. Mit zunehmender Anzahl variablen- oder fallbezogen fehlender Werte steigt durch den damit einhergehenden Informationsverlust grundsätzlich das Risiko schlechterer Schätzergebnisse (Sinharay et al., 2001). Als grobe Richtgröße kann hier festgehalten werden, dass durch die Berücksichtigung von Variablen oder Fällen mit mehr als 30 Prozent fehlenden Werten zumeist mehr Unsicherheiten und Fehler erkauft werden, als substantielle Informationen für die weiterführenden statistischen Analysen gewonnen werden (Wirtz, 2004).

Zusammenhang zwischen berücksichtigten Variablen

Eine weitere Voraussetzung für die adäquate Ersetzung fehlender Werte besteht ganz grundsätzlich in der Verfügbarkeit geeigneter Kovariaten, die *essentielle Zusammenhänge* zu den Variablen aufweisen, deren fehlende Werte ersetzt werden sollen. Nur wenn solche Variablen im Datensatz vorhanden sind, die jeweils möglichst hoch mit dem fehlwertbehafteten Merkmal korrelieren, ist es möglich, die fehlenden Werte auf der Basis der Information aus den entsprechend vorliegenden Werten zu schätzen (Allison, 2002; Little & Rubin, 2002; Sinharay et al., 2001). Daher wird empfohlen, im Zuge der Datenerhebung möglichst immer auch solche Merkmale zu erfassen, die nicht unbedingt selbst für die weiterführenden Analysen benötigt werden, jedoch als Informationsbasis für die Ersetzung fehlender Werte in relevanten Zielvariablen benutzt werden können (sog. „Hilfsvariablen“, vgl. Collins et al., 2001).

Anzahl berücksichtigter Kovariaten

Die Berücksichtigung einer *hohen Anzahl von Kovariaten* in einem Modell zur Ersetzung fehlender Werte erhöht einerseits die Wahrscheinlichkeit, eine vorliegende MAR-Bedingung angemessen abzubilden, indem wichtige Gründe für das Vorliegen fehlender Werte nicht leichtfertig außer Acht gelassen werden. Andererseits wird durch die Einbeziehung einer größeren Anzahl von Parametern jedoch auch die Variabilität der beobachteten Zusammenhänge erhöht, was zur Überschätzung der wahren Varianz fehlender Werte führen kann. Bezüglich der optimalen Anzahl der bei der Ersetzung fehlender Werte einzubeziehenden Kovariaten finden sich in der Literatur dementsprechend uneinheitliche Befunde. In einigen Studien wurde gefunden, dass die Varianz der fehlwertbehafteten Variablen unter bestimmten Voraussetzungen durch den Einbezug einer höheren Anzahl von Kovariaten im Vergleich zur sparsamen Verwendung von Kovariaten zum Teil deutlich überschätzt wurde, was letztlich wiederum in einer Unterschätzung der Korrelationen zu anderen Variablen resultierte (z.B. Sinharay et al., 2001). In anderen Untersuchungen konnte hingegen gezeigt werden, dass der Einbezug einer möglichst großen Anzahl von Hilfsvariablen gegenüber der restriktiven Verwendung von Kovariaten im schlechtesten Falle zu vergleichbar guten Resultaten führt, die Schätzungsgüte im besten Falle jedoch erheblich verbessern kann (z.B. Collins et al., 2001).

Stichprobenumfang

Nicht zuletzt spielt aber auch die *Größe der vorliegenden Untersuchungsstichprobe* eine bedeutsame Rolle. Da fehlende Werte mit einem Informationsverlust einhergehen und jede Schätzung der Ausprägung fehlender Werte auf den verbleibenden Informationen basiert, muss sichergestellt sein, dass die nach Datenausfall vorhandenen Daten, selbst im Falle hoher Anteile fehlender Werte, noch die für die Fehlwertersetzung benötigten Informationen bereitstellen können. Wird eine Untersuchungsstichprobe jedoch von vorneherein sehr klein angelegt, so kann es passieren, dass die nach Datenausfall verbleibenden Daten selbst im Falle „vollständig zufällig fehlender“ Werte (MCAR) aufgrund zufallsbedingter Selektionseffekte nicht mehr repräsentativ für die untersuchte Grundgesamtheit sind und damit im Zuge des Ersetzungsprozesses zu verzerrten Schätzergebnissen führen. Dementsprechend sollte

der zu erwartende Datenausfall jeweils bereits im Vorfeld der Datenerhebung im Rahmen der Untersuchungsplanung bei der Kalkulation der benötigten Stichprobenumfänge berücksichtigt werden.

3. Fragestellung

Die Güte der verschiedenen verfügbaren Verfahren zum Umgang mit fehlenden Werten wurde bislang zumeist in Simulationsstudien anhand von fiktiven Datensätzen, die nach bestimmten Kriterien konstruiert wurden, überprüft und demonstriert. In derartigen Studien werden, neben den den fehlenden Werten zugrunde liegenden Ausfallmechanismen, üblicherweise auch die weiteren bedeutsamen Faktoren, wie etwa der Zusammenhang zwischen den zu berücksichtigenden Variablen, systematisch variiert. Solche Simulationsstudien können dementsprechend aber auch lediglich Hinweise auf die Ersetzungsgüte der einzelnen Verfahren unter streng kontrollierten Rahmenbedingungen geben. Über die Robustheit der verschiedenen Ersetzungsmethoden unter naturalistischen Bedingungen, in denen die Rahmenbedingungen nur eingeschränkt kontrollierbar und die Voraussetzungen für den Einsatz bestimmter Ersetzungsverfahren häufig nur bedingt gegeben sind, ist bislang nur wenig bekannt.

Eine besondere Anforderung an Verfahren zum Umgang mit fehlenden Werten stellen Längsschnittstudien, da hier häufig zu einzelnen Messzeitpunkten komplette Datensätze fehlen („*Unit-Nonresponse*“). Während sich die Ersetzung fehlender Werte im Falle der *Item-Nonresponse* auf die ansonsten aus derselben Erhebung zum jeweiligen Probanden vorliegenden Informationen stützen kann, müssen die fehlenden Werte im Falle der *Unit-Nonresponse* basierend auf den Angaben des jeweiligen Probanden aus früheren oder späteren Erhebungen geschätzt werden. Insofern wird hier die Unsicherheit der Schätzung durch eine zusätzliche zeitliche Komponente erhöht. Dies gilt insbesondere für 2-Punkt-Erhebungen (z.B. in gängigen Prä-Post-Untersuchungen), da hier fehlende Werte im Falle einer *Unit-Nonresponse* aus lediglich einer weiteren Messung geschätzt werden müssen. Im Rahmen von Prozessstudien mit einer Vielzahl von Erhebungszeitpunkten lassen sich fehlende Werte zu einem einzelnen Messzeitpunkt, insbesondere beim Vorliegen der MCAR-Bedingung, hingegen wiederum deutlich besser ersetzen. Diese Option wird in Longitudinal-Studien mit mehrfachen Verlaufserhebungen zum Teil sogar systematisch genutzt, indem personenbezogen auf einzelne a priori zufällig be-

stimmte Verlaufsmessungen verzichtet wird, um den Erhebungsaufwand zu minimieren (sog. „*Random Effects Models*“, vgl. Hedeker & Gibbons, 1997).

Auch und insbesondere im Kontext der einrichtungsvergleichenden Qualitätssicherung (vgl. Kapitel 1) finden sich Rahmenbedingungen, die erhebliche Anforderungen an die einzusetzenden Verfahren zum Umgang mit fehlenden Werten stellen können. Dies betrifft vor allem den Aspekt der eingeschränkten Homogenität von Rahmenbedingungen *zwischen* den untersuchten Einrichtungen. Die verschiedenen im vorangegangenen Kapitel dargestellten Faktoren, die die Güte der Fehlwertersetzung maßgeblich beeinflussen, können nämlich in der Praxis von Einrichtung zu Einrichtung erheblich voneinander abweichen.

So ist beispielsweise davon auszugehen, dass - je nach Einrichtungsphilosophie - einige Einrichtungen versuchen, möglichst alle behandelten Patienten in eine entsprechende Studie einzubeziehen, während in anderen Einrichtungen besonders schwer beeinträchtigte Patienten eher von der Untersuchung ausgeschlossen werden, um sie nicht zusätzlich zu belasten. In diesem Falle würden sich somit bereits a priori die *Quoten fehlender Werte* (vor allem in Form der sog. „*Unit-Nonresponse*“) sowie die den fehlenden Werten zugrunde liegenden *Ausfallmechanismen* zwischen den miteinander zu vergleichenden Einrichtungen unterscheiden. Weiterhin zeigt die Erfahrung, dass in einigen Einrichtungen besonders sorgfältig auf die Vollständigkeit der erhobenen Daten geachtet wird, während sich andere Einrichtungen aufgrund knapper bemessener personeller Ressourcen oder geringeren Interesses an der Untersuchung weniger engagiert mit der Datenerhebung beschäftigen, was letztlich wiederum zu weiteren Unterschieden bezüglich der *Anteile fehlender Werte* (in diesem Falle in Form der sog. „*Item-Nonresponse*“) führen dürfte. Außerdem ist es häufig gar nicht möglich, in allen untersuchten Einrichtungen identisch umfangreiche *Stichproben* zu erheben. Während in größeren Einrichtungen in kurzer Zeit große Stichprobenumfänge erreichbar sind, muss die Datenerhebung in kleineren Einrichtungen aufgrund zeitlicher Begrenzungen häufig auf deutlich kleinere Stichproben beschränkt bleiben. Vor diesem Hintergrund basiert die spätere Ersetzung fehlender Werte je nach Einrichtungsgröße in der Praxis häufig auf unterschiedlich umfangreichen Basisinformationen. Vor allem bei Längsschnittuntersuchungen ergibt sich eine weitere Varianzquelle zwischen den vergli-

chenen Einrichtungen durch die unterschiedliche einrichtungsbezogene Homogenität in Bezug auf die zeitlichen Verläufe. Da einzelne Einrichtungen zumeist in mehrere Abteilungen und Unterabteilungen (z.B. Stationen) untergliedert sind, in denen wiederum verschiedene Individuen an der Behandlung beteiligt sind, ist das Behandlungsergebnis innerhalb einer Einrichtung nicht nur durch die Ausgangsbedingungen eines Patienten, sondern grundsätzlich auch durch eine Vielzahl weiterer möglicher Einflussgrößen und deren Kombinationen determiniert. Je nachdem, wie sehr diese Einflussgrößen innerhalb einer Einrichtung, z.B. durch entsprechend strukturierte Behandlungskonzepte, aufeinander abgestimmt sind, können mehr oder weniger homogene Behandlungsverläufe resultieren. Homogenere Behandlungsverläufe würden sich wiederum in einem höheren Zusammenhang zwischen den patientenbezogenen Ausgangsbedingungen und dem erzielten Behandlungsergebnis widerspiegeln, was die Ersetzbarkeit fehlender Daten zum Behandlungsergebnis im Falle vorliegender Basisdaten begünstigen würde.

Insofern könnte es im Extremfall also vorkommen, dass eine Methode zur Ersetzung fehlender Werte, die sich für die Daten einer speziellen Einrichtung als optimal erweist, für die Daten einer anderen Einrichtung, die im Rahmen desselben Qualitätssicherungsprogramms erhoben wurden, gänzlich ungeeignet ist. Würden in so einem Falle die fehlenden Daten aller verglichenen Einrichtungen mit der gleichen Methode ersetzt, so könnte dies in erheblichen Verzerrungen der Ergebnisse zugunsten bzw. zulasten einzelner Einrichtungen resultieren.

Im Rahmen der vorliegenden Arbeit sollte daher anhand eines Modelldatensatzes mit empirischen Daten zur Ergebnisqualität im Bereich der stationären Rehabilitation von psychischen und psychosomatischen Erkrankungen exemplarisch überprüft werden, inwieweit die verschiedenen in Frage kommenden Verfahren zum Umgang mit fehlenden Werten unter den Bedingungen der Praxis der einrichtungsvergleichenden Qualitätssicherung dazu geeignet sind, dem Problem fehlender Werte gerecht zu werden. Da das Problem der *Unit-Nonresponse*, d.h. fallbezogen komplett fehlende Patientendaten zu einem oder mehreren Erhebungszeitpunkten, in diesem Kontext ein besonders häufiges Phänomen von hoher praktischer Relevanz darstellt, wurde der Schwerpunkt dieser Arbeit auf diesen Spezialfall fehlender Werte gelegt.

Dabei sollte einerseits untersucht werden, wie gut die in Kapitel 2.2 beschriebenen Imputationsverfahren unter verschiedenen Rahmenbedingungen in der Lage sind, fehlende Entlassungsdaten zu ersetzen. Darüber hinausgehend sollte überprüft werden, inwieweit die auf der Basis vervollständigter Datensätze durchgeführten Einrichtungsvergleiche tatsächlich in anderen Ergebnissen bezüglich der Bewertung der einzelnen Einrichtungen resultieren als dies bei Anwendung der einfachsten Methoden zum Umgang mit fehlenden Daten, nämlich dem listenweisen Fallausschluss, der Fall wäre. Ziel dieser Analysen ist dabei die Ableitung von Empfehlungen zum Einsatz bestimmter Verfahren zum Umgang mit fehlenden Werten unter bestimmten Voraussetzungen, wobei bei äquivalenter Performanz das unter den jeweiligen Rahmenbedingungen unaufwändigste Verfahren zu bevorzugen wäre.

Konkret sollen die folgenden Fragestellungen beantwortet werden:

1. **Unterscheiden sich verschiedene Imputationsverfahren unter naturalistischen Bedingungen bezüglich ihrer Güte der Schätzung fehlender Werte?**

   Hier wäre gemäß der Ausführungen in Kapitel 2.2 zu erwarten, dass die Fehlwertersetzung mittels *Multipler Imputationen* zu den besten Resultaten führt. Die Ersetzung fehlender Werte auf der Grundlage des *EM-Algorithmus* sollte zu besseren Schätzungen führen als die Fehlwertimputation über *Regressionsschätzungen*.

2. **Inwieweit beeinflussen verschiedene Rahmenbedingungen die Güte der Schätzung fehlender Werte?**

   Gemäß der Ausführungen in Kapitel 2.3 ist hier zu erwarten, dass die Fehlwertschätzung bei *zufällig fehlenden* Werten zu besseren Resultaten führt als bei *systematischen* Datenausfällen.

   Außerdem ist davon auszugehen, dass sich die Güte der Fehlwertersetzung mit zunehmenden *Anteilen variablen- wie fallbezogen fehlender Werte* verringert.

Des Weiteren sollte sich die Ersetzungsgüte mit zunehmender *Anzahl berücksichtigter Kovariaten* und zunehmend engerem *Zusammenhang* zwischen berücksichtigten Kovariaten und fehlwertbehafteten Kriterien erhöhen.

Letztlich ist in diesem Zusammenhang auch damit zu rechnen, dass die Fehlwertschätzung in kleineren *Stichproben* aufgrund der eingeschränkten Varianzinformation zu schlechteren Resultaten führt als in größeren Stichproben.

3. Führt die Anwendung von Imputationsverfahren letztlich zu anderen Resultaten bezüglich der Bewertung verglichener Einrichtungen als die Anwendung von Eliminierungsverfahren?

Aufgrund ihrer Bindung an die MCAR-Voraussetzung sollten *Eliminierungsverfahren* insbesondere im Falle systematischer Datenausfälle zu schlechteren Schätzungen führen als die verschiedenen *Imputationsverfahren* (vgl. Kap. 2.2). Dementsprechend müsste die Anwendung von Eliminierungsverfahren in solchen Fällen letztlich auch in verzerrten Resultaten bezüglich der Bewertung verglichener Einrichtungen resultieren. Als Folge der mit dem Einsatz von Eliminierungsverfahren einhergehenden Stichprobenreduktion könnten zudem bestehende Unterschiede zwischen den verglichenen Einrichtungen, die bei Anwendung von Imputationsverfahren nachzuweisen wären, aufgrund der eingeschränkten statistischen Power nicht mehr aufgedeckt werden.

Im Anschluss an die Untersuchung der ersten beiden Fragestellungen sollen die ermittelten Befunde auf die Praxis der einrichtungsvergleichenden Qualitätssicherung übertragen werden. Dabei soll anhand von Daten aus dem Qualitätssicherungsprogramm der Gesetzlichen Krankenversicherung für den Indikationsbereich *psychische und psychosomatische Erkrankungen* demonstriert werden, welche Implikationen die Anwendung eines angemessenen Modells zum Umgang mit vorliegenden fehlenden Werten mit sich bringt (Fragestellung 3).

4. Material und Methoden

Als empirische Grundlage für die geplanten Untersuchungen dient ein Datensatz zur Ergebnisqualität von stationärer Rehabilitation, der in der Pilotphase des von den Spitzenverbänden der Krankenkassen initiierten Projektes zur „Qualitätssicherung durch die gesetzlichen Krankenkassen in der Medizinischen Rehabilitation“ (QS-Reha®-Verfahren) im Indikationsbereich *psychische und psychosomatische Erkrankungen* in elf Rehabilitations-Fachkliniken für Psychosomatische Medizin und Psychotherapie erhoben wurde (vgl. Kapitel 4.1 und 4.2). Die untersuchte Stichprobe wird in Kapitel 4.3 beschrieben.

Nach einer grundlegenden Analyse fehlender Werte (vgl. Kapitel 4.4) bildet die systematische Untersuchung zur Angemessenheit der verschiedenen in Kapitel 2.2 vorgestellten Verfahren zur Ersetzung fehlender Werte den zentralen empirischen Teil der vorliegenden Arbeit. Da die Überprüfung der Güte von Fehlwertersetzungen die Kenntnis der tatsächlichen Ausprägung der fehlenden Werte voraussetzt, wurden die entsprechenden Analysen anhand eines vollständigen Teildatensatzes durchgeführt, in dem im Rahmen einer Simulationsstudie künstliche Datenausfälle erzeugt wurden (vgl. Kapitel 4.5). Ergänzend sollten im Rahmen dieser Simulationsstudie auch die Auswirkungen der Eliminierung von fehlenden Werten überprüft werden.

Die konkreten Hypothesen, die sich in Bezug auf die untersuchten Fragestellungen ableiten lassen, sind in Kapitel 4.6 aufgeführt.

Im Anschluss an die Simulationsstudie soll abschließend demonstriert werden, wie sich die Anwendung derjenigen Methodik zum Umgang mit fehlenden Werten, die sich als die angemessenste erwiesen hat, auf das Resultat eines Einrichtungsvergleichs und somit auf die Bewertung der einzelnen verglichenen Einrichtungen auswirkt (vgl. Kapitel 4.7).

4.1. Untersuchungsrahmen und -design

Das QS-Reha®-Verfahren wurde zunächst von Vertretern der Spitzenverbände und der Abteilung für Qualitätsmanagement und Sozialmedizin (AQMS) des Universitätsklinikums Freiburg für die somatischen Indikationsbereiche (Muskuloskeletale Erkrankungen, Kardiologie, Neurologie u.a.) entwickelt und später durch das Institut und Poliklinik für Medizinische Psychologie des Universitätsklinikums Hamburg-Eppendorf für die Indikationsfelder psychische, psychosomatische und Abhängigkeits-Erkrankungen adaptiert (Farin et al., 2003; Kawski & Koch, 2004). Im Rahmen des QS-Reha®-Verfahrens werden unter Verwendung verschiedener Erhebungsinstrumente und Messmethoden Informationen erfasst, die die Bewertung der untersuchten Rehabilitations-Einrichtungen hinsichtlich verschiedener Qualitätsaspekte ermöglichen sollen. Nach Abschluss der Erhebungen erhält jede Klinik ein umfassendes Qualitätsprofil, dass Hinweise auf Stärken und Schwächen der Einrichtung bereitstellt (Farin et al., 2003). Die Pilotstudie zum QS-Reha®-Verfahren für den Indikationsbereich *psychische und psychosomatische Erkrankungen* wurde im Jahre 2003 als *naturalistische Untersuchung* in elf Rehabilitations-Fachkliniken zur Behandlung von Patienten mit psychischen und psychosomatischen Störungen durchgeführt.

Im QS-Reha®-Verfahren werden grundsätzlich in jeder Einrichtung, die sich am Qualitätssicherungsprogramm beteiligt, Daten zur Struktur-, Prozess- und Ergebnisqualität sowie zur Patientenzufriedenheit und optional zur Mitarbeiterzufriedenheit erfasst.

Die in der vorliegenden Arbeit interessierenden *Daten zur Ergebnisqualität* wurden im Rahmen einer *Längsschnittstudie* mit drei Messzeitpunkten erhoben: Zu Beginn der stationären Behandlung (T0), am Ende des stationären Aufenthalts (T1) sowie in einer 6-Monats-Katamnese (T2) beantworteten die einbezogenen Patienten umfangreiche Fragebogenbatterien. Ergänzend bearbeiteten die jeweiligen Bezugstherapeuten entsprechende Fragebögen zu Beginn (T0) und zum Ende (T1) der stationären Rehabilitationsmaßnahme. Zum ersten Messzeitpunkt (T0) wurden neben dem Einsatz einer Reihe standardisierter Erhebungsinstrumente (vgl. Kapitel 4.2) außerdem diverse klinische, soziodemographische und sozialmedizinische Parame-

ter sowie die Therapieziele für die Rehabilitation erhoben. Bei Entlassung (T1) wurden wiederum die zu T0 erhobenen Parameter sowie zusätzlich weitere Merkmale der Behandlung sowie Nachsorgeempfehlungen oder die Zufriedenheit mit der Behandlung erfasst (vgl. Kapitel 4.2). Sechs Monate nach Abschluss der stationären Rehabilitation (T2) wurden die Patienten erneut zu den bereits zu T0 und T1 erfassten Parametern sowie zu diversen Aspekten der Nachsorge befragt. Zur Untersuchung der in Kapitel 3 formulierten Fragestellungen sollen im Rahmen der vorliegenden Arbeit jedoch nur die zu den ersten beiden Messzeitpunkten (T0 und T1) erhobenen Daten berücksichtigt werden.

In jeder der elf beteiligten Kliniken sollten konsekutive Stichproben von jeweils 200 Patienten, deren Rehabilitation nach Möglichkeit von den gesetzlichen Krankenkassen oder der gesetzlichen Rentenversicherung finanziert wurde, in die Studie eingeschlossen werden. In einem Informationsschreiben wurden die Patienten über die Studie aufgeklärt und um ihre Einwilligung zur Teilnahme an der Untersuchung gebeten. Patienten, die aufgrund sprachlicher Verständnisschwierigkeiten Probleme hatten, die Fragebögen auszufüllen, mussten von der Untersuchung ausgeschlossen werden. Für diejenigen Patienten, die nicht an der Studie teilnehmen konnten oder wollten, war jedoch ein entsprechender Therapeutenbogen vom behandelnden Therapeuten auszufüllen. Die Datenerhebung begann im Februar 2003, die letzten eingeschlossenen Patienten wurden im November 2003 aufgenommen. Insgesamt wurden in den elf beteiligten Fachkliniken im Untersuchungszeitraum N=2.386 Rehabilitanden behandelt, von denen n=2.161 an der Studie teilnahmen (vgl. die Stichprobenbeschreibung in Kapitel 4.3).

4.2. Instrumente und berücksichtigte Variablen

Das im Rahmen der Pilotstudie des QS-Reha®-Verfahrens für den Indikationsbereich psychischer und psychosomatischer Erkrankungen zur Erfassung der Ergebnisqualität eingesetzte Erhebungsinstrumentarium ist in Tabelle 5 wiedergegeben. Die in der vorliegenden Arbeit zur Beantwortung der in Kapitel 3 formulierten Fragestellungen berücksich-

tigten Variablen bzw. Variablengruppen sind dort *kursiv* gedruckt und sollen im Folgenden näher erläutert werden.

Zentrale Maße der Ergebnisqualität

Obgleich die Ziele psychotherapeutischer Behandlungen, anders als in vielen Bereichen der somatischen Medizin, selten eindeutig vorgegeben, sondern zumeist auch eng mit individuellen ethischen Werten wie Autonomie, Selbstentfaltung und Beziehungsfähigkeit verwoben sind, können die Reduktion einer vorhandenen Symptombelastung sowie die Steigerung der Lebensqualität doch als kleinster gemeinsamer Nenner bezüglich der Zielsetzung einer jeden Psychotherapie angesehen werden. Die im Verlauf einer Behandlung in diesen Bereichen erzielten Verbesserungen sind ein wichtiger Indikator auch für den längerfristigen Behandlungserfolg.

Aufgrund der Schwierigkeiten, die mit der Einschätzung des Behandlungserfolges durch die behandelnden Therapeuten selbst verbunden sind, welche damit ja indirekt auch ihre eigene Arbeit zu beurteilen haben, spielen die Patienteneinschätzungen für die objektive Beurteilung der Ergebnisqualität in Klinikvergleichen eine tragende Rolle.

Als zentrale Maße der Ergebnisqualität sollen daher in der vorliegenden Arbeit exemplarisch zwei patientenseitig erfasste Parameter, nämlich der „Globale Symptomschwere-Index“ (GSI) der *Symptom-Check-Liste SCL-14* (Harfst et al., 2002) sowie die „Psychische Summenskala“ (PSK) der *Health Survey Short Form SF-8* (Ware et al., 2000), verwendet werden.

- Die *Symptom-Check-Liste SCL-14* ist eine Kurzform der international weit verbreiteten Symptom-Check-Liste SCL-90-R (Franke, 1995; Franke, 2002), die von Harfst et al. (Harfst et al., 2002) an einer umfangreichen klinischen Stichprobe entwickelt und an gesunden und klinischen Stichproben validiert wurde. Die SCL-14 erfasst die psychopathologische Symptomatik von Patienten in den Bereichen Depressivität, Somatisierung und Phobische Angst. Zur Reliabilität und Validität der SCL-14 werden zufrieden stellende Befunde berichtet (Harfst et al., 2002). In der vorliegenden Arbeit soll der über alle 14 Items der SCL-14 bestimmte „Globale Symptomschwere-Index“ (GSI) als allgemeiner Marker für die psychische und somatoforme Beeinträchtigung der untersuchten Patienten berücksichtigt werden.

- Mit der *Health Survey Short Form SF-8* (Ware et al., 2000), einer von den amerikanischen Originalautoren entwickelten Kurzform des Fragenbogens zur gesundheitsbezogenen Lebensqualität SF-36 (Bullinger, 2000; Bullinger & Kirchberger, 1998; Ware, 2000), werden über zwei Summenskalen (Psychische und Körperliche Summenskala, PSK bzw. KSK) Aspekte des psychischen und körperlichen Befindens und deren Auswirkungen auf das berufliche und private Leben erfasst. Wie auch schon die SF-36 liegt die SF-8 inzwischen in zahlreichen Übersetzungen vor und ist international weit verbreitet. Ihre psychometrische Güte konnte in zahlreichen Studien belegt werden (vgl. Ware et al., 2000). In der vorliegenden Arbeit soll die über vier Items der SF-8 bestimmte „Psychische Summenskala" (PSK) als Maß für die psychische Gesundheit der untersuchten Patienten verwendet werden.

Hilfsvariablen zur Schätzung fehlender Ergebniswerte

Neben den zentralen Maßen der Ergebnisqualität wurden patienten- wie therapeutenseitig eine Reihe weiterer Variablen erhoben, die erfahrungsgemäß mit dem Behandlungsergebnis assoziiert sind und daher im Falle fehlender Ergebniswerte als Hilfsvariablen zur Fehlwertersetzung dienen sollten. Zur Operationalisierung der interessierenden Merkmale wurden diverse standardisierte und hinsichtlich ihrer psychometrischen Eigenschaften gut überprüfte Selbst- und Fremdbeurteilungsinstrumente eingesetzt. Ergänzend wurden außerdem verschiedene soziodemografische, sozialmedizinische und klinische Parameter erfasst. Die zur Erfassung der Hilfsvariablen eingesetzten Instrumente sollen im Folgenden näher dargestellt werden.

Patientenseitig kamen an standardisierten Verfahren neben SCL-14 und SF-8 der Fragebogen zur Lebenszufriedenheit (FLZ; Fahrenberg et al., 2000), die Allgemeine Depressionsskala (ADS-K; Hautzinger & Bailer, 1993; Hautzinger & Bailer, 2002) sowie das Inventar zur Erfassung Interpersonaler Probleme (IIP-D; Horowitz et al., 1994) zum Einsatz:

- Der *Fragebogen zur Lebenszufriedenheit FLZ* (Fahrenberg et al., 2000) erfasst mit acht Items die Zufriedenheit mit verschiedenen Lebensbereichen und erfragt zusätzlich ein Globalurteil zur allgemeinen Lebenszufriedenheit. Die acht Items lassen sich zu einem Gesamtwert zusammenfassen, der das Ausmaß allgemeiner Lebenszufriedenheit repräsentiert.

- Bei der *Allgemeinen Depressionsskala ADS* (Hautzinger & Bailer, 2002) handelt es sich um die deutsche Übersetzung der „Center for Epidemiological Studies Depression Scale“ (CES-D; Radloff, 1977), die für den Einsatz in großen epidemiologischen Studien entwickelt wurde. Die in der vorliegenden Untersuchung eingesetzte Kurzversion (ADS-K) enthält 15 Items, welche umfangreich die vorhandene depressive Symptomatik erfassen. Als Index der depressiven Belastungsintensität kann ein Summenwert gebildet werden.
- Die deutsche Fassung des *Inventars zur Erfassung Interpersonaler Probleme IIP-D* (Horowitz et al., 1994) umfasst 64 Items, in denen interpersonale Verhaltensweisen erfragt werden, die einem Patienten entweder schwer fallen können oder die ein Patient im Übermaß zeigen kann. Die Items des IIP-D lassen sich acht Skalen zuordnen, die wiederum die Oktanten des interpersonalen Circumplex-Modells von Leary (Leary, 1957) repräsentieren. Neben diesen Skalenwerten lässt sich ein Gesamtwert bestimmen, der das Ausmaß interpersonaler Schwierigkeiten insgesamt reflektiert.
- Zusätzlich zu den standardisierten Selbstbeurteilungsinstrumenten wurden den Patienten 18 weitere Items vorgegeben, in denen *Globalurteile zur Beeinträchtigung* in verschiedenen Bereichen (z.B. körperliche Beschwerden, seelische Beschwerden, Schmerzen usw.) abzugeben waren.

Im Bereich der *Soziodemografie* wurden unter anderem Angaben zu Geschlecht, Alter, Nationalität, Partnersituation, Schulbildung und beruflicher Situation erfasst.

Als *sozialmedizinische Parameter* wurden patientenseitig die Arbeitsunfähigkeits- bzw. Krankheitszeiten in den letzten sechs Monaten sowie ein etwaig bestehender Berentungswunsch erhoben.

Auf Therapeutenseite kamen zwei standardisierte Verfahren, nämlich eine Fremdbeurteilungsversion der SF-8 (SF-8-F; Schulz et al., in Vorbereitung) sowie die deutsche Version der „Health of the Nation Outcome Scales“ (HoNOS-D; Andreas, 2005; Andreas et al., in press) zum Einsatz:

- Die *Health Survey Short Form SF-8-F* wurde als ergänzendes Fremdeinschätzungsverfahren zur therapeutenseitigen Erfassung von Beeinträchtigungen der gesundheitsbe-

zogenen Lebensqualität eingesetzt. Die SF-8-F ist eine Adaptation des Selbsteinschätzungsbogens SF-8 (s. o.) an die Erfordernisse eines Fremdeinschätzungsinstruments und lehnt sich an das bestehende Fremdeinschätzungsverfahren der SF-36 an (Schulz et al., in Vorbereitung). Wie bei der Selbstbeurteilungsversion lassen sich auch für die SF-8-F eine psychische und eine körperliche Summenskala bestimmen. Der Zusammenhang zwischen Selbst- und Fremdeinschätzung der gesundheitsbezogenen Lebensqualität liegt im Bereich einer großen Effektstärke (Schulz et al., in Druck).

- Als weiteres Fremdeinschätzungsverfahren wurden die international weit verbreiteten *Health of the Nation Outcome Scales* (*HoNOS*) eingesetzt (Wing et al., 1998), die der Einschätzung des Beeinträchtigungsschweregrades von Patienten mit psychischen Störungen dienen und u.a. in Großbritannien und Australien zur Evaluation und Qualitätssicherung der Behandlung von Patienten mit psychischen Störungen in der klinischen Routine verwendet werden (Audin et al., 2001). Bei der vorliegenden deutschen Version (HoNOS-D) handelt es sich um die durch die englischen Originalautoren autorisierte Übersetzung (Andreas, 2005; Andreas et al., in press).

Ergänzend wurden therapeutenseitig verschiedene *klinische, soziodemografische und sozialmedizinische Variablen* erhoben. Diese umfassen u.a. Angaben zur Diagnose (nach ICD-10), zur Chronifizierung, zur Behandlungs-Motivation, zum Renten-Status sowie zur Kostenträgerschaft. Außerdem wurde zu Behandlungsende die Dauer der Rehabilitationsmaßnahme erfasst.

In Abbildung 3 sind die einzelnen Variablengruppen nochmals im Überblick nach ihrer Funktion als Maß der Ergebnisqualität bzw. als mögliche Hilfsvariablen bei der Ersetzung fehlender Daten zum Behandlungsergebnis aufgeführt.

Tabelle 5: Instrumentarium zur Erfassung der Ergebnisqualität in der Pilotphase des QS-Reha®-Verfahrens (Indikationsbereich psychische/ psychosomatische Erkrankungen)

	Erfasste Merkmale	**Instrumente**	**T0 Reha-Beginn**	**T1 Reha-Ende**
PATIENT	*Soziodemografie*	Eigenentwicklung	✓	
	Allgemeine Beeinträchtigungen	Globalurteile, Eigenentwicklung	✓	✓
	Allgemeine Lebenszufriedenheit	Fragebogen zur Lebenszufriedenheit (FLZ)	✓	✓
	Psychische und körperliche Beschwerden	Symptom-Check-Liste, Kurzform 14 (SCL-14)	✓	✓
	Depressive Symptomatik	Allgemeine Depressionsskala (ADS-K)	✓	✓
	Gesundheitsbezogene Lebensqualität	Health Survey Short Form, Kurzform 8, Selbstbeurteilung (SF-8)	✓	✓
	Interpersonelle Probleme	Inventar zur Erfassung interpersonaler Probleme (IIP-D)	✓	✓
	Konsum von Medikamenten, Genussmitteln und Drogen	Eigenentwicklung	✓	✓
	Behandlungsziele	Profil Psychotherapeutischer Zielsetzungen, Patientenversion (PPZ-P)	✓	✓
	Inanspruchnahme medizinischer Leistungen und Sozialmedizin	Eigenentwicklung	✓	
	Behandlungsprozess, Nachsorgevorbereitung und Sozialmedizin	Eigenentwicklung		✓
THERAPEUT	*Soziodemografie und Sozialmedizin*	Eigenentwicklung	✓	
	Klinische Daten	Eigenentwicklung	✓	
	Gesundheitsbezogene Lebensqualität	Health Survey Short Form, Kurzform 8, Fremdbeurteilung (SF-8-F)	✓	✓
	Beeinträchtigungsschwere	Health of the Nation Outcome Scales (HoNOS-D)	✓	✓
	Allgemeine Angaben zur Behandlung	Eigenentwicklung		✓
	Behandlungsziele	Profil Psychotherapeutischer Zielsetzungen, Therapeutenversion (PPZ-T)	✓	✓

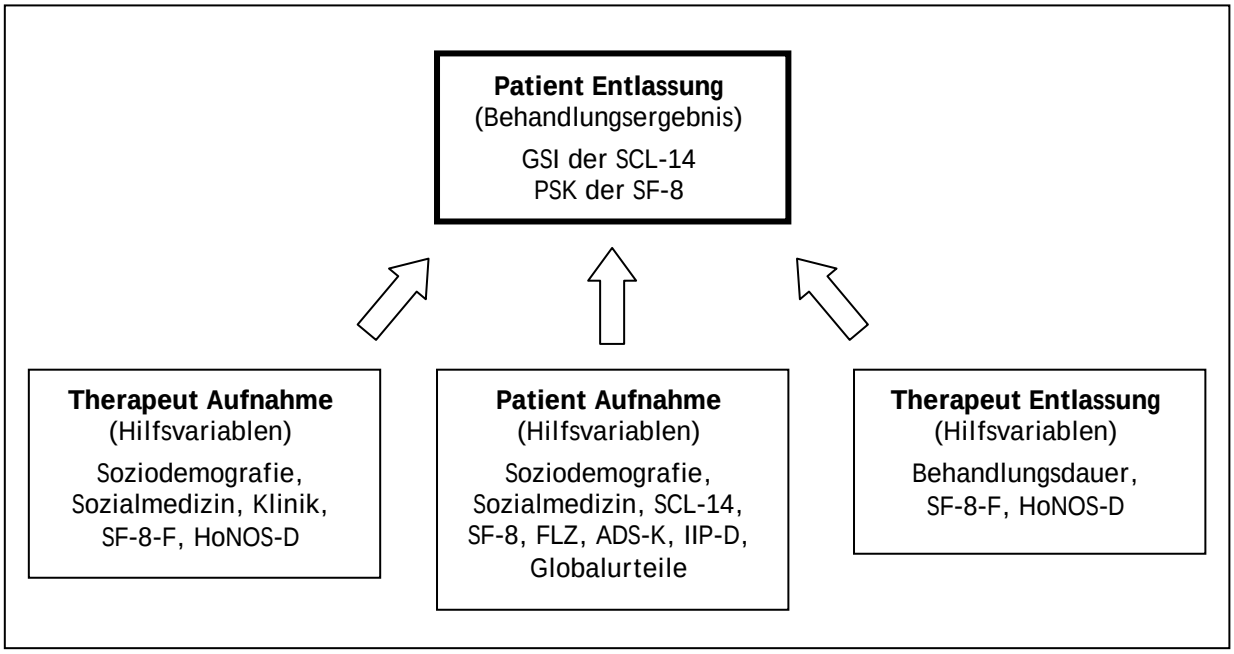

Abbildung 3: Ergebnismaße und Hilfsvariablen zur Schätzung fehlender Ergebniswerte

4.3. Stichprobe

Im Untersuchungszeitraum wurden in den elf an der Studie beteiligten Fachkliniken insgesamt N=2.386 Patienten behandelt, die grundsätzlich für eine Teilnahme an der Untersuchung in Frage gekommen wären. Von diesen beteiligten sich n=2.161 Patienten an der Studie, was einer Teilnehmerquote von 91 Prozent entspricht. Zwischen den einzelnen Kliniken schwanken die Teilnahmequoten zwischen 85 und 100 Prozent. Da nur für die Studienteilnehmer entsprechend vollständige Informationen vorliegen, beziehen sich die Angaben zur Charakterisierung der Untersuchungsstichprobe in Tabelle 6 auf diese 2.161 Patienten. Die Stichproben der teilnehmenden Patienten schwanken in den einzelnen Kliniken zwischen n=72 und n=250.

Das mittlere Alter der Studienteilnehmer liegt bei 42 Jahren, die Altersspanne reicht von 14 bis 98 Jahren. 77 Prozent der Studienteilnehmer waren Frauen, wobei der Anteil weiblicher Patienten in den einzelnen Kliniken zwischen 62 und 92 Prozent schwankt.

Jeweils knapp ein Drittel der untersuchten Patienten haben ihre Schulausbildung mit einem Hauptschulabschluss, der mittleren Reife oder der (Fach-)Hochschulreife abgeschlossen. Innerhalb der einzelnen Kliniken sind bezüglich der Schulbildung der behandelten Patienten jedoch erhebliche Unterschiede zu beobachten: So schwankt beispielsweise der Anteil von Patienten, die über die (Fach-) Hochschulreife verfügen je nach Klinik zwischen 13 und 50 Prozent. Etwa die Hälfte der Patienten war zum Zeitpunkt der Untersuchung berufstätig, allerdings finden sich auch hier erhebliche klinikspezifische Abweichungen (Spannweite: 29-66%). Mehr als die Hälfte der Patienten befanden sich zum Zeitpunkt der stationären Rehabilitation in einer festen Partnerschaft.

Am häufigsten wurden depressive Störungen als Haupterkrankung diagnostiziert, gefolgt von Belastungs- bzw. Anpassungsstörungen und Essstörungen. Weitere häufig gestellte Erstdiagnosen waren Angststörungen sowie somatoforme Störungen. Je nach Spezialisierung der einzelnen Kliniken finden sich bezüglich der Verteilung der behandelten Störungsbilder jedoch auch deutliche einrichtungsspezifische Besonderheiten. Das erste Auftreten der psychischen Erkrankung lag im Regelfall mehr als ein Jahr zurück, bei einem Viertel der Patienten war die Erkrankung bereits mehr als zehn Jahre chronifiziert (Median 2-5 Jahre; vgl. Tabelle 6).

Tabelle 6: Soziodemografische und klinische Stichprobencharakteristika*

	Kliniken											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Gesamt**
Stichprobenumfang												
N gesamt	217	283	231	273	171	272	168	211	232	244	84	2.386
N teilgenommen	217	250	210	234	148	247	166	200	196	221	72	2.161
% teilgenommen	*100,0*	*88,3*	*90,9*	*85,7*	*86,5*	*90,8*	*98,8*	*94,8*	*84,5*	*90,6*	*85,7*	*90,6*
Alter												
Mittelwert	23,9	41,3	40,7	51,5	45,0	46,4	36,7	43,2	44,8	47,8	44,6	42,2
Standardabweichung	*9,4*	*13,3*	*10,6*	*15,8*	*14,3*	*11,4*	*11,7*	*10,4*	*9,6*	*9,8*	*12,0*	*14,0*
Geschlecht												
% weiblich	91,6	68,1	79,9	87,2	77,4	78,0	71,0	62,4	68,9	72,3	80,3	76,7
Schulbildung												
% Hauptschule	12,1	52,3	21,3	32,6	29,9	36,9	17,8	28,0	24,9	48,3	8,7	30,4
% Realschule	24,7	25,5	35,3	29,0	32,6	38,9	24,4	43,0	37,6	25,6	37,7	31,6
% (Fach-) Abitur	33,5	13,2	34,8	27,1	27,1	20,5	48,9	26,2	31,7	17,4	49,3	27,9
% sonstige	29,7	9,0	8,6	11,3	10,4	3,7	8,9	2,8	5,8	8,7	4,3	10,1
Erwerbssituation												
% berufstätig	29,2	49,8	51,0	19,7	42,9	65,7	46,2	57,0	65,2	61,7	53,6	48,9
% arbeitslos	10,5	15,2	15,5	5,5	12,9	8,7	14,6	25,2	17,4	17,9	4,3	13,2
% Hausfrau/-mann	5,8	10,8	15,0	7,3	9,3	12,8	11,5	3,7	4,3	4,3	14,5	9,1
% Alters-/EU-Rente	3,5	9,5	4,0	56,4	15,7	5,0	10,0	0,9	6,5	6,8	20,3	13,2
% sonstige	51,0	14,7	14,5	11,1	19,2	7,8	17,7	13,2	6,6	9,3	7,3	15,6

(Fortsetzung siehe nächste Seite)

Tabelle 6: Soziodemografische und klinische Stichprobencharakteristika (Fortsetzung)*

	Kliniken											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Gesamt**
Partnersituation												
% langfristig ohne Partner (> 1 Jahr)	35,9	25,2	37,1	31,2	32,8	17,7	26,8	37,0	24,7	16,1	19,4	27,7
% kurzfristig ohne Partner (< 1 Jahr)	22,1	13,6	7,5	15,1	9,6	11,8	17,3	10,0	16,7	9,0	14,9	13,6
% wechselnde Partner	4,6	2,4	,5	1,0	1,6	1,4	2,4	1,0	2,3	1,9	4,5	2,1
% fester Ehepartner	10,8	40,3	46,2	41,7	40,8	52,7	32,3	36,0	44,3	56,8	47,8	40,7
% fester Partner (nicht Ehepartner)	26,7	18,4	8,6	11,1	15,2	16,4	21,3	16,0	12,1	16,1	13,4	16,0
Hauptdiagnose n. ICD-10												
% Depressive Störung (F32-33, F34.1)	0,0	43,6	48,1	41,5	31,8	40,5	44,6	32,5	46,9	33,9	63,9	37,3
% Essstörung (F50)	94,0	17,2	4,3	6,4	7,4	1,2	8,4	,0	1,0	1,4	5,6	14,3
% Belastungs-/Anpassungsstörung (F43)	0,0	12,8	16,7	19,7	10,8	21,1	14,5	10,5	14,3	29,0	15,3	15,2
% Angststörung (F40-41)	0,0	9,6	10,0	8,1	5,4	9,7	6,6	3,5	21,4	16,7	5,6	9,1
% Somatoforme Störung (F45)	0,0	5,2	3,8	7,3	2,0	8,1	2,4	2,5	6,1	7,2	4,2	4,7
% sonstige Störungen	6,0	11,6	17,1	17,1	42,6	19,4	23,5	51,0	10,2	11,8	5,5	19,4
Chronifizierung												
% < 1 Jahr	1,9	9,3	14,0	9,0	11,4	19,1	10,9	7,3	10,6	14,5	15,3	11,0
% 1-2 Jahre	21,0	21,4	18,8	14,2	21,4	27,5	13,1	18,0	18,6	28,5	26,4	20,8
% 2-5 Jahre	27,6	20,2	23,2	21,0	20,0	16,5	29,9	19,7	23,4	23,1	18,1	22,0
% 5-10 Jahre	26,2	19,8	15,5	20,2	16,4	12,3	14,6	18,5	16,5	11,8	19,4	17,4
% > 10 Jahre	22,4	25,0	26,6	34,3	27,9	22,5	29,9	36,0	25,0	21,3	18,1	26,5

* Angaben beziehen sich jeweils auf die Stichprobe der teilnehmenden Patienten

4.4. Analyse fehlender Werte

Im Rahmen der Missing-Data-Analysen wurden zunächst die absoluten und relativen Häufigkeiten sowohl fall- als auch itembezogen fehlender Werte in den in den verschiedenen beteiligten Einrichtungen erhobenen Datensätzen ermittelt, um erste Hinweise auf das Ausmaß und die Verteilung fehlender Werte in den einzelnen Erhebungseinheiten (Patient/Therapeut x Aufnahme/Entlassung) zu erhalten.

Außerdem wurde analysiert, ob die fehlenden Werte in einem systematischen Zusammenhang zu den Werten anderer Variablen auftreten oder ob umschriebene Muster fehlender Werte gehäuft vorliegen. Hierzu wurde für jedes erhobene Merkmal eine neue Indikatorvariable gebildet, in der das Vorliegen bzw. Fehlen des entsprechenden Datums kodiert wurde (d.h. fehlende Werte wurden mit „0“, vorliegende Werte mit „1“ kodiert). Für die beiden Stufen dieser Indikatorvariablen (0, 1) konnte dann über die Berechnung von Korrelationskoeffizienten überprüft werden, ob die vorhandenen Angaben in allen anderen Variablen je nach Ausprägungen der Indikatorvariablen differieren. Über die Korrelation zwischen den verschiedenen Indikatorvariablen konnte zudem untersucht werden, inwieweit sich Muster fehlender Werte nachweisen lassen: Höhere Korrelationen zwischen einer Indikatorvariable X und einer Indikatorvariable Y deuten darauf hin, dass tendenziell immer, wenn Variable X nicht beantwortet wurde, gleichzeitig auch die Angaben zu Variable Y fehlen. Aus den Analysen zur Fehlwertsystematik sollten sich Hinweise auf die zugrunde liegenden Ausfallmechanismen ableiten lassen, die wiederum, wie z.B. im Falle einer nicht gegebenen MCAR-Bedingung, erhebliche Konsequenzen in Bezug auf die Anwendbarkeit der verschiedenen Verfahren zum Umgang mit den fehlenden Werten haben können.

Abschließend wurde überprüft, inwieweit die zu berücksichtigenden Fallzahlen in Abhängigkeit verschiedener Kriterien zur Vollständigkeit vorhandener Daten (z.B. 70% fallbezogen gültige Werte) variieren, um die Reduktion der verwertbaren Stichproben und den damit verbundenen Informationsverlust unter bestimmten Rahmenbedingungen abschätzen zu können.

4.5. Simulationsstudie zum Umgang mit fehlenden Werten

Um die Angemessenheit eines Umgangs mit fehlenden Werten, wie er in der aktuellen (Forschungs-) Praxis selbstverständlich umgesetzt wird, empirisch zu überprüfen, wurde eine Simulationsstudie realisiert. Im Rahmen dieser Simulationsstudie sollte beispielhaft untersucht werden, inwieweit sich fehlende Ergebnisdaten, also fehlende Patientenangaben zum Entlassungszeitpunkt, unter verschiedenen Rahmenbedingungen adäquat ersetzen lassen. Hierzu wurden die verschiedenen in Kapitel 2.3 dargestellten Faktoren, die die Güte einer Fehlwertersetzung beeinflussen können, systematisch variiert (vgl. Tabelle 8). Ergänzend sollte überprüft werden, ob im Falle eines Fallausschlusses der Nachweis vergleichbarer Basismerkmale von Responder- und Dropout-Stichproben genügt, um damit die Repräsentativität des Behandlungsergebnisses sicherzustellen Das methodische Vorgehen im Rahmen der Simulationsstudie soll im Folgenden detailliert beschrieben werden.

Datenbasis für die Simulationsstudie

Die Beurteilung der Güte von Fehlwertersetzungen wie auch der Angemessenheit von sonstigen Varianten des Umgangs mit fehlenden Werten (z.B. listenweiser Fallausschluss) ist an die Kenntnis der tatsächlichen Ausprägung der fehlenden Werte gebunden. Daher sollten die entsprechenden Analysen auf der Grundlage eines vollständigen Teildatensatzes des vorliegenden Gesamtdatensatzes durchgeführt werden (vgl. Kapitel 4.1 und 4.2).

Da sich die verwertbare Stichprobe bei ausschließlicher Berücksichtigung von Fällen mit 100prozentig vollständigen Daten jedoch so stark reduzieren würde, dass sie für die geplanten Analysen nicht mehr zu verwerten wäre (n=378; s. Kapitel 5.1, Tabelle 10), wurde ein etwas toleranteres, wenngleich immer noch verhältnismäßig konservatives Vollständigkeitskriterium angelegt. Berücksichtigt wurden nämlich nur diejenigen Fälle, zu denen in jedem einzelnen der für die späteren Analysen benötigten Instrumente jeweils mindestens 80 Prozent gültiger Werte vorlagen (vgl. die Einschlusskriterien in Tabelle 7). In der resultierenden Stichprobe (n=1.248; vgl. Kapitel 5.1, Tabelle 10) fanden sich durchschnittlich jeweils lediglich 0,5 Prozent fehlende Therapeutenangaben zu

T0 (Range 0,0 – 11,1 %) und T1 (Range 0,0 – 10,0 %) sowie Patientenangaben zu T1 (Range 0,0 – 13,6 %). Die durchschnittliche Fehlwertquote bezüglich der Patientenangaben zu T0 betrug 0,8 Prozent (Range 0,0 – 8,1 %).

Die einzelnen fehlenden Werte wurden, jeweils getrennt nach Erhebungsmodulen und Kliniken, mittels der EM-Methode ersetzt, um so einen Ausgangsdatensatz mit 100prozentig vollständigen Daten zu generieren, der als Grundlage für die Simulationsstudie dienen sollte.

Tabelle 7: Kriterien für den Einschluss von Fällen in die Simulationsstichprobe

Erhebungsmodul	**Berücksichtigte Instrumente***	**Anzahl Variablen**	**max. tolerierte Anzahl fehlender Werte**
Patient Aufnahme (T0)	FLZ	9	1
	SCL-14	14	2
	ADS-K	15	3
	SF-8	8	1
	IIP-64	64	12
	Globalurteile	18	3
	sonstige Angaben	8	1
Patient Entlassung (T1)	SCL-14	14	2
	SF-8	8	1
Therapeut Aufnahme (T0)	HoNOS-D	12	2
	SF-8-F	8	1
	sonstige Angaben	7	1
Therapeut Entlassung (T1)	HoNOS-D	12	2
	SF-8-F	8	1
	sonstige Angaben	1	0

* Zu den berücksichtigten Instrumenten vgl. Kapitel 4.2 und Tabelle 5

Simulation fehlender Werte

Zur Simulation fehlender Werte wurden im vollständigen Simulationsdatensatz jeweils für nach verschiedenen Kriterien ausgewählte Fälle die Patientenangaben zum Entlassungszeitpunkt eliminiert. Der Entlassungszeitpunkt wurde exemplarisch ausgewählt, da den Entlassungsdaten einerseits eine zentrale Bedeutung zukommt, wenn es darum geht, das Ergebnis einer Behandlung zu beurteilen, und zugleich in der Forschungsrealität – verglichen mit den Aufnahmeerhebungen – zur Entlassung die größte Wahrscheinlichkeit für eine Unit-Nonresponse besteht (z.B. aufgrund von Therapieabbrüchen, Verlegungen usw.). Die Eliminierung der patientenbezogenen Entlassungsdaten erfolgte jeweils getrennt für die einzelnen Einrichtungen. Die Kriterien, nach denen die Dropoutfälle ausgewählt wurden, bezogen sich zum einen auf den Ausfallmechanismus, zum anderen auf die Anzahl fehlender Werte.

1. *Simulation unterschiedlicher Ausfallmechanismen*

Realisiert wurden zwei Ausfallmechanismen, nämlich *vollständig zufällig fehlende Werte (MCAR)* und *systematisch fehlende Werte (MAR)*. In der *MCAR-Bedingung* wurden die Fälle, für die die Patientenangaben zur Entlassung eliminiert werden sollten („Dropoutfälle"), vollkommen zufällig ausgewählt. In der *MAR-Bedingung* wurden hingegen solche Fälle ausgewählt, deren Symptombelastung sich im Behandlungsverlauf nicht signifikant gebessert hatte. Dieses Auswahlkriterium sollte der Annahme Rechnung tragen, dass in der empirischen Praxis tendenziell eher solche Patienten aus Studien ausscheiden, deren Behandlungsverlauf weniger erfolgreich ausfiel (z.B. aus motivationalen Gründen oder aufgrund organisatorischer Gegebenheiten wie Verlegungen, Therapieabbrüchen o.ä.). Operationalisiert wurde das Kriterium fehlender Symptombesserung über den „Reliable Change Index" (RCI), einem Maß für statistisch signifikante Veränderung in der Symptom-Checkliste (SCL), das ursprünglich für die 90-Item-Fassung SCL-90-R bestimmt wurde, in der vorliegenden Arbeit jedoch auf die SCL-14 angewendet wurde (Schauenburg & Strack, 1998; Schauenburg & Strack, 1999). Nach dem RCI-Kriterium wurden diejenigen Patienten identifiziert, deren Verbesserung auf dem Globalen Symptomschwereindex (GSI) der SCL-14 von der Aufnahme bis zur Entlassung weniger als 0,43 Einheiten betrug. Lagen in einer Klinik mehr Fälle ohne Symptomverbesserung vor, als im Rahmen der

Dropoutsimulation benötigt wurden (s.u.), so wurden die Dropoutfälle zufällig aus den in Frage kommenden Fällen ausgewählt. Lagen in einer Klinik weniger Fälle ohne Symptomverbesserung vor, als im Rahmen der Dropoutsimulation ausgewählt werden sollten, so wurden die zusätzlich benötigten Fälle zufällig aus der jeweiligen Stichprobe von Patienten mit positiver Symptomentwicklung gezogen.

2. *Simulation unterschiedlicher Umfänge fehlender Werte*

   Entsprechend der im Originaldatensatz zu beobachtenden Fehlwertquoten (vgl. Kapitel 5.1, Tabelle 9) wurden *(itembezogene) Datenausfälle im Umfang von 10, 20, 30, 40 und 50 Prozent* generiert. Dabei wurde jeweils in Abhängigkeit der einrichtungsbezogenen Stichprobengröße eine entsprechende Anzahl von Dropoutfällen zufällig aus der jeweiligen Substichprobe derjenigen Fälle gezogen, die gemäß den oben beschriebenen Kriterien zum zufälligen oder systematischen Datenausfall in Frage kamen. Für die identifizierten Dropoutfälle wurden dann jeweils die patientenbezogenen Entlassungsdaten vollständig eliminiert.

   Während durch dieses Vorgehen zunächst nur die Anteile itembezogen fehlender Werte direkt variiert wurden, sollte der Aspekt *fallbezogen fehlender Werte* zumindest ansatzweise durch eine unterschiedliche Festlegung der zu ersetzenden Variablen berücksichtigt werden. Hierzu sollten zum einen die einzelnen Items der Ergebnisskalen (GSI der SCL-14, PSK der SF-8), zum anderen aber auch die Skalenwerte selbst geschätzt werden. Dementsprechend beträgt die Anzahl fallbezogen fehlender Werte je nach Ersetzungsvariante 14 (Einzelitems des GSI der SCL-14), 4 (Einzelitems der PSK der SF-8) oder 1 (Skalenwerte GSI oder PSK).

Vorgeschaltete Analysen zum Resultat der Dropoutsimulation

Im Vorwege der Fehlwertersetzungen sollte zunächst für die verschiedenen Dropoutkonstellationen (MCAR/MAR x 10/20/30/40/50% Dropout) untersucht werden, welche Konsequenzen die simulierten Fehlwerte auf die Ausprägung der Ergebnisvariablen in den resultierenden Datensätzen zeigt. Dabei sollte insbesondere überprüft werden, unter welchen Dropout-Bedingungen sich signifikante Abweichungen zu den Verteilungsei-

genschaften der Originaldaten ergeben haben. Die so gewonnenen Informationen geben Hinweise darauf, wie gut die wahren einrichtungsbezogenen Ergebnisse bei Anwendung der Methode des listenweisen Fallausschlusses, also der einfachsten Variante der Eliminierung von Fällen mit fehlenden Werten, abgebildet würden. Analog hierzu wurde außerdem anhand der entsprechenden Aufnahmewerte im GSI der SCL-14 und der PSK der SF-8 überprüft, ob die zufällige respektive systematische Auswahl von Dropout-Patienten eventuell auch auf dieser Ebene in Unterschieden zwischen den Fällen, deren Entlassungsdaten eliminiert wurden, und den Fällen, deren Entlassungsdaten beibehalten wurden, resultiert. Eine Korrespondenz zwischen den Aufnahme- und Entlassungsbezogenen Abweichungs-Befunden wäre als Anhaltspunkt für die Angemessenheit von Dropoutanalysen zur Sicherstellung der Repräsentativität ermittelter Ergebnisse zu bewerten.

Ersetzung fehlender Werte

Im Rahmen der Ersetzung der systematisch generierten Fehlwerte wurden zum einen die Ersetzungsmethoden, zum anderen die berücksichtigten Kovariaten variiert.

1. Überprüfte Methoden der Fehlwertersetzung

In der Simulationsstudie wurden drei verschiedene Methoden zur Ersetzung fehlender Werte eingesetzt: Die einfachen Imputationsverfahren der *Regressionsschätzung (RA)* und *EM-Ersetzung (EM)* wurden über das „Missing Values Analysis“- (MVA) Modul des Statistikpaketes SPSS 12.0 (Norusis, 2004) realisiert. Die Fehlwertersetzung durch *Multiple Imputationen (MI)* erfolgte über die Prozedur „ICE“, die in der Statistiksoftware STATA 9.0 (Stata Corporation, 2005) verfügbar ist.

Im Rahmen der *RA* werden in SPSS fehlende Werte mittels mehrfacher linearer Regression geschätzt. Die Regressionsschätzer werden dabei jeweils um eine Zufallskomponente ergänzt, wobei die Fehlerterme zufällig aus den beobachteten Residuen vollständiger Fälle ausgewählt und den Regressionsschätzungen hinzugefügt werden.

Bei der *EM* werden in SPSS zunächst die Mittelwerte, die Kovarianzmatrix und die Korrelation der Variablen mit fehlenden Werten unter Verwendung eines iterativen Prozesses unter Annahme der Normalverteilung geschätzt. Die maximale Anzahl der Iterationen wurde dabei jeweils auf 25 festgelegt. Fehlende Werte werden schließlich durch die in der EM-Methode geschätzten Werte ersetzt. Zu den methodischen Grundlagen der Anwendung des EM-Algorithmus sei auf die entsprechende Darstellung in Kapitel 2.2 verwiesen.

Die zur *MI* verwendete Prozedur ICE wurde von Royston (2005) als Add-on für STATA entwickelt. Da Multiple Imputationen grundsätzlich kein einheitliches Verfahren darstellen, sondern auf verschiedensten methodischen Zugängen basieren können, soll die ICE-Prozedur im Folgenden etwas ausführlicher erläutert werden. ICE realisiert im Wesentlichen zwei Analyseschritte: In einem ersten Schritt werden die fehlenden Werte einer Variablen in einem Unterprogramm namens „*uvis*“ anhand der Informationen aus den anderen verfügbaren Prädiktorvariablen regressionsanalytisch geschätzt, wobei fehlende Werte in den Kovariaten zunächst durch zufällig ausgewählte Ausprägungen anderer Fälle in derselben Variablen aufgefüllt werden. Im zweiten Schritt werden dann nach einem als „regression switching“ bezeichneten Analyseschema nach und nach alle weiteren Variablen mit fehlenden Werten mittels *uvis* anhand der jeweils vorhandenen Kovariaten ersetzt. Indem ICE *uvis* immer wieder ausführt, wird der Imputationsschritt also für alle fehlwertbehafteten Variablen wiederholt durchlaufen (sog. „cycling“), was letztlich in einem optimierten Ersetzungsergebnis resultiert. In der vorliegenden Arbeit wurden jeweils zehn solche „cycles“ durchlaufen. Der gesamte ICE-Prozess wurde für jedes simulierte Ersetzungsszenarium zehnmal wiederholt, so dass sich jeweils m=10 verschiedene Datensätze mit imputierten Werten ergaben. Gemäß der Logik der Multiplen Imputation sollte jeder dieser Datensätze als Grundlage für weiterführende Analysen dienen, deren Resultate dann abschließend integriert werden könnten. Um das Ersetzungsresultat der *MI* jedoch direkt mit den Imputationsergebnissen von *RA* und *EM* vergleichen zu können wurden die zehn aus der *MI* resultierenden Datensätze jeweils im Anschluss an die Fehlwertersetzung durch fallbezogene Mittelwertsbildung zu einem einzelnen Ergebnisdatensatz aggregiert.

2. Berücksichtigte Kovariaten

Als Grundlage für die Schätzung fehlender Entlassungswerte stehen grundsätzlich drei verschiedene *Sets von Kovariaten* zur Verfügung: Die vorliegenden Patientenangaben zum Aufnahmezeitpunkt (T0), die Therapeutenangaben zum Aufnahmezeitpunkt (T0) und die Therapeutenangaben zum Entlassungszeitpunkt (T1; vgl. auch Tabelle 7 und Tabelle 27). Diese Variablensets wurden in fünf unterschiedlichen Kombinationen als Kovariaten in die Imputationsanalysen einbezogen:

- nur Patientenangaben T0 (141 Variablen, „P“)
- nur Therapeutenangaben T0 (30 Variablen, „T“)
- Patienten- und Therapeutenangaben T0 (171 Variablen, „PT“)
- Patienten- und Therapeutenangaben T0 und Therapeutenangaben T1 (192 Variablen, „PTT“)
- Therapeutenangaben T0 und T1 (51 Variablen, „TT“)

Berücksichtigung weiterer Rahmenbedingungen

Neben den systematisch variierten Rahmenbedingungen wurden außerdem noch zwei Faktoren in den Analysen berücksichtigt, die sich eher indirekt aus den vorliegenden Daten ergeben. Dabei handelt es sich zum einen um den *Zusammenhang zwischen den berücksichtigten Variablen*, zum anderen um die verschiedenen der Fehlwertersetzung zugrunde liegenden *Stichprobenumfänge*.

1. Zusammenhänge zwischen Kovariaten und zu schätzendem Kriterium

Die *Zusammenhänge zwischen den in den Fehlwertersetzungen berücksichtigten Variablen* wurden in der vorliegenden Simulationsstudie nicht systematisch variiert, um die externe Validität der Untersuchung nicht unnötig einzuschränken. Unterschiedlich ausgeprägte Assoziationen zwischen den jeweils berücksichtigten Kovariaten und den zu ersetzenden Kriterienvariablen ergeben sich jedoch aus den real bestehenden Zusammenhängen zwischen den einzelnen Variablen. Hier ist insbesondere davon auszugehen, dass sich für die patientenseitig erfassten Variablen der SF-8 (Selbstbeurteilung) im Vergleich zu den Items der SCL-14 engere Zusammenhänge zu den the-

rapeutenseitig erhobenen Variablen der SF-8-F (Fremdbeurteilung) nachweisen lassen könnten, da hier von Patienten und Therapeuten ein identisches Instrument beantwortet wurde.

2. *Einrichtungsbezogene Stichprobengrößen*

   Die *Stichprobenumfänge* wurden nicht künstlich homogenisiert, um den damit verbundenen Informationsverlust zu vermeiden und die in der Praxis anzutreffende Bandbreite an Stichproben abzubilden. Die erreichten Stichprobenumfänge variieren jedoch zwischen den einzelnen Einrichtungen so erheblich, dass sich hieraus gegebenenfalls Informationen zum Einfluss der Stichprobengröße auf die Güte der Ersetzung fehlender Werte ableiten lassen könnten.

Kombination der variierten Rahmenbedingungen

Die einzelnen Ausprägungen der im Zuge der Simulationsstudie zur Ersetzbarkeit fehlender Werte berücksichtigten Rahmenbedingungen sind in Tabelle 8 nochmals im Überblick dargestellt. Alleine durch die Untersuchung der zentralen Frage, welche Ersetzungsmethode in Abhängigkeit einer vorliegenden Dropoutquote und der Verfügbarkeit bestimmter Prädiktoren in der angemessensten Fehlwertersetzung resultiert, ergeben sich durch die Kombination der drei angewandten Imputationsverfahren (RA/EM/MI), der fünf simulierten Dropoutquoten (10/20/30/40/50% variablenbezogen fehlende Werte) und der fünf überprüften Kovariaten-Sets (P/PT/PTT/TT/T) bereits 75 Ersetzungsvarianten (vgl. den Würfel in Abbildung 4). Bezieht man zusätzlich die beiden simulierten Ausfallmechanismen (MCAR/MAR), die beiden Ergebniskriterien (GSI/PSK) sowie die beiden Ersetzungs-Ebenen (Items/Skala) mit ein, so resultieren daraus schon 600 Ersetzungsvarianten (vgl. Abbildung 5). Da die Fehlwertersetzung außerdem getrennt für jede der elf beteiligten Einrichtungen erfolgte, ergeben sich aus der Kombination sämtlicher systematisch oder natürlich variierten Rahmenbedingungen letztlich insgesamt 6.600 verschiedene Ersetzungsszenarien, die im Rahmen der vorliegenden Arbeit überprüft wurden.

Tabelle 8: Rahmenbedingungen der Überprüfung potentieller Einflussfaktoren auf die Güte der Ersetzung fehlender Patientenangaben zum Entlassungszeitpunkt

Einflussfaktoren	**Variationen**
Ersetzungsmethode	Regressionsschätzung
	EM-Ersetzung
	Multiple Imputation
Ausfallmechanismus	zufällig (MCAR)
	systematisch (MAR)
Berücksichtigte Kovariaten	Patientenangaben T0
	Patientenangaben T0 + Therapeutenangaben T0
	Patientenangaben T0 + Therapeutenangaben T0 + T1
	Therapeutenangaben T0
	Therapeutenangaben T0 + T1
Ausmaß fehlender Werte (variablenbezogen)	10 Prozent
	20 Prozent
	30 Prozent
	40 Prozent
	50 Prozent
Ausmaß fehlender Werte (fallbezogen)	einzeln (Skalenwerte)
	multipel (Items)
Zusammenhang zwischen berücksichtigten	GSI der SCL-14
Variablen je nach Ergebniskriterium	PSK der SF-8
Stichprobenumfang je nach Klinik	11 Kliniken

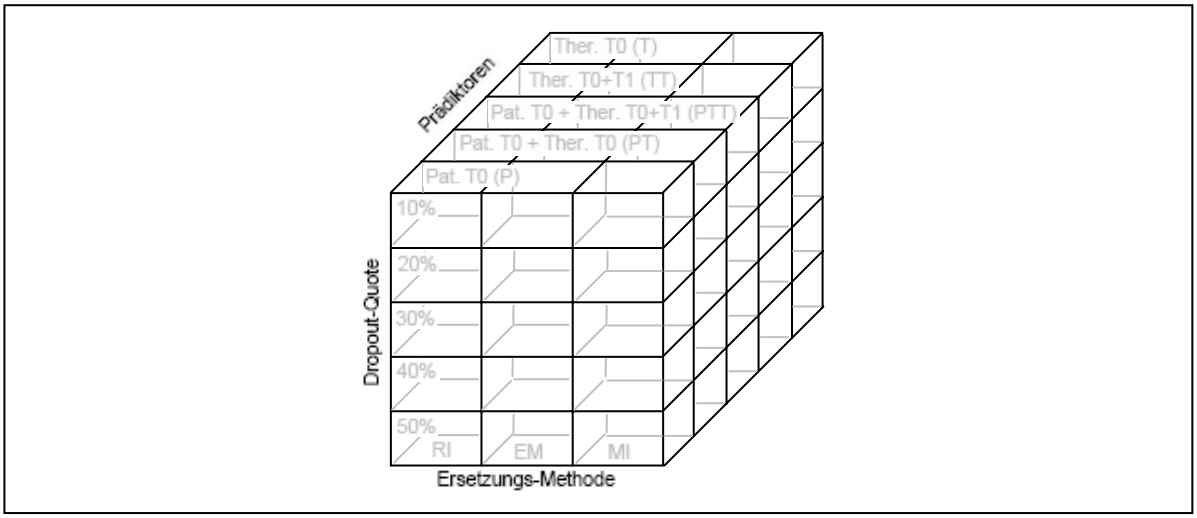

Abbildung 4: Ersetzungsvarianten durch Kombination von Dropoutquoten, Ersetzungsmethoden und Prädiktoren

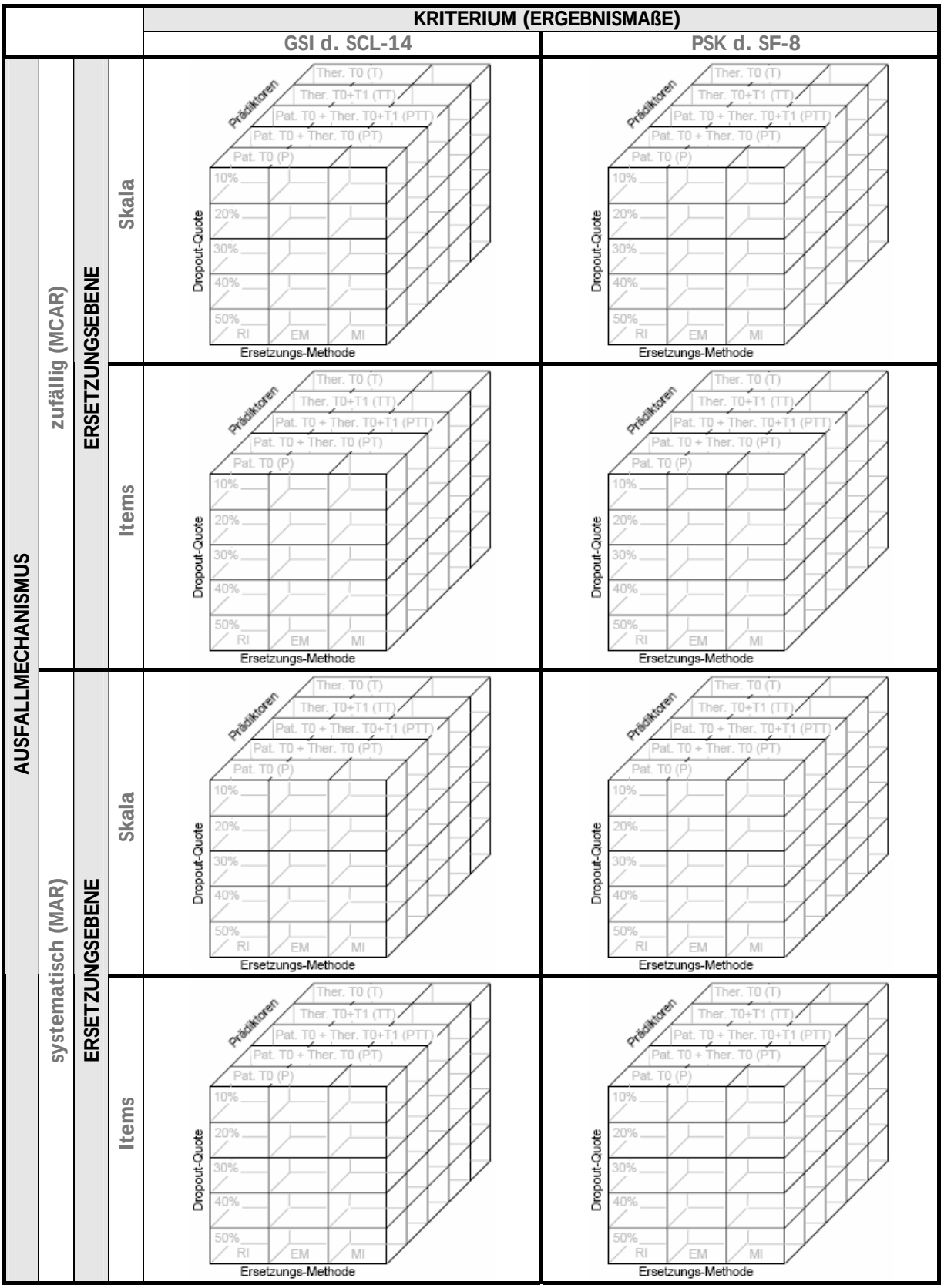

Abbildung 5: Einrichtungsbezogene Ersetzungsvarianten durch Kombination aller Rahmenbedingungen

Beurteilung der Güte von Fehlwertersetzungen

Die Güte von Fehlwertersetzungen bemisst sich an der Übereinstimmung respektive Ähnlichkeit der einzelnen geschätzten und „wahren“ Werte. Übertragen auf die Ersetzung mehrerer fehlender Werte drückt sich eine hohe Übereinstimmung demnach sowohl in einem engen Zusammenhang, also hohen Korrelationen zwischen geschätzten und wahren Werten, als auch in einer ähnlichen mittleren Ausprägung von geschätzten und wahren Werten aus. Die reliable Ersetzung fehlender Werte sollte somit letztlich in vergleichbaren Verteilungseigenschaften (Mittelwerte und Streuung) von geschätzten und wahren Werten resultieren.

Einen geeigneten statistischen Kennwert zur Beurteilung der Güte von Fehlwertersetzungen stellt im vorliegenden Fall intervallskalierter Variablen die Intraklassenkorrelation (ICC) dar (eine ausführliche Darstellung der ICC findet sich z.B. bei Wirtz & Caspar, 2002). Die ICC kann ähnlich wie eine Produkt-Moment-Korrelation interpretiert werden: Ein Wert von 0 bedeutet dementsprechend, dass kein Zusammenhang zwischen geschätzten und wahren Werten besteht, die Güte der Fehlwertersetzung also minimal ausfällt. Je näher die ICC sich einem Wert von 1 nähert, umso höher ist der Zusammenhang zwischen geschätzten und wahren Werten und somit auch die Reliabilität der Fehlwertersetzung.

Da nicht nur der Zusammenhang (die Korrelation) zwischen wahren und geschätzten Werten, sondern auch deren relative Lage bedeutsam ist, stellen *unjustierte ICC* das angemessene Reliabilitätsmaß dar. Eine unjustierte ICC (ICC_{unjust}) kann nur dann hohe Werte annehmen, wenn die Mittelwerte zweier verglichener Variablen (Original und Schätzung) ähnlich sind und die Werte der beiden Variablen zugleich eine hohe Korrelation aufweisen. Aus der zusätzlichen Modellannahme der Varianzhomogenität folgt dann, dass die absoluten Werte der verglichenen Variablen ähnlich sein müssen.

Gemessene wie geschätzte Werte lassen sich nach dem varianzanalytischen Modell jeweils als Summe der wahren Merkmalsausprägung und einer Fehlerkomponente darstellen. Da die einzelnen Werte im vorliegenden Fall von jeweils unterschiedlichen Patienten angegeben und nicht von jeweils identischen Raterpaaren eingeschätzt wurden, lässt sich diese Fehlerkomponente nicht weiter systematisch zerlegen, so dass hier das *einfaktorielle* Modell der ICC anzuwenden ist ($ICC_{unjust\ einfakt}$).

Nach ihrer Definition als Varianzaufklärungsmaß müssten ICC theoretisch immer Werte zwischen 0 und 1 annehmen. In der Praxis können bei der Berechnung von ICC jedoch auch negative Werte resultieren, wenn die erwarteten Varianzen (Mean square, MS) innerhalb der Personen (MS_{inn}) größer ausfallen als diejenigen zwischen den Personen (MS_{zw}). Dies kann im Kontext der Fehlwertersetzung entweder der Fall sein, wenn sich bezüglich der zu schätzenden Werte von vorneherein kaum Varianz zwischen den Personen findet oder aber wenn die einzelnen geschätzten Werte sehr stark von den wahren Werten abweichen. Da negative Reliabilitätsschätzungen jedoch nicht sinnvoll interpretiert werden können, sollten grundsätzlich nur solche ICC verwertet werden, für die sich signifikante Effekte nachweisen lassen. Sollen jedoch, wie in der vorliegenden Simulationsstudie, ICC aus verschiedenen Stichproben und Ersetzungsvarianten miteinander verglichen werden, würde die tatsächliche Reliabilität der Fehlwertersetzungen durch den Ausschluss negativer ICC deutlich überschätzt. Um mögliche Verzerrungen durch negative ICC zu umgehen, wurde in der vorliegenden Arbeit der von Hartung als Alternative zu den klassischen ICC vorgeschlagene korrigierte ICC verwendet ($ICC_{unjust,einfakt,korr}$; Hartung, 1981). Zur Veranschaulichung der Unterschiede bezüglich der Bestimmung klassischer vs. korrigierter ICC sind die entsprechenden Formeln in Abbildung 6 wiedergegeben.

Klassische ICC lassen sich nach Klein et al. (2000) als Reliabilitätskennwerte interpretieren, so dass Werte ab 0,70 als akzeptabel, Werte zwischen 0,50 und 0,70 als grenzwertig und schließlich Werte kleiner 0,50 als dürftig gelten können (vgl. auch Greve & Wentura, 1997). Obgleich die Beträge korrigierter ICC - insbesondere im Bereich geringer Varianzaufklärung - tendenziell höher ausfallen als die Beträge klassischer ICC, sollen in der vorliegenden Arbeit die von Klein et al. formulierten Grenzwerte als Richtgrößen für die Beurteilung der Ersetzungsgüte angelegt werden.

Korrigierte ICC wurden getrennt für die Resultate aller 6.600 realisierten Ersetzungsvarianten berechnet. Um die konkrete Güte der Fehlwertersetzung zu überprüfen, wurden dabei jeweils zum einen nur diejenigen Fälle berücksichtigt, für die zuvor fehlende Werte simuliert und ersetzt worden sind. Zum anderen wurden ICC jeweils aber auch für die Gesamtstichprobe der jeweiligen Einrichtung bestimmt, um ein Maß für die Auswirkungen der Fehlwertersetzung auf die jeweiligen Stichprobenkennwerte zu generieren, da diese im Falle eines Einrichtungsvergleiches die zentrale Ergebnisgröße darstellen.

Klassische ICC:

$$ICC_{unjust,einfakt} = \frac{MS_{zw} - MS_{inn}}{MS_{zw} + (k-1) \cdot MS_{inn}}$$

Korrigierte ICC:

$$ICC_{unjust,einfakt,korr} = \frac{\frac{k}{1+k^2} \cdot MS_{zw}}{\frac{k}{1+k^2} \cdot MS_{zw} + MS_{inn}}$$

MS: erwartete Varianz (Mean square), **zw**: zwischen Personen, **inn**: innerhalb Personen, **k**: Anzahl verglichener Variablen

Abbildung 6: Formeln zur Bestimmung klassischer und korrigierter Intraklassen-Korrelations-Koeffizienten (ICC) – unjustiertes einfaktorielles Modell (n. Wirtz & Caspar, 2002)

Ein Problem der ICC besteht in ihrem eingeschränkten Variationsbereich: Klassische ICC sind auf einen Wertebereich von $[1 \geq ICC_{klass} \geq -1/(k-1)]$, korrigierte ICC auf einen Wertebereich von $[1 \geq ICC_{korr} \geq 0]$ normiert. Durch die Normierung ihres Wertebereichs sind ICC-Werte, wie alle anderen Korrelationskoeffizienten, jedoch nicht intervallskaliert (Fisher, 1956). Je mehr sich der wahre Zusammenhang ρ dem Wert +1 (bzw. -1) annähert, desto rechts- (bzw. links-) schiefer wird die Kennwerteverteilung: Der Unterschied zwischen zwei ICC von 0,8 und 0,9 ist also wesentlich größer als der Unterschied zwischen zwei ICC von 0,1 und 0,2. Um die Abstände zwischen zwei oder mehreren ICC miteinander vergleichen zu können, müssen ihre Werte zunächst in eine Verhältnisskala transformiert werden. Fisher hat hierzu eine nichtlineare Transformation vorgeschlagen (vgl. Bortz, 1999), die nach ihm benannte „Fisher-Z-Transformation“ (vgl. Abbildung 7). Die Fisher-Z-Werte sind nicht mehr auf einen bestimmten Wertebereich beschränkt, sie ermöglichen – im Gegensatz zu ICC – Vergleiche von Differenzen und Mittelwertsbildungen und können damit auch mit den gängigen inferenzstatistischen Verfahren verarbeitet werden (Wirtz & Caspar, 2002).[2] Zur Aggregierung mehrerer ICC über verschiedene

[2] Ein alternativer Vorschlag, der sich konkret auf die zufallskritische Absicherung der Unterschiede zwischen ICC bezieht und beim Vergleich *einzelner* ICC der Verwendung der Fisher-Z-Methode, die ja ursprünglich für den Korrelationskoeffizienten r und nicht für ICC entwickelt wurde, als angemessener belegen ließ, findet sich z.B. bei McGraw & Wong (1996). Da in der vorliegenden Arbeit jedoch jeweils nicht einzelne, sondern größere Gruppen von ICC miteinander verglichen werden sollen, erscheint die Verwendung von Fisher-Z-Werten hier durchaus angemessen.

Ersetzungsvarianten sowie zur Überprüfung von Unterschieden zwischen verschiedenen Gruppen von ICC wurden dementsprechend jeweils die korrespondierenden Fisher-Z-Werte verwendet.

Fisher-Z-Transformation: $$Z_{Fisher} = 0{,}5 \cdot \ln\left(\frac{1+ICC}{1-ICC}\right)$$

Abbildung 7: Formel zur Transformation von ICC in Fisher-Z-Werte (n. Bortz, 1999)

4.6. Hypothesen und Hypothesenprüfung

Vor dem Hintergrund der in Kapitel 2 dargestellten Grundlagen der Fehlwertersetzung und der im vorangegangen Abschnitt (Kapitel 4.5) erläuterten Methodik der durchgeführten Simulationsstudie lassen sich in Bezug auf die in Kapitel 3 benannten Fragestellungen der vorliegenden Arbeit Vorhersagen und Hypothesen formulieren, die im Folgenden aufgeführt werden.

Methoden der Fehlwertersetzung

In Bezug auf die eingesetzten Verfahren zur Fehlwertersetzung wird angenommen, dass

a) die Multiple Imputation (MI) im Vergleich zu den beiden einfachen Imputationsverfahren (EM-Ersetzung und Regressionsschätzung) bessere Ersetzungsresultate generiert, da sie die Unsicherheit einzelner Schätzungen durch mehrfache Schätzungen ausgleichen sollte, und

b) die EM-Ersetzung (EM) bessere Ersetzungsresultate erbringt als die Regressionsschätzung (RA), da letztere zur Überschätzung der wahren Varianz tendiert.

Diese Annahmen lassen sich über die folgenden statistischen Hypothesen prüfen:

Hypothese 1a:

$\mathbf{H_0}$: $\overline{ICC}_{MI} - 0{,}5 \cdot \left(\overline{ICC}_{EM} + \overline{ICC}_{RA} \right) \leq 0$

$\mathbf{H_1}$: $\overline{ICC}_{MI} - 0{,}5 \cdot \left(\overline{ICC}_{EM} + \overline{ICC}_{RA} \right) > 0$

Hypothese 1b:

$\mathbf{H_0}$: $\overline{ICC}_{EM} - \overline{ICC}_{RA} \leq 0$

$\mathbf{H_1}$: $\overline{ICC}_{EM} - \overline{ICC}_{RA} > 0$

Systematik fehlender Werte

Bezüglich der verschiedenen simulierten Ausfallmechanismen (zufälliger vs. systematischer Dropout bei negativem Therapieverlauf, MCAR/MAR) ist davon auszugehen, dass die Fehlwertersetzung bei zufällig fehlenden Werten wegen der höheren Wahrscheinlichkeit zufällig richtiger Schätzungen zu besseren Ersetzungsergebnissen führt als bei systematisch fehlenden Werten. Zusätzlich sollten sich auf der Ebene der jeweiligen Einrichtungsstichproben bei zufällig fehlenden Werte selbst bei unbefriedigender Ersetzungsgüte geringere Verzerrungen ergeben als bei systematisch fehlenden Werten.

Aus dieser Vorhersage lässt sich die folgende statistische Hypothese ableiten:

Hypothese 2:

$\mathbf{H_0}$: $\overline{ICC}_{MCAR} - \overline{ICC}_{MAR} \leq 0$

$\mathbf{H_1}$: $\overline{ICC}_{MCAR} - \overline{ICC}_{MAR} > 0$

Anteile variablenbezogen fehlender Werte

Mit zunehmenden Anteilen fehlender Werte verringern sich die Datenbasis und damit das Ausmaß verfügbarer Informationen, auf deren Grundlage die Ausprägungen fehlender Werte geschätzt werden können. Im Falle systematischer Datenausfälle ist zudem die Repräsentativität der vorliegenden Datenbasis eingeschränkt, was bei der Ersetzung fehlender Werte zu Verzerrungen führen kann. Die aus diesen Effekten resultierende Unsicherheit der Fehlwertersetzung sollte umso größer ausfallen, je höher der Anteil fehlender Werte ist.

Dementsprechend lässt sich zum Einfluss fehlender Werte auf die Güte der Fehlwertersetzung folgende statistische Hypothese aufstellen:

Hypothese 3:

$$\mathbf{H_0:}\quad \overline{ICC}_{10\%} - \overline{ICC}_{20\%} \leq 0 \;\cup\; \overline{ICC}_{20\%} - \overline{ICC}_{30\%} \leq 0 \;\cup\; \overline{ICC}_{30\%} - \overline{ICC}_{40\%} \leq 0 \;\cup\; \overline{ICC}_{40\%} - \overline{ICC}_{50\%} \leq 0$$

$$\mathbf{H_1:}\quad \overline{ICC}_{10\%} - \overline{ICC}_{20\%} > 0 \;\cap\; \overline{ICC}_{20\%} - \overline{ICC}_{30\%} > 0 \;\cap\; \overline{ICC}_{30\%} - \overline{ICC}_{40\%} > 0 \;\cap\; \overline{ICC}_{40\%} - \overline{ICC}_{50\%} > 0$$

Berücksichtigte Kovariaten

Im Hinblick auf die zur Schätzung fehlender Werte herangezogenen Kovariaten ist davon auszugehen, dass

a) die Berücksichtigung der Patientenangaben zu T0 (P) aufgrund der anzunehmend engeren Zusammenhänge zu den fehlwertbehafteten Kriteriumsvariablen zu besseren Ersetzungsresultaten führt als die bloße Berücksichtigung der Therapeutenangaben zu T0 (T) bzw. T0 und T1 (TT) und

b) der Einbezug größerer Gruppen von Kovariaten wegen der damit verbundenen höheren Kovarianzinformation in besseren Ersetzungsergebnissen resultiert als die Berücksichtigung einer geringeren Anzahl von Kovariaten.

Für diese Annahmen lassen sich die folgenden statistischen Hypothesen formulieren:

Hypothese 4a:

$$\mathbf{H_0:} \quad \left(\overline{ICC}_P + \overline{ICC}_{PT} + \overline{ICC}_{PTT}\right) - 1{,}5 \cdot \left(\overline{ICC}_T + \overline{ICC}_{TT}\right) \leq 0$$

$$\mathbf{H_1:} \quad \left(\overline{ICC}_P + \overline{ICC}_{PT} + \overline{ICC}_{PTT}\right) - 1{,}5 \cdot \left(\overline{ICC}_T + \overline{ICC}_{TT}\right) > 0$$

Hypothese 4b:

$$\mathbf{H_0:} \quad \overline{ICC}_{PTT} - \overline{ICC}_{PT} \leq 0 \ \cup \ \overline{ICC}_{PT} - \overline{ICC}_P \leq 0 \ \cup \ \overline{ICC}_P - \overline{ICC}_{TT} \leq 0 \ \cup \ \overline{ICC}_{TT} - \overline{ICC}_T \leq 0$$

$$\mathbf{H_1:} \quad \overline{ICC}_{PTT} - \overline{ICC}_{PT} > 0 \ \cap \ \overline{ICC}_{PT} - \overline{ICC}_P > 0 \ \cap \ \overline{ICC}_P - \overline{ICC}_{TT} > 0 \ \cap \ \overline{ICC}_{TT} - \overline{ICC}_T > 0$$

Anzahl patientenbezogen fehlender Werte

Die Anzahl patientenbezogen fehlender Werte wurde nicht systematisch variiert, schwankt jedoch, je nachdem ob die fehlenden Ergebnisdaten auf Item- (mehrere fehlende Werte) oder auf Skalenebene (ein fehlender Wert) ersetzt werden. Aufgrund des relativ erhöhten Anteils fehlender Informationen im Falle der itemweisen Fehlwertersetzung ist davon auszugehen, dass sich bei der itembezogenen Ersetzung fehlender Werte größere Unsicherheiten der Schätzung ergeben, die kumuliert in einer schlechteren Ersetzungsgüte resultieren.

Vor diesem Hintergrund ist die folgende statistische Hypothese zu überprüfen:

Hypothese 5:

$$\mathbf{H_0:} \quad \overline{ICC}_{Skala} - \overline{ICC}_{Items} \leq 0$$

$$\mathbf{H_1:} \quad \overline{ICC}_{Skala} - \overline{ICC}_{Items} > 0$$

Ergebnismaße

Da der SF-8 (psychische Summenskala, PSK) sowohl patienten- als auch therapeutenseitig eingesetzt wurde, lässt sich vermuten, dass patientenseitig fehlende SF-8-Daten aufgrund des damit verbunden engeren Zusammenhangs zu den als Kovariaten berücksichtigten therapeutenseitigen Angaben besser geschätzt werden können als fehlende Angaben in der SCL-14 (Globaler Symptomschwereindex, GSI).

Dementsprechend lässt sich folgende statistische Hypothese aufstellen:

Hypothese 6:

H_0: $\overline{ICC}_{PSK} - \overline{ICC}_{GSI} \leq 0$

H_1: $\overline{ICC}_{PSK} - \overline{ICC}_{GSI} > 0$

Stichprobengröße

Die Stichprobengröße wurde nicht systematisch variiert, es finden sich jedoch natürliche Schwankungen zwischen den Umfängen der einzelnen Klinikstichproben. Diese könnten Hinweise auf einen Einfluss der zugrunde liegenden Stichprobengröße auf die Güte der Ersetzung fehlender Werte liefern. Obgleich mit zunehmenden Stichprobenumfängen von besseren Ersetzungsergebnissen auszugehen wäre, kann hier keine eindeutig gerichtete Hypothese formuliert werden, da zwischen den einzelnen Einrichtungen zugleich relevante Unterschiede bezüglich der Variation der interessierenden Kriterien und Kovariaten anzutreffen sind, die ihrerseits das Ergebnis der Fehlwertersetzung beeinflussen können. Es ist demzufolge lediglich davon auszugehen, dass sich zwischen einzelnen Kliniken Unterschiede bezüglich der mittleren Ersetzungsgüte nachweisen lassen.

Die entsprechend ungerichtete statistische Hypothese lautet daher:

Hypothese 7:

H_0: $\overline{ICC}_{Klinik\,i} - \overline{ICC}_{Klinik\,j} = 0$

H_1: $\overline{ICC}_{Klinik\,i} - \overline{ICC}_{Klinik\,j} \neq 0$

Hypothesenprüfung

Alle statistischen Hypothesen beziehen sich auf Unterschiede hinsichtlich der mittleren Ersetzungsgüte verschiedener Gruppen von Ersetzungsvarianten, die über Intraklassenkorrelationskoeffizienten (ICC) zur Übereinstimmung zwischen geschätzten und Original-Werten operationalisiert wurden. Aufgrund des oben beschriebenen Problems der eingeschränkten Variationsbreite von ICC wurden diese Vergleiche jeweils anhand der korrespondierenden Fisher-Z-transformierten ICC-Werte durchgeführt. Zur Überprüfung der verschiedenen gerichteten Hypothesen (Hypothesen 1-6) wurden Varianzanalysen mit geplanten (a priori) Kontrasten berechnet. Zur Überprüfung der ungerichteten Hypothese (Hypothese 7) wurde eine Varianzanalyse mit post hoc Einzelvergleichen nach Scheffé (bei Varianzgleichheit) bzw. Tamhane (bei nicht gegebener Varianzgleichheit) angewendet.

Um Unterschiede zwischen den Ersetzungsvarianten bezüglich der konkreten Güte der Fehlwertersetzung (bezogen auf den Einzelfall) aufzudecken, erfolgte die Prüfung der Hypothesen jeweils auf der Ebene derjenigen Fälle, für die zuvor fehlende Werte simuliert und ersetzt worden waren (Dropout-Stichproben). Zur Aufdeckung unterschiedlicher Auswirkungen der Ersetzungsvarianten auf die resultierenden Ergebniskennwerte in den jeweiligen Gesamtstichproben wurden die Hypothesen außerdem auf der Ebene der Einrichtungsstichproben überprüft, da diese im Falle eines Einrichtungsvergleiches letztlich die zentrale Größe darstellen.

Um einer möglichen Kumulierung des Alpha-Fehlers zu begegnen, wurde das Signifikanzniveau bei der Überprüfung der neun untersuchten Hypothesen in Anlehnung an die Bonferroni-Korrektur auf p=0,005 festgelegt.

Als Grundlage für die Beurteilung des Ausmaßes ermittelter Unterschiede wurden jeweils Effektstärken in Form von eta^2 bestimmt. Gemäß der Konventionen von Cohen (1988) lassen sich dabei kleine, mittlere und große Effektstärken unterscheiden. Die jeweiligen unteren Grenzen liegen für eta^2 bei 0,0099, 0,0588 und 0,1379 (vgl. Cohen, 1988).

4.7. Exemplarische Übertragung der Befunde zur Ersetzbarkeit fehlender Entlassungswerte auf den Kontext des Einrichtungsvergleichs

Im Anschluss an die Simulationsstudie sollten die ermittelten Befunde zur Ersetzbarkeit fehlender Werte exemplarisch auf den konkreten Anwendungskontext des Einrichtungsvergleiches übertragen werden. Im QS-Reha®-Verfahren wird die Ergebnisqualität der untersuchten Einrichtungen grundsätzlich über den *zum Entlassungszeitpunkt erhobenen Gesundheitsstatus* der in den verschiedenen Einrichtungen behandelten Patienten miteinander verglichen. Exemplarisch werden hierfür in der vorliegenden Arbeit der bei Entlassung erhobene „Globale Symptomschwere-Index“ (GSI) der SCL-14 sowie die „Psychische Summenskala“ (PSK) der SF-8 herangezogen (vgl. Kapitel 4.2).

Dementsprechend wurden im Gesamtdatensatz (N=2.386) zunächst diejenigen Fälle ermittelt, für die wenigstens vollständige Entlassungsangaben in den 14 Items des GSI und den 4 Items der PSK der SF-8 vorlagen (im Folgenden als „Responder“ benannt). Um zu prüfen, ob die Befunde zur Ergebnisqualität, die auf der Basis der eingeschränkten Stichproben von Patienten mit vollständig vorhandenen Entlassungsdaten ermittelt würden, auf die jeweiligen Gesamtstichproben der jeweils insgesamt in den verschiedenen Kliniken behandelten Patienten generalisierbar wären, wurden entsprechende Dropout-Analysen durchgeführt. Im Rahmen dieser Dropout-Analysen wurde anhand der jeweils sowohl für die Responder wie auch für die Dropouts vorliegenden Daten einrichtungsbezogen überprüft, ob zwischen den beiden Patientengruppen Unterschiede hinsichtlich potentieller „Risikofaktoren“ bestehen. Als Risikofaktoren bzw. Konfounder sind solche Variablen definiert, die potentiell in einem (kausalen) Zusammenhang mit dem Behandlungsergebnis stehen. Konkret wurden in der vorliegenden Arbeit die Variablen Geschlecht, Alter, Nationalität, Partnersituation, Schulbildung, aktuelle Erwerbssituation, AU- bzw. Krankheitszeiten in den letzten 6 Monaten, Rentenantrag, Chronifizierung der Erkrankung, Behandlungsmotivation und die psychischen und somatischen Diagnosen als Risikofaktoren berücksichtigt. Außerdem wurde die Ausgangsbelastung, repräsentiert durch die Skalenwerte in den verschiedenen zu Behandlungsbeginn patientenseitig eingesetzten psychometrischen Verfahren, als weiterer Konfounder in die Dropoutanalysen einbezogen. Zur Absicherung von Gruppenunterschieden wurden je nach Skalenniveau der berücksichtigten Variablen t-Tests oder Chi-Quadrat-Tests durchgeführt.

Danach wurden die fehlenden Entlassungswerte in SCL-14 und SF-8 gemäß der im vorangegangenen Kapitel beschriebenen Methodik ersetzt. Einrichtungsbezogene Stichprobenkennwerte für Entlassungs-GSI und –PSK wurden sodann zum einen für die jeweiligen Responder-Stichproben, zum anderen für die nach der optimalen Fehlwertersetzung resultierenden Gesamtstichproben (Responder + Dropouts) bestimmt und einander gegenüber gestellt. Dabei wurde einerseits überprüft, ob die nach Fehlwertersetzung resultierenden Stichprobenkennwerte signifikant von den entsprechenden ohne Fehlwertersetzung ermittelten Stichprobenkennwerten abweichen. Außerdem wurde untersucht, inwieweit sich die Fehlwertersetzung auf die Bewertung der einzelnen Einrichtungen im Vergleich zu den jeweiligen Referenzeinrichtungen auswirkt.

Abschließend wurde noch untersucht, inwieweit die Abweichungen zwischen Respondern und Dropouts, die sich in den Dropoutanalysen ergeben hatten, auch in den Abweichungen zwischen vor und nach Fehlwertersetzung ermittelten Entlassungswerten widerspiegeln, um Hinweise auf die Eignung von Dropoutanalysen, die sich im Wesentlichen auf initiale Unterschiede beziehen, zur Sicherstellung der Repräsentativität in Bezug auf ermittelte Behandlungsergebnisse zu generieren.

Für die Signifikanzprüfung wurden wiederum je nach Skalenniveau Chi-Quadrat- oder t-Tests durchgeführt, das Signifikanzniveau wurde auf $p \leq 0,05$ festgelegt. Auf eine Anpassung des Signifikanzniveaus wurde in diesem Zusammenhang aufgrund des explorativen Charakters der Überprüfung verzichtet. Das Ausmaß des Unterschiedes wurde jeweils in Form von Effektgrößen nach Cohen (*Phi, d*) berechnet (Cohen, 1988). Gemäß der Konventionen von Cohen (1988) lassen sich auch hier kleine, mittlere und große Effektstärken unterscheiden. Die jeweiligen unteren Grenzen liegen für *d* bei 0,20, 0,50 und 0,80, für *Phi* liegen sie bei 0,10, 0,30 und 0,50.

5. Ergebnisse

Im folgenden Ergebnisteil werden zunächst die Befunde der für den Originaldatensatz durchgeführten Fehlwertanalysen berichtet (Kapitel 5.1). Im Anschluss daran findet sich sodann die Darstellung der Ergebnisse aus der durchgeführten Simulationsstudie zur Überprüfung der Rahmenbedingungen für eine adäquate Ersetzbarkeit fehlender Werte (Kapitel 5.2). Den Abschluss des Ergebniskapitels bildet schließlich die Illustration einer Anwendung der in der Simulationsstudie ermittelten Resultate auf den Kontext des konkreten Einrichtungsvergleichs (Kapitel 5.3).

5.1. Fehlende Werte

Anteile fehlender Werte

Für alle in der vorliegenden Arbeit berücksichtigten Variablen (vgl. hierzu die Aufstellung im Anhang, Kap. 12.1, Tabelle 27) wurden die Quoten patienten- und itembezogen fehlender Werte getrennt für die elf beteiligten Kliniken ermittelt. Die durchschnittlichen Anteile fehlender Werte in den einzelnen Erhebungsmodulen (Patient/Therapeut x Aufnahme/Entlassung) sind in Tabelle 9, ergänzt um die jeweiligen Spannweiten fehlender Werte, wiedergegeben. Die Verteilungen der Anteile fehlender Werte in den einzelnen Erhebungseinheiten sind in Abbildung 8 für die Gesamtstichprobe dargestellt.

Im Mittel liegt die Quote fallbezogen wie itembezogen fehlender Werte mit 8,5 Prozent in einem akzeptablen Bereich. Dabei fällt auf, dass die Anteile fehlender Angaben in den Patientenfragebögen deutlich höher ausfallen als in den Therapeutenfragebögen. Zudem zeigen sich die Fehlwertquoten zum Behandlungsende wiederum deutlich größer als zu Behandlungsbeginn (vgl. Tabelle 9).

Zwischen den untersuchten Kliniken zeigen sich in allen Erhebungseinheiten deutliche Abweichungen bezüglich der Vollständigkeit der erhobenen Daten. Während die durchschnittliche Quote fehlender Werte für die Patientenfragebögen in den meisten Kliniken

Tabelle 9: Durchschnittliche Anteile fall- und itembezogen fehlender Werte

	Kliniken											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Gesamt**
Fallbezogen fehlende Werte												
% fehlender Werte Patient T0	2,8	3,7	2,3	6,5	3,2	2,4	18,8	33,2	3,7	23,2	4,7	9,5
Range % fehlender Werte Patient T0	*0-100*	*0-100*	*0-100*	*0-100*	*0-100*	*0-100*	*0-100*	*0-100*	*0-99*	*0-100*	*0-100*	*0-100*
% fehlender Werte Therapeut T0	3,9	0,6	0,4	1,7	4,9	4,1	16,8	3,2	2,0	0,5	1,8	3,4
Range % fehlender Werte Therapeut T0	*0-96*	*0-42*	*0-8*	*0-46*	*0-96*	*0-100*	*0-100*	*0-100*	*0-85*	*0-31*	*0-8*	*0-100*
% fehlender Werte Patient T1	2,3	7,9	4,2	12,0	5,4	8,4	37,6	10,8	3,5	21,5	7,0	10,8
Range % fehlender Werte Patient T1	*0-100*	*0-100*	*0-100*	*0-100*	*0-100*	*0-100*	*0-100*	*0-100*	*0-100*	*0-100*	*0-100*	*0-100*
% fehlender Werte Therapeut T1	3,3	1,1	1,6	2,6	8,6	5,3	25,2	4,9	4,8	0,8	1,7	5,1
Range % fehlender Werte Therapeut T1	*0-95*	*0-67*	*0-95*	*0-95*	*0-95*	*0-100*	*0-100*	*0-100*	*0-100*	*0-38*	*0-10*	*0-100*
Itembezogen fehlende Werte												
% fehlender Werte Patient T0	2,8	3,7	2,3	6,5	3,2	2,4	18,8	33,2	3,7	23,2	4,7	9,5
Range % fehlender Werte Patient T0	*1-38*	*1-35*	*0-33*	*3-48*	*1-36*	*1-28*	*2-38*	*2-50*	*0-24*	*22-39*	*1-25*	*5-35*
% fehlender Werte Therapeut T0	4,4	1,4	0,7	1,8	6,0	4,0	16,9	3,4	2,7	1,4	1,8	3,8
Range % fehlender Werte Therapeut T0	*0-25*	*0-10*	*0-5*	*0-7*	*0-32*	*0-9*	*0-34*	*0-11*	*0-12*	*0-8*	*0-31*	*0-10*
% fehlender Werte Patient T1	2,3	7,9	4,2	12,0	5,4	8,4	37,6	10,8	3,5	21,5	7,0	10,8
Range % fehlender Werte Patient T1	*2-5*	*6-18*	*3-15*	*11-20*	*3-12*	*8-15*	*36-46*	*9-19*	*2-20*	*20-31*	*6-22*	*10-20*
% fehlender Werte Therapeut T1	3,4	2,0	1,9	2,6	9,7	5,3	25,3	5,0	5,5	1,9	1,7	5,5
Range % fehlender Werte Therapeut T1	*0-8*	*0-10*	*0-7*	*0-8*	*1-15*	*3-11*	*1-30*	*3-15*	*1-13*	*0-7*	*0-29*	*1-12*

T0: Aufnahme, T1: Entlassung

unter 10 Prozent liegt, gibt es jedoch auch einzelne Kliniken, in denen deutlich höhere Anteile fehlender Werte zu beobachten sind (Kliniken 7, 8 und 10). In Bezug auf die Vollständigkeit der Therapeutenangaben ergibt sich ein homogeneres Bild, hier fällt lediglich eine Klinik (Klinik 7) durch eine deutlich erhöhte Fehlwertquote auf.

Fallbezogen findet sich in nahezu allen Erhebungsmodulen und Kliniken die volle Bandbreite fehlender Werte (0-100%). Dies bedeutet, dass es pro Erhebungseinheit also zumeist sowohl Patienten gibt, für die komplett vollständige Daten vorliegen, als auch Patienten, für die die Daten komplett fehlen, wobei die Fälle mit vollständig vorliegenden Daten deutlich dominieren (vgl. Abbildung 8). Itembezogen ergibt sich ein homogeneres Bild, hier finden sich im Extremfall klinikspezifische Datenausfälle von höchstens 50 Prozent. Die höchsten Anteile fehlender Werte finden sich in den Therapeutenfragebögen für das Item 8 der HoNOS-D (durchschnittlich 10% bei Aufnahme bzw. 12% bei Entlassung), was auf grundsätzliche Schwierigkeiten bei der Beantwortung dieses Items zurückzuführen sein dürfte (vgl. hierzu Andreas et al., in press). Bezogen auf die Patientenfragebögen finden sich die höchsten Fehlwertquoten in den Angaben zu einem laufenden Rentenverfahren (durchschnittlich 21%), in einer Frage nach der Partnerschaftszufriedenheit (25%) sowie in den Angaben zu Krankheitszeiten (35%).

Systematik fehlender Werte

Die Systematik fehlender Werte wurde insbesondere für die zentralen Maße der Ergebnisqualität (Items der SCL-14 und SF-8 zum Entlassungszeitpunkt) überprüft. Im Einzelnen finden sich keine auffällig hohen systematischen Zusammenhänge zwischen fehlenden Angaben in den Ergebnismaßen und der Ausprägung der verschiedenen erhobenen Hilfsvariablen (alle Korrelationen $r < 0{,}10$). Die beiden höchsten Zusammenhänge zum Datenausfall in den Ergebnismaßen finden sich mit mittleren Korrelationen von $r = 0{,}09$ bzw. $r = 0{,}07$ (beide $p < 0{,}001$) für die Diagnosegruppen „Psychische und Verhaltensstörungen durch psychotrope Substanzen“ (ICD-10: F1) sowie „Migräne und sonstige Kopfschmerzsyndrome“ (ICD-10: G43-44), die Zusammenhänge zu anderen potentiellen Einflussgrößen liegen durchwegs deutlich unter $r = 0{,}05$. Regressionsanalytisch lässt sich das Fehlen von Werten in den Items von SCL-14 und SF-8 zum Entlassungszeitpunkt bei gleichzeitiger Berücksichtigung aller übrigen erhobenen Variablen jedoch zu durch-

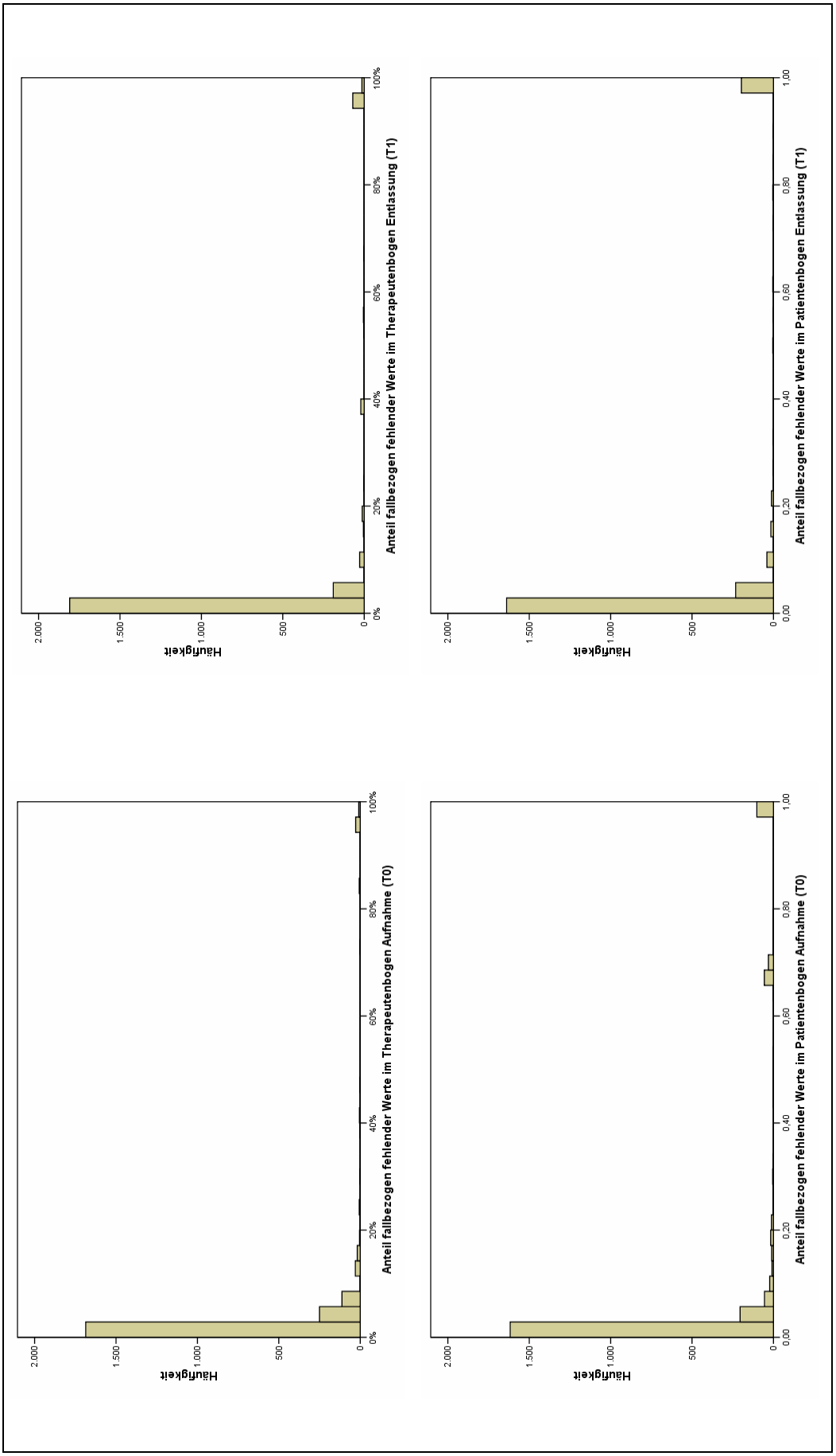

Abbildung 8: Anteile fallbezogen fehlender Werte in den einzelnen Erhebungsmodulen

schnittlich knapp 80 Prozent vorhersagen (R^2=0,79), was für das Vorliegen bedingt zufällig fehlender Werte spricht (MAR-Bedingung). Zusätzlich finden sich Hinweise auf systematische Häufungen fehlender Werte (sog. „Fehlwert-Muster"): Das Fehlen von Werten in den zentralen Ergebnismaßen geht nämlich einher mit einer erhöhten Wahrscheinlichkeit ebenfalls fehlender Angaben zum Rentenstatus (r = 0,17), zur Chronifizierung (r = 0,15), zur Nationalität (r = 0,27), zur Partnersituation (r = 0,19), zur Schulbildung (r = 0,26), zur beruflichen Situation (r = 0,22), zum laufenden Rentenantrag (r = 0,18) sowie nicht zuletzt ebenfalls fehlender Angaben in den entsprechenden SCL-14- und SF-8-Items zu Behandlungsbeginn (0,26 < r <0,33; alle p < 0,001). Die höchsten Zusammenhänge zeigen sich jedoch erwartungsgemäß zwischen dem Fehlen eines Ergebniswertes und dem Fehlen von Angaben in den anderen Ergebnismaßen (durchschnittlich r = 0,89), worin sich ein hoher Anteil von komplett fehlenden Ergebnisdaten widerspiegelt („Unit-Nonresponse").

Stichprobenreduktion durch fehlende Werte

Durch die vorhandenen fehlenden Werte können sich je nach Auswertungsmethodik zum Teil erhebliche Verringerungen der verwertbaren Stichproben ergeben. Die Auswirkungen verschiedener Anforderungen an die Vollständigkeit der einzubeziehenden Daten sind in Tabelle 10 für die Gesamtstichprobe dargestellt.

Bei gleichzeitiger Berücksichtigung aller relevanten Variablen ergäbe sich demnach für den vorliegenden Datensatz bei listenweisem Fallausschluss eine resultierende Untersuchungsstichprobe von nur noch n=389 Patienten, für die zu 100 Prozent vollständig verwertbare Daten in allen Erhebungsmodulen vorliegen. Bei Anwendung des gebräuchlichen Mindestkriteriums von 70 Prozent gültigen Werten (Wirtz, 2004), ergibt sich hingegen eine Stichprobe von immerhin 1.747 Patienten.

Bei Anwendung des in Kapitel 4.5 definierten Auswahlkriteriums zur Bildung der Ausgangsstichprobe für die im Zentrum der vorliegenden Arbeit stehende Simulationsstudie (vgl. Kapitel 5.2), demzufolge für jeden berücksichtigten Fall in jedem einzelnen der zu den verschiedenen Erhebungszeitpunkten eingesetzten Instrumente (SCL-14, SF-8, IIP-D, FLZ, HoNOS-D usw.) jeweils mindestens 80 Prozent gültige Werte vorliegen müssen, resultiert ein Stichprobenumfang von n=1.248.

Tabelle 10: Stichprobenumfänge nach Vollständigkeit der Daten

N	**Patient T0**	**Therapeut T0**	**Patient T1**	**Therapeut T1**	**Gesamt T0-T1**
Kriterium					
ohne Einschränkung	2.161	2.161	2.161	2.161	2.161
mind. 70% gültige Werte	1.955	2.106	1.942	2.047	1.747
mind. 80% gültige Werte	1.937	2.100	1.927	2.044	1.710
mind. 90% gültige Werte	1.886	2.048	1.871	2.027	1.586
mind. 100% gültige Werte	624	1.686	1.639	1.809	389
jeweils mind. 80% gültige Werte in jedem einzelnen Instrument (SCL-14, SF-8, HoNOS-D usw.)	1.573	1.958	1.826	2.011	1.248

T0: Aufnahme, T1: Entlassung

Weitere Angaben zu den Fallzahlen, die sich je nach definiertem Vollständigkeitskriterium für die einzelnen Kliniken ergeben würden, finden sich in den Abschnitten 5.2. (Simulationsstudie; Tabelle 10) und 5.3 (Einrichtungsvergleich; Tabelle 25).

5.2. Simulationsstudie zur Ersetzbarkeit fehlender Werte

Im Rahmen der durchgeführten Simulationsstudie wurden in einem vollständigen Datensatz nach zwei definierten Kriterien (Zufall vs. Systematik; vgl. Kapitel 4.5) unterschiedliche Dropoutquoten bezüglich des Vorliegens der relevanten Entlassungsdaten simuliert. Die fehlenden Werte wurden daraufhin unter Berücksichtigung verschiedener Rahmenbedingungen ersetzt. Bevor auf die Güte der in den insgesamt 6.600 verschiedenen Ersetzungsszenarien realisierten Fehlwertersetzungen eingegangen wird, werden im vorliegenden Abschnitt zunächst die Auswirkungen der Dropoutsimulation auf die resultierenden Stichprobengrößen sowie die zentralen Stichprobenkennwerte berichtet.

Stichprobengrößen in der Fehlwertsimulation

Wie in Kapitel 4.5 beschrieben, bildeten die vollständigen Daten einer Teilstichprobe von N=1.248 Patienten die Basis für die vorliegende Simulationsstudie. Die klinikbezogenen Stichprobengrößen schwanken zwischen n=54 (Klinik 11) und n=189 (Klinik 6; vgl. Tabelle 11). Die je nach simulierter Dropoutquote resultierenden einrichtungsbezogenen Fallzahlen sind in Tabelle 11 aufgeführt.

Tabelle 11: Klinikbezogene Fallzahlen mit vollständigen Daten nach Dropoutsimulation

Klinik / Dropout	1	2	3	4	5	6	7	8	9	10	11	GES
0%	119	146	155	121	83	189	66	75	128	112	54	1248
10%	107	131	139	109	75	170	59	67	115	101	49	1122
20%	95	117	124	97	66	151	53	60	102	90	43	998
30%	83	102	108	85	58	132	46	52	90	78	38	872
40%	71	88	93	73	50	113	40	45	77	67	32	749
50%	59	73	77	60	41	94	33	37	64	56	27	621

Auswirkungen der Fehlwertsimulation auf die zentralen Stichprobenkennwerte

In Tabelle 12 sind die durch den simulierten Dropout verursachten Abweichungen in den Ergebnisdaten zum Entlassungszeitpunkt im Vergleich zur Originalstichprobe im Überblick dargestellt. Die detaillierten Verteilungskennwerte der Ergebnismaße für die einzelnen Dropoutvarianten finden sich im Anhang (Kapitel 12.2) in Tabelle 28 (Simulation zufällig fehlender Werte) und Tabelle 29 (Simulation systematisch fehlender Werte).

Durch die Simulation zufällig fehlender Werte ergeben sich für die einzelnen Einrichtungen bezüglich der Verteilungskennwerte der verbleibenden Entlassungsdaten nahezu keine signifikanten Abweichungen von der Originalverteilung (vgl. Tabelle 12). Im Extremfall resultiert das zufällige Fehlen von 50 Prozent der vorhandenen Fälle in der Klinik mit der kleinsten Stichprobengröße (Klinik 11) allerdings bereits in (nicht signifikanten) Abweichungen von immerhin kleiner Effektstärke.

Im Falle systematisch eliminierter Werte finden sich demgegenüber deutlich häufiger Abweichungen von der Originalwerteverteilung. Lediglich in einer Klinik zeigen sich diese Abweichungen weitgehend unabhängig vom Ausmaß des Datenausfalls (Klinik 3), in anderen Kliniken ergeben sich erst mit zunehmenden Dropoutquoten signifikante Abweichungen (z.B. Klinik 1 und 2). Für zahlreiche Kliniken finden sich in den Entlassungswerten jedoch trotz systematischen Datenausfalls keine Abweichungen zur Originalwerteverteilung. Der Eliminierungssystematik entsprechend (systematische Auswahl von Fällen mit nicht-positivem Behandlungsverlauf), fallen alle signifikanten Abweichungen jeweils in Richtung einer geringeren Beeinträchtigung zum Entlassungszeitpunkt aus. Sollten die untersuchten Kliniken also anhand der vorliegenden Entlassungswerte beurteilt werden, so würde die Bewertung der Einrichtungen in diesen Fällen jeweils durch die systematischen Datenausfälle zugunsten der jeweiligen Einrichtung verzerrt.

Entsprechende Statistiken zu den Auswirkungen der Dropoutsimulation auf die Verteilungskennwerte der Aufnahmedaten finden sich in Tabelle 13 bzw. Tabelle 30 (Simulation zufällig fehlender Werte) und Tabelle 31 (Simulation systematisch fehlender Werte) im Anhang (Kapitel 12.2). Bezüglich der Aufnahmedaten zeigt sich, dass die zufällige Eliminierung von Fällen zu keinen signifikanten Verteilungsunterschieden gegenüber den Originaldaten führt (vgl. Tabelle 13). Demgegenüber ergeben sich, insbesondere im Vergleich zu den Auswirkungen der Fehlwertsimulation auf die Repräsentativität der Entlassungsdaten, bei systematischem Datenausfall hinsichtlich der Aufnahmedaten deutlich häufiger signifikante Abweichungen von der Originalwerteverteilung. Allerdings fällt auch hier auf, dass sich im Falle geringer Fehlwertquoten (10-20%) nahezu keine nachweisbaren Abweichungen zeigen.

Würde man die Repräsentativität der vorliegenden Aufnahmedaten – wie in den gängigen Dropoutanalysen üblich – als Maßstab für die Bewertung der Repräsentativität der auf Basis der vollständig vorliegenden Entlassungsdaten ermittelten Befunde heranziehen, hätte dies im Falle systematischer Datenausfälle somit erhebliche Konsequenzen: Zum einen würde die Repräsentativität ermittelter Entlassungskennwerte bei höheren Dropoutquoten (30-50%) vielfach zu Unrecht in Frage gestellt, weil sich zwar in den Aufnahmedaten Abweichungen zwischen Respondern und Dropouts ergeben, bezüglich der entsprechenden Entlassungsdaten jedoch keine Unterschiede bestehen. Zum anderen würde die Repräsentativität der ermittelten Behandlungsverläufe (Prä-post-Differenzen)

Tabelle 12: Unterschiede zwischen Original- und Dropout-Stichproben bezüglich der Ausprägung der Ergebnismaße (Patientenangaben zur Entlassung)

Klinik **Dropoutvarianten**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**GES**
ZUFÄLLIG												
GSI (SCL-14)												
10%	o	o	o	o	o	o	o	o	o	o	o	o
20%	o	o	o	o	o	-	o	o	o	o	o	o
30%	o	o	o	o	o	o	o	o	o	o	o	o
40%	o	o	o	o	o	o	o	o	o	o	o	o
50%	o	o	o	o	o	o	o	o	o	o	(++)	o
PSK (SF-8)												
10%	o	o	o	o	o	o	o	o	o	o	o	o
20%	o	o	o	o	o	o	o	o	o	o	o	o
30%	o	o	o	o	o	o	o	o	o	o	o	o
40%	o	o	o	o	o	o	o	o	o	o	o	+
50%	o	o	o	o	o	o	o	o	o	o	(++)	o
SYSTEMATISCH												
GSI (SCL-14)												
10%	o	o	--	o	o	-	o	o	o	o	o	o
20%	o	o	--	o	o	o	o	o	o	o	o	o
30%	--	o	--	o	o	o	o	o	o	o	o	-
40%	--	o	--	o	o	o	o	o	o	o	o	-
50%	--	--	--	o	o	o	o	o	o	--	o	-
PSK (SF-8)												
10%	o	o	o	o	o	o	o	o	o	o	o	o
20%	o	o	o	o	o	o	o	o	o	o	o	o
30%	o	o	--	o	o	o	o	o	o	o	o	o
40%	o	o	--	o	o	o	--	o	o	o	o	-
50%	o	(--)	--	o	o	o	(--)	(++)	o	o	o	-

++: signifikant höhere Ausprägung in Dropout-Stichprobe mit mindestens kleiner Effektstärke, (**+**+): nicht signifikant höhere Ausprägung in Dropout-Stichprobe mit mindestens kleiner Effektstärke, **+**: signifikant höhere Ausprägung in Dropout-Stichprobe mit geringerer als kleiner Effektstärke, **o**: kein signifikanter Unterschied zwischen Original- und Dropout-Stichprobe, **-**: signifikant geringere Ausprägung in Dropout-Stichprobe mit geringerer als kleiner Effektstärke, (**--**): nicht signifikant geringere Ausprägung in Dropout-Stichprobe mit mindestens kleiner Effektstärke, **--**: signifikant geringere Ausprägung in Dropout-Stichprobe mit mindestens kleiner Effektstärke

Tabelle 13: Unterschiede zwischen Original- und Dropout-Stichproben bezüglich der Ausprägung der Basisdaten (Patientenangaben zur Aufnahme)

Klinik **Dropoutvarianten**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**GES**
ZUFÄLLIG												
GSI (SCL-14)												
10%	o	o	o	o	o	o	o	o	o	o	o	o
20%	o	o	o	o	o	o	o	o	o	o	o	o
30%	o	o	o	o	o	o	o	o	o	o	o	o
40%	o	o	o	o	o	o	o	o	o	o	o	o
50%	o	o	o	o	o	o	o	o	o	o	(++)	o
PSK (SF-8)												
10%	o	o	o	o	o	o	o	o	o	o	o	o
20%	o	o	o	o	o	o	o	o	o	o	o	o
30%	o	o	o	o	o	o	o	o	o	o	o	o
40%	o	o	o	o	o	o	o	o	o	o	o	o
50%	o	o	o	o	o	o	o	o	o	o	o	o
SYSTEMATISCH												
GSI (SCL-14)												
10%	o	o	o	o	o	o	o	o	o	o	o	o
20%	o	o	o	o	o	o	o	o	o	o	o	+
30%	++	++	++	++	o	+	(++)	o	o	o	o	+
40%	++	++	++	++	o	++	(++)	(++)	++	(++)	o	++
50%	++	++	++	++	(++)	++	++	++	++	++	(++)	++
PSK (SF-8)												
10%	o	o	o	o	o	o	o	o	o	o	o	o
20%	o	o	o	o	o	o	o	o	o	o	o	+
30%	o	o	o	(++)	o	o	o	o	o	o	o	+
40%	++	o	o	o	o	++	o	o	o	(++)	o	+
50%	++	o	o	++	(++)	++	++	(++)	++	(++)	++	++

++: signifikant höhere Ausprägung in Dropout-Stichprobe mit mindestens kleiner Effektstärke, (**++**): nicht signifikant höhere Ausprägung in Dropout-Stichprobe mit mindestens kleiner Effektstärke, **+**: signifikant höhere Ausprägung in Dropout-Stichprobe mit geringerer als kleiner Effektstärke, **o**: kein signifikanter Unterschied zwischen Original- und Dropout-Stichprobe, **-**: signifikant geringere Ausprägung in Dropout-Stichprobe mit geringerer als kleiner Effektstärke, (**--**): nicht signifikant geringere Ausprägung in Dropout-Stichprobe mit mindestens kleiner Effektstärke, **--**: signifikant geringere Ausprägung in Dropout-Stichprobe mit mindestens kleiner Effektstärke

bei geringen Fehlwertquoten (10-20%) fälschlicherweise angenommen, da sich trotz systematischen Datenausfalls in den verfügbaren Aufnahmewerten keine Unterschiede zwischen Respondern und Dropouts nachweisen lassen.

Güte der Ersetzung fehlender Werte

Zur Überprüfung der Güte der Fehlwertsersetzungen wurden getrennt für die Resultate aller 6.600 realisierten Ersetzungsvarianten korrigierte Intraklassen-Korrelations-Koeffizienten (ICC) für die Übereinstimmung zwischen den geschätzten und den Original-Werten berechnet. Dabei wurden jeweils zum einen nur diejenigen Fälle berücksichtigt, für die zuvor fehlende Werte simuliert und ersetzt worden sind, um die konkrete Güte der Fehlwertersetzung zu überprüfen. Um ein Maß für die Auswirkungen der Fehlwertersetzung auf die jeweiligen Stichprobenkennwerte zu generieren, wurden außerdem ICC für die Gesamtstichprobe der jeweiligen Einrichtung bestimmt, da diese Stichprobenkennwerte im Falle eines Einrichtungsvergleiches die zentrale Ergebnisgröße darstellen.

Die mittlere korrigierte ICC über alle 6.600 Ersetzungsvarianten beträgt bezogen auf die Fälle mit fehlenden Werten 0,38 bzw. 0,82 bezogen auf die einzelnen Klinikstichproben. Die Verteilungen der ermittelten ICC sind in Abbildung 9 (bezogen auf die Dropout-Stichproben) und Abbildung 10 (bezogen auf die Einrichtungsstichproben) dargestellt. Dabei sind den im Folgenden berücksichtigten korrigierten ICC jeweils auch die korrespondierenden klassischen ICC gegenübergestellt, um die Auswirkungen der ICC-Korrektur nach Hartung (1981) zu veranschaulichen. Es fällt auf, dass die korrigierten ICC im Falle geringer Ersetzungsgüte deutlich höher ausfallen als die klassischen ICC (vgl. Abbildung 9). Im Falle guter Übereinstimmung zwischen geschätzten und Originalwerten ergeben sich demgegenüber nur geringe Abweichungen zwischen den klassischen und den korrigierten ICC, im Falle sehr hoher Ersetzungsgüte wird die Übereinstimmung zwischen geschätzten und Originalwerten durch die korrigierten ICC sogar tendenziell unterschätzt (vgl. Abbildung 10). Dementsprechend empfiehlt sich bei der Beurteilung einer einzelnen Ersetzungsvariante gegebenenfalls die simultane Berücksichtigung der jeweiligen korrigierten *und* klassischen ICC.

Auf der Ebene der Dropout-Stichproben kann in nur 45 der 6.600 überprüften Ersetzungsvarianten (0,7%) eine gute Übereinstimmung mit den Originaldaten nachgewiesen werden (ICC ≥ 0,70). Allerdings findet sich darunter keine einzige (!) Ersetzungsvariante, die für alle 11 untersuchten Einrichtungen in guten Ersetzungsergebnissen resultieren würde. In immerhin 850 Ersetzungsvarianten (12,9%) werden grenzwertig gute Ersetzungsergebnisse erzielt (0,50 ≤ ICC ≤ 0,70). Der größte Anteil der Ersetzungsvarianten (86,4%) resultiert jedoch in dürftigen Übereinstimmungen zwischen geschätzten und wahren Werten (ICC < 0,50). Trotz dieser mäßigen Ersetzungsgüte resultieren auf der Ebene der verschiedenen Einrichtungsstichproben immer noch 72,8 Prozent (4.805) der überprüften Ersetzungsvarianten in guten Übereinstimmungskennwerten (ICC ≥ 0,70). Dürftige Übereinstimmungsergebnisse (ICC < 0,50) finden sich auf dieser Ebene in nur 2,3 Prozent (149) der Ersetzungsvarianten.

In Abschnitt 12.3 im Anhang finden sich drei Tabellen, in denen die ermittelten ICC getrennt nach den verschiedenen Ersetzungsvarianten aufgeführt werden. Aus Gründen der Übersichtlichkeit beschränkt sich diese Aufstellung jedoch auf die Darstellung der 600 einrichtungsunabhängigen Varianten, deren Befunde jeweils über die Ergebnisse der elf untersuchten Einrichtungen aggregierten wurden. Eine Datei mit den Befunden aller 6.600 Ersetzungsvarianten kann bei Interesse über den Verfasser dieser Arbeit bezogen werden.

Güte der Fehlwertersetzung nach Ersetzungsmethoden (Hypothese 1)

Die Kennwerte zur Ersetzungsgüte der drei untersuchten Ersetzungsmethoden (Regressionsschätzung, RA; EM-Ersetzung, EM; Multiple Imputation, MI) sind in Tabelle 14 wiedergegeben.

Der Vergleich der drei Methoden ergibt auf der Ebene der Fälle mit fehlenden Werten (Dropout-Substichproben) signifikante Unterschiede zwischen den einzelnen Verfahren ($F=249,879$; $p<0,001$; $eta^2=0,070$, mittlere Effektgröße). In den Einzelvergleichen (geplante Kontraste) erweist sich die Ersetzungsgüte der MI größer als die mittlere Ersetzungsgüte von RA und EM (Hypothese 1a/Dropout-Stichproben: $T_{MI>EM,RA}=10,727$; $p<0,001$). Entgegen der Erwartung führt die EM jedoch zu deutlich

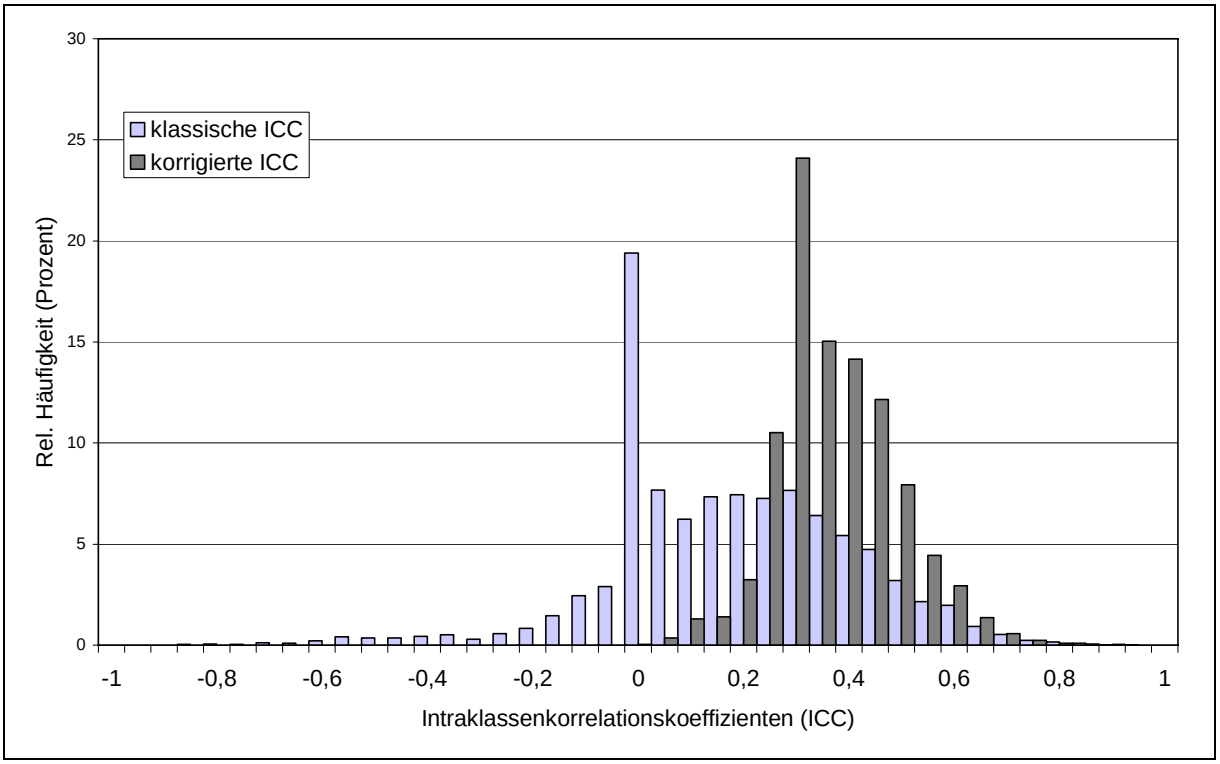

Abbildung 9: Verteilung klassischer und korrigierter Intraklassen-Korrelations-Koeffizienten (ICC) in den 6.600 Ersetzungsvarianten – bezogen auf die Dropout-Stichproben

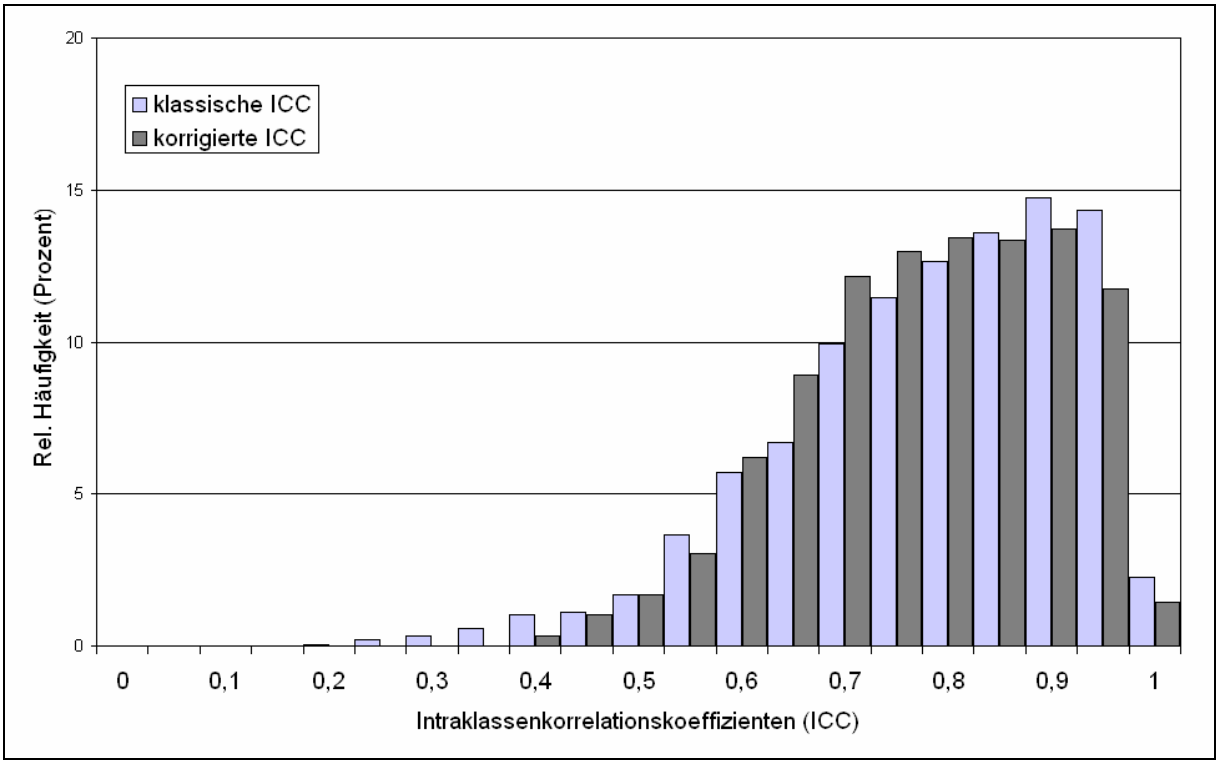

Abbildung 10: Verteilung klassischer und korrigierter Intraklassen-Korrelations-Koeffizienten (ICC) in den 6.600 Ersetzungsvarianten - bezogen auf die Gesamtstichproben

schlechteren Ersetzungsresultaten als die RA (Hypothese 1b/Dropout-Stichproben: $T_{EM>RA}$=-21,955, p<0,001), die sogar vergleichbar gute Ersetzungsresultate liefert wie die MI.

Auch auf der Ebene der jeweiligen Gesamtstichproben lassen sich noch signifikante Unterschiede zwischen den einzelnen Verfahren nachweisen (F=22,801; p<0,001), die jedoch eine geringere als kleine Effektstärke aufweisen (eta^2=0,007). Auch hier zeigen die Einzelvergleiche für MI signifikant höhere ICC als für RA und EM (Hypothese 1a/Einrichtungsstichproben: $T_{MI>EM,RA}$=5,241; p<0,001) und wiederum erwartungswidrig signifikant niedrigere ICC für EM als für RA (Hypothese 1b/Einrichtungsstichproben: $T_{EM>RA}$=-4,297; p<0,001).

Die Betrachtung der Effektgrößen (eta^2) weist darauf hin, dass sich die verschiedenen Ersetzungsmethoden teilweise zwar bedeutsam hinsichtlich ihrer Ersetzungsgüte unterscheiden (Ebene der Dropout-Stichproben), sich diese Unterschiede auf der Ebene der resultierenden Gesamtstichproben jedoch in geringerem Ausmaß niederschlagen. Dieser Befund spiegelt sich beispielsweise auch in den Anteilen guter oder zumindest grenzwertig guter Ersetzungsresultate wieder: Obwohl die MI gemessen an der konkreten Ersetzungsgüte (Ebene der Dropout-Stichproben) vier Mal so viele (zumindest grenzwertig) gute Ersetzungsresultate hervorbringt wie die EM (22,0% vs. 5,5% ICC ≥ 0,5), verschwindet dieser Unterschied auf der Ebene der jeweiligen Gesamtstichproben nahezu vollständig (98,5% vs. 96,5% ICC ≥ 0,5).

Güte der Fehlwertersetzung nach Systematik des Datenausfalls (Hypothese 2)

Die Kennwerte zur Ersetzungsgüte der beiden simulierten Ausfallmechanismen (zufälliger, MCAR, vs. systematischer Datenausfall bei ungünstigem Therapieverlauf, MAR) sind in Tabelle 15 wiedergegeben.

Der Vergleich der beiden Ausfallvarianten ergibt auf der Ebene der Fälle mit fehlenden Werten (Dropout-Substichproben) keine signifikanten Unterschiede bezüglich der Ersetzungsgüte (Hypothese 2/Dropout-Stichproben: F=0,133; $T_{MCAR>MAR}$= -0,365; p=0,715).

Auf der Ebene der jeweiligen Gesamtstichproben lassen sich in Abhängigkeit der Ausfallmechanismen signifikante Unterschiede nachweisen, wobei die ICC bei zufällig fehlenden Werten höher ausfallen als bei systematisch fehlenden Werten (Hypothese 2/Einrichtungsstichproben: $F=61{,}308$; $T_{MCAR>MAR}=7{,}830$; $p<0{,}001$). Diese Unterschiede erreichen annähernd die Größe einer kleinen Effektstärke ($eta^2=0{,}009$).

Die Verbindung der Einzelergebnisse zeigt, dass sich die beiden realisierten Ausfallmechanismen zwar nicht bedeutsam auf die durchschnittliche Ersetzungsgüte auswirken (Ebene der Dropout-Stichproben), die Ungenauigkeit der Fehlwertersetzung jedoch auf der Ebene der resultierenden Gesamtstichproben im MAR-Fall zu stärkeren Abweichungen von der wahren Werteverteilung führen als unter MCAR-Bedingungen. Dementsprechend finden sich auf der Ebene der Einrichtungsstichproben in der MCAR-Bedingung für 77,6 Prozent der Ersetzungsvarianten gute Übereinstimmungskennwerte (ICC ≥ 0,7), in der MAR-Bedingung hingegen nur für 68,0 Prozent.

Güte der Fehlwertersetzung nach Dropoutquoten (Hypothese 3)

Die Kennwerte zur Ersetzungsgüte in Abhängigkeit der unterschiedlichen simulierten variablenbezogenen Fehlwertquoten (10/20/30/40/50%) sind in Tabelle 16 aufgeführt.

Der Vergleich der Ersetzungsgüte nach den unterschiedlichen Fehlwertquoten ergibt auf der Ebene der Fälle mit fehlenden Werten (Dropout-Substichproben) keine signifikanten Unterschiede zwischen den einzelnen Dropoutszenarien (Hypothese 3/Dropout-Stichproben: $F=0{,}720$; $p=0{,}578$; $T_{10>20}=-1{,}283$; $T_{20>30}=0{,}597$; $T_{30>40}=0{,}571$; $T_{40>50}=-1{,}108$; alle $p(T)\geq 0{,}199$).

Auf der Ebene der jeweiligen Gesamtstichproben lassen sich in Abhängigkeit der verschiedenen Dropoutquoten jedoch signifikante Unterschiede nachweisen ($F=4164{,}840$; $p<0{,}001$), die sehr hohe Effektstärken aufweisen ($eta^2=0{,}716$). Die Überprüfung der geplanten Kontraste bestätigt die Annahme abnehmender ICC mit steigender Dropoutquote (Hypothese 3/Einrichtungsstichproben: $T_{10>20}=41{,}580$; $T_{20>30}=32{,}536$; $T_{30>40}=24{,}206$; $T_{40>50}=15{,}625$; alle $p<0{,}001$).

Tabelle 14: Güte der Fehlwertersetzung nach Ersetzungsmethoden (korrigierte ICC)

	Regressionsschätzung	EM-Ersetzung	Multiple Imputation
ICC in Dropout-Substichproben			
Mittelwert*	,398	,329	,401
Minimum	,036	,053	,016
25. Perzentil	,326	,278	,306
50. Perzentil	,386	,293	,395
75. Perzentil	,452	,359	,484
Maximum	,922	,847	,896
ICC in Klinik-Gesamtstichproben			
Mittelwert*	,818	,801	,827
Minimum	,403	,401	,372
25. Perzentil	,695	,662	,716
50. Perzentil	,788	,769	,809
75. Perzentil	,883	,873	,888
Maximum	,998	,984	,993

* Die mittleren ICC wurden jeweils auf der Basis Fisher-Z-transformierter Werte bestimmt.

Tabelle 15: Güte der Fehlwertersetzung nach Systematik des Datenausfalls (korrigierte ICC)

	zufällig (MCAR)	systematisch (MAR)
ICC in Dropout-Substichproben		
Mittelwert*	,376	,377
Minimum	,016	,035
25. Perzentil	,291	,287
50. Perzentil	,358	,352
75. Perzentil	,441	,444
Maximum	,922	,896
ICC in Klinik-Gesamtstichproben		
Mittelwert*	,828	,803
Minimum	,392	,372
25. Perzentil	,709	,666
50. Perzentil	,799	,777
75. Perzentil	,884	,878
Maximum	,763	,993

* Die mittleren ICC wurden jeweils auf der Basis Fisher-Z-transformierter Werte bestimmt.

Die Integration dieser Befunde offenbart, dass sich die unterschiedlichen Dropoutquoten zwar nicht direkt auf die Güte der eigentlichen Fehlwertersetzung auswirken (Ebene der Dropout-Substichproben). Allerdings wirken sich die zunächst von den Dropoutquoten unabhängigen Schätzfehler bei größeren Fehlwertquoten stärker verzerrend auf die resultierenden Stichprobenkennwerte aus als bei geringeren Fehlwertquoten. Übertragen auf die Anteile guter Übereinstimmung (ICC ≥ 0,7) drückt sich dies auf der Ebene der Einrichtungsstichproben in einer beträchtlichen Spanne von 100 Prozent guter Ersetzungsergebnisse (bei 10% fehlenden Werten) bis hinab zu 25,5 Prozent (bei 50% fehlenden Werten) aus.

Güte der Fehlwertersetzung nach berücksichtigten Kovariaten (Hypothese 4)

Die Kennwerte zur Ersetzungsgüte in Abhängigkeit der verschiedenen Gruppen berücksichtigter Kovariaten (Patientenangaben zu T0 [P], Therapeutenangaben zu T0 und T1 [T/TT]) sind in Tabelle 17 angegeben.

Der Vergleich der Ersetzungsgüte nach den unterschiedlichen Sets einbezogener Kovariaten ergibt auf der Ebene der Fälle mit fehlenden Werten (Dropout-Substichproben) signifikante Unterschiede zwischen den verschiedenen Kovariatengruppen ($F=55,907$; $p<0,001$; $eta^2=0,033$, kleine Effektstärke). Die Überprüfung der geplanten Kontraste bestätigt die Annahme höherer ICC bei Einbezug der patientenseitigen Angaben zu T0 (Hypothese 4a/Dropout-Stichproben: $T_{P,PT,PTT>T,TT}=10,657$; $p<0,001$). Die Annahme kontinuierlich höherer ICC mit steigender Anzahl berücksichtigter Kovariaten lässt sich hingegen nicht stringent belegen (Hypothese 4b/Dropout-Stichproben: $T_{PTT>PT}=2,576$, $p=0,010$; $T_{PT>P}=-0,290$, $p=0,772$; $T_{P>TT}=1,033$, $p=0,302$; $T_{TT>T}=10,298$, $p<0,001$). Die Ergebnisse der Einzelvergleiche deuten allerdings darauf hin, dass der Einbezug der therapeutenseitigen Angaben zu T1 (Entlassungszeitpunkt) jeweils in signifikant höheren ICC resultiert als die Fehlwertschätzung ohne Berücksichtigung dieser Angaben.

Auf der Ebene der jeweiligen Gesamtstichproben lassen sich ebenfalls signifikante Unterschiede in Abhängigkeit der verschiedenen berücksichtigten Kovariatensets nachweisen ($F=35,896$; $p<0,001$; $eta^2=0,021$, kleiner Effekt). Die Überprüfung der geplanten Kontraste bestätigt auch hier die Annahme höherer ICC bei Berücksichti-

gung der patientenseitigen Angaben (Hypothese 4a/Einrichtungsstichproben: $T_{P,PT,PTT>T,TT}$=11,587; p<0,001). Die Annahme höherer ICC-Ausprägungen mit steigender Anzahl berücksichtigter Kovariaten lässt sich wiederum nicht durchgängig belegen (Hypothese 4b/Einrichtungsstichproben: $T_{PTT>PT}$=1,400, p=0,162; $T_{PT>P}$=0,938, p=0,348; $T_{P>TT}$=5,412, p<0,001; $T_{TT>T}$=1,949, p=0,051).

Übertragen auf die Quoten guter Übereinstimmung zwischen ersetzten und Original-Daten (ICC ≥ 0,7) ergibt sich auf der Ebene der Einrichtungsstichproben eine Spanne von lediglich 63,5 Prozent guter Ersetzungsresultate im Falle der Fehlwertersetzung auf der eingeschränkten Basis der Therapeutenangaben zu T0 bis hin zu 79,7 Prozent bei Einbezug aller verfügbaren Kovariaten (Patientenangaben zu T0 plus Therapeutenangaben zu T0 und T1, PTT).

Güte der Fehlwertersetzung nach Anzahl fallbezogen fehlender Werte (Hypothese 5)

Die Kennwerte zur Ersetzungsgüte der beiden im Hinblick auf die Anzahl zu ersetzender Variabeln überprüften Verfahrensvarianten (item- vs. skalenbezogene Fehlwertersetzung) sind in Tabelle 18 wiedergegeben.

Der Vergleich der beiden Varianten zur Anzahl zu ersetzender Variablen zeigt auf der Ebene der Fälle mit fehlenden Werten (Dropout-Substichproben) signifikante Unterschiede bezüglich der Ersetzungsgüte (Hypothese 5/Dropout-Stichproben: F=42,063; $T_{Skala>Items}$=6,486; p<0,001). Die Unterschiede weisen jedoch eine geringere als kleine Effektstärke auf (eta^2=0,006).

Auf der Ebene der jeweiligen Gesamtstichproben lassen sich in Abhängigkeit der Anzahl zu ersetzender Variablen keine signifikanten Unterschiede bezüglich der erreichten ICC nachweisen (Hypothese 5/Einrichtungsstichproben: F=0,639; $T_{Skala>Items}$=-0,799; p=0,424).

Die Verbindung der Einzelergebnisse ergibt, dass sich die beiden realisierten Ersetzungsvarianten zwar geringfügig auf die durchschnittliche Ersetzungsgüte auswirken, wobei sich diese Effekte auf der Ebene der resultierenden Stichprobenkennwerte jedoch nicht niederschlagen.

Tabelle 16: Güte der Fehlwertersetzung nach Fehlwertquoten (korrigierte ICC)

	10%	**20%**	**30%**	**40%**	**50%**
ICC in Dropout-Substichproben					
Mittelwert*	,373	,380	,377	,374	,379
Minimum	,016	,058	,046	,062	,059
25. Perzentil	,276	,294	,288	,289	,289
50. Perzentil	,338	,360	,358	,359	,363
75. Perzentil	,444	,445	,442	,434	,449
Maximum	,922	,769	,789	,728	,733
ICC in Klinik-Gesamtstichproben					
Mittelwert*	,937	,860	,781	,707	,654
Minimum	,746	,604	,487	,409	,372
25. Perzentil	,905	,822	,730	,645	,594
50. Perzentil	,935	,861	,782	,709	,649
75. Perzentil	,956	,890	,825	,757	,701
Maximum	,998	,972	,925	,900	,888

* Die mittleren ICC wurden jeweils auf der Basis Fisher-Z-transformierter Werte bestimmt.

Tabelle 17: Güte der Fehlwertersetzung nach Prädiktoren (Kovariaten) (korrigierte ICC)

	P	**PT**	**PTT**	**TT**	**T**
ICC in Dropout-Substichproben					
Mittelwert*	,386	,385	,397	,381	,322
Minimum	,078	,121	,114	,016	,025
25. Perzentil	,291	,291	,291	,300	,269
50. Perzentil	,361	,363	,373	,372	,328
75. Perzentil	,449	,450	,471	,451	,388
Maximum	,922	,816	,854	,847	,786
ICC in Klinik-Gesamtstichproben					
Mittelwert*	,825	,830	,836	,797	,786
Minimum	,424	,424	,424	,403	,372
25. Perzentil	,704	,711	,720	,662	,647
50. Perzentil	,797	,804	,812	,766	,754
75. Perzentil	,884	,889	,894	,865	,859
Maximum	,998	,989	,990	,985	,987

P: Patientenangaben T0, PT: Patienten- und Therapeutenangaben T0, PTT: Patientenangaben T0 plus Therapeutenangaben T0+T1, TT: Therapeutenangaben T0+T1, T: Therapeutenangaben T0

* Die mittleren ICC wurden jeweils auf der Basis Fisher-Z-transformierter Werte bestimmt.

Güte der Fehlwertersetzung nach Ergebnismaßen (Hypothese 6)

Die Kennwerte zur Güte der Ersetzung der beiden überprüften Ergebnismaße (Psychische Summenskala der SF-8, PSK, und Globaler Symptomschwere-Index der SCL-14, GSI) sind in Tabelle 19 wiedergegeben.

Der Vergleich der beiden Kriteriumsvarianten ergibt auf der Ebene der Fälle mit fehlenden Werten (Dropout-Substichproben) keine signifikanten Unterschiede bezüglich der Ersetzungsgüte (Hypothese 6/Dropout-Stichproben: F=1,854; $T_{PSK>GSI}$= -1,361; p=0,173).

Auf der Ebene der jeweiligen Gesamtstichproben lässt sich bezüglich der erreichten ICC zwar ein signifikanter Unterschied zugunsten der PSK nachweisen (Hypothese 6/Einrichtungsstichproben: F=8,404; $T_{PSK>GSI}$=2,899; p=0,004), die Effektgröße dieses Unterschieds liegt jedoch nahe an Null (eta^2=0,001).

Güte der Fehlwertersetzung nach Einrichtungen (Hypothese 7)

Die Kennwerte zur Ersetzungsgüte im Vergleich der elf untersuchten Einrichtungen sind in Tabelle 20 wiedergegeben.

Der Einrichtungsvergleich ergibt auf der Ebene der Fälle mit fehlenden Werten (Dropout-Substichproben) signifikante Unterschiede bezüglich der Ersetzungsgüte (Hypothese 7/Dropout-Stichproben: F=27,585; p<0,001; eta^2=0,040, kleine Effektstärke). Die durchgeführten post hoc Einzelvergleiche ergeben keine Hinweise auf einen systematischen Einfluss der einrichtungsbezogenen Stichprobengrößen auf das Ergebnis der Fehlwertersetzung.

Unterschiede zwischen den Einrichtungen finden sich auch auf der Ebene der jeweiligen Gesamtstichproben (Hypothese 7/Einrichtungsstichproben: F=10,087; p<0,001; eta^2=0,015, kleine Effektstärke). Auch hier ergeben die Einzelvergleiche keine Hinweise auf einen systematischen Effekt der einrichtungsbezogenen Stichprobengrößen.

Dementsprechend schwanken die Anteile guter Ersetzungsresultate (ICC ≥ 0,7) auf der Ebene der verschiedenen Gesamtstichproben je nach Einrichtung zwischen 63,2 Prozent (Klinik 1) und 84,0 Prozent (Klinik 4).

Tabelle 18: Güte der Fehlwertersetzung nach Komplexität des Kriteriums (korrigierte ICC)

	multipel (Items)	**einfach (Skala)**
ICC in Dropout-Substichproben		
Mittelwert*	,367	,386
Minimum	,016	,035
25. Perzentil	,286	,291
50. Perzentil	,350	,361
75. Perzentil	,434	,450
Maximum	,847	,922
ICC in Klinik-Gesamtstichproben		
Mittelwert*	,817	,814
Minimum	,372	,401
25. Perzentil	,694	,688
50. Perzentil	,792	,786
75. Perzentil	,882	,879
Maximum	,991	,998

* Die mittleren ICC wurden jeweils auf der Basis Fisher-Z-transformierter Werte bestimmt.

Tabelle 19: Güte der Fehlwertersetzung nach Ergebnismaßen (korrigierte ICC)

	GSI der SCL-14	**PSK der SF-8**
ICC in Dropout-Substichproben		
Mittelwert*	,379	,374
Minimum	,016	,036
25. Perzentil	,286	,291
50. Perzentil	,356	,356
75. Perzentil	,457	,431
Maximum	,896	,922
ICC in Klinik-Gesamtstichproben		
Mittelwert*	,811	,820
Minimum	,372	,403
25. Perzentil	,683	,698
50. Perzentil	,787	,792
75. Perzentil	,877	,885
Maximum	,993	,998

* Die mittleren ICC wurden jeweils auf der Basis Fisher-Z-transformierter Werte bestimmt.

Tabelle 20: Güte der Fehlwertersetzung nach Kliniken (korrigierte ICC)

	1	2	3	4	5	6	7	8	9	10	11
ICC in Droout-Substichprobe											
Mittelwert*	,349	,386	,337	,383	,371	,384	,347	,368	,387	,427	,400
Minimum	,046	,066	,039	,062	,031	,072	,025	,036	,072	,108	,016
25. Perzentil	,284	,293	,278	,291	,286	,329	,285	,274	,293	,292	,280
50. Perzentil	,330	,373	,323	,373	,361	,373	,331	,329	,367	,413	,351
75. Perzentil	,402	,454	,392	,450	,449	,430	,403	,439	,450	,509	,495
Maximum	,737	,842	,616	,823	,743	,634	,896	,789	,847	,786	,922
ICC in Klinik-Gesamtstichpr.											
Mittelwert*	,791	,827	,784	,825	,819	,795	,810	,821	,831	,834	,829
Minimum	,407	,505	,392	,382	,452	,372	,464	,403	,468	,485	,401
25. Perzentil	,654	,720	,656	,735	,692	,677	,669	,709	,687	,750	,677
50. Perzentil	,769	,803	,753	,806	,777	,758	,760	,793	,810	,819	,787
75. Perzentil	,874	,893	,860	,873	,879	,868	,867	,884	,899	,889	,889
Maximum	,963	,991	,969	,985	,987	,973	,993	,984	,979	,974	,998

* Die mittleren ICC wurden jeweils auf der Basis Fisher-Z-transformierter Werte bestimmt.

Zusammenfassung der hypothesengeleitet ermittelten Befunde zur Ersetzungsgüte

Die Ergebnisse bezüglich der einzelnen getesteten Hypothesen zur Güte der Fehlwertersetzung und ihrer Auswirkungen auf der Ebene der resultierenden Gesamtstichproben sind in Tabelle 21 nochmals zusammenfassend wiedergegeben. In der Übersicht der Befunde fällt auf, dass die variablenbezogenen Fehlwertquoten den größten Einfluss auf die Qualität der resultierenden Stichprobenkennwerte zeigen, obgleich sie die eigentliche Güte der Fehlwertersetzung nicht systematisch beeinflussen. Der Einfluss der überprüften Imputationsverfahren auf das Ersetzungsergebnis erreicht zwar die Größe einer mittleren Effektstärke, ihr Einfluss auf die Qualität der resultierenden Stichprobenkennwerte fällt demgegenüber jedoch nur sehr gering aus. Einen konstanten Einfluss, sowohl auf die Güte der Fehlwertersetzung, als auch auf die Güte der resultierenden Stichprobenkennwerte, zeigt die Auswahl der einbezogenen Kovariaten - wenn auch nur mit kleiner Effektstärke. Gleiches gilt für die Effekte, die auf die unterschiedlichen Einrichtungen zurückzu-

führen sind, wobei jedoch nicht bekannt ist, wodurch diese Unterschiede konkret verursacht werden. Die unterschiedlichen Ausfallmechanismen wirken sich geringfügig auf die Qualität der resultierenden Stichprobenkennwerte aus, die Ersetzungsgüte wird durch sie jedoch nicht beeinflusst. Demgegenüber wirkt sich die Anzahl fallbezogen zu ersetzender Werte zwar geringfügig auf die Ersetzungsgüte aus, es lassen sich jedoch keine Effekte auf die aus der Ersetzung resultierende Qualität der Stichprobenkennwerte nachweisen.

Tabelle 21: Übersicht zu den Ergebnissen der Hypothesentestung

Überprüfte Alternativhypothesen (H_1) **Nr.**	**Hypothesen**	**Bestätigung der H_1 auf Ebene der Dropout-Stichproben**	**Bestätigung der H_1 auf Ebene der Klinik-Stichproben**
1a	$\overline{ICC}_{MI} - 0{,}5 \cdot \left(\overline{ICC}_{EM} + \overline{ICC}_{RA}\right) > 0$	✓✓✓	✓
1b	$\overline{ICC}_{EM} - \overline{ICC}_{RA} > 0$	ø*	ø*
2	$\overline{ICC}_{MCAR} - \overline{ICC}_{MAR} > 0$	ø	✓
3	$\overline{ICC}_{10\%} - \overline{ICC}_{20\%} > 0 \cap \overline{ICC}_{20\%} - \overline{ICC}_{30\%} > 0 \cap \overline{ICC}_{30\%} - \overline{ICC}_{40\%} > 0 \cap \overline{ICC}_{40\%} - \overline{ICC}_{50\%} > 0$	ø	✓✓✓✓
4a	$\left(\overline{ICC}_{P} + \overline{ICC}_{PT} + \overline{ICC}_{PTT}\right) - 1{,}5 \cdot \left(\overline{ICC}_{T} + \overline{ICC}_{TT}\right) > 0$	✓✓	✓✓
4b	$\overline{ICC}_{PTT} - \overline{ICC}_{PT} > 0 \cap \overline{ICC}_{PT} - \overline{ICC}_{P} > 0 \cap \overline{ICC}_{P} - \overline{ICC}_{TT} > 0 \cap \overline{ICC}_{TT} - \overline{ICC}_{T} > 0$	ø	ø
5	$\overline{ICC}_{Skala} - \overline{ICC}_{Items} > 0$	✓	ø
6	$\overline{ICC}_{PSK} - \overline{ICC}_{GSI} > 0$	ø	✓
7	$\overline{ICC}_{Klinik\,i} - \overline{ICC}_{Klinik\,j} \neq 0$	✓✓	✓✓

ERSETZUNGSMETHODEN: **MI** = Multiple Imputation, **EM** = Expectation Maximization, **RA** = Regressionsschätzung; *AUSFALLMECHANISMEN*: **MCAR** = zufällig fehlende Werte, **MAR** = systematisch fehlende Werte; *DROPOUTQUOTEN*: **10/20/30/40/50%** = Anteile variablenbezogen fehlender Werte; *KOVARIATEN*: **P** = Patientenangaben zu T0 (Aufnahme), **T** = Therapeutenangaben zu T0, **TT** = Therapeutenangaben zu T0 und T1 (Aufnahme und Entlassung); *ERSETZUNGSEBENE*: **Skala** = Schätzung des Skalenwertes, **Items** = Schätzung der einzelnen Itemwerte; *KRITERIUMSVARIABLEN*: **PSK** = Psychische Summenskala der SF-8, **GSI** = Globaler Symptomschwere-Index der SCL-14; *EINRICHTUNGEN*: **Klinik$_{i,j}$** = untersuchte Einrichtungen.

✓/✓✓/✓✓✓/✓✓✓✓: Bestätigung der Alternativhypothese mit geringerer als kleiner/kleiner/mittlerer/großer Effektstärke; **ø**: Ablehnung der Alternativhypothese; **ø***: Unterschied in entgegengesetzter Richtung

Integration der Befunde zur Ersetzungsgüte

Kombiniert man die einzelnen Befunde zur Güte der Fehlwertersetzung in Abhängigkeit der verschiedenen überprüften Rahmenbedingungen, so wäre davon auszugehen, dass sich die besten Ersetzungsresultate unter der Bedingung zufällig fehlender Werte, einer geringen variablen- wie fallbezogenen Fehlwertquote sowie einer möglichst großen Anzahl geeigneter Kovariaten bei Anwendung der Multiplen Imputation ermitteln lassen. Demgegenüber wäre bei systematisch fehlenden Werten, hohen Fehlwertquoten, begrenzt zweckmäßigen Kovariaten und der Anwendung der EM-Methode mit einem vergleichsweise schlechten Ersetzungsresultat zu rechnen.

In Tabelle 22 sind die Ersetzungsresultate dieser beiden Extremszenarien aufgeführt. Im Mittel resultiert die Fehlwertersetzung unter den günstigsten Rahmenbedingungen bereits auf der Ebene der Fälle mit fehlenden Werten (Dropout-Stichproben) in annähernd grenzwertig guten Übereinstimmungen mit den Originaldaten (ICC = 0,46). Auf Gesamtstichprobenebene liegen die Übereinstimmungskoeffizienten durchwegs im guten Bereich (ICC ≥ 0,7) und sogar im Durchschnitt nahe an maximaler Übereinstimmung (ICC = 0,95). Im Falle der ungünstigsten Rahmenbedingungen hingegen erreicht auf der Ebene der Dropoutfälle keine der Ersetzungsvarianten auch nur ein grenzwertig gutes Ergebnis (alle ICC < 0,5). Und sogar auf der Ebene der Gesamtstichproben erreicht die Übereinstimmungsgüte im Durchschnitt nur knapp ein grenzwertig gutes Ausmaß (ICC ≥ 0,5), keiner der Übereinstimmungskoeffizienten liegt im guten Bereich (alle ICC < 0,7).

Während Anteile fehlender Werte und Ausfallmechanismen in der Forschungspraxis nur schwerlich zu kontrollieren sind, lassen sich die übrigen Rahmenbedingungen durch eine geeignete Untersuchungsplanung sehr wohl beeinflussen. Im Folgenden sollte daher überprüft werden, wie sich die Ersetzungsgüte im Falle der optimalen Gestaltung kontrollierbarer Rahmenbedingungen für die verschiedenen möglichen Varianten der unkontrollierbaren Rahmenbedingungen entwickelt. Konkret wurde daher untersucht, zu welchen Ersetzungsresultaten die Multiple Imputation bei Einbezug aller verfügbaren Kovariaten und skalenweiser Fehlwertersetzung in Abhängigkeit der verschiedenen simulierten Dropoutquoten und Fehlwertmechanismen führt (auf die zusätzliche Berücksichtigung des zu ersetzenden Kriteriums wurde

Tabelle 22: Güte der Fehlwertersetzung nach günstigsten vs. ungünstigsten Rahmenbedingungen (korrigierte ICC)

	„Best Case“	**„Worst Case“**
Methode	Multiple Imputation	EM-Ersetzung
Ausfallmechanismus	MCAR	MAR
Fehlwertquote	10%	50%
Ersetzungsebene	Skala	Items
Kovariaten	PTT	T
Kriterium	PSK	GSI
ICC in Dropout-Substichproben		
Mittelwert*	0,464	0,318
Minimum	0,294	0,164
Median	0,427	0,292
Maximum	0,756	0,448
ICC in Klinik-Gesamtstichproben		
Mittelwert*	0,951	0,546
Minimum	0,906	0,401
Median	0,947	0,537
Maximum	0,977	0,691

***** Die mittleren ICC wurden jeweils auf der Basis Fisher-Z-transformierter Werte bestimmt. **MCAR** = zufällig fehlende Werte, **MAR** = systematisch fehlende Werte; **PTT** = Patientenangaben zu T0 (Aufnahme) und Therapeutenangaben zu T0 und T1 (Aufnahme und Entlassung), **T** = Therapeutenangaben zu T0; **Skala** = Schätzung des Skalenwertes, **Items** = Schätzung der einzelnen Itemwerte; **PSK** = Psychische Summenskala der SF-8, **GSI** = Globaler Symptomschwere-Index der SCL-14

angesichts des geringen Effektes auf das Ersetzungsergebnis verzichtet). Die entsprechenden Kennwerte lassen sich der Tabelle 23 entnehmen.

Es zeigt sich, dass die MI bei skalenweiser Fehlwertersetzung unter Einbezug aller verfügbaren Kovariaten auf der Ebene der Dropout-Stichproben weitestgehend unabhängig von den weiteren Rahmenbedingungen zu Ersetzungsresultaten führt, deren Güte im Mittel im unteren Bereich grenzwertig guter Übereinstimmung liegt (ICC um 0,5). Auf der Ebene der Einrichtungsstichproben ergeben sich zwar deutliche Effekte der Fehlwertquote, die Übereinstimmung mit den Originaldaten fällt durchschnittlich jedoch sogar bei 50 Prozent fehlenden Werten noch gut aus (ICC > 0,7).

Tabelle 23: Güte der Fehlwertersetzung nach Ausfallmechanismen und Fehlwertquoten bei skalenweiser Ersetzung unter Einbezug aller verfügbaren Kovariaten mittels Multipler Imputation (mittlere korrigierte ICC [mit 95%-Konfidenzintervall])

	MCAR	**MAR**
ICC in Dropout-Substichproben		
10% Fehlwertquote	0,456 [0,370;0,535]	0,550 [0,478;0,614]
20% Fehlwertquote	0,521 [0,467;0,571]	0,529 [0,492;0,564]
30% Fehlwertquote	0,511 [0,475;0,545]	0,507 [0,456;0,554]
40% Fehlwertquote	0,479 [0,443;0,514]	0,517 [0,467;0,564]
50% Fehlwertquote	0,466 [0,431;0,499]	0,500 [0,459;0,538]
ICC in Klinik-Gesamtstichproben		
10% Fehlwertquote	0,953 [0,944;0,961]	0,954 [0,941;0,964]
20% Fehlwertquote	0,899 [0,887;0,910]	0,905 [0,889;0,919]
30% Fehlwertquote	0,852 [0,837;0,865]	0,825 [0,797;0,849]
40% Fehlwertquote	0,787 [0,764;0,809]	0,769 [0,735;0,800]
50% Fehlwertquote	0,736 [0,714;0,757]	0,715 [0,683;0,745]

***** Die mittleren ICC und die Konfidenzintervalle wurden jeweils auf der Basis Fisher-Z-transformierter Werte bestimmt.
MCAR = zufällig fehlende Werte, **MAR** = systematisch fehlende Werte

Da der EM-Algorithmus in der aktuellen Literatur noch häufig als bevorzugte Methode zum Umgang mit fehlenden Werten empfohlen wird (z.B. Little & Rubin, 2002), die sich aufgrund der weiten Verbreitung der Statistiksoftware SPSS (z.B. Norusis, 2004) auch durch eine hohe Verfügbarkeit auszeichnet, die für das Verfahren der Multiplen Imputation (noch) nicht gegeben ist, soll im Folgenden ergänzend berichtet werden, wie sich die Ersetzungsgüte unter identisch optimalen Rahmenbedingungen bei Anwendung der EM darstellt (vgl. Tabelle 24).

Der Vergleich von EM und MI ergibt, dass die EM-Ersetzung auf der Ebene der Dropout-Stichproben in erheblich schlechteren Übereinstimmungskoeffizienten resultiert, die sich, insbesondere im Falle höherer Fehlwertquoten bei systematischem Datenausfall, auch noch auf der Ebene der Einrichtungsstichproben in deutlich geringeren ICC niederschlagen.

Tabelle 24: Güte der Fehlwertersetzung nach Ausfallmechanismen und Fehlwertquoten bei skalenweiser Ersetzung unter Einbezug aller verfügbaren Kovariaten mittels EM-Ersetzung (mittlere korrigierte ICC [mit 95%-Konfidenzintervall])

	MCAR	**MAR**
ICC in Dropout-Substichproben		
10% Fehlwertquote	0,288 [0,273;0,304]	0,283 [0,258;0,309]
20% Fehlwertquote	0,295 [0,287;0,303]	0,283 [0,267;0,299]
30% Fehlwertquote	0,289 [0,277;0,300]	0,273 [0,262;0,284]
40% Fehlwertquote	0,289 [0,283;0,295]	0,275 [0,261;0,288]
50% Fehlwertquote	0,286 [0,276;0,295]	0,271 [0,264;0,279]
ICC in Klinik-Gesamtstichproben		
10% Fehlwertquote	0,952 [0,936;0,963]	0,926 [0,906;0,941]
20% Fehlwertquote	0,861 [0,847;0,874]	0,857 [0,830;0,879]
30% Fehlwertquote	0,803 [0,782;0,822]	0,746 [0,714;0,776]
40% Fehlwertquote	0,732 [0,710;0,753]	0,653 [0,619;0,684]
50% Fehlwertquote	0,662 [0,640;0,684]	0,584 [0,555;0,612]

***** Die mittleren ICC und die Konfidenzintervalle wurden jeweils auf der Basis Fisher-Z-transformierter Werte bestimmt.
MCAR = zufällig fehlende Werte, **MAR** = systematisch fehlende Werte

Ergänzend sei angemerkt, dass die RA unter identischen Bedingungen zwar zu besseren Ersetzungsergebnissen führt als der EM-Algorithmus, sich der MI jedoch als unterlegen erweist (auf eine ausführlichere Darstellung der Befunde soll an dieser Stelle verzichtet werden).

Auswirkungen der Ersetzung fehlender Werte auf Lage- und Dispersionsmaße

Da Intraklassenkorrelationskoeffizienten (ICC) für sich genommen eine verhältnismäßig abstrakte Größe zur Beurteilung der Ersetzungsgüte darstellen, soll abschließend noch beispielhaft demonstriert werden, welche Auswirkungen die Fehlwertersetzung konkret auf die jeweiligen Stichprobenkennwerte (Mittelwerte und Standardabweichungen) zeigt. Hierzu sind in Abbildung 11 und Abbildung 12 für die zuvor beschriebenen Ersetzungsszenarien unter optimalen Rahmenbedingungen (skalenweise Fehlwertersetzung unter Einbezug aller verfügbaren Kovariaten) exemplarisch die nach Fehlwertersetzung in Abhängigkeit der verschiedenen Dropoutquoten und Fehlwertmechanismen resultierenden Stichprobenkennwerte für den Globalen Symptomschwereindex der SCL-14 abgebildet.

In der grafischen Darstellung der Stichprobenkennwerte wird deutlich, dass die drei eingesetzten Imputationsverfahren (Multiple Imputation, MI; Regressionsschätzung, RA; EM-Algorithmus, EM) unter den gegebenen Rahmenbedingungen grundsätzlich alle in einer verhältnismäßig hohen Übereinstimmung zwischen den fehlwertersetzten und den wahren Stichprobenkennwerten resultieren. Mit zunehmenden Fehlwertquoten finden sich tendenziell immer höhere Abweichungen von den wahren Stichprobenkennwerten. In den meisten Fällen wird die wahre Ausprägung der Stichprobenkennwerte dabei durch die Fehlwertersetzung eher unter- als überschätzt. Bei Anwendung eines Bonferoni-korrigierten α-Niveaus von 0,05 ergibt sich jedoch in lediglich 11 der 330 abgebildeten Ersetzungsvarianten eine signifikante Abweichung zwischen den aus der Fehlwertersetzung resultierenden und den originalen einrichtungsbezogenen Mittelwerten. Es fällt aber auf, dass sich die Einrichtungsmittelwerte zumeist exakter als die entsprechenden Dispersionsmaße (Standardabweichungen) replizieren lassen. Insbesondere die EM resultiert mit zunehmenden Fehlwertquoten – und besonders im Falle systematisch fehlender Werte (MAR) - in einer erheblichen Unterschätzung der wahren Varianz der Stichprobenkennwerte. Die RA führt zwar im Falle zufällig fehlender Werte (MCAR) zu einer sehr guten Abbildung der Stichprobenkennwerte, unter MAR-Bedingungen neigt sie jedoch ebenfalls zu einer deutlicheren Unterschätzung der wahren Stichprobenkennwerte.

Aus der Zusammenschau der Abbildung 11 und Abbildung 12 lässt sich konstatieren, dass die Anwendung der MI im Einzelfall zwar nicht unbedingt die beste Abbildung der wahren Stichprobenkennwerte gewährleistet, über alle dargestellten Varianten hinweg

jedoch die besten und in Bezug auf die Homogenität der Ersetzungsqualität konsistentesten Ergebnisse liefert.

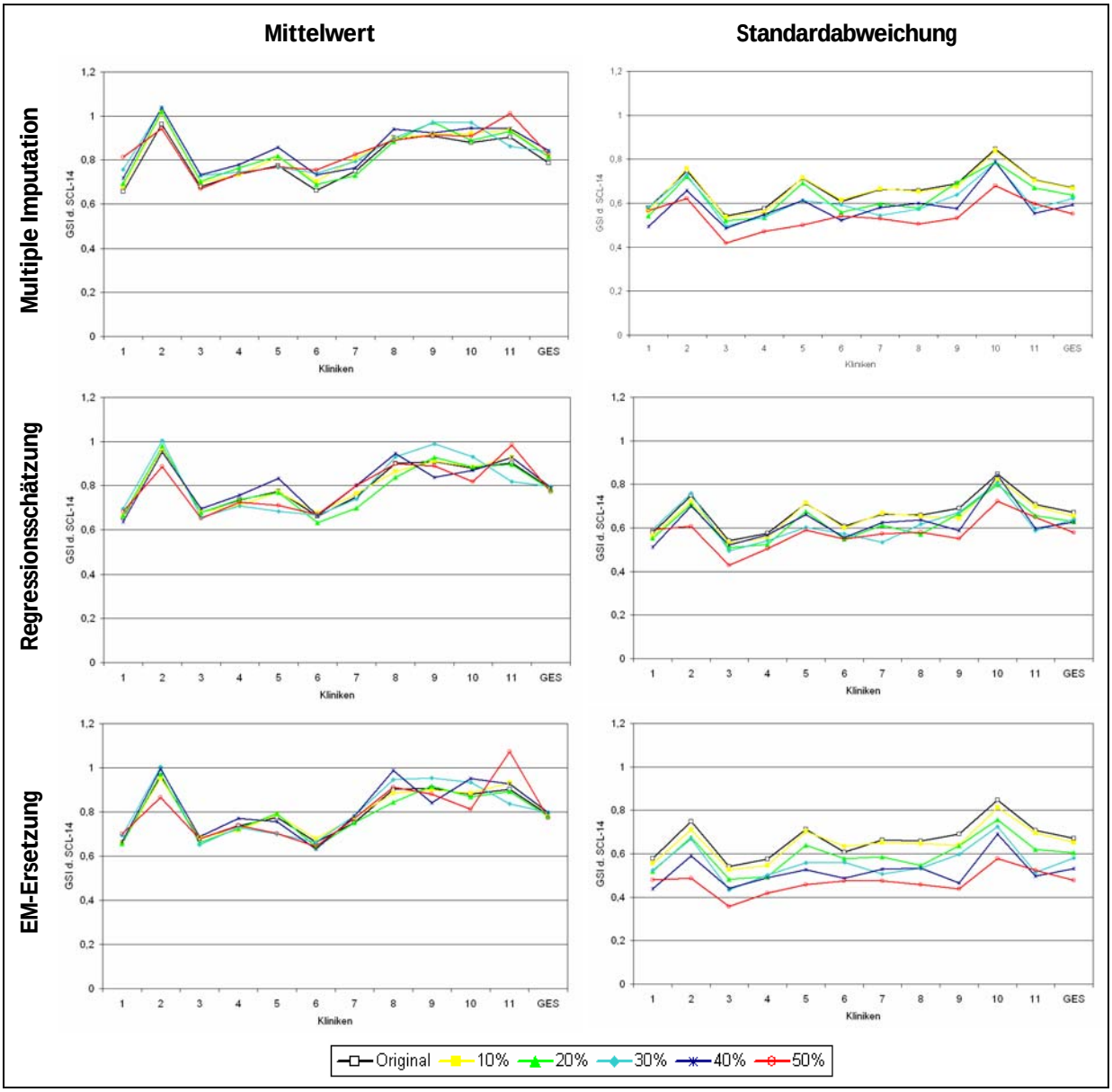

Abbildung 11: Stichprobenkennwerte nach Fehlwertersetzung bei zufällig fehlenden Werten (MCAR - skalenweise Ersetzung fehlender Werte des Globalen Symptomschwereindex der SCL-14 unter Einbezug aller verfügbaren Kovariaten)

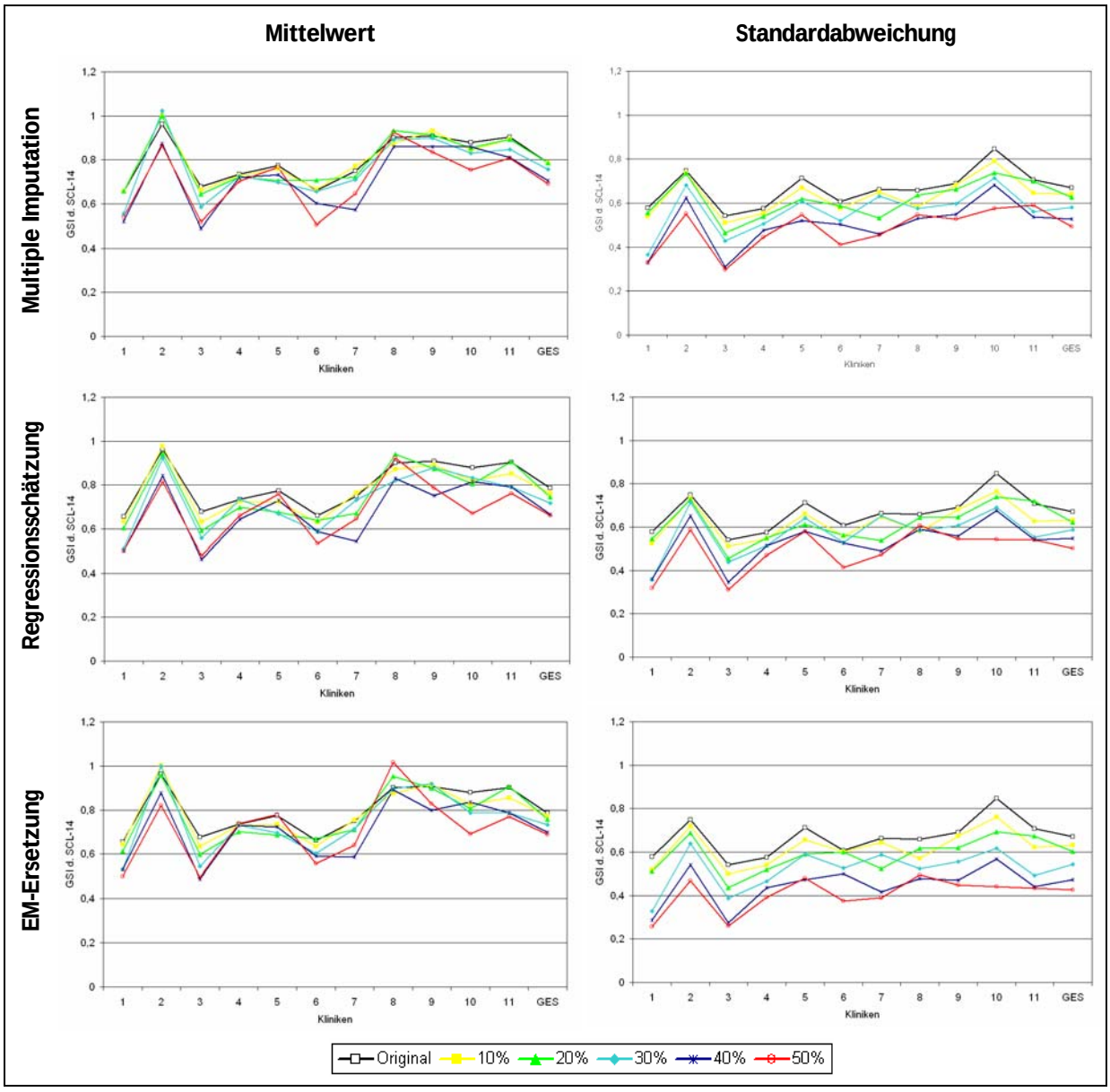

Abbildung 12: Stichprobenkennwerte nach Fehlwertersetzung bei systematisch fehlenden Werten (MAR - skalenweise Ersetzung fehlender Werte des Globalen Symptomschwereindex der SCL-14 unter Einbezug aller verfügbaren Kovariaten)

5.3. Exemplarische Darstellung eines Einrichtungsvergleiches mit Multipler Imputation fehlender Werte

In einen Einrichtungsvergleich anhand der Entlassungswerte im Globalen Symptomschwereindex (GSI) der SCL-14 und der Psychischen Summenskala (PSK) der SF-8 könnten ohne Fehlwertersetzung 1.639 Patienten einbezogen werden, für die die entsprechenden Angaben vollständig vorliegen („Responder"). Die durchschnittliche Responsequote liegt damit, gemessen an der Anzahl der aktiv an der Studie teilnehmenden Patienten (n=2.161), bei 76 Prozent. Bezogen auf die Anzahl der insgesamt im Untersuchungszeitraum in den elf beteiligten Fachkliniken behandelten Patienten (N=2.386) liegt die Responsequote bei durchschnittlich 69 Prozent, klinikbezogen schwankt sie zwischen 52 und 93 Prozent (vgl. Tabelle 25). Im Vergleich der Einrichtungen erweist sich die Responsequote der Klinik 1 als signifikant überdurchschnittlich, die Responsequoten der Kliniken 7 und 10 hingegen als unterdurchschnittlich (alle $p < 0{,}001$).

Zu den Patienten mit fehlenden Entlassungswerten (Dropouts) lagen im Mittel 160 Angaben zu den für eine Fehlwertersetzung benötigten Kovariaten aus dem Therapeutenbogen oder dem Aufnahme-Patientenbogen vor (SD: 49). Die mittlere Anzahl vorhandener Angaben zu den Kovariaten schwankt je nach Einrichtung zwischen 181 (Klinik 1) und 135 (Klinik 8). Für die Hälfte aller Dropout-Fälle liegen Angaben zu mindestens 166 Kovariaten vor, was der Anzahl der Variablen in den Patienten- und Therapeuten-Aufnahmefragebögen entspräche (vgl. Tabelle 25). Für lediglich 17 Patienten (2,3%) liegen weniger als 30 gültige Angaben in den für eine Fehlwertersetzung benötigten Kovariaten vor, was einer geringeren Zahl von Items entspricht, als sie im Therapeuten-Aufnahmebogen erhoben werden. Da diese 17 Patienten somit das Mindestkriterium für die Ersetzbarkeit fehlender Werte nicht erfüllen, mussten sie von der Fehlwertersetzung ausgeschlossen werden.

Tabelle 25: Anteile von Patienten mit fehlenden Werten in den Items des Globalen Symptomschwereindex (GSI) der SCL-14 und der Psychischen Summenskala (PSK) der SF-8 zur Entlassung (T1) und Anteile verfügbarer Kovariaten zur Fehlwertersetzung

Klinik		1	2	3	4	5	6	7	8	9	10	11	GES
Behandelte Patienten	n	217	283	231	273	171	272	168	211	232	244	84	2.386
	%	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*
GSI/PSK T1 vorhanden (100% gültige Werte)	n	201	192	171	175	116	200	88	146	151	144	55	1.639
	%	*92,6*	*67,8*	*74,0*	*64,1*	*67,8*	*73,5*	*52,4*	*69,2*	*65,1*	*59,0*	*65,5*	*68,7*
GSI/PSK T1 unvollständig (mind. 1 fehlender Wert)	n	16	91	60	98	55	72	80	65	81	100	29	747
	%	*7,4*	*32,2*	*26,0*	*35,9*	*32,2*	*26,5*	*47,6*	*30,8*	*34,9*	*41,0*	*34,5*	*31,3*
- davon mit < 30 vorhandenen Kovariaten (weniger als T)	n	0	2	1	1	0	0	10	1	0	0	2	17
	%	*0,0*	*2,2*	*1,7*	*1,0*	*0,0*	*0,0*	*12,5*	*1,5*	*0,0*	*0,0*	*6,9*	*2,3*
- davon mit ≥ 30 vorhandenen Kovariaten (≅ mind. T)	n	0	4	8	16	14	12	9	9	21	11	7	111
	%	*0,0*	*4,4*	*13,3*	*16,3*	*25,5*	*16,7*	*11,3*	*13,8*	*25,9*	*11,0*	*24,1*	*14,9*
- davon mit ≥ 51 vorhandenen Kovariaten (≅ mind. TT)	n	2	29	13	27	12	21	14	36	16	39	5	214
	%	*12,5*	*31,9*	*21,7*	*27,6*	*21,8*	*29,2*	*17,5*	*55,4*	*19,8*	*39,0*	*17,2*	*28,6*
- davon mit ≥ 141 vorhandenen Kovariaten (≅ mind. P)	n	0	5	2	1	3	2	16	2	1	2	0	34
	%	*0,0*	*5,5*	*3,3*	*1,0*	*5,5*	*2,8*	*20,0*	*3,1*	*1,2*	*2,0*	*0,0*	*4,6*
- davon mit ≥ 171 vorhandenen Kovariaten (≅ mind. PT)	n	14	51	36	53	26	37	31	17	43	48	15	371
	%	*87,5*	*56,0*	*60,0*	*54,1*	*47,3*	*51,4*	*38,8*	*26,2*	*53,1*	*48,0*	*51,7*	*49,7*

P = Anzahl der Patientenangaben zu T0 (Aufnahme), **T** = Anzahl der Therapeutenangaben zu T0, **TT** = Anzahl der Therapeutenangaben zu T0 und T1 (Aufnahme und Entlassung)

Dropoutanalysen

Um zu überprüfen, ob sich die für die einzelnen Kliniken auf der Grundlage der Responder-Stichproben ermittelten Befunde jeweils auch auf die insgesamt im Untersuchungszeitraum in der jeweiligen Klinik behandelten Patienten generalisieren ließen, wurden Dropoutanalysen durchgeführt, in denen die in einem Einrichtungsvergleich ohne Fehlwertersetzung berücksichtigten Responderstichproben anhand verschiedener Merkmale, die sich als signifikant mit dem Behandlungsergebnis konfundiert erwiesen haben, mit den jeweiligen Stichproben nicht verwertbarer Fälle (Dropouts) verglichen wurden. Bezüglich der therapeutenseitig erhobenen Daten konnten in diesen Vergleichen jeweils

potentiell alle Dropouts, also sowohl die Nonresponder, die überhaupt nicht an der Studie teilgenommen hatten, als auch die Partial-Responder, die zwar an der Studie teilgenommen haben, jedoch aufgrund einzelner oder mehrerer fehlender Werte von der einrichtungsvergleichenden Analyse ausgeschlossen werden mussten, berücksichtigt werden, soweit die entsprechenden Daten vorlagen. Bezüglich der patientenseitig erfassten Merkmale musste sich die Repräsentativitätsanalyse auf einen Vergleich von Respondern und Partial-Respondern beschränken, da die benötigten Angaben nur von letzteren vorliegen konnten.

Die Befunde der durchgeführten Dropoutanalysen sind in Tabelle 26 im Überblick aufgeführt. Es zeigte sich, dass lediglich in einer einzigen Einrichtung (Klinik 2) in sämtlichen untersuchten Risiko-Merkmalen *keine signifikanten Unterschiede* zwischen den im Einrichtungsvergleich zu berücksichtigenden und den aufgrund fehlender Werte von den vergleichenden Analysen auszuschließenden Patienten nachzuweisen waren. Dementsprechend würde man für Klinik 2 von einer uneingeschränkten Repräsentativität der ermittelten Befunde ausgehen. In der Mehrzahl der Kliniken finden sich *einzelne signifikante Unterschiede* zwischen den Stichproben der Responder und der Dropouts, die zum Teil ein erhöhtes Risiko für ein schlechteres Behandlungsergebnis darstellen, teilweise jedoch auch mit einem verringerten Risiko für ein schlechtes Behandlungsergebnis assoziiert sind. Sind diese unterschiedlichen positiven und negativen Risiken zwischen Responder- und Dropoutpatienten weitestgehend ausgeglichen, wie es beispielsweise für die Kliniken 7 und 11 der Fall ist, so könnte von weitgehend repräsentativen Befunden zur Ergebnisqualität ausgegangen werden. So findet sich beispielsweise in Klinik 7 in der Responderstichprobe im Vergleich zur Dropout-Stichprobe ein verringertes Risiko für ein schlechtes Behandlungsergebnis, da für diese Patienten eine höhere Reha-Motivation angegeben wurde. Gleichzeitig weisen die Patienten der Responderstichprobe jedoch auch einen erhöhten Chronifizierungs-Grad auf, was das Risiko für ein schlechtes Behandlungsergebnis wiederum erhöht. In den meisten Einrichtungen (Kliniken 1, 3, 4, 5, 6, 8, 9 und 10) deuten die Stichprobenunterschiede zwischen Respondern und Dropouts jedoch durchwegs auf ein *erhöhtes „Risiko" eines besseren Behandlungsergebnisses in der Responderstichprobe* hin. Demzufolge wäre es möglich, dass die wahren Behandlungseffekte für die betroffene Einrichtung durch die vergleichenden Analysen ohne Fehlwertersetzung mehr oder weniger deutlich überschätzt würden (vgl. Tabelle 26).

Tabelle 26: Repräsentativität der Responderstichproben der einzelnen Kliniken

Klinik **Risikofaktor**		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**GES**
Therapeutenangaben T0 (Responder vs. Dropouts)													
Reha-Motivation	☺	o	o	o	--	o	o	--	--	--	--	--	--
Chronifizierung	☹	o	o	o	o	o	o	++	o	o	o	o	o
Angststörungen	☺	o	o	o	o	o	o	o	o	o	o	o	o
PTBS	☹	o	o	o	o	o	o	o	o	o	o	o	o
Essstörungen	☺	o	o	o	o	o	o	o	o	o	o	o	--
Persönlichkeitsstörungen	☹	o	o	o	o	o	o	o	o	o	o	o	o
Komorbidität	☹	o	o	--	o	o	o	o	o	--	o	o	-
Beeinträchtigungsschwere (HoNOS-D)	☹	o	o	o	--	--	o	o	o	--	o	o	-
Eingeschränkte psychische Lebensqualität (SF-8-F)	☹	o	o	o	o	--	o	o	o	o	o	o	-
Eingeschränkte somatische Lebensqual. (SF-8-F)	☹	o	o	o	o	o	o	o	o	o	o	o	-
Patientenangaben T0 (Responder vs. Partial-Responder)													
Depressivität (ADS-K)	☹	--	o	o	--	--	o	o	o	o	o	o	-
Depressivität (SCL-14)	☹	--	o	o	o	--	o	o	o	o	o	++	o
Somatisierung (SCL-14)	☹	o	o	o	o	--	--	o	o	o	o	o	--
Phobische Angst (SCL-14)	☹	o	o	o	o	--	o	o	o	o	o	o	o
Eingeschränkte psychische Lebensqualität (SF-8)	☹	--	o	o	--	--	o	o	o	o	o	o	-
Eingeschränkte somatische Lebensqualität (SF-8)	☹	o	o	o	--	--	o	o	o	o	o	o	--
Interpersonelle Probleme (IIP-64)	☹	--	o	o	o	o	o	o	o	o	o	++	o
Hauptschulabschluss	☹	o	o	--	o	--	o	o	o	o	o	o	o
Sonderschul- oder kein Schulabschluss	☹	o	o	o	o	o	--	o	o	o	o	o	o
Hausfrau/-mann	☺	o	o	o	o	o	o	o	o	o	o	o	o
Rentner	☺	o	o	o	o	o	o	o	o	o	o	--	o
Rentenantrag	☹	o	o	o	o	o	o	o	o	o	o	o	o
Dropoutquoten (%)		*7,4*	*32,2*	*26,0*	*35,9*	*32,2*	*26,5*	*47,6*	*30,8*	*34,9*	*41,0*	*34,5*	*31,3*

(Anmerkungen siehe nächste Seite)

Anmerkungen zu Tabelle 26:

☹: Risikofaktor für schlechteres Behandlungsergebnis, ☺: Risikofaktor für besseres Behandlungsergebnis; **++**: signifikant höheres Risiko für schlechteren Outcome in Responderstichprobe mit mindestens kleiner Effektstärke, **+**: signifikant höheres Risiko für schlechteren Outcome in Responderstichprobe mit geringerer als kleiner Effektstärke, **o**: kein signifikanter Unterschied zwischen Respondern und Dropouts, **-**: signifikant geringeres Risiko für schlechteren Outcome in Responderstichprobe mit geringerer als kleiner Effektstärke, **--**: signifikant geringeres Risiko für schlechteren Outcome in Responderstichprobe mit mindestens kleiner Effektstärke

Einrichtungsvergleich mit und ohne Imputation fehlender Werte

Die Resultate des Einrichtungsvergleichs werden in Abbildung 13 und Abbildung 14 grafisch veranschaulicht. Dabei sind die auf der Basis der vollständig vorhandenen Entlassungsdaten bestimmten Einrichtungsmittelwerte im Globalen Symptomschwere Index (GSI) der SCL-14 (Abbildung 13) bzw. in der Psychischen Summenskala (PSK) der SF-8 (Abbildung 14) jeweils als dunkelblaue Quadrate dargestellt. Die auf der Grundlage von mittels Multipler Imputationen vervollständigten Entlassungswerten ermittelten Einrichtungskennwerte sind jeweils als hellblaue Rauten markiert. Zu jedem Einrichtungskennwert ist außerdem das entsprechende 95%-Konfidenzintervall für den Mittelwert abgebildet. Im oberen Teil der Grafiken sind die Einrichtungen jeweils aufsteigend nach den auf der Basis der vollständig vorhandenen Entlassungsdaten bestimmten Kennwerten sortiert, im unteren Teil hingegen nach den aus der Fehlwertersetzung resultierenden Einrichtungskennwerten. Liegt der auf der Basis imputierter Werte ermittelte Einrichtungsmittelwert außerhalb des 95%-Konfidenzintervalls des ohne Fehlwertersetzung bestimmten Einrichtungsmittelwertes (und umgekehrt), so spricht dies für eine signifikante Abweichung der beiden Mittelwerte auf einem Signifikanzniveau von p=0,05.

Signifikante Abweichungen zwischen den auf der Basis der vollständigen und den auf der Basis der mittels Multipler Imputation vervollständigten Entlassungsdaten ermittelten Einrichtungskennwerten ergeben sich für den GSI der SCL-14 in den Kliniken 4 (t=4,1226, p<0,001, d=0,39), 5 (t=2,8013, p=0,005, d=0,35) und 10 (t=1,9586, p=0,05, d=0,25), für die PSK der SF-8 lediglich in der Klinik 4 (t=2,2746, p=0,02, d=0,20). Dabei fallen die nach Fehlwertersetzung ermittelten Entlassungswerte jeweils höher aus als die ursprünglich auf der Basis der vollständig vorliegenden Daten bestimmten Beeinträchtigungsscores (vgl. Abbildung 13 und Abbildung 14).

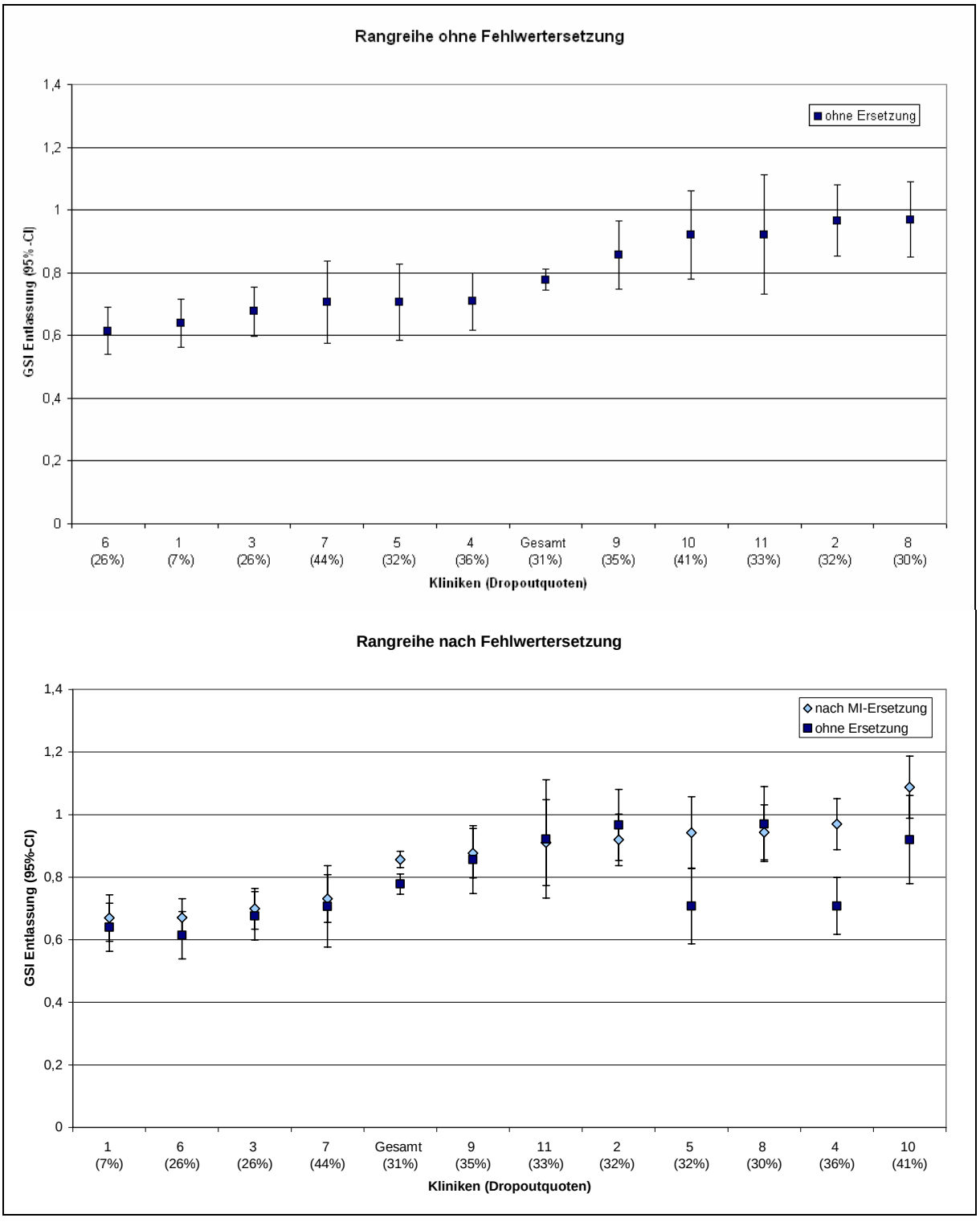

Abbildung 13: Klinikrangreihen nach Globalem Symptomschwereindex (GSI) der SCL-14 - mit und ohne Fehlwertersetzung

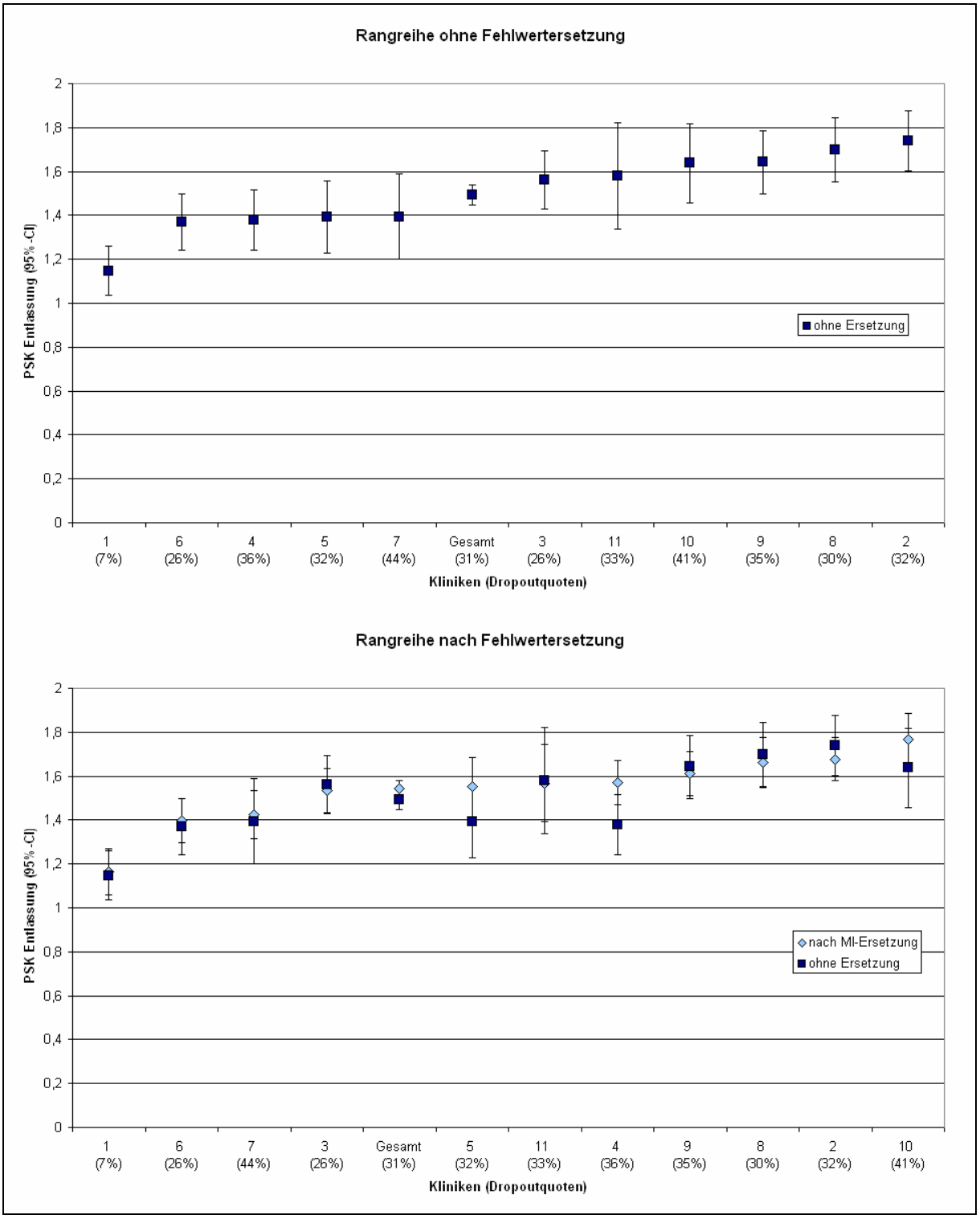

Abbildung 14: Klinikrangreihen nach Psychischer Summenskala (PSK) der SF-8 - mit und ohne Fehlwertersetzung

Betrachtet man die anhand der Einrichtungswerte gebildeten Rangreihen, so ergeben sich durch die Fehlwertersetzung teilweise erhebliche Verschiebungen: Bezogen auf die Entlassungswerte im GSI der SCL-14 „verschlechtern“ sich die drei Kliniken, bei denen die Fehlwertersetzung in signifikant veränderten Klinikkennwerten resultierte, um bis zu vier Ränge (Klinik 4: -4 Ränge, Klinik 5: -3 Ränge, Klinik 10: -3 Ränge; vgl. Abbildung 13). In der Rangreihe auf Grundlage der Entlassungswerte in der PSK der SF-8 verliert die Klinik 4, für die die Fehlwertersetzung auch hier signifikant verschlechterte Kennwerte ergab, nach der Imputation fehlender Werte ebenfalls vier Ränge. Aber auch die Klinik 10, für die sich bezüglich der PSK kein signifikanter Unterschied zwischen vorhandenen und fehlwertersetzten Entlassungsdaten nachweisen ließ, verschlechtert sich noch um drei Ränge (vgl. Abbildung 14). Die übrigen Veränderungen in der Rangfolge der Einrichtungen sind entweder auf diese systematischen Verschiebungen zurückzuführen oder bewegen sich im Bereich zufälliger Abweichungen (+/- 1 Rang).

Korrespondenz von Dropoutanalysen und Fehlwert-bereinigtem Einrichtungsvergleich

Setzt man nun die Ergebnisse der Dropoutanalysen (vgl. Tabelle 26) in Relation zu den Effekten, die sich durch die Ersetzung fehlender Werte im Einrichtungsvergleich ergeben haben (vgl. Abbildung 13 und Abbildung 14), so lässt sich zunächst festhalten, dass die Fehlwertersetzung für diejenigen Einrichtungen, in denen sich in den Dropoutanalysen keine (bzw. zumindest ausgeglichene positive wie negative) Abweichungen zwischen Respondern und Dropouts gezeigt hatten (Kliniken 2, 7 und 11), einheitlich auch keine Veränderungen bezüglich der Stichprobenkennwerte zur Folge hatte.

Ein deutlich heterogeneres Bild ergibt sich jedoch beim Blick auf diejenigen Einrichtungen, für die die Dropoutanalysen signifikante Unterschiede zwischen Respondern und Dropouts aufgedeckt hatten.

Je nach Korrespondenz zwischen Dropoutanalysen und Effekten der Fehlwertersetzung lassen sich die Einrichtungen in vier Kategorien einteilen:

1. Einrichtungen, für die sich in den Dropoutanalysen deutliche Unterschiede zwischen Respondern und Dropouts nachweisen ließen und für die sich dementsprechend auch deutliche Effekte durch die Ersetzung der fehlenden Werte ergeben haben (z.B. Klinik 4).
2. Einrichtungen, für die sich in den Dropoutanalysen nur geringe Unterschiede zwischen Respondern und Dropouts nachweisen ließen und für die sich dementsprechend auch nur geringe Effekte durch die Ersetzung der fehlenden Werte ergeben haben (z.B. Klinik 6).
3. Einrichtungen, für die sich in den Dropoutanalysen deutliche Unterschiede zwischen Respondern und Dropouts nachweisen ließen, für die sich jedoch keine oder nur geringe Effekte durch die Ersetzung der fehlenden Werte ergeben haben (z.B. Klinik 1, z.T. auch Klinik 5).
4. Einrichtungen, für die sich in den Dropoutanalysen nur geringe Unterschiede zwischen Respondern und Dropouts nachweisen ließen, für die sich jedoch deutliche Effekte durch die Ersetzung der fehlenden Werte ergeben haben (z.B. Klinik 10).

In den Kategorien 1 und 2 spiegeln sich die Befunde der Dropoutanalysen direkt in den Konsequenzen der Fehlwertersetzung wider, in den Kategorien 3 und 4 ergeben sich jedoch deutliche Abweichungen zwischen den Resultaten der beiden Varianten zum Umgang mit fehlenden Werten, weshalb auf letztere noch etwas ausführlicher eingegangen werden soll.

Beispielsweise ergeben sich für die Klinik 1 in den Dropoutanalysen in vier Bereichen (Depressivität in ADS-K und SCL-14, Psychische Lebensqualität in SF-8, Interpersonelle Probleme im IIP-64) signifikante Unterschiede zwischen Respondern und Dropouts, deren Ausmaß überwiegend sogar mittlere Effektstärken erreicht, weshalb bei ausschließlicher Berücksichtigung der Responder von einer Überschätzung der Behandlungseffekte ausgegangen werden müsste (vgl. Tabelle 26). Durch die Ersetzung der fehlenden Werte ergeben sich jedoch keinerlei Veränderungen der für diese Einrichtung zur Entlassung ermittelten Ergebniswerte im GSI der SCL-14 und in der PSK der SF-8, was wiederum für die Repräsentativität der Responder für die insgesamt in dieser Einrichtung behandelten

Patienten spricht (vgl. Abbildung 13 und Abbildung 14). Zu erklären ist die Diskrepanz zwischen Dropoutanalyse und Fehlwertersetzung in diesem Falle durch die verhältnismäßig geringe Dropoutquote in der betreffenden Einrichtung (7%), die dafür verantwortlich sein dürfte, dass die abweichende Merkmalsausprägung der Dropouts auf Gesamtstichprobenebene kaum ins Gewicht fällt.

Eine ähnliche Diskrepanz zwischen den Resultaten der Dropoutanalysen und den ermittelten Konsequenzen der Fehlwertersetzung ergibt sich – allerdings nur bezogen auf die PSK der SF-8 – auch für die Klinik 5. Hier finden sich in den Dropoutanalysen zwar in neun Bereichen signifikante Abweichungen zwischen Respondern und Dropouts (vgl. Tabelle 26), die Fehlwertersetzung resultiert jedoch nicht in signifikant abweichenden Entlassungswerten in der PSK und auch die Position der Klinik in der entsprechenden Rangordnung der elf untersuchten Einrichtungen verändert sich nur unwesentlich (Rang -1; vgl. Abbildung 14). Da die Dropoutquote in Klinik 5 jedoch bei 32 Prozent liegt, kann diese Diskrepanz nicht durch einen zu vernachlässigenden Anteil von Dropouts erklärt werden.

Der umgekehrte Sachverhalt findet sich für die Klinik 10. Hier ergibt sich in den Dropoutanalysen lediglich eine signifikante Abweichung zwischen Respondern und Dropouts (höhere Reha-Motivation bei Respondern; vgl. Tabelle 26), die Fehlwertersetzung resultiert jedoch in ausgeprägten Veränderungen der Einrichtungskennwerte, durch die sich die Position der Klinik im Einrichtungsvergleich jeweils um drei Ränge auf den letzten Rang verschlechtert (vgl. Abbildung 13 und Abbildung 14). Inwieweit sich die Diskrepanz zwischen einer als relativ gering ermittelten Abweichung zwischen Respondern und Dropouts und der erheblichen Konsequenzen einer Fehlwertersetzung auf die Einrichtungskennwerte durch die verhältnismäßig hohe Dropoutquote erklären lässt, bleibt hierbei offen.

6. Diskussion

Einrichtungsvergleiche gewinnen in der Qualitätssicherung medizinischer Maßnahmen mehr und mehr an Bedeutung (Schulz & Koch, 2002). Da absolute Standards zur Bewertung der Behandlungsqualität häufig fehlen, soll der Vergleich verschiedener Einrichtungen die Beurteilung der Behandlungsqualität einzelner Einrichtungen ermöglichen. Der Schluss von der relativen Lage einer Einrichtung bezüglich eines definierten Kriteriums, z.B. der Höhe einer Effektstärke in einer definierten Outcomevariable, auf die Qualität der Behandlung in dieser Einrichtung ist jedoch nur sinnvoll, wenn dabei sichergestellt wird, dass alle verglichenen Einrichtungen grundsätzlich dieselbe Chance haben, das definierte Kriterium zu erreichen. In der Literatur zur einrichtungsvergleichenden Qualitätssicherung nimmt dementsprechend der Aspekt ungleich verteilter Risikofaktoren, d.h. Merkmale der Patienten, die – vor allem negativ - mit dem Outcome in Beziehung stehen, eine zentrale Stellung ein (z.B. Farin et al., 2004; Iezzoni, 1997; Schulz et al., 2004; Wegscheider, 2004). Dabei wird zu Recht auf die Notwendigkeit einer Risikoadjustierung hingewiesen, um faire Einrichtungsvergleiche zu gewährleisten. Daneben wird jedoch häufig der – eigentlich triviale - Umstand vernachlässigt, dass Einrichtungsvergleiche jeweils nur dann zu gültigen Aussagen führen können, wenn die Daten, auf denen sie basieren, die Versorgungsrealität innerhalb der verglichenen Einrichtungen zuverlässig abbilden. Spiegeln die vorliegenden Daten die Behandlungsqualität der verglichenen Einrichtung jedoch nicht ausreichend wider, so kann auch eine noch so umfassende oder elaborierte Risikoadjustierung nicht sicherstellen, dass die resultierenden Bewertungen der Einrichtungen fair ausfallen.

Der Anspruch der vorliegenden Arbeit war es vor diesem Hintergrund, *fehlende Werte* als eine Hauptursache für die Verletzung der grundlegenden Voraussetzung valider Einrichtungsdaten in den Fokus der Aufmerksamkeit zu rücken. Neben inadäquaten methodischen Zugängen zur Datenerhebung und der – willentlichen oder unwillentlichen - Verfälschung von Daten stellen Datenausfälle, wie sie in nahezu allen empirischen Untersuchungen auftreten, nämlich ein großes Risiko für das Zustandekommen einer nur eingeschränkt generalisierbaren Datenbasis dar. Im Rahmen der vorliegenden Arbeit wurden daher am Beispiel der externen einrichtungs-

vergleichenden Qualitätssicherung von Behandlungen der medizinischen Rehabilitation im Indikationsbereich „psychische und psychosomatische Erkrankungen“ Aspekte (a) der Notwendigkeit einer umfassenden Berücksichtigung fehlender Daten, (b) der Ersetzbarkeit fehlender Daten und (c) der Konsequenzen einer möglichen Ersetzung von fehlenden Daten für das Ranking der Einrichtungen behandelt. Als empirische Grundlage diente dabei beispielhaft ein Datensatz zur Ergebnisqualität stationärer Rehabilitation, der im Rahmen der Pilotphase zum Qualitätssicherungsprogramm der Gesetzlichen Krankenversicherung in elf Rehabilitationsfachkliniken für Psychosomatische Medizin und Psychotherapie zu Beginn und am Ende der Rehabilitationsmaßnahmen an konsekutiven Stichproben von insgesamt 2.386 behandelten Patienten erhoben wurde.

6.1. Notwendigkeit der Berücksichtigung fehlender Werte

Datenausfälle bedeuten in jedem Falle einen Informationsverlust. Durch fehlende Daten reduzieren sich zum einen die verwertbaren Stichproben, was eine verringerte statistische Power zum Nachweis vorhandener Unterschiede zur Folge hat. Im Extremfall ist durch fehlende Werte eine Reduktion der Stichprobengröße möglich, die Auswertungen aufgrund einer Power unterhalb der Zufallswahrscheinlichkeit obsolet macht. Vor allem gefährden systematische, aber auch zufällige Datenausfälle jedoch immer auch die Repräsentativität der verbleibenden Daten, was eine eingeschränkte Generalisierbarkeit ermittelter Befunde zur Folge hat. Im schlimmsten Falle können systematische Datenausfälle zu stark verzerrten Ergebnissen führen, ohne dass diese als solche erkannt und behandelt würden.

Im untersuchten Beispieldatensatz lagen für lediglich 389 der 2.386 im Untersuchungszeitraum behandelten Patienten komplett vollständige Daten vor (16,3%). Bei gleichzeitiger Berücksichtigung aller erhobenen Variablen im Rahmen multivariater Analysen ergäbe sich also eine immense Stichprobenreduktion, die die Generalisierbarkeit ermittelter Befunde mehr als deutlich in Frage stellen würde. Die in der vorliegenden Arbeit exemplarisch als zentrales Maß der Ergebnisqualität defi-

nierten Werte im Globalen Symptomschwereindex (GSI) der von den Patienten zum Entlassungszeitpunkt bearbeiteten Symptomcheckliste SCL-14 (Harfst et al., 2002) und der Psychischen Summenskala (PSK) der SF-8 (Ware et al., 2000) wurden von immerhin 1.639 Patienten vollständig angegeben (68,7%). Die Analyse fehlender Werte zeigte jedoch, dass sich zwischen den einzelnen Einrichtungen erhebliche Unterschiede bezüglich der Vollständigkeit der patientenseitig erhobenen Daten zur Entlassung nachweisen ließen. Die Anteile vollständig vorliegender Patientendaten schwankten je nach Einrichtung zum Entlassungszeitpunkt zwischen 52,4 und 92,6 Prozent. Die entsprechenden mittleren patientenbezogenen Fehlwertquoten variieren zwischen den Einrichtungen von 2,3 bis 37,6 Prozent (Durchschnitt: 10,8%). Während also im besten Falle zu neun von zehn behandelten Patienten einer Einrichtung vollständige Daten zur Beurteilung der Ergebnisqualität vorliegen, müsste sich die Bewertung der Ergebnisqualität einer Einrichtung im schlechtesten Falle auf die Angaben nur jedes zweiten behandelten Patienten stützen. Inwieweit die in letzterem Falle ermittelten Befunde ohne weitere Absicherung als repräsentativ für die Grundgesamtheit der in der entsprechenden Einrichtung behandelten Patienten angesehen werden dürften, erscheint zumindest fraglich.

Zur Annäherung an die Frage der Repräsentativität der vollständig vorliegenden Daten wurden gemäß gängiger Forschungspraxis entsprechende Dropoutanalysen durchgeführt, in denen die Vergleichbarkeit derjenigen Patienten, von denen die benötigten Entlassungsdaten vollständig vorlagen (sog. „Responder"), mit denjenigen Patienten, von denen die entsprechenden Angaben fehlten (sog. „Dropouts"), jeweils anhand von Daten überprüft wurde, die für beide Patientengruppen vorlagen. Berücksichtigt wurden dabei insbesondere solche Variablen, für die ein (kausaler) Zusammenhang zum erzielten Behandlungsergebnis angenommen werden kann (sog. „Konfounder", z.B. Reha-Motivation, Beeinträchtigungsschwere bei Aufnahme usw.). Es zeigte sich, dass bezüglich der überprüften Konfounder in nur einer der elf untersuchten Einrichtungen keine signifikanten Unterschiede zwischen Respondern und Dropouts nachzuweisen waren. In allen anderen Einrichtungen ergaben sich unterschiedlich umfangreiche Abweichungen zwischen Respondern und Dropouts, die zumeist in Richtung eines als besser zu erwartenden Behandlungsergebnisses in der Responderstichprobe interpretiert werden müssten. Auf der Grundlage der Befunde aus den Dropoutanalysen wäre also davon auszugehen, dass die Beur-

teilung der Ergebnisqualität der untersuchten Einrichtungen auf der Basis der jeweils vollständig vorhandenen Entlassungsdaten in den meisten Kliniken in einer *Überschätzung* der tatsächlichen Qualität der insgesamt im Untersuchungszeitraum durchgeführten Behandlungen resultieren würde.

Inwieweit Dropoutanalysen jedoch überhaupt geeignet sind, um die Repräsentativität ermittelter Befunde zum Behandlungsergebnis sicherzustellen, konnte im Rahmen der durchgeführten Simulationsstudie überprüft werden. In der Fehlwertsimulation wurden die interessierenden Entlassungsdaten (Items des GSI und der PSK) dafür nämlich jeweils entweder zufällig oder systematisch (d.h. nur bei ungünstigen Behandlungsverläufen) eliminiert. Die entsprechenden Dropoutanalysen anhand der Aufnahmedaten bestätigten, dass sich im Falle zufällig fehlender Werte grundsätzlich auch keine Hinweise auf Einschränkungen der Repräsentativität der vorhandenen Entlassungsdaten ergeben. Im Falle der Simulation systematisch fehlender Werte ergab sich jedoch ein inkonsistentes Bild: Bei geringen Fehlwertquoten (10-20%) erbrachte der Vergleich von Respondern und Dropouts nämlich trotz systematischen Datenausfalls keine Hinweise auf die eingeschränkte Repräsentativität der vollständig vorliegenden Daten. Bei höheren Fehlwertquoten (30-50%) fanden sich zwar gehäuft Hinweise auf die eingeschränkte Repräsentativität der vorliegenden Daten, welche sich jedoch nur in Ausnahmefällen in der Ausprägung der Entlassungsdaten widerspiegelte. Diese Befunde stellen die Eignung von Dropoutanalysen zur Sicherstellung der Repräsentativität vorhandener Daten somit erheblich in Zweifel. Ergeben sich in Dropoutanalysen bezüglich ausgewählter Konfounder keine Unterschiede zwischen Respondern und Dropouts, so gewährleistet dies nicht notwendigerweise die Repräsentativität der vollständig vorliegenden Daten. Dieses ist auch theoretisch erklärlich, da es bisher nur in Ansätzen gelungen ist, alle potentiellen Konfounder auf Seiten der Patienten zu identifizieren und damit auch erheben zu können, und zum anderen, weil immer auch weitere, nicht unbedingt patientenbezogene Faktoren auf die Behandlung und damit auch auf das Behandlungsergebnis Einfluss nehmen. Finden sich in Dropoutanalysen jedoch Unterschiede zwischen Respondern und Dropouts, so ist daraus noch nicht abzuleiten, in welchen Outcomekriterien und vor allem in welchem Ausmaß sich die erwiesenermaßen eingeschränkte Repräsentativität der Daten niederschlägt, was im Zweifelsfalle zu fal-

schen Schlussfolgerungen bezüglich der Bewertung der Behandlungserfolge einer Einrichtung resultieren kann.

Zusammenfassend lässt sich in Bezug auf die Notwendigkeit der Berücksichtigung fehlender Werte somit zunächst festhalten, dass fehlende Werte insbesondere im Kontext der einrichtungsvergleichenden Qualitätssicherung eine nicht zu vernachlässigende Größe darstellen. Es zeigte sich nämlich, dass die Vollständigkeit der zur Bewertung der Behandlungsqualität der verschiedenen Einrichtungen benötigten Daten trotz verhältnismäßig geringer durchschnittlicher Fehlwertquoten zwischen den einzelnen Kliniken sehr heterogen verteilt war. In den Dropoutanalysen fanden sich außerdem in nahezu allen Einrichtungen Hinweise auf eine eingeschränkte Repräsentativität der vollständig vorliegenden Daten, was eine Generalisierbarkeit der auf Grundlage dieser Daten ermittelten Befunde in Frage stellt. Da sich Dropoutanalysen zudem im Rahmen der durchgeführten Simulationsstudie als nur bedingt geeignet erwiesen haben, um die Repräsentativität vorliegender Daten differenziert zu beurteilen, sind geeignete Verfahren zum Umgang mit fehlenden Werten dringend erforderlich.

6.2. Ersetzbarkeit fehlender Werte

Seit den späten 50er Jahren des vergangenen Jahrhunderts wurde eine Vielzahl von Methoden zum Umgang mit fehlenden Werten entwickelt (vgl. Afifi & Elashoff, 1966; Little & Rubin, 1987; Schafer, 1999; Scheffer, 2002). Die bisherige Anwendung von Verfahren zur Ersetzung fehlender Werte bezog sich bislang jedoch zumeist auf die Problematik einzelner fehlender Werte innerhalb umschriebener Erhebungseinheiten, das Problem komplett fehlender Datensätze zu einzelnen Erhebungseinheiten (sog. „Unit-Nonresponse“, z.B. komplett fehlende Patientendaten zur Entlassung), wie es sich in Längsschnittstudien häufig ergibt, blieb dabei eher unberücksichtigt. Lediglich im Bereich der Prozessforschung wurden Ansätze entwickelt, die die Ersetzbarkeit zufällig fehlender Werte im Rahmen von Prozessstudien mit einer Vielzahl von Erhebungszeitpunkten sogar gezielt nutzen, indem fallbezo-

gen auf einzelne a priori zufällig bestimmte Verlaufsmessungen verzichtet wird, um den Erhebungsaufwand zu minimieren (sog. „*Random Effects Models*“, vgl. Hedeker & Gibbons, 1997). Im Kontext der Outcomeforschung finden sich hingegen erst in letzter Zeit Studien, in denen beispielsweise fehlende Follow-up-Daten auf der Grundlage von Daten aus der Baseline-Erhebung mittels Multipler Imputationen geschätzt wurden (Sbarra & Emery, 2005). Allerdings stützen die Autoren besagter Arbeit die Anwendung einer derartigen Ersetzungsmethodik lediglich auf die Aussage, dass „suboptimale Multiple Imputationen immer noch besser seien als in weiten Teilen unvollständige Daten“ (S. 68, Übersetzung d. Verf.). Eine systematische Überprüfung dieser Annahme stand bislang allerdings noch aus. Da bisherige Versuche, das Behandlungsergebnis von stationärer Psychotherapie vorherzusagen, zudem eher moderaten Erfolg zeigten (vgl. z.B. Hannöver & Kordy, 2005), erschien es eher fraglich, ob die Ersetzung fehlender Daten zum Behandlungsergebnis anhand vorliegender Basisdaten überhaupt möglich ist.

Im Rahmen der in der vorliegenden Arbeit dargestellten Simulationsstudie zur Ersetzbarkeit fehlender Werte sollte vor diesem Hintergrund überprüft werden, inwieweit sich patientenseitig vollständig fehlende Entlassungsdaten anhand der aus den patientenseitig beantworteten Aufnahmefragebögen vorliegenden Basisdaten und ergänzend herangezogener Therapeutenangaben adäquat ersetzten lassen. Entsprechend der Fehlwertcharakteristika im Originaldatensatz (einrichtungsbezogen bis zu 50 Prozent weitestgehend bedingt zufällig fehlende Werte) wurden im Simulationsdatensatz ebenfalls bis zu 50 Prozent (10-50%) vollständig zufällig (MCAR) oder bedingt zufällig (MAR) fehlende Entlassungswerte generiert. Zur Ersetzung der fehlenden Werte wurden ein regressionsanalytisches Verfahren (RA), ein Verfahren auf Grundlage des „Expectation Maximization“-Algorithmus (EM) sowie ein Multiples Imputationsverfahren (MI) überprüft. Als weitere Rahmenbedingungen wurden die Art und Anzahl berücksichtigter Kovariaten (Patienten- und/oder Therapeutenangaben), das zu ersetzende Kriterium (GSI der SCL-14 oder PSK der SF-8) und die Ersetzungsebene (skalen- oder item-bezogene Fehlwertersetzung) variiert. Durch die zusätzliche Berücksichtigung der Einrichtungsebene (d.h. der elf Kliniken) resultierten somit letztlich 6.600 Ersetzungsvarianten. Bei der Bewertung der Ersetzungsgüte wurden jeweils zwei Ebenen unterschieden: Zum einen wurde für die Fälle mit simuliert fehlenden Werten (Dropout-Stichproben) überprüft, wie gut die

geschätzten Werte mit der tatsächlichen Ausprägung der fehlenden Werte übereinstimmen. Zum zweiten wurde jedoch auch überprüft, inwiefern die aus der Fehlwertersetzung resultierenden Stichprobenkennwerte (Mittelwerte und Standardabweichungen) in den jeweiligen Gesamtstichproben mit den ursprünglichen Stichprobenkennwerten übereinstimmen. Als Maß für die Übereinstimmung zwischen Original- und ersetzten Werten wurden unjustierte, einfaktorielle, korrigierte Intraklassenkorrelationskoeffizienten verwendet ($ICC_{unjust,einfakt,korr}$, vgl. Hartung, 1981).

In Bezug auf die Ersetzbarkeit fehlender Werte ergibt sich zunächst ein sehr heterogenes Bild. Die Ersetzungsgüte schwankt auf der Ebene der Dropout-Stichproben nämlich je nach Ersetzungsvariante zwischen ICC=0,02 (überhaupt keine Übereinstimmung) und ICC=0,92 (sehr hohe Übereinstimmung). Auf der Ebene der Gesamtstichproben reicht die Spannweite der ermittelten ICC von 0,37 (mäßige Übereinstimmung) bis 1,00 (perfekte Übereinstimmung). Die mittlere Ausprägung der ICC über alle 6.600 realisierten Ersetzungsvarianten beträgt auf der Ebene der Dropout-Stichproben 0,38 (mäßige Übereinstimmung), auf der Ebene der Gesamtstichproben 0,82 (gute Übereinstimmung). Diese Befunde spiegeln zuallererst die Tatsache wider, dass sich die Ersetzungsgüte natürlich dann als deutlich schlechter darstellt, wenn der Anteil fehlender Werte sehr hoch ist, wie dies im extremen Falle der Dropout-Stichproben (= 100 Prozent fehlende Werte) im Vergleich zu den Gesamtstichproben, wo ein größerer Teil der Werte vorliegt (= nur 10-50 Prozent fehlende Werte) und somit schlecht ersetzte Werte in der Gesamtbeurteilung weniger stark gewichtet werden. Weiterhin lässt sich festhalten, dass die Güte einer Fehlwertersetzung in Abhängigkeit der jeweils gegebenen Rahmenbedingungen sehr unterschiedlich ausfallen kann. Dabei ist zudem zu berücksichtigen, dass die Übereinstimmung zwischen geschätzten und Originalwerten im Bereich mäßiger Ersetzungsgüte durch die ICC-Korrektur nach Hartung (1981) tendenziell sogar noch überschätzt wird.

Da die Güte einer Fehlwertersetzung also maßgeblich von den Rahmenbedingungen, unter denen sie durchgeführt wird, bestimmt zu werden scheint, sollen die Effekte der einzelnen überprüften Einflussfaktoren nochmals zusammenfassend erörtert werden. In Bezug auf die *konkrete Ersetzungsgüte* auf der Ebene der Dropout-

Stichproben zeigten sich die größten Unterschiede für die verwendete Ersetzungsmethode, wobei sich die Multiple Imputation den beiden anderen angewandten Verfahren (EM und RA) als überlegen erwies. Überraschenderweise zeigte sich die RA dabei wiederum der EM als deutlich überlegen (eine mögliche Erklärung für diesen Befund wird in Kapitel 6.5 diskutiert). Kleine Effektstärken ließen sich bezüglich der konkreten Ersetzungsgüte außerdem für die berücksichtigten Kovariaten nachweisen: Hier führte der Einbezug der patientenseitig erfassten Aufnahmedaten zu deutlich besseren Ersetzungsresultaten als die alleinige Berücksichtigung der Therapeutenangaben. Die Ersetzungsebene offenbarte schließlich nur noch sehr geringe Unterschiede in der Güte der Fehlwertersetzung, wobei die direkte Ersetzung fehlender Skalenwerte einer Ersetzung der einzelnen fehlenden Items einer Skala vorzuziehen ist. Keine Unterschiede bezüglich der Übereinstimmung zwischen ersetzten und Original-Werten ließen sich entgegen der Erwartung für die Art des Ausfallmechanismus, das Ausmaß fehlender Werte oder das zu ersetzende Kriterium zeigen. Im Hinblick auf die *resultierenden Stichprobenkennwerte* der jeweiligen Gesamtstichproben ergab sich demgegenüber ein in einigen Bereichen deutlich abweichendes Bild. Hier zeigte – erwartungsgemäß - das Ausmaß fehlender Werte die größten Effekte auf die Güte des Ersetzungsresultates. Mit zunehmender Fehlwertquote sinkt die Übereinstimmung zwischen den geschätzten und den Original-Kennwerten deutlich ab (im Mittel von ICC=0,94 bei 10 Prozent fehlenden Werten auf ICC=0,65 bei 50 Prozent fehlenden Werten). Weitere Effekte zeigen sich, wie auch schon auf der Ebene der Dropout-Stichproben, für die einbezogenen Kovariaten und die Ersetzungsmethode. Allerdings fällt die Überlegenheit der Multiplen Imputation in Bezug auf die Güte der resultierenden Stichprobenkennwerte deutlich geringer aus als im Hinblick auf die Güte der konkreten Fehlwertersetzung. Die auf die Ersetzungsmethodik zurückführbaren Unterschiede erreichen hier nur noch geringere als kleine Effektstärken gegenüber mittleren Effektstärken auf der Ebene der Dropout-Stichproben. Während sich die Überlegenheit einer skalenbezogenen Fehlwertersetzung auf der Ebene der Gesamtstichproben nicht mehr nachweisen lässt, zeigen sich hier jedoch – im Gegensatz zur konkreten Ersetzungsebene (Ebene der Dropout-Stichproben) - zumindest signifikante Unterschiede, wenn auch mit geringerer als kleiner Effektstärke, in Abhängigkeit des dem Datenausfall zugrunde liegenden Mechanismus sowie des zu ersetzenden Kriteriums: Bei vollständig zufäl-

lig fehlenden Werten (MCAR) erweist sich das Ersetzungsergebnis in Bezug auf die resultierenden Stichprobenkennwerte der jeweiligen Gesamtstichproben nämlich als besser als bei bedingt zufällig fehlenden Werten (MAR). Außerdem führte die Ersetzung der fehlenden Werte in der PSK der SF-8 – gemessen an den resultierenden Stichprobenkennwerten – zu geringeren Abweichungen als die Ersetzung fehlender Werte im GSI der SCL-14. Neben den verschiedenen kontrolliert variierten Rahmenbedingungen zeigten sich auf beiden Untersuchungsebenen (Dropout-Stichproben und Gesamtstichproben) jeweils auch Unterschiede kleiner Effektstärke zwischen den einzelnen Einrichtungen, die jedoch nicht allein auf die unterschiedlich ausgeprägten Stichprobengrößen zurückzuführen waren.

Die Gegenüberstellung der auf der Ebene der Dropout-Stichproben und der auf der Ebene der Gesamtstichproben ermittelten Befunde verdeutlicht sehr gut, dass sich die verschiedenen untersuchten Rahmenbedingungen an unterschiedlichen Stellen auf das letztlich relevante Ergebnis der Fehlwertersetzung, nämlich die Angemessenheit der aus der Fehlwertersetzung resultierenden Stichprobenkennwerte, auswirken: Die konkrete Ersetzungsgüte, also die einzelfallbezogene Übereinstimmung zwischen der wahren und der geschätzten Ausprägung fehlender Werte, wird vor allem durch die angewandte Ersetzungsmethode und die dabei berücksichtigten Kovariaten determiniert. Da die Übereinstimmung zwischen der wahren und der geschätzten Ausprägung fehlender Werte jedoch nie vollkommen ist, können sich die durch eine unvollkommene Schätzung verursachten Abweichungen je nach den weiteren vorliegenden Rahmenbedingungen unterschiedlich gravierend auf die resultierenden Kennwerte in der jeweiligen Gesamtstichprobe auswirken. Hierbei spielen insbesondere die Fehlwertquote und der Fehlwertmechanismus eine entscheidende Rolle. Liegen in einem Datensatz nur wenige fehlende Werte vor, so können sich diese in Relation zu der großen Anzahl vorhandener Daten auch im Falle einer sehr schlechten Fehlwertersetzung nie so gravierend auf die Stichprobenkennwerte auswirken, wie dies der Fall wäre, wenn durch ein hohes Ausmaß fehlender Werte bereits ein großer Teil der stichprobenbezogenen Information fehlt. Ein etwas abweichender Mechanismus ist bezüglich der den fehlenden Werten zugrunde liegenden Systematik anzunehmen. Durch zufällig fehlende Werte werden die resultierenden Stichprobenkennwerte nämlich von vorneherein nicht oder nur geringfügig verfälscht, so dass a priori eine höhere Wahrscheinlichkeit für eine gute

Übereinstimmung zwischen den wahren und den nach der Fehlwertersetzung resultierenden Stichprobenkennwerten besteht. Umgekehrt werden die Stichprobenkennwerte durch systematisch fehlende Werte verfälscht, was nur durch eine adäquate Fehlwertersetzung ausgeglichen werden könnte. Ist eine optimale Ersetzung der fehlenden Werte jedoch nicht zu realisieren, so bleiben die durch den systematischen Datenausfall verursachten Abweichungen der Stichprobenkennwerte trotz Fehlwertersetzung (zumindest teilweise) bestehen.

Auch wenn die Gesamtschau der ermittelten Befunde zunächst den Schluss nahe legt, dass die Ersetzung fehlender Werte nur in seltenen Fällen adäquat möglich ist, bleibt doch festzustellen, dass sich bei optimaler Gestaltung der kontrollierbaren Rahmenbedingungen durchaus akzeptable Ersetzungsresultate erzielen lassen. Die Ersetzung fehlender Skalenwerte mittels Multipler Imputation resultierte bei Einbezug aller verfügbaren Kovariaten nämlich unabhängig von den vorliegenden Fehlwertquoten und Fehlwertmechanismen durchwegs in immerhin annähernd guten Übereinstimmungen zwischen der geschätzten und der wahren Ausprägung fehlender Werte (ICC um 0,5). Übertragen auf die jeweiligen Gesamtstichproben ergaben sich je nach Fehlwertquote gute bis sehr gute Übereinstimmungen zwischen den wahren und geschätzten Stichprobenkennwerten (ICC über 0,7).

6.3. Konsequenzen der Ersetzung fehlender Werte

Die möglichen Auswirkungen einer umfassenden Ersetzung fehlender Werte wurden exemplarisch anhand des vorliegenden Gesamtdatensatzes demonstriert. In einem Anwendungsbeispiel wurden die elf untersuchten Einrichtungen jeweils anhand der patientenseitig erhobenen Entlassungsdaten im GSI der SCL-14 und in der PSK der SF-8 verglichen. Der Einrichtungsvergleich erfolgte dabei zunächst auf der eingeschränkten Basis der vollständig vorliegenden Entlassungsdaten und in einem zweiten Schritt auf der Basis von mittels Multipler Imputation unter Einbezug aller jeweils verfügbaren Kovariaten vervollständigten Entlassungsdaten. Um die Auswirkungen der Fehlwertersetzung zusätzlich zu veranschaulichen, wurden die Einrich-

tungen jeweils nach der Beeinträchtigung, die die behandelten Patienten durchschnittlich zum Entlassungszeitpunkt angegeben hatten, in Rangreihen angeordnet.

Durch die Fehlwertersetzung ergaben sich bezüglich des GSI in drei der elf Einrichtungen und bezüglich der PSK in einer Einrichtung im Vergleich zu den jeweils vollständig vorliegenden Daten signifikant erhöhte Entlassungskennwerte, wobei die Unterschiede zum Teil das Ausmaß einer kleinen Effektstärke erreichten. Dies bedeutet, dass sich die Patienten der betroffenen Einrichtungen, in denen die Dropoutquoten zwischen 32 und 41 Prozent betrugen, nach der Fehlwertersetzung zum Entlassungszeitpunkt als deutlich beeinträchtigter darstellten als dies auf der Grundlage der fehlwertbehafteten Daten der Fall gewesen wäre. Übertragen auf die Ränge, die die Einrichtungen gemäß der Beeinträchtigung ihrer Patienten zum Entlassungszeitpunkt im Einrichtungsvergleich einnähmen, würden sich die betroffenen Kliniken durch die Fehlwertersetzung um drei bis vier Ränge verschlechtern. Als Folge der Rangverschiebungen dieser Einrichtungen, hätten sich die Positionen anderer Einrichtungen im Gegenzug teilweise um bis zu drei Ränge verbessert. Ergänzend sei an dieser Stelle nur kurz angemerkt, dass sich die durchgeführten Dropoutanalysen, in denen die Repräsentativität der vollständig vorliegenden Daten überprüft werden sollte, auch in diesem Falle als nicht geeignet erwiesen haben, um die durch die Fehlwertersetzung ermittelten Abweichungen vorherzusagen.

Würden die einzelnen Einrichtungen nun also nach dem Ausmaß der Beeinträchtigung ihrer Patienten zum Entlassungszeitpunkt beurteilt, so hätte die Ersetzung der fehlenden Werte erhebliche Konsequenzen auf die Bewertung der verglichenen Einrichtungen. Während drei Kliniken ohne Fehlwertersetzung (zu Unrecht) auf guten Plätzen im Mittelfeld eingestuft worden wären, müssten sie auf der Grundlage der fehlwertersetzten Daten jeweils erheblich schlechter bewertet werden. Im umgekehrten Falle könnten andere Einrichtungen eine bessere Beurteilung erzielen, obwohl sich ihre Entlassungsdaten als repräsentativ erwiesen haben.

Das vorgestellte Anwendungsbeispiel verdeutlicht somit sehr anschaulich, dass sich durch fehlende Werte verursachte Verzerrungen von Befunden im Kontext von Einrichtungsvergleichen nicht nur auf die Bewertung der jeweiligen Einrichtungen, deren Daten systematisch fehlwertbehaftet sind, auswirken, sondern auch die Bewertung anderer Einrichtungen, die gegebenenfalls überhaupt keine Fehlwerte

aufweisen müssen, beeinflussen können. Da in den meisten Fällen nicht davon auszugehen ist, dass ausgerechnet solche Daten fehlen, die zu einer positiveren Bewertung einer Einrichtung geführt hätten, könnte im Kontext von einrichtungsvergleichenden Qualitätssicherungsprogrammen, in denen fehlende Werte nicht angemessen berücksichtigt werden, schnell der Eindruck entstehen, dass sprichwörtlich „der Ehrliche der Dumme ist“. Um faire Einrichtungsvergleiche zu gewährleisten, sollte also in jedem Falle ein geeigneter Umgang mit fehlenden Werten sichergestellt sein.

6.4. Empfehlungen zum Umgang mit fehlenden Werten

Wie in der vorliegenden Arbeit gezeigt werden konnte, stellen fehlende Werte eine nicht zu vernachlässigende Größe dar, wenn es darum geht, faire Einrichtungsvergleiche zu gewährleisten. Dementsprechend sollte der Umgang mit fehlenden Werten nach Möglichkeit immer bereits im Vorfeld jeder Durchführung von einrichtungsvergleichenden Untersuchungen geplant werden.

Da sich fehlende Werte in den seltensten Fällen wirklich optimal ersetzen lassen und fehlende Informationen stets das Risiko verzerrter Befunde bergen, sollte im Vorfeld und während der Datenerhebung höchste Sorgfalt darauf verwandt werden, fehlende Werte – soweit irgend möglich – zu vermeiden. Hierzu ist es zum Beispiel notwendig, bereits bei der Auswahl der einzusetzenden Erhebungsinstrumente ein besonderes Augenmerk auf die Praktikabilität der jeweiligen Verfahren zu richten. Während der Datenerhebung kann die Sicherstellung möglichst vollständiger Datensätze dann beispielsweise durch entsprechende Prüfroutinen im Studienprotokoll (z.B. Kontrolle der Vollständigkeit von Angaben beim Einsammeln von Fragebögen oder Programmierung entsprechender Kontrollmechanismen im Falle der computergestützten Datenerhebung) unterstützt werden.

Da sich fehlende Werte jedoch nie vollständig vermeiden lassen werden, müssen im Rahmen der Studienplanung geeignete Vorkehrungen getroffen werden, die die spätere Kontrolle verzerrender Effekte durch auftretende Fehlwerte ermöglichen.

Hierzu kann es beispielsweise notwendig sein, ergänzende Informationen zu erfassen, die nicht direkt dem eigentlichen Untersuchungsgegenstand dienen, sondern gegebenenfalls lediglich als Hilfsmittel zur Sicherstellung einer adäquaten Datenverarbeitung fungieren. Im vorliegenden Beispiel des Qualitätssicherungsprogramms der Gesetzlichen Krankenkassen für den Indikationsbereich psychischer und psychosomatischer Erkrankungen nehmen insbesondere die therapeutenseitig erfassten Variablen eine solche Funktion ein: Die Therapeutenangaben werden nämlich nicht direkt als Grundlage für die Beurteilung der Ergebnisqualität der untersuchten Einrichtungen herangezogen, sondern dienen stattdessen vor allem dem Zweck, eine faire Beurteilung der patientenseitig erfassten Informationen zur Behandlungsqualität zu gewährleisten. Dazu werden sie im Rahmen von Dropoutanalysen verwendet, um die Repräsentativität der vorliegenden Patientendaten abzuschätzen, sie werden außerdem im Rahmen der Risikoadjustierung einbezogen, um das erzielte Behandlungsergebnis zu relativieren, und können letztlich, wie in der vorliegenden Arbeit demonstriert, auch im Rahmen von Fehlwertersetzungen als Kovariaten einen wichtigen Beitrag zur Ersetzungsgüte leisten. Allerdings ist in diesem Zusammenhang auf die Problematik hinzuweisen, dass die befragten Therapeuten mit ihren Angaben letztlich auch ihre eigene Leistung beurteilen müssen. Dieser Umstand birgt – insbesondere im Kontext der externen Qualitätssicherung – jedoch immer auch das Risiko positiver Verzerrungen. Insofern können therapeutenseitig erfasste Kovariaten immer nur mit entsprechender Vorsicht verwendet werden. In jedem Falle sollte der zu erwartende Datenausfall jedoch immer schon im Vorfeld der Datenerhebung im Rahmen der Untersuchungsplanung bei der Kalkulation der benötigten Stichprobenumfänge berücksichtigt werden.

Bezüglich des konkreten Umgangs mit fehlenden Werten konnte in der vorliegenden Arbeit gezeigt werden, dass letztlich nur eine adäquate Fehlwertersetzung dazu geeignet erscheint, faire Einrichtungsvergleiche zu gewährleisten. Die bis heute weit verbreitete Methode einer Eliminierung von Fällen mit fehlenden Daten in Kombination mit entsprechenden Dropoutanalysen zur Sicherstellung der Repräsentativität der verwertbaren Daten konnte nicht als geeignete Alternative zur Fehlwertersetzung bestätigt werden. Zum einen konnte in diesem Zusammenhang belegt werden, dass ein „Negativ-Befund“ bezüglich signifikanter Unterschiede zwischen Responder- und Dropout-Stichproben nicht unbedingt sicherstellt, dass auch

die Kriterien, anhand derer die jeweiligen Einrichtungen miteinander verglichen und beurteilt werden sollen, als repräsentativ für die jeweils interessierende Grundgesamtheit behandelter Patienten behandelt werden dürfen. Vor allem stellen Dropoutanalysen jedoch auch überhaupt keine Lösung zur Kontrolle verzerrter Befunde für den Fall nachweislich nicht gegebener Repräsentativität der verwertbaren Daten bereit. Da jedoch gezeigt werden konnte, dass die Berücksichtigung von nicht oder nur eingeschränkt repräsentativen Daten in Einrichtungsvergleichen nicht nur zu verzerrten Beurteilungen der jeweils von der eingeschränkten Datenrepräsentativität betroffenen Einrichtung führt, sondern zusätzlich auch die Beurteilung der übrigen beteiligten Einrichtungen verfälscht, müssten Einrichtungen mit nachweislich nicht repräsentativen Datensätzen streng genommen grundsätzlich von Einrichtungsvergleichen ausgeschlossen werden.

Um einen Ausschluss von Einrichtungen mit nachweislich nicht repräsentativen Daten von den Einrichtungsvergleichen zu vermeiden, sollten fehlende Daten also möglichst immer ersetzt werden. In der vorliegenden Arbeit haben sich Multiple Imputationen grundsätzlich als die Methode der Wahl erwiesen. Im konkreten Anwendungskontext sollte vor dem geplanten Einsatz eines bestimmten Ersetzungsverfahrens jedoch nach Möglichkeit überprüft werden, inwieweit die gewählte Ersetzungsmethode unter den gegebenen Rahmenbedingungen (Fehlwertquoten, Fehlwertmechanismen, Fehlwertmuster usw.) in adäquaten Schätzungen für die fehlenden Werte resultiert. Hierzu können Fehlwertsimulationen anhand vollständig vorliegender Teildatensätze, wie sie in der vorliegenden Arbeit vorgestellt wurden, dienen. Bestehen Zweifel an der Güte der gewählten Ersetzungsmethodik, so wäre – falls keine geeignetere Alternative zur Verfügung steht – zu erwägen, den Einrichtungsvergleich sowohl auf der Basis der fehlwertbehafteten als auch auf der Grundlage der fehlwertbereinigten Daten durchzuführen und beide Resultate zu berichten, um die Unsicherheit bezüglich der Verlässlichkeit der ermittelten Befunde zum Ausdruck zu bringen.

Wenn es jedoch gelänge, die Fehlwertquoten in vergleichenden Untersuchungen durch ein geeignetes Studienprotokoll auf maximal 20 bis 30 Prozent zu begrenzen, könnten Verzerrungen durch fehlende Werte gemäß der Befunde der vorliegenden

Arbeit durch eine geeignete Fehlwertersetzung nahezu vollständig ausgeschlossen werden.

6.5. Validität der ermittelten Befunde

Obgleich die vorliegende Arbeit eine Vielzahl bis dato offener Fragen beantworten konnte, ergibt sich aus den gewonnenen Einsichten und den Beschränkungen der durchgeführten Untersuchung wiederum eine Fülle neuer Fragestellungen, die zumeist die Validität der im Rahmen der durchgeführten Simulationsstudie ermittelten Befunde betreffen.

Im Hinblick auf die interne Validität der Untersuchung stellt sich zunächst die Frage, inwieweit die simulierten Datenausfälle tatsächlich einen wesentlichen Aspekt der im Originaldatensatz vorliegenden Bedingungen fehlender Daten abbilden konnten. Auf der einen Seite stellt das vollständige Fehlen von Patientenangaben zu einem Messzeitpunkt in der Forschungspraxis (z.B. in Studien, in denen das Ergebnis stationärer Behandlung über indirekte Veränderungsmessungen erfasst werden soll) ein zentrales Problem dar, das im Vergleich zur Problematik einzelner fehlender Werte innerhalb umschriebener Erhebungseinheiten durch die zusätzliche zeitliche Komponente erhöhte methodische Anforderungen an eine Fehlwertersetzung stellt. Auf der anderen Seite bilden fehlende Werte in einer begrenzten Anzahl von Variablen bei ansonsten vollständig vorliegenden Daten in allen übrigen erfassten Variablen in der Realität eher die Ausnahme, so dass die in der Simulationsstudie realisierten Ausfallbedingungen dahingehend als stark vereinfachend charakterisiert werden müssen. Da unklar bleibt, inwieweit sich diese beiden Aspekte verschärfter versus vereinfachter Rahmenbedingungen gegenseitig die Waage halten, sind weitere Untersuchungen vonnöten, in denen neben den Fehlwertquoten und Fehlwertmechanismen auch die entsprechenden Fehlwertmuster simuliert werden müssten. Erst auf der Basis solcher Studien ließe sich die Frage beantworten, inwieweit sich die in der Simulation ermittelten Befunde zur Ersetzbarkeit fehlender Entlassungsdaten tatsächlich auf die real vorliegenden Bedingungen fehlender Wer-

te verallgemeinern lassen. Weiterer Forschungsbedarf ergibt sich in diesem Zusammenhang außerdem bezüglich der Identifikation von Mindestvoraussetzungen (z.B. bezüglich der Qualität und Quantität verfügbarer Kovariaten) für eine adäquate Fehlwertersetzung sowie bezüglich der Überprüfung der Ersetzbarkeit von fehlenden Follow-up-Daten.

Ein weiterer Gesichtspunkt der Gültigkeit der ermittelten Befunde betrifft nicht zuletzt die Spezifität der fehlwertbehafteten Variablen. Im Rahmen der vorliegenden Simulationsstudie wurde die Ersetzbarkeit fehlender Werte zunächst exemplarisch für die Items zweier ausgewählter Skalen überprüft (GSI der SCL-14 und PSK der SF-8). Inwieweit sich die ermittelten Befunde auch auf andere Ergebnismaße oder sonstige erfasste Merkmale übertragen lassen, müsste dementsprechend erst noch in weiteren Studien untersucht werden.

Ein weiterer Aspekt der internen Validität der vorliegenden Untersuchung betrifft die Operationalisierung der verschiedenen Ersetzungsmethoden. Zur Imputation fehlender Werte steht grundsätzlich eine Vielzahl kommerzieller und nichtkommerzieller Statistikprogramme zur Verfügung. In der vorliegenden Arbeit wurde die EM-Ersetzung über das „Missing Values Analysis“-Modul (MVA) des Statistikpaketes SPSS (Norusis, 2004) und die Multiplen Imputationen (MI) über das STATA-Add-on „ICE“ (Royston, 2005; Stata Corporation, 2005) realisiert. Dabei stellt sich nun die Frage, inwieweit diese beiden Programme als repräsentative Vertreter für „die“ EM-Ersetzung oder „die“ Multiple Imputation angesehen werden können. Bezüglich der EM-Ersetzung ergeben sich an einer solchen Annahme nicht zuletzt angesichts der im Vergleich zur MI - und insbesondere auch zur Regressionsschätzung - unerwartet deutlich geringeren Ersetzungsgüte berechtigte Zweifel. Hinweise auf die eingeschränkte Eignung des MVA-Moduls zur EM-Imputation finden sich auch bei von Hippel (2004), der die fehlende Berücksichtigung residualer Varianz bei der Fehlwertersetzung als Ursache für teilweise stark verzerrte Schätzwerte anführt. In Bezug auf die Multiple Imputation ist eventuell einschränkend anzumerken, dass die im Programm ICE realisierte MI nicht auf der Basis von EM-Schätzern vollzogen wird. Möglicherweise könnten somit andere Statistik-Lösungen, wie z.B. das von Schafer entwickelte Programm NORM (Schafer, 1997), in dem fehlende Werte zu-

nächst mittels der EM-Ersetzung geschätzt und daraufhin multipel imputiert werden, zu noch besseren Ersetzungsresultaten führen.

In Bezug auf die externe Validität der ermittelten Befunde ergibt sich die zentrale Frage, inwieweit die in der vorliegenden Arbeit anhand eines Datensatzes zur kurzfristigen Ergebnisqualität aus dem Bereich der Qualitätssicherung stationärer Rehabilitation von Patienten mit psychischen und psychosomatischen Erkrankungen ermittelten Befunde sich auch auf andere Anwendungsfelder übertragen lassen. Dabei wäre zum einen die Übertragbarkeit der ermittelten Befunde auf alternative Versorgungsbereiche desselben Indikationsfeldes (z.B. ambulante bzw. teilstationäre psychosomatisch/psychotherapeutische Rehabilitation, aber auch die stationäre Behandlung in psychosomatischen Akutkrankenhäusern) zu überprüfen. Vor allem gilt es jedoch, die adäquate Ersetzbarkeit fehlender Werte auch in anderen Indikationsbereichen der medizinischen Rehabilitation sicherzustellen, um der Forderung des Gesetzgebers nach einer einrichtungsvergleichenden Qualitätssicherung in Form von fairen Einrichtungsvergleichen nachkommen zu können.

7. Fazit

Die angemessene Berücksichtigung fehlender Daten stellt im Kontext der einrichtungsvergleichenden Qualitätssicherung eine grundlegende Voraussetzung zur Sicherstellung fairer Einrichtungsvergleiche dar. Im Rahmen der vorliegenden Arbeit konnte gezeigt werden, dass die adäquate Ersetzung fehlender Werte bei entsprechender Berücksichtigung der zu kontrollierenden Rahmenbedingungen grundsätzlich auch bei höheren Fehlwertquoten und systematischem Datenausfall möglich ist und somit in Bezug auf die Vermeidung verzerrender Auswirkungen von fehlenden Werten in Einrichtungsvergleichen die Methode der Wahl darstellen sollte. Die grundlegenden Voraussetzungen für eine optimale Fehlwertersetzung sollten in weiteren Studien vertiefend untersucht werden.

8. Zusammenfassung

Einrichtungsvergleiche stellen in der medizinischen Versorgung ein zentrales Instrument der Qualitätssicherung dar, das im Bereich der Rehabilitation sogar explizit vom Gesetzgeber gefordert wird (vgl. §20 Abs. 1, SGB IX). Der direkte Vergleich verschiedener Einrichtungen eines Indikationsbereichs soll dabei als Grundlage für die Beurteilung der Qualität der erbrachten medizinischen Leistungen einzelner Kliniken dienen. Allerdings ist bei derartigen Vergleichen immer zu berücksichtigen, dass die jeweils vorliegenden Daten nicht unbedingt direkt die Behandlungsqualität der verschiedenen untersuchten Einrichtungen widerspiegeln müssen. Neben der tatsächlichen Qualität der erbrachten Leistungen determinieren nämlich vor allem die spezifischen Eigenschaften der behandelten Patienten das Therapieergebnis, ohne dass diese durch die leistungserbringenden Einrichtungen zu beeinflussen wären. Zur Kontrolle derartig mit dem Behandlungsergebnis konfundierter Faktoren haben sich statistische Verfahren der Risikoadjustierung bewährt.

Während die Notwendigkeit einer Risikoadjustierung in der aktuellen Literatur – zu Recht – als unerlässlich für die Durchführung fairer Einrichtungsvergleiche proklamiert wird, rückt eine andere – nur scheinbar triviale – Problematik in den Hintergrund. In empirischen Untersuchungen ergeben sich nämlich immer mehr oder weniger hohe Quoten fehlender Daten, die die Repräsentativität ermittelter Befunde erheblich einschränken können. Einrichtungsvergleiche können jedoch nur dann zu validen Aussagen führen, wenn die Daten, auf denen sie basieren, die Versorgungsrealität innerhalb der verglichenen Einrichtungen zuverlässig abbilden. Spiegeln die vorliegenden Daten die Behandlungsqualität der verglichenen Einrichtung jedoch nicht ausreichend wider, so kann auch eine noch so elaborierte Risikoadjustierung nicht sicherstellen, dass die resultierenden Bewertungen der Einrichtungen fair ausfallen.

Der Anspruch der vorliegenden Arbeit war es daher, verschiedene Möglichkeiten zum Umgang mit fehlenden Informationen im Rahmen einer Simulationsstudie systematisch hinsichtlich ihrer Voraussetzungen respektive ihrer Konsequenzen auf die Bewertung verglichener Einrichtungen zu überprüfen. Untersucht wurde hierzu eine konsekutive Stichprobe von insgesamt N=2.386 Patienten, die in 11 Rehabilitations-

Fachkliniken für Psychosomatische Medizin und Psychotherapie stationär behandelt wurden. Informationen zur Ergebnisqualität wurden zu Beginn und Ende der Rehabilitationsmaßnahme über standardisierte Selbst- und Fremdbeurteilungsinstrumente erfasst.

An einem vollständigen Teildatensatz von n=1.248 Patienten wurden nach verschiedenen Kriterien künstliche Datenausfälle erzeugt. Variiert wurden dabei die Fehlwertmechanismen (MCAR vs. MAR) sowie die Quoten fehlender Werte (10-50 Prozent). Die simuliert fehlenden Daten wurden daraufhin in verschiedenen methodischen Varianten ersetzt. Hierbei wurden die drei Ersetzungsmethoden „Regressionsschätzung“ (RA), „EM-Imputation“ (EM) und „Multiple Imputation“ (MI) überprüft. Weitere Variationen bezogen sich beispielsweise auf die Anzahl und Qualität berücksichtigter Kovariaten. Insgesamt wurden 6.600 verschiedene Ersetzungsvarianten realisiert. Die imputierten Daten wurden jeweils mittels Intraklassen-Korrelations-Koeffizienten (ICC) auf Übereinstimmung mit den Original-Daten geprüft.

Die Güte der Fehlwertersetzung schwankt im Einzelfall je nach eingesetzten Imputationsverfahren und überprüften Rahmenbedingungen erheblich, die Spannweite der ermittelten ICC reicht dabei von 0,02 (überhaupt keine Übereinstimmung) bis 1,00 (perfekte Übereinstimmung). Die Ersetzung fehlender Skalenwerte mittels Multipler Imputation resultiert bei Einbezug aller verfügbaren Kovariaten unabhängig von den vorliegenden Fehlwertquoten und Fehlwertmechanismen durchwegs in immerhin annähernd guten Übereinstimmungen zwischen der geschätzten und der wahren Ausprägung fehlender Werte (ICC um 0,5). Übertragen auf die jeweiligen Gesamtstichproben ergeben sich je nach Fehlwertquote gute bis sehr gute Übereinstimmungen zwischen den wahren und geschätzten Stichprobenkennwerten (ICC über 0,7).

Die adäquate Ersetzung fehlender Daten erweist sich unter bestimmten Voraussetzungen als möglich und sollte daher die Methode der Wahl darstellen, wenn es darum geht, verzerrende Auswirkungen von fehlenden Werten in Einrichtungsvergleichen zu vermeiden. Die angemessene Berücksichtigung fehlender Daten ist im Kontext der einrichtungsvergleichenden Qualitätssicherung unerlässlich, da sie unter Umständen erhebliche Konsequenzen bezüglich der „fairen“ Bewertung der Qualität untersuchter Einrichtungen zeigen können.

9. Literaturverzeichnis

Afifi, A., Elashoff, R.M. (1966). Missing Oberservations in multivariate statistics I: Review of the Literature. *Journal of the American Statistical Association, 61*, 595-605.

Allison, P.D. (2002). *Missing Data*. Thousand Oaks, CA: Sage Publications.

Andreas, S. (2005). *Fallgruppen in der Versorgung von Patienten mit psychischen Störungen - Überprüfung der Eignung eines Fremdeinschätzungsinstrumentes "Die Health of the Nation Outcome Scales, HoNOS-D" zur differenzierten Erfassung des Schweregrades im Rahmen der Entwicklung eines Klassifikationssystems.* Hamburg: Staats- und Universitätsbibliothek.

Andreas, S., Harfst, T., Dirmaier, J., Kawski, S., Koch, U., Schulz, H. (in press). A psychometric evaluation of the German version of the "Health of the Nation Outcome Scales, HoNOS-D": On the feasibility and reliability of a clinician-rated measure of severity in patients with mental disorders. *Psychopathology*.

Audin, K., Marginson, F.R., Clark, J.M., Barkham, M. (2001). Value of HoNOS in assessing patient change in NHS psychotherapy and psychological treatment services. *British Journal of Psychiatry, 178*, 561-566.

Beale, E.M.L., Little, R.J.A. (1975). Missing Data in Multivariate Analysis. *Journal of the Royal Statistical Society, Series B, 37*, 129-146.

Bortz, J. (1999). *Statistik für Sozialwissenschaftler (5. Auflage)*. Berlin: Springer.

Buck, S.F. (1960). A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer. *Journal of the Royal Statistical Society, Series B, 22*, 302-306.

Bullinger, M. (2000). Erfassung der gesundheitsbezogenen Lebensqualität mit dem SF-36-Health Survey. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz, 43*, 190-197.

Bullinger, M., Kirchberger, I. (1998). *SF-36. Fragebogen zum Gesundheitszustand*. Göttingen: Hogrefe.

Cochran, W.G. (1977). *Sampling Techniques*. New York: Wiley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Collins, L.M., Schafer, J.L., Kam, C.M. (2001). A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. *Psychological Methods, 6*, 330-351.

Fahrenberg, J., Myrtek, M., Schumacher, J., Brähler, E. (2000). *Fragebogen zur Lebenszufriedenheit (FLZ): Handanweisung*. Göttingen: Hogrefe.

Farin, E., Gerdes, N., Jäckel, W.H., Follert, P., Klein, K., Glattacker, M. (2003). "Qualitätsprofile" von Rehabilitationskliniken als Modell der Qualitätsmessung in Einrichtungen des Gesundheitswesens. *Gesundheitsökonomie und Qualitätsmanagement, 8*, 191-204.

Farin, E., Glattacker, M., Follert, P., Kuhl, H.C., Klein, K., Jäckel, W.H. (2004). Einrichtungsvergleiche in der medizinischen Rehabilitation. *Zeitschrift für ärztliche Fortbildung und Qualitätssicherung im Gesundheitswesen, 98*, 655-662.

Fisher, R.A. (1956). *Statistical methods of research workers*. London: Oliver and Boyd.

Franke, G. (1995). *SCL-90-R. Die Symptom-Checkliste von Derogatis - Deutsche Version*. Göttingen: Beltz.

Franke, G. (2002). *SCL-90-R. Die Symptom Checkliste von L.R. Derogatis - Manual zur Deutschen Version*. Beltz.

Greve, W., Wentura, D. (1997). *Wissenschaftliche Beobachtungen. Eine Einführung*. Weinheim: PVU.

Hannöver, W., Kordy, H. (2005). Predicting outcomes of inpatient psychotherapy using quality management data: comparing classification and regression trees with logistic regression and linear discriminant analysis. *Psychotherapy Research, 15*, 236-257.

Harfst, T., Koch, U., Kurtz von Aschoff, C., Nutzinger, D.O., Rüddel, H., Schulz, H. (2002). Entwicklung und Validierung einer Kurzform der Symptom Checklist-90-R. *DRV-Schriften, 33*, 71-73.

Hartung, J. (1981). Non-negative minimum biased invariant estimation in variance component models. *Annals of Statistics, 9*, 278-292.

Hautzinger, M., Bailer, M. (1993). *Allgemeine Depressionsskala (ADS). Die deutsche Version des CES-D*. Weinheim: Beltz.

Hautzinger, M., Bailer, M. (2002). Die Allgemeine Depressions Skala (ADS). In: E. Brähler, J. Schumacher, B. Strauß (Hrsg.). *Diagnostische Verfahren in der Psychotherapie*. Göttingen: Hogrefe Verlag, 25-28.

Hedeker, D., Gibbons, R.D. (1997). Application of Random-Effects Pattern-Mixture Models to Missing Data in Longitudinal Studies. *Psychological Methods, 2*, 64-78.

Horowitz, L.M., Strauß, B., Kordy, H. (1994). *Inventar zur Erfassung interpersonaler Probleme - Deutsche Version - (IIP-D)*. Weinheim: Beltz.

Iezzoni, L.I. (1997). Risk adjustment for measuring healthcare outcomes. Chicago: Health Administration Press.

Kawski, S., Koch, U. (1999). Qualitätssicherung in der Psychosomatischen Rehabilitation. *Psychotherapie, Psychosomatik, Medizinische Psychologie, 49*, 316-325.

Kawski, S., Koch, U. (2002). Zum Stand der Qualitätssicherung in der Rehabilitation. Zur Entwicklung der medizinischen Rehabilitation in den 90er-Jahren. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz, 45*, 260-266.

Kawski, S., Koch, U. (2004). Qualitätssicherung in der medizinischen Rehabilitation in Deutschland. Entwicklungsstand und Perspektiven. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz, 47*, 111-117.

Klein, K.J., Bliese, P.D., Kozlowski, S.W., Dansereau, F., Gavin, M.B., Griffin, M.A., Hofmann, D.A., James, L.R., Yammarino, F.J., Bligh, M.C. (2000). Multilevel analytical techniques: Commonalities, differences, and continuing questions. In: K.J. Klein, S.W. Kozlowski (Hrsg.). *Multilevel theory, research, and methods in organizations*. San Francisco: Jossey-Bass, 512-553.

Leary, T. (1957). *Interpersonal diagnosis of personality*. New York: Ronald.

Little, R.J.A. (1983). The Ignorable Case. In: W.G. Madow, I. Olkin, D.B. Rubin (Hrsg.). *Incomplete Data in Sample Surveys, Vol.2*. New York: Academic Press.

Little, R.J.A., Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Little, R.J.A., Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: Wiley.

McGraw, K.O., Wong, S.P. (1996). Forming Inferences About Some Intraclass Correlation Coefficients. *Psychological Methods, 1*, 30-46.

Musil, C.M., Warner, C.B., Yobas, P.K., Jones, S.L. (2002). A Comparison of Imputation Techniques for Handling Missing Data. *Western Journal of Nursing Research, 24*, 815-829.

Norusis, M.J. (2004). *SPSS 12 Guide to Data Analysis*. Upper Saddle River, NJ: Prentice Hall.

Radloff, L. (1977). The CES-D scale: A self-report depression scale for the research in the general population. *Applied Psychological Measurement, 1*, 385-401.

Royston, P. (2005). Multiple imputation of missing values: update. *Stata Journal, 5*, 188-201.

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika, 63*, 581-592.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D.B., Schenker, N. (1986). Multiple Estimation for Interval Imputation from Simple Random Samples with Ignorable Nonresponse. *Journal of the American Statistical Association, 81*, 366-374.

Sbarra, D.A., Emery, R.E. (2005). Coparenting Conflict, Nonacceptance, and Depression Among Divorced Adults: Results From a 12-Year Follow-Up Study of Child Custody Mediation Using Multiple Imputation. *American Journal of Orthopsychiatry, 75*, 63-75.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J.L. (1999). Multiple Imputation: a primer. *Statistical Methods in Medical Research, 8*, 3-15.

Schafer, J.L., Graham, J.W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods, 7*, 147-177.

Schauenburg, H., Strack, M. (1998). Die Symptom Checklist-90-R (SCL-90-R) zur Darstellung von statistisch und klinischsignifikanten Psychotherapieergebnissen. *Psychotherapie Psychosomatik Medizinische Psychologie, 48*, 257-264.

Schauenburg, H., Strack, M. (1999). Measuring psychotherapeutic change with the Symptom Checklist 90 R - SCL 90 R. *Psychotherapy and Psychosomatics, 68*, 199-207.

Scheffer, J. (2002). Dealing with Missing Data. *Res. Lett. Inf. Math. Sci., 3*, 153-160.

Schulz, H., Andreas, S., Dirmaier, J., Kawski, S., Harfst, T., Rabung, S., Watzke, B., Koch, U. (in Druck). Erfassung der Lebensqualität. In: A. von Leupoldt, T. Ritz (Hrsg.). *Verhaltensmedizin - Perspektiven aus Psychobiologie, Psychopathologie und klinischer Anwendung*.

Schulz, H., Barghaan, D., Watzke, B., Koch, U., Harfst, T. (2004). Klinikvergleiche als Instrument der Qualitätssicherung in der Rehabilitation von Patienten mit psychischen/psychosomatischen Störungen: Bedeutung von Risikoadjustierung. *Zeitschrift für ärztliche Fortbildung und Qualitätssicherung im Gesundheitswesen, 98*, 663-672.

Schulz, H., Harfst, T., Dirmaier, J., Watzke, B., Andreas, S., Kawski, S., Rabung, S., Koch, U. (in Vorbereitung). Konstruktion und erste psychometrische Überprüfung einer Fremdeinschätzungsversion des SF-8-Fragebogens.

Schulz, H., Koch, U. (2002). Qualitätssicherung in der psychotherapeutischen Medizin. In: S. Ahrens, W. Schneider (Hrsg.). *Lehrbuch der Psychotherapie und Psychosomatischen Medizin*. Stuttgart: Schattauer, 17-27.

Shwartz, M., Ash, A.S., Iezzoni, L.I. (1997). Comparing outcomes across providers. In: L.I. Iezzoni (Hrsg.). *Risk adjustment for measuring health care outcomes*. Chicago: Health Administration Press, 471-516.

Sinharay, S., Stern, H.S., Russel, D. (2001). The Use of Multiple Imputation for the Analysis of Missing Data. *Psychological Methods, 6*, 317-329.

Stata Corporation (2005). *Stata User's Guide. Release 9*. College Station, Tex.: Stata Press.

von Hippel, P.T. (2004). Biases in SPSS 12.0 Missing Value Analysis. *The American Statistician, 58*, 160-164.

Ware, J.E. (2000). SF- 36 Health survey update. *Spine, 25*, 3130- 3139.

Ware, J.E., Kosinski, M., Dewey, J.E., Gandek, B. (2000). How to score and interpret single-item health status measures: A manual for users of the SF-8 Health Survey. Lincoln: Quality Metric.

Wegscheider, K. (2004). Methodische Anforderungen an Einrichtungsvergleiche ('Profiling') im Gesundheitswesen. *Zeitschrift für ärztliche Fortbildung und Qualitätssicherung im Gesundheitswesen, 98*, 647-654.

Wing, J.K., Beevor, A.S., Curtis, R.H., Park, S.B.G., Hadden, S., Burns, A. (1998). Health of the Nation Outcome Scales (HoNOS): Research and development. *British Journal of Psychiatry, 172*, 11-18.

Wirtz, M. (2004). Über das Problem fehlender Werte: Wie der Einfluss fehlender Informationen auf Analyseergebnisse entdeckt und reduziert werden kann. *Rehabilitation, 43*, 109-115.

Wirtz, M., Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.

10. Abbildungsverzeichnis

Abbildung 1: Determinanten der Ergebnisqualität

Abbildung 2: Determinanten der gemessenen Ergebnisqualität

Abbildung 3: Ergebnismaße und Hilfsvariablen zur Schätzung fehlender Ergebniswerte

Abbildung 4: Ersetzungsvarianten durch Kombination von Dropoutquoten, Ersetzungsmethoden und Prädiktoren

Abbildung 5: Einrichtungsbezogene Ersetzungsvarianten durch Kombination aller Rahmenbedingungen

Abbildung 6: Formeln zur Bestimmung klassischer und korrigierter Intraklassen-Korrelations-Koeffizienten (ICC) – unjustiertes einfaktorielles Modell (n. Wirtz & Caspar 2002)

Abbildung 7: Formel zur Transformation von ICC in Fisher-Z-Werte (n. Bortz 1999)

Abbildung 8: Anteile fallbezogen fehlender Werte in den einzelnen Erhebungsmodulen

Abbildung 9: Verteilung klassischer und korrigierter Intraklassen-Korrelations-Koeffizienten (ICC) in den 6.600 Ersetzungsvarianten – bezogen auf die Dropout-Stichproben

Abbildung 10: Verteilung klassischer und korrigierter Intraklassen-Korrelations-Koeffizienten (ICC) in den 6.600 Ersetzungsvarianten - bezogen auf die Gesamtstichproben

Abbildung 11: Stichprobenkennwerte nach Fehlwertersetzung bei zufällig fehlenden Werten (MCAR - skalenweise Ersetzung fehlender Werte des Globalen Symptomschwereindex der SCL-14 unter Einbezug aller verfügbaren Kovariaten)

Abbildung 12: Stichprobenkennwerte nach Fehlwertersetzung bei systematisch fehlenden Werten (MAR - skalenweise Ersetzung fehlender Werte des Globalen Symptomschwereindex der SCL-14 unter Einbezug aller verfügbaren Kovariaten)

Abbildung 13: Klinikrangreihen nach Globalem Symptomschwereindex (GSI) der SCL-14 - mit und ohne Fehlwertersetzung

Abbildung 14: Klinikrangreihen nach Psychischer Summenskala (PSK) der SF-8 - mit und ohne Fehlwertersetzung

11. Tabellenverzeichnis

Tabelle 1: Unterscheidung von Ausfallmechanismen nach den Ursachen für fehlende Daten

Tabelle 2: Gängige Verfahren zum Umgang mit fehlenden Werten

Tabelle 3: Vor- und Nachteile gängiger Verfahren zum Umgang mit fehlenden Werten

Tabelle 4: Handhabbarkeit der verschiedenen Mechanismen fehlender Werte

Tabelle 5: Instrumentarium zur Erfassung der Ergebnisqualität in der Pilotphase des QS-Reha®-Verfahrens (Indikationsbereich psychische/psychosomatische Erkrankungen)

Tabelle 6: Soziodemografische und klinische Stichprobencharakteristika*

Tabelle 7: Kriterien für den Einschluss von Fällen in die Simulationsstichprobe

Tabelle 8: Rahmenbedingungen der Überprüfung potentieller Einflussfaktoren auf die Güte der Ersetzung fehlender Patientenangaben zum Entlassungszeitpunkt

Tabelle 9: Durchschnittliche Anteile fall- und itembezogen fehlender Werte

Tabelle 10: Stichprobenumfänge nach Vollständigkeit der Daten

Tabelle 11: Klinikbezogene Fallzahlen mit vollständigen Daten nach Dropoutsimulation

Tabelle 12: Unterschiede zwischen Original- und Dropout-Stichproben bezüglich der Ausprägung der Ergebnismaße (Patientenangaben zur Entlassung)

Tabelle 13: Unterschiede zwischen Original- und Dropout-Stichproben bezüglich der Ausprägung der Basisdaten (Patientenangaben zur Aufnahme)

Tabelle 14: Güte der Fehlwertersetzung nach Ersetzungsmethoden (korrigierte ICC)

Tabelle 15: Güte der Fehlwertersetzung nach Systematik des Datenausfalls (korrigierte ICC)

Tabelle 16: Güte der Fehlwertersetzung nach Fehlwertquoten (korrigierte ICC)

Tabelle 17: Güte der Fehlwertersetzung nach Prädiktoren (Kovariaten) (korrigierte ICC)

Tabelle 18: Güte der Fehlwertersetzung nach Komplexität des Kriteriums (korrigierte ICC)

Tabelle 19: Güte der Fehlwertersetzung nach Ergebnismaßen (korrigierte ICC)

Tabelle 20: Güte der Fehlwertersetzung nach Kliniken (korrigierte ICC)

Tabelle 21: Übersicht zu den Ergebnissen der Hypothesentestung

Tabelle 22: Güte der Fehlwertersetzung nach günstigsten vs. ungünstigsten Rahmenbedingungen (korrigierte ICC)

Tabelle 23: Güte der Fehlwertersetzung nach Ausfallmechanismen und Fehlwertquoten bei skalenweiser Ersetzung unter Einbezug aller verfügbaren Kovariaten mittels Multipler Imputation (mittlere korrigierte ICC [mit 95%-Konfidenzintervall])

Tabelle 24: Güte der Fehlwertersetzung nach Ausfallmechanismen und Fehlwertquoten bei skalenweiser Ersetzung unter Einbezug aller verfügbaren Kovariaten mittels EM-Ersetzung (mittlere korrigierte ICC [mit 95%-Konfidenzintervall])

Tabelle 25: Anteile von Patienten mit fehlenden Werten in den Items des Globalen Symptomschwereindex (GSI) der SCL-14 und der Psychischen Summenskala (PSK) der SF-8 zur Entlassung (T1) und Anteile verfügbarer Kovariaten zur Fehlwertersetzung

Tabelle 26: Repräsentativität der Responderstichproben der einzelnen Kliniken

Tabelle 27: Berücksichtigte Variablen in den verschiedenen Erhebungsmodulen

Tabelle 28: Verteilungseigenschaften der Ergebnismaße nach Quoten zufällig fehlender Werte (Angaben zu T1 - Entlassung)

Tabelle 29: Verteilungseigenschaften der Ergebnismaße nach Quoten systematisch fehlender Werte (Angaben zu T1 - Entlassung)

Tabelle 30: Verteilungseigenschaften der Basisdaten nach Quoten zufällig fehlender Werte (Angaben zu T0 - Aufnahme)

Tabelle 31: Verteilungseigenschaften der Basisdaten nach Quoten systematisch fehlender Werte (Angaben zu T0 - Aufnahme)

Tabelle 32: Übereinstimmung zwischen geschätzten und Originaldaten bei Fehlwertersetzung mittels Regressionsschätzung auf Ebene der Dropoutfälle [normaler Druck] und auf Ebene der Einrichtungsstichproben [*kursiv*] (korrigierte ICC, aggregiert über Einrichtungsergebnisse)

Tabelle 33: Übereinstimmung zwischen geschätzten und Originaldaten bei Fehlwertersetzung mittels EM-Imputation auf Ebene der Dropoutfälle [normaler Druck] und auf Ebene der Einrichtungsstichproben [*kursiv*] (korrigierte ICC, aggregiert über Einrichtungsergebnisse)

Tabelle 34: Übereinstimmung zwischen geschätzten und Originaldaten bei Fehlwertersetzung mittels Multipler Imputation auf Ebene der Dropoutfälle [normaler Druck] und auf Ebene der Einrichtungsstichproben [*kursiv*] (korrigierte ICC, aggregiert über Einrichtungsergebnisse)

12. Anhang

12.1. Berücksichtigte Variablen

Tabelle 27: Berücksichtigte Variablen in den verschiedenen Erhebungsmodulen

Patientenangaben zu T0 (Aufnahme)	
Geschlecht	(1 Item)
Alter	(1 Item)
Nationalität	(1 Item)
Partnersituation	(1 Item)
höchster Schulabschluss	(1 Item)
aktuelle berufliche Situation	(1 Item)
Antrag auf Berentung	(1 Item)
Globalurteile zu verschiedenen Beeinträchtigungen	(18 Items)
Fragebogen zur Lebenszufriedenheit (FLZ)	(8 Items)
Globalurteil zur Lebenszufriedenheit	(1 Item)
Symptom Checklist (SCL-14)	(14 Items)
Allgemeine Depressionsskala (ADS-K)	(15 Items)
Fragen zum allgemeinen Gesundheitszustand (SF-8)	(8 Items)
Inventar Interpersonaler Probleme (IIP-64)	(64 Items)
AU-/Krankheitszeiten in den letzten 6 Monaten	(1 Item)
Patientenangaben zu T1 (Entlassung)	
Symptom Checklist (SCL-14)	(14 Items)
Fragen zum allgemeinen Gesundheitszustand (SF-8)	(4 Items)
Therapeutenangaben zu T0 (Aufnahme)	
Berentung	(1 Item)
Antrag auf Berentung	(1 Item)
Kostenträger	(1 Item)
Diagnosen nach ICD-10	(5 Items)
Chronifizierung der Haupterkrankung	(1 Item)
Reha-Motivation	(1 Item)
Fragen zum allgemeinen Gesundheitszustand (SF-8-Fremd)	(8 Items)
Health of the Nation Outcome Scales (HoNOS-D)	(12 Items)
Therapeutenangaben zu T1 (Entlassung)	
Fragen zum allgemeinen Gesundheitszustand (SF-8-Fremd)	(8 Items)
Health of the Nation Outcome Scales (HoNOS-D)	(12 Items)
Dauer der Rehabilitation	(1 Item)

12.2. Verteilungseigenschaften der Ergebnismaße nach Dropoutsimulation

Tabelle 28: Verteilungseigenschaften der Ergebnismaße nach Quoten zufällig fehlender Werte (Angaben zu T1 - Entlassung)

		Kliniken											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Gesamt**
Globaler Symptomschwereindex der SCL-14													
Original	Mittelwert	0,7	1,0	0,7	0,7	0,8	0,7	0,7	0,9	0,9	0,9	0,9	0,8
	Standardabweichung	*0,6*	*0,7*	*0,5*	*0,6*	*0,7*	*0,6*	*0,7*	*0,7*	*0,7*	*0,8*	*0,7*	*0,7*
10% Dropout	Mittelwert	0,7	1,0	0,7	0,7	0,8	0,7	0,8	0,9	0,9	0,9	0,9	0,8
	Standardabweichung	*0,6*	*0,8*	*0,6*	*0,6*	*0,7*	*0,6*	*0,7*	*0,7*	*0,7*	*0,9*	*0,7*	*0,7*
20% Dropout	Mittelwert	0,7	1,0	0,7	0,7	0,8	0,6	0,8	0,8	0,9	0,9	0,9	0,8
	Standardabweichung	*0,6*	*0,8*	*0,5*	*0,6*	*0,7*	*0,6*	*0,7*	*0,6*	*0,7*	*0,8*	*0,7*	*0,7*
30% Dropout	Mittelwert	0,7	1,0	0,7	0,7	0,7	0,7	0,8	0,9	1,0	0,9	0,8	0,8
	Standardabweichung	*0,6*	*0,8*	*0,5*	*0,6*	*0,7*	*0,6*	*0,6*	*0,6*	*0,7*	*0,9*	*0,6*	*0,7*
40% Dropout	Mittelwert	0,7	1,0	0,7	0,8	0,8	0,6	0,8	1,0	0,8	1,0	0,9	0,8
	Standardabweichung	*0,6*	*0,8*	*0,6*	*0,6*	*0,7*	*0,6*	*0,7*	*0,7*	*0,6*	*0,9*	*0,7*	*0,7*
50% Dropout	Mittelwert	0,7	0,9	0,7	0,7	0,7	0,6	0,8	0,9	0,9	0,8	1,1	0,8
	Standardabweichung	*0,7*	*0,7*	*0,5*	*0,6*	*0,7*	*0,6*	*0,7*	*0,7*	*0,6*	*0,8*	*0,7*	*0,7*
Psychische Summenskala der SF-8													
Original	Mittelwert	1,2	1,7	1,6	1,4	1,6	1,4	1,5	1,7	1,7	1,5	1,7	1,5
	Standardabweichung	*0,8*	*0,9*	*0,9*	*0,9*	*0,9*	*1,0*	*0,9*	*,9*	*0,9*	*1,1*	*0,9*	*0,9*
10% Dropout	Mittelwert	1,2	1,8	1,5	1,4	1,6	1,4	1,5	1,7	1,7	1,6	1,7	1,5
	Standardabweichung	*0,9*	*0,9*	*0,9*	*0,9*	*0,9*	*1,0*	*0,9*	*,9*	*0,9*	*1,1*	*0,9*	*0,9*
20% Dropout	Mittelwert	1,1	1,8	1,6	1,5	1,6	1,4	1,5	1,7	1,6	1,5	1,7	1,5
	Standardabweichung	*0,8*	*1,0*	*0,9*	*0,9*	*0,9*	*1,0*	*0,9*	*,9*	*0,9*	*1,1*	*0,9*	*0,9*
30% Dropout	Mittelwert	1,2	1,8	1,5	1,4	1,5	1,4	1,5	1,8	1,8	1,6	1,7	1,5
	Standardabweichung	*0,9*	*1,0*	*0,9*	*0,9*	*0,9*	*1,0*	*0,9*	*,9*	*0,9*	*1,1*	*0,8*	*0,9*
40% Dropout	Mittelwert	1,1	1,8	1,6	1,5	1,5	1,4	1,5	1,8	1,7	1,7	1,8	1,6
	Standardabweichung	*0,8*	*1,0*	*0,9*	*1,0*	*0,9*	*0,9*	*0,9*	*,9*	*0,9*	*1,1*	*0,8*	*0,9*
50% Dropout	Mittelwert	1,2	1,6	1,6	1,4	1,5	1,3	1,5	1,7	1,6	1,5	1,9	1,5
	Standardabweichung	*0,9*	*0,9*	*0,9*	*0,9*	*0,9*	*0,9*	*0,9*	*1,0*	*0,8*	*1,1*	*0,9*	*0,9*

Tabelle 29: Verteilungseigenschaften der Ergebnismaße nach Quoten systematisch fehlender Werte (Angaben zu T1 – Entl.)

		Kliniken											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Gesamt**
Globaler Symptomschwereindex der SCL-14													
Original	Mittelwert	0,7	1,0	0,7	0,7	0,8	0,7	0,7	0,9	0,9	0,9	0,9	0,8
	Standardabweichung	*0,6*	*0,7*	*0,5*	*0,6*	*0,7*	*0,6*	*0,7*	*0,7*	*0,7*	*0,8*	*0,7*	*0,7*
10% Dropout	Mittelwert	0,6	1,0	0,6	0,7	0,7	0,6	0,8	0,9	0,9	0,8	0,9	0,8
	Standardabweichung	*0,5*	*0,8*	*0,5*	*0,6*	*0,7*	*0,6*	*0,7*	*0,6*	*0,7*	*0,8*	*0,7*	*0,7*
20% Dropout	Mittelwert	0,6	1,0	0,6	0,7	0,7	0,7	0,7	1,0	0,9	0,8	0,9	0,8
	Standardabweichung	*0,6*	*0,8*	*0,5*	*0,6*	*0,7*	*0,6*	*0,6*	*0,7*	*0,7*	*0,8*	*0,8*	*0,7*
30% Dropout	Mittelwert	0,5	1,0	0,5	0,7	0,7	0,6	0,7	0,9	0,9	0,8	0,8	0,7
	Standardabweichung	*0,4*	*0,8*	*0,5*	*0,6*	*0,7*	*0,6*	*0,7*	*0,6*	*0,7*	*0,7*	*0,6*	*0,6*
40% Dropout	Mittelwert	0,5	0,9	0,5	0,7	0,7	0,6	0,6	0,9	0,8	0,8	0,8	0,7
	Standardabweichung	*0,4*	*0,7*	*0,4*	*0,6*	*0,6*	*0,6*	*0,5*	*0,6*	*0,6*	*0,7*	*0,6*	*0,6*
50% Dropout	Mittelwert	0,5	0,8	0,5	0,7	0,8	0,6	0,6	1,0	0,8	0,7	0,8	0,7
	Standardabweichung	*0,4*	*0,7*	*0,4*	*0,6*	*0,7*	*0,5*	*0,6*	*0,7*	*0,6*	*0,6*	*0,6*	*0,6*
Psychische Summenskala der SF-8													
Original	Mittelwert	1,2	1,7	1,6	1,4	1,6	1,4	1,5	1,7	1,7	1,5	1,7	1,5
	Standardabweichung	*0,8*	*0,9*	*0,9*	*0,9*	*0,9*	*1,0*	*0,9*	*,9*	*0,9*	*1,1*	*0,9*	*0,9*
10% Dropout	Mittelwert	1,2	1,8	1,5	1,4	1,5	1,4	1,5	1,7	1,7	1,5	1,6	1,5
	Standardabweichung	*0,8*	*0,9*	*0,9*	*0,9*	*0,9*	*0,9*	*1,0*	*0,9*	*0,9*	*1,0*	*0,9*	*0,9*
20% Dropout	Mittelwert	1,1	1,7	1,5	1,4	1,5	1,5	1,5	1,8	1,7	1,4	1,7	1,5
	Standardabweichung	*0,8*	*1,0*	*0,9*	*0,9*	*0,9*	*1,0*	*0,9*	*0,9*	*0,9*	*1,0*	*0,9*	*0,9*
30% Dropout	Mittelwert	1,1	1,7	1,4	1,5	1,5	1,3	1,4	1,7	1,7	1,4	1,7	1,5
	Standardabweichung	*0,8*	*0,9*	*0,8*	*0,9*	*0,9*	*0,9*	*0,9*	*0,9*	*0,9*	*1,0*	*0,8*	*0,9*
40% Dropout	Mittelwert	1,1	1,6	1,3	1,5	1,5	1,4	1,2	1,7	1,6	1,5	1,6	1,4
	Standardabweichung	*0,8*	*0,9*	*0,8*	*0,9*	*0,8*	*0,9*	*0,8*	*0,9*	*0,9*	*1,0*	*0,9*	*0,9*
50% Dropout	Mittelwert	1,1	1,5	1,3	1,5	1,7	1,4	1,3	1,9	1,6	1,3	1,6	1,4
	Standardabweichung	*0,8*	*0,9*	*0,8*	*0,9*	*0,9*	*0,9*	*0,8*	*0,8*	*0,8*	*1,0*	*0,8*	*0,9*

Tabelle 30: Verteilungseigenschaften der Basisdaten nach Quoten zufällig fehlender Werte (Angaben zu T0 - Aufnahme)

		Kliniken											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Gesamt**
Globaler Symptomschwereindex der SCL-14													
Original	Mittelwert	1,3	1,4	1,3	1,2	1,1	1,2	1,4	1,2	1,3	1,3	1,4	1,3
	Standardabweichung	*0,6*	*0,8*	*0,6*	*0,7*	*0,7*	*0,7*	*0,7*	*0,7*	*0,7*	*0,8*	*0,7*	*0,7*
10% Dropout	Mittelwert	1,3	1,4	1,3	1,2	1,1	1,2	1,4	1,2	1,3	1,3	1,5	1,3
	Standardabweichung	*0,6*	*0,8*	*0,6*	*0,7*	*0,7*	*0,7*	*0,7*	*0,7*	*0,7*	*0,9*	*0,7*	*0,7*
20% Dropout	Mittelwert	1,3	1,4	1,3	1,2	1,1	1,2	1,4	1,1	1,4	1,2	1,4	1,3
	Standardabweichung	*0,6*	*0,8*	*0,6*	*0,7*	*0,7*	*0,7*	*0,7*	*0,7*	*0,8*	*0,9*	*0,7*	*0,7*
30% Dropout	Mittelwert	1,3	1,5	1,2	1,2	1,0	1,2	1,4	1,3	1,3	1,3	1,4	1,3
	Standardabweichung	*0,6*	*0,8*	*0,5*	*0,8*	*0,7*	*0,7*	*0,7*	*0,7*	*0,8*	*0,9*	*0,7*	*0,7*
40% Dropout	Mittelwert	1,3	1,5	1,3	1,3	1,0	1,1	1,5	1,3	1,2	1,3	1,4	1,3
	Standardabweichung	*0,6*	*0,8*	*0,6*	*0,8*	*0,7*	*0,7*	*0,6*	*0,7*	*0,6*	*0,9*	*0,8*	*0,7*
50% Dropout	Mittelwert	1,4	1,4	1,2	1,3	1,1	1,2	1,4	1,2	1,3	1,2	1,6	1,3
	Standardabweichung	*0,7*	*0,8*	*0,6*	*0,8*	*0,8*	*0,8*	*0,6*	*0,7*	*0,7*	*0,8*	*0,7*	*0,7*
Psychische Summenskala der SF-8													
Original	Mittelwert	2,2	2,4	2,4	2,1	2,1	2,2	2,5	2,1	2,4	2,3	2,5	2,3
	Standardabweichung	*0,9*	*0,8*	*0,8*	*0,9*	*0,9*	*0,9*	*0,9*	*0,9*	*0,8*	*1,0*	*0,9*	*0,9*
10% Dropout	Mittelwert	2,2	2,4	2,4	2,0	2,1	2,2	2,5	2,0	2,5	2,3	2,6	2,3
	Standardabweichung	*0,9*	*0,8*	*0,8*	*0,9*	*0,9*	*0,9*	*0,9*	*0,9*	*0,8*	*1,0*	*0,8*	*0,9*
20% Dropout	Mittelwert	2,2	2,4	2,3	2,1	2,1	2,1	2,5	2,0	2,5	2,3	2,5	2,3
	Standardabweichung	*0,8*	*0,8*	*0,8*	*0,9*	*0,9*	*0,9*	*0,9*	*0,9*	*0,8*	*1,0*	*0,8*	*0,9*
30% Dropout	Mittelwert	2,2	2,5	2,3	2,0	2,0	2,2	2,5	2,1	2,4	2,4	2,5	2,3
	Standardabweichung	*0,9*	*0,8*	*0,8*	*0,9*	*0,9*	*0,9*	*0,9*	*0,9*	*0,9*	*0,9*	*0,8*	*0,9*
40% Dropout	Mittelwert	2,1	2,5	2,4	2,1	2,1	2,1	2,5	2,2	2,4	2,4	2,5	2,3
	Standardabweichung	*0,8*	*0,9*	*0,8*	*1,0*	*0,8*	*0,9*	*0,7*	*0,8*	*0,8*	*0,8*	*0,8*	*0,9*
50% Dropout	Mittelwert	2,3	2,4	2,4	2,1	2,2	2,1	2,5	2,2	2,6	2,3	2,6	2,3
	Standardabweichung	*0,8*	*0,9*	*0,8*	*0,9*	*0,9*	*0,9*	*0,9*	*0,8*	*0,8*	*1,0*	*0,9*	*0,9*

Tabelle 31: Verteilungseigenschaften der Basisdaten nach Quoten systematisch fehlender Werte (Angaben zu T0 – Aufn.)

		Kliniken											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Gesamt**
Globaler Symptomschwereindex der SCL-14													
Original	Mittelwert	1,3	1,4	1,3	1,2	1,1	1,2	1,4	1,2	1,3	1,3	1,4	1,3
	Standardabweichung	*0,6*	*0,8*	*0,6*	*0,7*	*0,7*	*0,7*	*0,7*	*0,7*	*0,7*	*0,8*	*0,7*	*0,7*
10% Dropout	Mittelwert	1,4	1,5	1,3	1,3	1,1	1,2	1,4	1,2	1,4	1,3	1,4	1,3
	Standardabweichung	*0,6*	*0,8*	*0,6*	*0,7*	*0,7*	*0,7*	*0,7*	*0,7*	*0,7*	*0,8*	*0,7*	*0,7*
20% Dropout	Mittelwert	1,4	1,5	1,4	1,3	1,1	1,3	1,5	1,3	1,4	1,3	1,5	1,4
	Standardabweichung	*0,6*	*0,8*	*0,6*	*0,8*	*0,7*	*0,7*	*0,7*	*0,7*	*0,7*	*0,8*	*0,8*	*0,7*
30% Dropout	Mittelwert	1,4	1,6	1,4	1,4	1,1	1,3	1,5	1,3	1,5	1,3	1,5	1,4
	Standardabweichung	*0,6*	*0,8*	*0,6*	*0,7*	*0,7*	*0,7*	*0,7*	*0,7*	*0,7*	*0,8*	*0,7*	*0,7*
40% Dropout	Mittelwert	1,5	1,6	1,5	1,5	1,2	1,4	1,6	1,4	1,5	1,4	1,6	1,5
	Standardabweichung	*0,5*	*0,8*	*0,5*	*0,7*	*0,7*	*0,7*	*0,7*	*0,7*	*0,7*	*0,7*	*0,6*	*0,7*
50% Dropout	Mittelwert	1,5	1,7	1,5	1,6	1,3	1,5	1,7	1,5	1,6	1,4	1,7	1,5
	Standardabweichung	*0,5*	*0,8*	*0,5*	*0,7*	*0,7*	*0,6*	*0,6*	*0,6*	*0,7*	*0,7*	*0,6*	*0,7*
Psychische Summenskala der SF-8													
Original	Mittelwert	2,2	2,4	2,4	2,1	2,1	2,2	2,5	2,1	2,4	2,3	2,5	2,3
	Standardabweichung	*0,9*	*0,8*	*0,8*	*0,9*	*0,9*	*0,9*	*0,9*	*0,9*	*0,8*	*1,0*	*0,9*	*0,9*
10% Dropout	Mittelwert	2,2	2,5	2,4	2,1	2,1	2,2	2,5	2,1	2,5	2,3	2,5	2,3
	Standardabweichung	*0,9*	*0,8*	*0,8*	*0,9*	*0,9*	*0,9*	*0,9*	*0,9*	*0,8*	*1,0*	*0,9*	*0,9*
20% Dropout	Mittelwert	2,3	2,5	2,4	2,1	2,1	2,3	2,6	2,2	2,5	2,3	2,5	2,3
	Standardabweichung	*0,8*	*0,8*	*0,8*	*0,9*	*0,9*	*0,8*	*0,9*	*0,9*	*0,8*	*0,9*	*0,9*	*0,9*
30% Dropout	Mittelwert	2,3	2,4	2,4	2,3	2,1	2,2	2,6	2,1	2,6	2,4	2,6	2,4
	Standardabweichung	*0,8*	*0,8*	*0,8*	*0,9*	*0,9*	*0,8*	*0,9*	*0,9*	*0,8*	*0,9*	*0,8*	*0,8*
40% Dropout	Mittelwert	2,4	2,5	2,5	2,2	2,2	2,4	2,6	2,2	2,6	2,5	2,5	2,4
	Standardabweichung	*0,7*	*0,9*	*0,8*	*0,9*	*0,7*	*0,8*	*0,9*	*1,0*	*0,8*	*0,9*	*0,9*	*0,8*
50% Dropout	Mittelwert	2,4	2,5	2,5	2,4	2,3	2,5	2,8	2,3	2,7	2,5	2,8	2,5
	Standardabweichung	*0,7*	*0,9*	*0,8*	*0,8*	*0,8*	*0,7*	*0,7*	*0,9*	*0,8*	*0,8*	*0,6*	*0,8*

12.3. Intraklassenkorrelationskoeffizienten für die einzelnen Ersetzungsvarianten

Tabelle 32: Übereinstimmung zwischen geschätzten und Originaldaten bei Fehlwertersetzung mittels Regressionsschätzung auf Ebene der Dropoutfälle [normaler Druck] und auf Ebene der Einrichtungsstichproben [*kursiv*] (korrigierte ICC, aggregiert über Einrichtungsergebnisse)

						KRITERIUM (ERGEBNISMAßE)									
						GSI d. SCL-14					PSK d. SF-8				
						KOVARIATEN					KOVARIATEN				
						P	PT	PTT	TT	T	P	PT	PTT	TT	T
AUSFALLMECHANISMUS	zufällig (MCAR)	ERSETZUNGSEBENE	Skala	DROPOUTQUOTE	10%	0,345	0,347	0,378	0,387	0,353	0,440	0,383	0,380	0,516	0,322
						0,953	*0,951*	*0,958*	*0,939*	*0,935*	*0,947*	*0,943*	*0,945*	*0,943*	*0,903*
					20%	0,435	0,389	0,450	0,402	0,333	0,390	0,336	0,405	0,394	0,315
						0,866	*0,858*	*0,875*	*0,838*	*0,803*	*0,855*	*0,848*	*0,866*	*0,842*	*0,809*
					30%	0,352	0,377	0,425	0,424	0,305	0,330	0,355	0,385	0,410	0,309
						0,778	*0,775*	*0,801*	*0,789*	*0,703*	*0,757*	*0,763*	*0,787*	*0,786*	*0,718*
					40%	0,416	0,423	0,439	0,379	0,319	0,377	0,419	0,417	0,382	0,354
						0,729	*0,730*	*0,745*	*0,707*	*0,648*	*0,686*	*0,719*	*0,716*	*0,690*	*0,667*
					50%	0,482	0,474	0,511	0,424	0,385	0,391	0,415	0,422	0,403	0,322
						0,730	*0,725*	*0,745*	*0,683*	*0,643*	*0,665*	*0,681*	*0,686*	*0,665*	*0,592*
			Items	DROPOUTQUOTE	10%	0,431	0,424	0,390	0,429	0,403	0,427	0,427	0,454	0,518	0,389
						0,967	*0,967*	*0,967*	*0,961*	*0,959*	*0,952*	*0,953*	*0,957*	*0,953*	*0,935*
					20%	0,420	0,434	0,460	0,513	0,374	0,416	0,377	0,382	0,436	0,381
						0,883	*0,887*	*0,895*	*0,897*	*0,864*	*0,883*	*0,875*	*0,879*	*0,873*	*0,856*
					30%	0,401	0,427	0,465	0,475	0,367	0,353	0,323	0,368	0,422	0,327
						0,821	*0,832*	*0,846*	*0,840*	*0,801*	*0,791*	*0,781*	*0,802*	*0,806*	*0,770*
					40%	0,447	0,450	0,508	0,414	0,396	0,382	0,425	0,430	0,406	0,369
						0,782	*0,784*	*0,810*	*0,759*	*0,753*	*0,727*	*0,750*	*0,755*	*0,728*	*0,714*
					50%	0,498	0,506	0,552	0,422	0,388	0,404	0,428	0,467	0,422	0,355
						0,752	*0,758*	*0,781*	*0,695*	*0,681*	*0,689*	*0,706*	*0,733*	*0,694*	*0,643*
	systematisch (MAR)	ERSETZUNGSEBENE	Skala	DROPOUTQUOTE	10%	0,442	0,432	0,357	0,374	0,320	0,366	0,363	0,389	0,386	0,323
						0,929	*0,927*	*0,918*	*0,900*	*0,886*	*0,936*	*0,940*	*0,941*	*0,926*	*0,902*
					20%	0,407	0,409	0,398	0,383	0,359	0,381	0,402	0,358	0,369	0,366
						0,864	*0,868*	*0,858*	*0,839*	*0,801*	*0,856*	*0,862*	*0,852*	*0,830*	*0,818*
					30%	0,410	0,408	0,411	0,380	0,367	0,430	0,362	0,382	0,389	0,314
						0,761	*0,762*	*0,764*	*0,735*	*0,714*	*0,771*	*0,745*	*0,752*	*0,744*	*0,693*
					40%	0,369	0,380	0,393	0,357	0,329	0,399	0,348	0,370	0,369	0,343
						0,632	*0,642*	*0,650*	*0,618*	*0,599*	*0,689*	*0,652*	*0,673*	*0,664*	*0,640*
					50%	0,431	0,446	0,465	0,330	0,314	0,367	0,403	0,419	0,378	0,329
						0,644	*0,657*	*0,670*	*0,562*	*0,536*	*0,623*	*0,651*	*0,660*	*0,617*	*0,568*
			Items	DROPOUTQUOTE	10%	0,472	0,434	0,426	0,432	0,338	0,406	0,364	0,322	0,347	0,303
						0,938	*0,933*	*0,933*	*0,925*	*0,913*	*0,944*	*0,940*	*0,939*	*0,928*	*0,920*
					20%	0,439	0,399	0,428	0,447	0,359	0,387	0,367	0,385	0,419	0,351
						0,890	*0,884*	*0,892*	*0,884*	*0,864*	*0,871*	*0,870*	*0,877*	*0,867*	*0,845*
					30%	0,415	0,423	0,429	0,431	0,362	0,361	0,398	0,449	0,407	0,368
						0,785	*0,789*	*0,791*	*0,791*	*0,757*	*0,764*	*0,784*	*0,807*	*0,783*	*0,756*
					40%	0,391	0,402	0,433	0,399	0,350	0,361	0,369	0,427	0,399	0,325
						0,669	*0,678*	*0,698*	*0,677*	*0,646*	*0,689*	*0,696*	*0,729*	*0,712*	*0,674*
					50%	0,455	0,466	0,471	0,363	0,348	0,390	0,388	0,420	0,364	0,337
						0,672	*0,682*	*0,684*	*0,600*	*0,593*	*0,651*	*0,653*	*0,676*	*0,622*	*0,601*

KRITERIUMSVARIABLEN: **PSK** = Psychische Summenskala der SF-8, **GSI** = Globaler Symptomschwere-Index der SCL-14; *KOVARIATEN*: **P** = Patientenangaben zu T0 (Aufnahme), **T** = Therapeutenangaben zu T0, **TT** = Therapeutenangaben zu T0 und T1 (Aufnahme und Entlassung); *ERSETZUNGSEBENE*: **Skala** = Schätzung des Skalenwertes, **Items** = Schätzung der einzelnen Itemwerte; *DROPOUTQUOTEN*: **10/20/30/40/50%** = Anteile variablenbezogen fehlender Werte

Tabelle 33: Übereinstimmung zwischen geschätzten und Originaldaten bei Fehlwertersetzung mittels EM-Imputation auf Ebene der Dropoutfälle [normaler Druck] und auf Ebene der Einrichtungsstichproben [*kursiv*] (korrigierte ICC, aggregiert über Einrichtungsergebnisse)

					KRITERIUM (ERGEBNISMAßE)									
					GSI d. SCL-14					**PSK d. SF-8**				
					KOVARIATEN					**KOVARIATEN**				
					P	**PT**	**PTT**	**TT**	**T**	**P**	**PT**	**PTT**	**TT**	**T**
AUSFALLMECHANISMUS	**zufällig (MCAR)**	**ERSETZUNGSEBENE**	**Skala**	**DROPOUTQUOTE** 10%	0,294	0,281	0,278	0,478	0,353	0,315	0,300	0,299	0,503	0,358
					0,951	*0,954*	*0,958*	*0,941*	*0,940*	*0,938*	*0,941*	*0,944*	*0,936*	*0,933*
				20%	0,315	0,294	0,297	0,431	0,391	0,300	0,299	0,294	0,407	0,369
					0,840	*0,853*	*0,858*	*0,835*	*0,857*	*0,840*	*0,861*	*0,864*	*0,837*	*0,859*
				30%	0,297	0,288	0,290	0,428	0,382	0,303	0,286	0,288	0,417	0,354
					0,785	*0,796*	*0,803*	*0,771*	*0,793*	*0,787*	*0,796*	*0,804*	*0,762*	*0,778*
				40%	0,291	0,296	0,289	0,402	0,387	0,292	0,295	0,289	0,400	0,369
					0,722	*0,736*	*0,737*	*0,703*	*0,726*	*0,712*	*0,725*	*0,727*	*0,689*	*0,712*
				50%	0,290	0,287	0,285	0,406	0,383	0,291	0,287	0,286	0,414	0,366
					0,656	*0,661*	*0,663*	*0,629*	*0,638*	*0,655*	*0,660*	*0,662*	*0,634*	*0,627*
			Items	**DROPOUTQUOTE** 10%	0,294	0,281	0,278	0,478	0,353	0,315	0,300	0,299	0,503	0,358
					0,951	*0,954*	*0,958*	*0,941*	*0,940*	*0,938*	*0,941*	*0,944*	*0,936*	*0,933*
				20%	0,315	0,294	0,297	0,431	0,391	0,300	0,299	0,294	0,407	0,369
					0,840	*0,853*	*0,858*	*0,835*	*0,857*	*0,840*	*0,861*	*0,864*	*0,837*	*0,859*
				30%	0,297	0,288	0,290	0,428	0,382	0,303	0,286	0,288	0,417	0,354
					0,785	*0,796*	*0,803*	*0,771*	*0,793*	*0,787*	*0,796*	*0,804*	*0,762*	*0,778*
				40%	0,291	0,296	0,289	0,402	0,387	0,293	0,295	0,289	0,400	0,370
					0,722	*0,736*	*0,737*	*0,703*	*0,726*	*0,712*	*0,725*	*0,727*	*0,689*	*0,711*
				50%	0,290	0,287	0,285	0,406	0,383	0,291	0,287	0,286	0,414	0,366
					0,656	*0,661*	*0,663*	*0,629*	*0,638*	*0,655*	*0,660*	*0,662*	*0,634*	*0,627*
	systematisch (MAR)	**ERSETZUNGSEBENE**	**Skala**	**DROPOUTQUOTE** 10%	0,306	0,299	0,289	0,408	0,373	0,288	0,265	0,278	0,479	0,340
					0,899	*0,912*	*0,912*	*0,912*	*0,913*	*0,926*	*0,934*	*0,937*	*0,935*	*0,934*
				20%	0,297	0,293	0,283	0,396	0,351	0,294	0,297	0,284	0,427	0,376
					0,841	*0,853*	*0,858*	*0,824*	*0,852*	*0,834*	*0,855*	*0,856*	*0,830*	*0,851*
				30%	0,294	0,279	0,273	0,401	0,343	0,286	0,282	0,273	0,456	0,361
					0,727	*0,738*	*0,739*	*0,716*	*0,730*	*0,736*	*0,754*	*0,753*	*0,762*	*0,760*
				40%	0,274	0,268	0,273	0,386	0,355	0,286	0,276	0,277	0,408	0,326
					0,609	*0,612*	*0,618*	*0,627*	*0,632*	*0,672*	*0,681*	*0,685*	*0,672*	*0,657*
				50%	0,277	0,270	0,269	0,334	0,318	0,290	0,276	0,274	0,376	0,325
					0,564	*0,563*	*0,563*	*0,532*	*0,546*	*0,604*	*0,604*	*0,604*	*0,586*	*0,576*
			Items	**DROPOUTQUOTE** 10%	0,306	0,299	0,289	0,408	0,373	0,288	0,265	0,278	0,479	0,340
					0,899	*0,912*	*0,912*	*0,912*	*0,913*	*0,926*	*0,934*	*0,937*	*0,934*	*0,934*
				20%	0,297	0,293	0,283	0,396	0,351	0,294	0,297	0,284	0,427	0,376
					0,841	*0,853*	*0,858*	*0,824*	*0,852*	*0,834*	*0,855*	*0,856*	*0,830*	*0,851*
				30%	0,294	0,279	0,273	0,401	0,344	0,286	0,282	0,273	0,456	0,361
					0,727	*0,738*	*0,739*	*0,716*	*0,730*	*0,736*	*0,754*	*0,753*	*0,762*	*0,760*
				40%	0,275	0,268	0,273	0,386	0,355	0,287	0,276	0,277	0,408	0,326
					0,609	*0,612*	*0,618*	*0,627*	*0,632*	*0,672*	*0,681*	*0,685*	*0,672*	*0,657*
				50%	0,277	0,270	0,269	0,334	0,318	0,290	0,276	0,274	0,376	0,325
					0,564	*0,563*	*0,563*	*0,532*	*0,546*	*0,604*	*0,604*	*0,604*	*0,586*	*0,575*

KRITERIUMSVARIABLEN: **PSK** = Psychische Summenskala der SF-8, **GSI** = Globaler Symptomschwere-Index der SCL-14; *KOVARIATEN***: P** = Patientenangaben zu T0 (Aufnahme), **T** = Therapeutenangaben zu T0, **TT** = Therapeutenangaben zu T0 und T1 (Aufnahme und Entlassung); *ERSETZUNGSEBENE***: Skala** = Schätzung des Skalenwertes, **Items** = Schätzung der einzelnen Itemwerte; *DROPOUTQUOTEN***: 10/20/30/40/50%** = Anteile variablenbezogen fehlender Werte

Tabelle 34: Übereinstimmung zwischen geschätzten und Originaldaten bei Fehlwertersetzung mittels Multipler Imputation auf Ebene der Dropoutfälle [normaler Druck] und auf Ebene der Einrichtungsstichproben [*kursiv*] (korrigierte ICC, aggregiert über Einrichtungsergebnisse)

						KRITERIUM (ERGEBNISMAßE)									
						GSI d. SCL-14					PSK d. SF-8				
						KOVARIATEN					KOVARIATEN				
						P	PT	PTT	TT	T	P	PT	PTT	TT	T
AUSFALLMECHANISMUS	zufällig (MCAR)	ERSETZUNGSEBENE	Skala	DROPOUTQUOTE	10%	0,342	0,429	0,448	0,349	0,372	0,396	0,471	0,464	0,382	0,391
						0,942	*0,955*	*0,955*	*0,934*	*0,940*	*0,944*	*0,954*	*0,951*	*0,937*	*0,937*
					20%	0,551	0,541	0,564	0,351	0,367	0,450	0,469	0,475	0,390	0,380
						0,900	*0,895*	*0,905*	*0,827*	*0,834*	*0,880*	*0,889*	*0,893*	*0,873*	*0,867*
					30%	0,515	0,516	0,529	0,328	0,308	0,474	0,485	0,492	0,362	0,337
						0,844	*0,851*	*0,855*	*0,743*	*0,744*	*0,842*	*0,847*	*0,848*	*0,792*	*0,776*
					40%	0,515	0,502	0,518	0,371	0,345	0,443	0,439	0,439	0,367	0,336
						0,805	*0,800*	*0,808*	*0,723*	*0,696*	*0,767*	*0,765*	*0,766*	*0,728*	*0,717*
					50%	0,491	0,499	0,492	0,400	0,330	0,413	0,426	0,439	0,389	0,327
						0,745	*0,750*	*0,749*	*0,672*	*0,626*	*0,705*	*0,711*	*0,723*	*0,682*	*0,647*
			Items	DROPOUTQUOTE	10%	0,219	0,254	0,279	0,090	0,101	0,357	0,390	0,429	0,332	0,352
						0,918	*0,926*	*0,931*	*0,843*	*0,848*	*0,945*	*0,953*	*0,954*	*0,944*	*0,944*
					20%	0,383	0,392	0,427	0,156	0,143	0,398	0,429	0,448	0,342	0,330
						0,853	*0,858*	*0,869*	*0,722*	*0,709*	*0,883*	*0,890*	*0,895*	*0,875*	*0,869*
					30%	0,378	0,403	0,404	0,142	0,126	0,422	0,438	0,448	0,320	0,294
						0,790	*0,804*	*0,806*	*0,612*	*0,593*	*0,837*	*0,842*	*0,846*	*0,802*	*0,788*
					40%	0,392	0,408	0,414	0,152	0,131	0,387	0,401	0,424	0,320	0,300
						0,746	*0,757*	*0,758*	*0,549*	*0,522*	*0,753*	*0,758*	*0,770*	*0,724*	*0,716*
					50%	0,412	0,415	0,432	0,180	0,143	0,404	0,414	0,417	0,320	0,280
						0,705	*0,707*	*0,720*	*0,513*	*0,472*	*0,713*	*0,719*	*0,721*	*0,669*	*0,645*
	systematisch (MAR)	ERSETZUNGSEBENE	Skala	DROPOUTQUOTE	10%	0,615	0,542	0,581	0,307	0,366	0,485	0,484	0,516	0,365	0,306
						0,955	*0,943*	*0,951*	*0,892*	*0,906*	*0,952*	*0,947*	*0,956*	*0,935*	*0,928*
					20%	0,517	0,516	0,554	0,355	0,326	0,451	0,465	0,503	0,402	0,350
						0,903	*0,900*	*0,910*	*0,844*	*0,830*	*0,886*	*0,889*	*0,899*	*0,873*	*0,853*
					30%	0,583	0,532	0,560	0,445	0,356	0,460	0,444	0,449	0,382	0,331
						0,853	*0,833*	*0,847*	*0,792*	*0,743*	*0,807*	*0,798*	*0,801*	*0,783*	*0,749*
					40%	0,518	0,521	0,547	0,388	0,323	0,464	0,465	0,485	0,395	0,333
						0,751	*0,753*	*0,768*	*0,671*	*0,623*	*0,762*	*0,761*	*0,771*	*0,726*	*0,689*
					50%	0,530	0,513	0,524	0,353	0,321	0,474	0,476	0,474	0,379	0,309
						0,725	*0,713*	*0,722*	*0,598*	*0,581*	*0,711*	*0,710*	*0,708*	*0,648*	*0,602*
			Items	DROPOUTQUOTE	10%	0,438	0,454	0,459	0,172	0,215	0,387	0,425	0,475	0,345	0,320
						0,929	*0,930*	*0,933*	*0,840*	*0,856*	*0,948*	*0,950*	*0,956*	*0,943*	*0,940*
					20%	0,410	0,432	0,462	0,158	0,146	0,426	0,424	0,464	0,321	0,317
						0,880	*0,887*	*0,895*	*0,747*	*0,735*	*0,892*	*0,891*	*0,899*	*0,867*	*0,863*
					30%	0,505	0,508	0,516	0,188	0,167	0,452	0,446	0,476	0,332	0,320
						0,835	*0,838*	*0,842*	*0,653*	*0,637*	*0,821*	*0,817*	*0,829*	*0,785*	*0,776*
					40%	0,487	0,486	0,513	0,239	0,209	0,461	0,467	0,481	0,323	0,308
						0,751	*0,751*	*0,765*	*0,596*	*0,572*	*0,774*	*0,776*	*0,784*	*0,713*	*0,708*
					50%	0,518	0,506	0,523	0,235	0,199	0,477	0,456	0,469	0,319	0,296
						0,731	*0,725*	*0,735*	*0,535*	*0,503*	*0,724*	*0,710*	*0,717*	*0,638*	*0,622*

KRITERIUMSVARIABLEN: **PSK** = Psychische Summenskala der SF-8, **GSI** = Globaler Symptomschwere-Index der SCL-14; *KOVARIATEN*: **P** = Patientenangaben zu T0 (Aufnahme), **T** = Therapeutenangaben zu T0, **TT** = Therapeutenangaben zu T0 und T1 (Aufnahme und Entlassung); *ERSETZUNGSEBENE*: **Skala** = Schätzung des Skalenwertes, **Items** = Schätzung der einzelnen Itemwerte; *DROPOUTQUOTEN*: **10/20/30/40/50%** = Anteile variablenbezogen fehlender Werte

Lebenslauf Sven Rabung

6.7.1973 geboren in München

Ausbildung

1979-1983	Grundschule an der Nadistraße, München
1983-1992	Theresien-Gymnasium, München
1993-1994	Studium im Magisterstudiengang Pädagogik (Nebenfach Psychologie) an der Ludwig-Maximilians-Universität München
1994-1996	Grundstudium im Lizentiatsstudiengang Psychologie an der Universität Fribourg, Schweiz
1996-1999	Hauptstudium im Diplomstudiengang Psychologie am Georg-Elias-Müller-Institut für Psychologie der Universität Göttingen
2000-2004	Therapieausbildung im Weiterbildenden Studiengang Psychologische Psychotherapie der Hochschulen Braunschweig und Göttingen (WSPP)

Berufstätigkeit

1998-1999	Studentische Hilfskraft in der Abteilung Sozial- und Kommunikationspsychologie des Georg-Elias-Müller-Instituts für Psychologie, Göttingen (Prof. Dr. M. Boos)
1998-1999	Studentische Hilfskraft in der Abteilung Psychosomatik und Psychotherapie der Universität Göttingen (Prof. Dr. U. Rüger)
1999-2000	Wissenschaftliche Hilfskraft in der Abteilung Psychosomatik und Psychotherapie der Universität Göttingen
2000-2003	Wissenschaftlicher Mitarbeiter in der Abteilung Psychosomatik und Psychotherapie der Universität Göttingen in einem Forschungsprojekt zur "Bedeutung emotionaler und neuroendokriner Faktoren für den Krankheitsverlauf bei Patienten mit atopischer Dermatitis" (Leitung: Prof. Dr. H. Schauenburg, Prof. Dr. G. Hüther) und später teilweise in einem von der DFG geförderten Forschungsprojekt zur Behandlung der Generalisierten Angststörung (Leitung: Prof. Dr. F. Leichsenring, Prof Dr. E. Leibing)
2000-2002	Freie Mitarbeit in einer Multicenter-Studie zu "Krankheitserleben und Verlauf bei Patienten mit einem implantierten Defibrillator" (German-Austrian ICD Multicenter Study GAIMS; Koordination: Prof. Dr. G. Bergmann, Heidelberg)
2001-2004	Qualitätsmanagementkoordinator und wissenschaftlicher Mitarbeiter im Funktionsbereich Dokumentation und Qualitätssicherung des Niedersächsischen Landeskrankenhauses Tiefenbrunn (Ärztlicher Direktor: Prof. Dr. U. Streeck, Bereichsleiter: Prof. Dr. F. Leichsenring)
seit 2005	Wissenschaftlicher Mitarbeiter im Institut für Medizinische Psychologie des Universitätsklinikums Hamburg-Eppendorf (Prof. Dr. Dr. U. Koch), Arbeitsgruppe für Psychotherapie- und Versorgungsforschung (AGPV) und Forschungsgruppe Qualitätsmanagement (FGQM)