

Hinweis: Die folgenden Kapitel betrachten im Wesentlichen Ergebnisse, die sich nur auf die Experimentalgruppe beziehen. In Analogie zur Auswertung der Projektfragebögen für die Teammitglieder der besuchten Tagesklinik (siehe 5.8.) ist auch in der Auswertung der Projektfragebögen für Besucher zu berücksichtigen, dass teilweise mehrere Peers gleichzeitig ein Team besuchten. Über die einzelnen Einschätzungen der Peers eines Falls wurde zunächst das arithmetische Mittel berechnet, erst dann wurden weitere Berechnungen (z.B. Projekthälftenvergleiche) angestellt. Auf notwendige Abweichungen von diesem Vorgehen bei einzelnen Fragestellungen wird im Text gesondert eingegangen.

5.5. Dokumentation der Therapieziele, Dokumentation und Angemessenheit der geplanten Therapiedauer

Es war angenommen worden, dass sich im Verlauf des Projektes die Dokumentation von Therapiezielen und geplanter Therapiedauer verbessert und dass bezüglich der Aufenthaltsdauer eine größere Übereinstimmung zwischen Therapeuten und Peers erreicht wird (siehe Abschnitt 4.2.). Im Folgenden werden Methodik und Ergebnisse aufgezeigt und kritisch hinterfragt.

5.5.1. Dokumentation der Therapieziele

Zur Beantwortung der Frage, ob sich die Dokumentation der Therapieziele im Verlauf des Projektes verbessert, wurde den Besuchern mit dem Projektfragebogen eine Skala vorgelegt, auf der sie in einem Intervall zwischen den Werten 1 und 10 zur Frage "Finde ich Therapieziele in den Unterlagen?" eine Markierung setzen sollten. Der Wert 1 steht für "nein, überhaupt nicht", 10 für "ja, komplett und aktuell".

Die Auswertung aller Fragebögen ergibt für den gesamten Zeitraum einen Mittelwert von 6,6 (siehe Tab. 8).

Antworten auf die Frage: „Finde ich Therapieziele in den Unterlagen?“			
	Mittelwert	N	s
1. Projekthälfte	6,43	17	2,18
2. Projekthälfte	6,68	17	3,17
Gesamt	6,55	34*	2,68

* entspricht 97,1% der Fälle der Gesamtstichprobe, zu einem Fall der 1. Projekthälfte lagen keine verwertbaren Daten vor.

Tab. 8: Therapiezieldokumentation - Mittelwerte

N: Anzahl der gewerteten Fälle, s: Standardabweichung

Zu 14 der in diese Untersuchung aufgenommen Fälle lagen 2 bzw. 3 Projektfragebögen vor. In welchem Maß verschiedene Beurteiler zu derselben Einschätzung kommen, wurde in einer Korrelationsanalyse untersucht: Die Korrelation nach Spearman ist zwischen einem 1. und einem 2. Beurteiler mit $r = .79$ hoch und auf dem Niveau von 0,01 (2-seitig) signifikant (siehe Streudiagramm Abb. 15).⁵

Die ermittelte Korrelation spricht für eine gute Interrater-Reliabilität der verwendeten Skala.

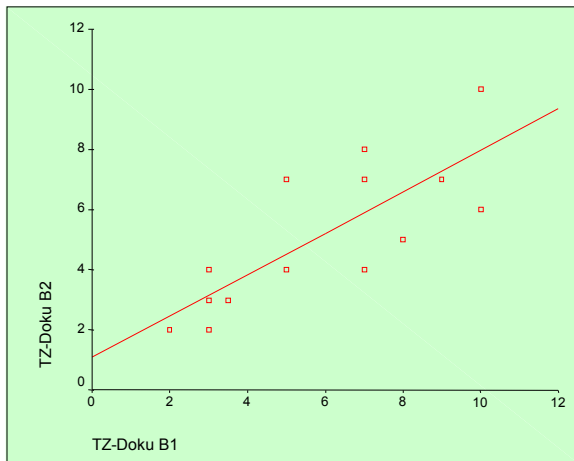


Abb. 15: Streudiagramm – Therapiezieldokumentation

Im Vergleich der 1. und 2. Projekthälfte unterscheiden sich die Mittelwerte für die Skala zur Therapieziel-Dokumentation nur minimal zugunsten der 2. Projekthälfte. Dieser Unterschied erwies sich statistisch als nicht signifikant (T-Test für unabhängige Stichproben nicht gezeigt). Eine wesentliche Verbesserung der Therapieziel-Dokumentation im Verlauf dieses Projektes konnte so nicht nachgewiesen werden.

5.5.2. Dokumentation der geplanten Therapiedauer

Die Daten zur Dokumentation der Therapiedauer wurden durch Auswertung der Projektfrage an die Besucher "Finde ich in den Unterlagen eine Angabe über die geplante Therapiedauer?" ermittelt. Als Antwortmöglichkeiten waren lediglich "ja" oder "nein" vorgegeben.

⁵ Bei insgesamt 2 Fällen mit einem 3. Besucher waren deren Einschätzungen mit dem 1. Besucher identisch und unterschieden sich zum 2. Beurteiler jeweils um den Wert 1.

Besuchten zwei oder drei Kollegen einen Patienten wurden die Angaben auf den einzelnen Fragebögen hinsichtlich ihrer Übereinstimmung überprüft. Dabei fielen lediglich in einem Fall differente Angaben auf, dieser wurde nicht gewertet.

Die Daten wurden so komprimiert, dass für jeden Patienten der Experimentalgruppe nur *eine* Antwort zugelassen wurde. In drei Fälle lagen keine Daten vor.

Ein Vergleich der relativen Häufigkeiten unter Ausschluss des nicht gewerteten Falles und der fehlenden Daten zeigt bei einem insgesamt niedrigen "Dokumentationsniveau" in der 2. Projekthälfte immerhin einen Anstieg der Fälle, bei denen in den Unterlagen eine Angabe zur geplanten Therapiedauer gefunden wurde, um 6 auf 35% (siehe Tab. 9). Spekulativ bleibt die Frage, warum mehrfach keine Angaben von den Besuchern gemacht worden sind und inwieweit dies das Ergebnis beeinflusst haben könnte, da alle fehlenden Daten in den Zeitraum der 1. Projekthälfte fallen.

Antworten auf die Frage: „Finde ich eine Angabe über die geplante Therapiedauer?“						
			Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
1. Projekthälfte	Gültig	Ja	4	22,2	28,6	28,6
		Nein	10	55,6	71,4	100,0
		Gesamt	14	77,8	100,0	
	Fehlend	Nicht gewertet	1	5,6		
		Fehlende Daten	3	16,7		
		Gesamt	4	22,2		
Gesamt		18	100,0			
2. Projekthälfte	Gültig	Ja	6	35,3	35,3	35,3
		Nein	11	64,7	64,7	100,0
		Gesamt	17	100,0	100,0	
Gesamt		17	100,0			

Tab. 9: Dokumentation der Therapiedauer – Häufigkeiten und Prozente

5.5.3. Angemessenheit der geplanten Therapiedauer

Neben der Frage, *ob* sich Angaben zur geplanten Therapiedauer in den Unterlagen finden, interessierte auch die Einschätzung des Besuchers, in welchem Umfang er diese für angemessen hielt. Im Projekt-Fragebogen wurden die Beurteilenden bei der Frage "Finde ich die geplante Therapiedauer angemessen?" ausdrücklich darauf hingewiesen, auch die vom Therapeuten des Patienten *erfragte* geplante Therapiedauer zu berücksichtigen, so dass bei jedem Besuch diese

Frage beantwortbar war. Die Einschätzung erfolgte wieder auf einer metrischen Skala im Intervall von 1 bis 10. Der Wert 1 steht für "nein, überhaupt nicht", 10 für "ja, vollständig". Die Auswertung der Mittelwerte zeigt Tab. 10.

Die Ergebnisse zeigen bei einem Gesamt-Mittelwert von 7,2 in der Aufteilung der Experimental-Gruppe in Projekthälften – entgegen der Erwartung – eine Abnahme von im Mittel 0,5 Punkten in der 2. Projekthälfte. Diese Tendenz ist statistisch nicht signifikant (T-Test für unabhängige Stichproben nicht gezeigt), könnte jedoch ein Hinweis darauf sein, dass die Angemessenheit der geplanten Therapiedauer – beurteilt durch die Besucher – in der 2. Projekthälfte tendenziell geringer ist.

Antworten auf die Frage: „Finde ich die geplante Therapiedauer angemessen?“			
	Mittelwert	N	s
1. Projekthälfte	7,48	15	1,97
2. Projekthälfte	6,97	15	2,29
Gesamt	7,22	30*	2,12

* entspricht 85,7% der Fälle der Gesamtstichprobe, zu drei Fällen der 1. Projekthälfte und zwei Fällen der 2. Projekthälfte lagen keine verwertbaren Daten vor.

Tab. 10: Angemessenheit der Therapiedauer – Mittelwerte

N: Anzahl der gewerteten Fälle, s: Standardabweichung

Zu 9 in diese Untersuchung aufgenommenen Fällen lagen 2 bzw. in einem Fall 3 Einschätzungen der Angemessenheit der geplanten Therapiedauer vor. Auch hier wurden vergleichende Analysen durchgeführt: Der Korrelationskoeffizient nach Spearman liegt bei $r = .22$ in der Auswertung von 1. und 2. Besucher, die Korrelation ist statistisch nicht signifikant.⁶

Die Interraterreliabilität ist damit – wie zu erwarten – deutlich schlechter als zu den Fragen der Therapieziel- oder Therapiedauer-Dokumentation.

Es fällt weiterhin auf, dass diese Frage häufig nicht beantwortet wurde: Von 51 in die Untersuchung eingegangenen Projektfragebögen zu 35 Fällen finden sich hinsichtlich der Frage nach der Angemessenheit der geplanten Therapiedauer 40 Beurteilungen.

⁶ In zwei Fällen mit einem 3. Besucher stimmen dessen Einschätzungen mit denen des 1. und/oder 2. Besuchers überein.

5.5.4. Diskussion und Fazit

Die Fragen nach der Dokumentation von Therapiezielen und geplanter Therapiedauer scheinen eindeutig (Augenscheinvalidität) und gut beantwortbar, für die eingesetzten Instrumente konnten hohe Interraterreliabilitäten nachgewiesen werden. Die Annahme jedoch, dass sich die Dokumentation der Therapieziele im Verlauf des Projekts verbessert, war so nicht beweisbar. Hinsichtlich der Dokumentation der geplanten Therapiedauer sollten die geringen Verbesserungen in der zweiten Projekthälfte aufgrund der Problematik kleiner Fallzahlen und fehlender Daten zurückhaltend interpretiert werden.

Während das Dokumentationsniveau bei den Therapiezielen befriedigend erscheint, decken die Ergebnisse einen Mangel bezüglich der Dokumentation der geplanten Therapiedauer auf.

Die Angemessenheit der geplanten Therapiedauer wurde von den Besuchern besonders positiv bewertet. Die Konstruktvalidität des eingesetzten Instruments scheint jedoch problematisch: Gemäß der Natur des Abstraktums „Angemessenheit“ ist kaum von einer einheitlichen Beurteilung durch verschiedene Besucher nach deren impliziten Kriterien (vgl. Abschnitt 1.1.2.) auszugehen. Die geringe Interraterreliabilität überrascht daher nicht.

Dennoch sollen an dieser Stelle Möglichkeiten diskutiert werden, warum sich die Angemessenheit der geplanten Therapiedauer – beurteilt durch die Besucher – im Projektverlauf entgegen der Erwartung tendenziell etwas verlor:

Vor dem Hintergrund eines wachsenden Drucks seitens der Kostenträger auch während des Erhebungszeitraums von 1998 bis 2000 könnte es bei den teilnehmenden Teams zunehmend zu Befürchtungen bezüglich der Intention des Projektes gekommen sein (vgl. Abschnitt 5.8.). In Sorge um die Patientenversorgung könnten von den Behandlern – quasi präventiv – zunehmend längere Behandlungsdauern veranschlagt worden sein, während die nicht direkt involvierten Besucher weiter einen „objektiveren“ Standpunkt vertraten. Eine andere Erklärung bestünde darin, dass gerade die Besucher dem Druck der Kassen besonders ausgesetzt waren und in der Folge kürzere Therapiedauern als geplant favorisierten. Tatsächlich bestand die Gruppe der Besucher zu einem weitaus größeren Anteil aus gegenüber den Kassen rechenschaftsschuldigen Ärzten als es im Gesamt der Teams der Fall war.

5.6. Relevanz und Realismus von Therapiezielen, Einbeziehung von Patientenwünschen

5.6.1. Projektfragebogen für Besucher

5.6.1.1. Methode

Nachdem im Vorangegangenen die Einschätzung der Besucher, ob und in welchem Ausmaß und welcher Aktualität sich Therapieziele in den Unterlagen zu den Patienten finden, ausgewertet wurde, sollen in diesem Abschnitt weitere Aspekte zu den Therapiezielen beleuchtet werden.

Mit dem Projektfragebogen wurden den Besuchern 4 weitere Fragen vorgelegt, die durch das Setzen einer Markierung im Intervall einer metrischen Skala von 1 bis 10 zu beantworten waren.

Die Fragen im Einzelnen:

- "Sind die Therapieziele realistisch?"
- "Sind die Ziele von zutreffendem Gewicht?"

Der Wert 1 steht bei diesen beiden Fragen für "nein, überhaupt nicht", der Wert 10 für "ja, vollständig".

- "War bei diesem Patienten eine Übereinstimmung der Therapieziele zwischen Patienten und Therapeuten zu erwarten?"

Der Wert 1 steht hier für "nein, überhaupt nicht" der Wert 10 für "ja, sehr weitgehend".

- "Wurde eine mögliche Übereinstimmung der Therapieziele vom Patienten und Therapeuten erreicht?"

Der Wert 1 steht in diesem Fall für "nein, völlig verfehlt", der Wert 10 für "ja, völlig erreicht."

5.6.1.2. Ergebnisse

Fehlende Daten: In einer durchgeführten Analyse der fehlenden Daten lag die Beantwortung der einzelnen Fragen für alle ausgewerteten Fragebögen zwischen 92 und 100%, wobei die Frage nach dem Erreichen einer möglichen Übereinstimmung der Therapieziele am häufigsten nicht beantwortet wurde.

Es war angenommen worden, dass die durch die Behandler aufgestellten Therapieziele aus Sicht der Besucher im Verlauf des Projektes realistischer und relevanter werden und dass Patientenwünsche stärker in der Therapiezielplanung berücksichtigt werden. Die Auswertung erfolgte im Vergleich der Projekthälften wie unter Abschnitt 4.2. beschrieben.

In Tabelle 11 und Abbildung 16 sind die Ergebnisse dargestellt.

Therapiezielrealismus und –relevanz – Projektfragebogen für Besucher					
		TZ Real.	TZ Gew.	TZ Überein. erwartet	TZ Überein. erreicht
1. Projekthälfte	Mittelwert	7,06	7,22	5,31	5,77
	N	18	18	18	17
	s	1,70	1,73	2,14	1,68
2. Projekthälfte	Mittelwert	7,09	7,29	5,50	6,06
	N	17	17	17	17
	S	2,26	2,22	2,72	1,78
Gesamt	Mittelwert	7,07	7,26	5,40	5,91
	N	35	35	35	34
	s	1,96	1,95	2,40	1,71

Tab. 11: Therapieziele: Realismus und –Relevanz, Übereinstimmung zwischen Patienten und Therapeuten - Mittelwerte

N: Anzahl der gewerteten Fälle, s: Standardabweichung

TZ Real.: Auswertung der Frage „Sind die Therapieziele realistisch?“

TZ Gew.: Auswertung der Frage „Sind die Ziele von zutreffendem Gewicht?“

TZ Überein. erwartet: Auswertung der Frage „War bei diesem Patienten eine Übereinstimmung der Therapieziele mit dem Therapeuten zu erwarten?“

TZ Überein. erreicht: Auswertung der Frage „Wurde eine mögliche Übereinstimmung der Therapieziele vom Patienten und Therapeuten erreicht?“

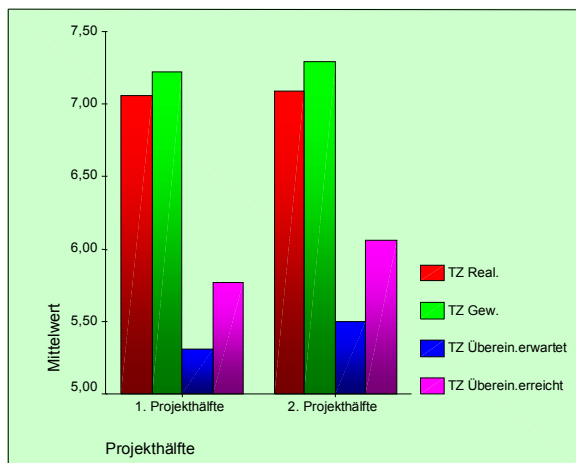


Abb. 16: Therapiezielrealismus und –relevanz – Mittelwerte

Zur Legende siehe Tab. 18

Die Auswertung aller Fragebögen ergibt über den gesamten Zeitraum für die Fragen an die Besucher, ob die aufgestellten Therapieziele realistisch und von zutreffendem Gewicht sind, Mittelwerte von 7,1 respektive 7,3. Diese liegen gerade im oberen Drittel der Skalen und sprechen somit für eine eher positive Einschätzung.

Beim Vergleich zwischen 1. und 2. Projekthälfte zeigt sich praktisch keine Veränderung dieser Werte, was als Hinweis auf eine gute Bedingungsstabilität gewertet werden kann. Die Annahme, dass die Therapieziele realistischer und relevanter werden, konnte so nicht bewiesen werden (siehe unten).

Auf die Frage, ob eine Übereinstimmung zwischen Patienten und Therapeuten zu erwarten war, fielen die Einschätzungen bei einem Mittelwert von 5,4 schon deutlich weniger positiv aus. Dieses Ergebnis überrascht nicht, zumal auch in anderen Arbeiten schon wiederholt auf die Problematik eines – besonders bei Erkrankungen aus dem schizophrenen Formenkreis – verzerrten Selbst- und Krankheitsverständnisses mit konsekutiver Inkongruenz hinsichtlich der Therapieziele zwischen Patient und Therapeut hingewiesen wurde [z.B. 11].

Im Vergleich der Projekthälften blieb auch dieser Wert stabil. Es war hier auch keine Änderung als Effekt des Programms zu erwarten.

Auch die Einschätzungen der Besucher zur Erreichung einer möglichen Übereinstimmung zwischen Patienten und Therapeuten fielen bei einem Mittelwert von 5,9 über das ganze Projekt ebenfalls nicht durchweg positiv aus.

Eine tendenzielle Zunahme dieses Wertes um 0,3 Punkte in der 2. Projekthälfte erwies sich als statistisch nicht signifikant (T-Test für unabhängige Stichproben nicht gezeigt).

Die Erwartung, dass durch das Projekt eine Übereinstimmung der Therapieziele von Patienten und Therapeuten in höherem Maße ermöglicht wird, konnte so nicht bestätigt werden.

5.6.1.3. Güte der eingesetzten Instrumente

Alle in diesem Abschnitt beschriebenen vier Items des Projektfragebogens für Besucher scheinen evident und inhaltlich valide, jedoch gibt es keine allgemein anerkannte wissenschaftliche Definition, wann z.B. Therapieziele als realistisch oder relevant zu bewerten sind. Auch hier wird wieder auf die Erfahrung und Ausbildung der Besucher zurückgegriffen (vgl. Abschnitte 1.1.2. und 5.5.). Um so mehr überraschten die folgenden Ergebnisse zur Interraterreliabilität:

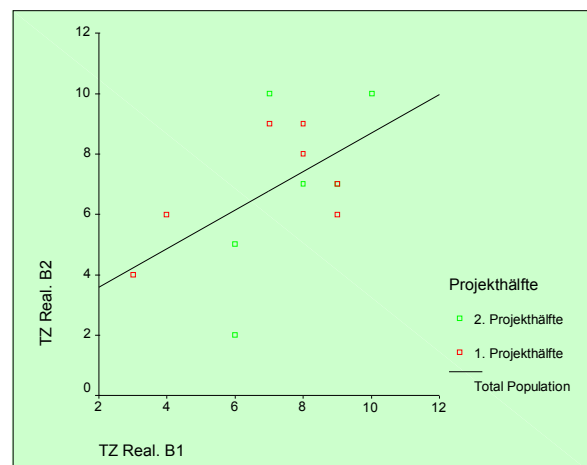
Ein 2. bzw. auch 3. Beurteiler fand sich in 11 bis 14 Fällen zu jeder Fragestellung. In welchem Maß diese unterschiedlichen Beurteiler zu ähnlichen Einschätzungen kommen zeigen Korrelationsanalysen und Streudiagramme (Tab. 12-15 und Abb. 17-20): Die Korrelationskoeffizienten nach Spearman-Rho zwischen einem 1. und einem 2. Beurteiler liegen je nach Item zwischen $r = .52$ und $.83^7$.

Signifikant korrelieren die Einschätzungen unterschiedlicher Besucher zu den Fragen nach dem zutreffenden Gewicht ($r = .76$; $p < 0,01$), der erwarteten ($r = .83$; $p < 0,01$) und der erreichten Übereinstimmung ($r = .73$; $p < 0,05$) der Therapieziele. Die Daten sprechen für eine gute Interrater-Reliabilität dieser Skalen.

Nicht signifikant und nur mäßig korrelieren die unterschiedlichen Einschätzungen hinsichtlich der Frage, ob und in welchem Ausmaß die Therapieziele realistisch sind, auch wenn den Behandlern im Zweitmeinungsverfahren ein im Mittel guter Therapiezielrealismus attestiert wird (siehe oben). Diese Diskrepanz könnte mit einem differenten Erfahrungs- und Ausbildungshintergrund unterschiedlicher Berufsgruppen zusammenhängen.

⁷ Korrelationen hinsichtlich eines 3. Beurteilers werden aufgrund der sehr geringen Fallzahl (2 Fälle pro zu beurteilendem Item) in diesen Überlegungen nicht berücksichtigt, finden sich jedoch in den dargestellten Tabellen.

Item „Therapiezielrealismus“ – Interkorrelationen Besucher				
		B1	B2	B3
Korrelationseffizient	B1	1,00		
	B2	,520	1,00	
	B3	1,00	1,00	1,00
Signifikanz (2-seitig)	B1			
	B2	,068		
	B3	1,00	1,00	
N	B1	35		
	B2	13	13	
	B3	2	2	2



Tab. 12 und Abb. 17: Interkorrelationen nach Spearman-Rho und Streudiagramm – Therapiezielrealismus

Legende siehe Tab. 18. B1, B2, B3: verschiedene Besucher

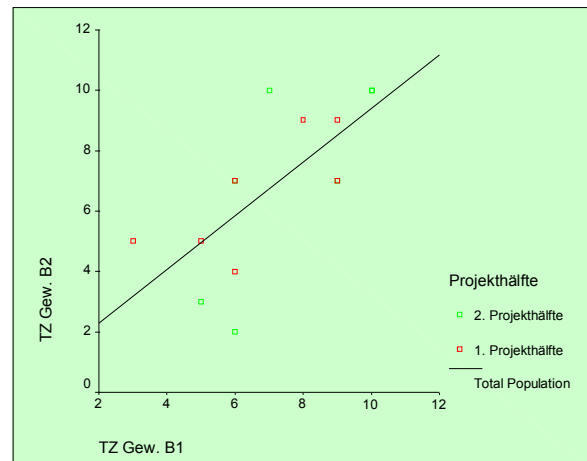
Deskriptiv:

TZ Real. B1: $M = 7,14$, $s = 2,01$, $N = 35$

TZ Real. B2: $M = 6,92$, $s = 2,36$, $N = 13$

TZ Real. B3: $M = 4,75$, $s = ,35$, $N = 2$

Item „Therapiezielrelevanz“ – Interkorrelationen Besucher				
		B1	B2	B3
Korrelationseffizient	B1	1,00		
	B2	,764**	1,00	
	B3	1,00		1,00
Signifikanz (2-seitig)	B1			
	B2	,001		
	B3	1,00		
N	B1	35		
	B2	14	14	
	B3	2	2	2



Tab. 13 und Abb. 18: Interkorrelationen nach Spearman-Rho und Streudiagramm – Therapiezielrelevanz

Legende siehe Tab. 18. B1, B2, B3: verschiedene Besucher
 ** Korrelation ist auf dem Niveau von 0,01 signifikant (2-seitig)

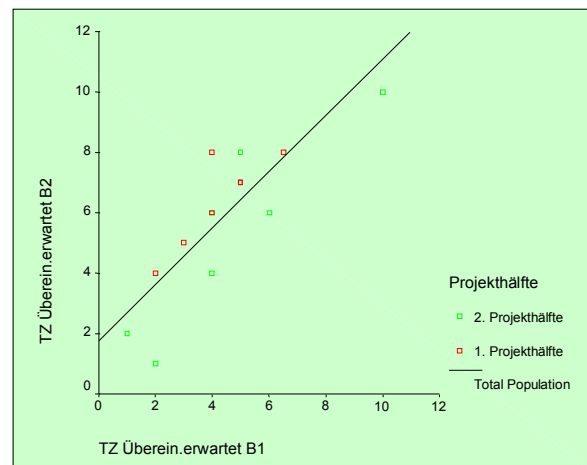
Deskriptiv:

TZ Gew. B1: M = 7,31, s = 1,91, N = 35

TZ Gew. B2: M = 6,79, s = 2,67, N = 14

TZ Gew. B3: M = 4,50, s = ,71, N = 2

Item „Therapiezielübereinstimmung erwartet“ – Interkorrelationen Besucher				
		B1	B2	B3
Korrelationseffizient	B1	1,00		
	B2	,829**	1,00	
	B3			
Signifikanz (2-seitig)	B1			
	B2	,000		
	B3			
N	B1	35		
	B2	14	14	
	B3	2	2	2



Tab. 14 und Abb. 19: Interkorrelationen nach Spearman-Rho und Streudiagramm – „Therapiezielübereinstimmung erwartet“

Legende siehe Tab. 18. B1, B2, B3: verschiedene Besucher
 ** Korrelation ist auf dem Niveau von 0,01 signifikant (2-seitig)

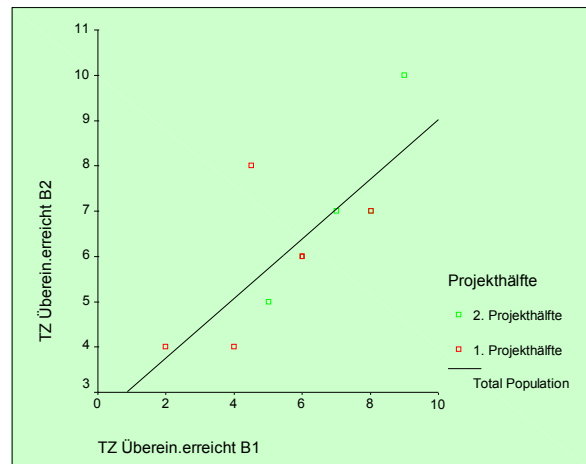
Deskriptiv:

TZ Überein. erwartet B1: M = 5,10, s = 2,47, N = 35

TZ Überein. erwartet B2: M = 5,86, s = 2,48, N = 14

TZ Überein. erwartet B3: M = 5,00, s = ,00, N = 2

Item „Therapiezielübereinstimmung erreicht“ – Interkorrelationen Besucher				
		B1	B2	B3
Korrelationseffizient	B1	1,00		
	B2	,728*	1,00	
	B3	-1,00	-1,00	1,00
Signifikanz (2-seitig)	B1			
	B2	,011		
	B3	1,00	1,00	
N	B1	32		
	B2	11	13	
	B3	2	2	2



Tab. 15 und Abb. 20: Interkorrelationen nach Spearman-Rho und Streudiagramm – „Therapiezielübereinstimmung erreicht“

Legende siehe Tab. 18. B1, B2, B3: verschiedene Besucher
 * Korrelation ist auf dem Niveau von 0,05 signifikant (2-seitig)
 Deskriptiv:
 TZ Überein. erreicht B1: M = 5,95, s = 1,78, N = 32
 TZ Überein. erreicht B2: M = 6,00, s = 1,87, N = 13
 TZ Überein. erreicht B3: M = 5,50, s = ,71, N = 2

5.6.1.4. Diskussion der Ergebnisse zum Projektfragebogen

Zunächst war untersucht worden, inwieweit das Projekt an sich einen Einfluss auf die Relevanz und die sachlich-nüchterne Einschätzung von Ressourcen und Möglichkeiten von Patient und therapeutischem Setting – also den Realismus bei der Auswahl von Therapiezielen hatte (siehe oben). Die Annahme bestand darin, dass die Auseinandersetzung im Team mit externen Besuchern und die intensive Beschäftigung mit individuellen Therapiezielen, wie sie die Skalenkonstruktion des GAS vorsieht, den Blick für relevante und realistisch-erreichbare Ziele schärft. Die Erwartung, dass dieser generelle Effekt durch die Auswertung der Fragen an die Besucher zu Relevanz und Realismus der Therapieziele messbar wird, hat sich – wie oben dargestellt – nicht erfüllt.

Da sowohl Besucher als auch Besuchte in wechselnden Rollen in das Programm involviert waren, haben sich möglicherweise auch deren "Ansprüche" an die Therapieziele erhöht und so eine Messbarkeit in diesem Ansatz unmöglich gemacht. Dafür sprechen auch die Überlegungen des folgenden Abschnitts.

5.6.2. GAS und Realismus von Therapiezielen

5.6.2.1. Einführung des GAS als zweitem Aspekt zum Therapiezielrealismus

Anders als die Einschätzung von Relevanz und Realismus der Therapieziele durch die Besucher, die – vor Konstruktion der individuellen Zielerreichungsskalen abgegeben – ja nichts anderes als eine zweite Meinung ist, wird mit der Antwort auf die Frage, wie nah lag der in den GAS-Skalen gemessene Behandlungserfolg am Wert "wie erwartet", abgebildet, wie realistisch die Therapieziele nach dem Prozess der peer review mit intensiver Auseinandersetzung zu den Therapiezielen zwischen Patient, Team und Besucher sowie der Konstruktion individueller Zielerreichungsskalen waren. Die Ergebnisse im Vergleich der Projekthälften werden im Folgenden dargestellt.

5.6.2.2. Ergebnisse

In die Auswertung gingen insgesamt 131 GAS-Skalen ein. Die Analyse fehlender Werte ergibt, dass bis auf einen Fall mit nur 2 GAS-Skalen zu jedem der 35 in die Untersuchung aufgenommenen Fälle 3-5 Skalen gewertet werden konnten.

Über die 2-5 Skalenwerte, die markierten, wo sich der Patient zum Ende des Beobachtungszeitraums befand, wurde das arithmetische Mittel gebildet. Die Experimentalgruppe wurde in Projekthälften (siehe Abschnitt 4.2.) geteilt.

Therapiezielrealismus – GAS		
		mGASend
1. Projekthälfte	Mittelwert	0,43
	N	18
	S	,53
2. Projekthälfte	Mittelwert	-0,10
	N	17
	S	,79
Gesamt	Mittelwert	0,17
	N	35
	S	,71

**Tab. 16: Therapiezielrealismus:
Zielerreichung mittels GAS**

N: Anzahl der gewerteten Fälle

s: Standardabweichung

mGASend : Arithmetisches Mittel der Endwerte auf der GAS je Fall

Der Mittelwertvergleich (Tab. 16) der gegen Ende der Beobachtungszeiträume erreichten GAS-Werte zeigt, dass, nachdem die Erwartungen der Skalenkonstrukteure in Ressourcen und Therapie in der 1. Projekthälfte eher übertroffen wurden, sie sich in der 2. Projekthälfte häufiger mit dem in der Therapie Erreichten decken. Die Häufigkeitsverteilungen (Abb. 21 a + b) demonstrieren dies eindrucksvoll.

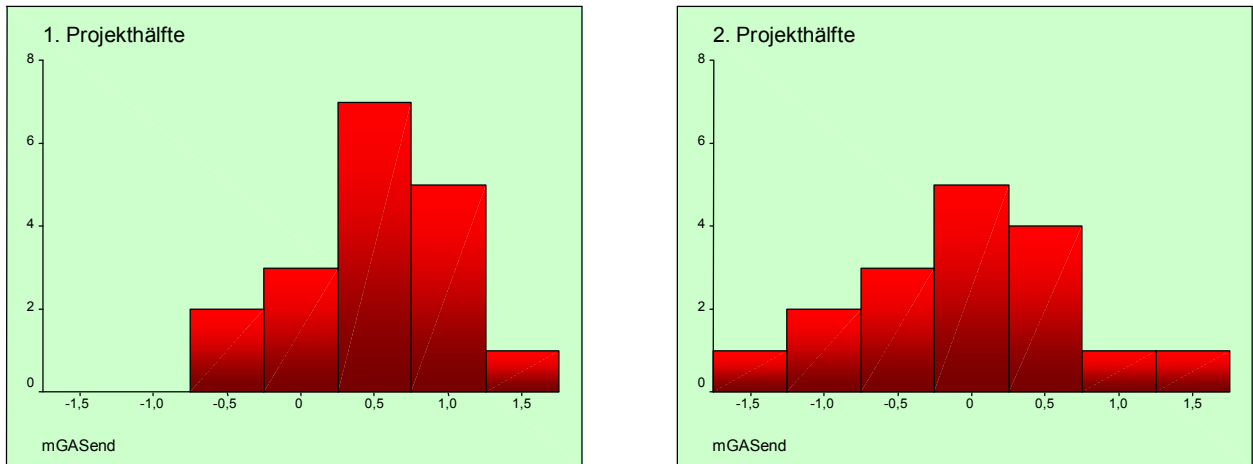


Abb. 21 a +b: Häufigkeitsverteilungen der GAS-Endwerte je Fall im Vergleich der Projekthälften

mGASend : Arithmetisches Mittel der Endwerte auf der GAS je Fall

Der T-Test (siehe Tab. 17) der je Fall gemittelten GAS-Endwerte deckt für die 1. Projekthälfte eine signifikante Abweichung vom Testwert 0 ("Ergebnis wie erwartet") auf ($p = 0,003$), dahingegen zeigt sich keine signifikante Abweichung in der 2. Projekthälfte.

Test bei einer Stichprobe						
Testwert = 0						
Projekthälften	T	df	Sig. (2-seitig)	Mittlere Differenz	95% Konfidenzintervall der Differenz	
					Untere	Obere
1. Projekthälfte MGASEND	3,44	17	,003	,43	,17	,69
2. Projekthälfte MGASEND	-,51	16	,620	-9,71E-02	-,51	,31

Tab. 17: GAS: Abweichung vom erwarteten Ergebnis – T-Test bei einer Stichprobe

Folgt man den oben gemachten Ausführungen zeigen die vorgelegten Daten, dass die Therapieziele durch das Programm sehr wohl realistischer geworden sind. Dabei wich im Projektverlauf ein für den therapeutischen Prozess als ungünstig einzuschätzender Pessimismus [vgl. z.B. 11, 60] realistischen bis diskret optimistischen Erwartungen.

5.6.2.3. GAS und Prognose – ein Vergleich mit Studien Bolms

In diesem Abschnitt werden weitere Ergebnisse dargestellt, die zeigen, inwieweit sich die in die Skalenkonstruktion zu U1 eingegangenen Erwartungen zum Zeitpunkt U2 erfüllt haben. Es geht also um die Prognosefähigkeit des GAS.

Im Gegensatz zu den in Abschnitt 5.6.2.2. gezeigten Daten, wo ein Summenwert der GAS-Endwerte je Patient gebildet worden war, wurde in den folgenden Daten der Fallbezug aufgegeben und jede GAS-Skala gleichwertig berücksichtigt.

Die Ergebnisse werden anschließend mit entsprechenden Voruntersuchungen Bolms verglichen.

Insgesamt zeigt sich in dieser Arbeit eine große prognostische Treffsicherheit: Über alle 131 Skalen der Zielerreichung wurde in knapp 43 % der Skalen (56) eine Punktlandung mit einem Ergebnis wie zu Therapiebeginn erwartet erreicht, bei noch mal 40 % (53 Skalen) wich das Ergebnis nur etwas vom wahrscheinlich Erreichbaren ab und nur bei 17 % (22 Skalen) lag das erreichte Ergebnis deutlich neben dem, was Patient und Therapie zugetraut worden war (siehe Abb. 22). Tendenziell ist prognostisch dabei ein nur diskreter therapeutischer Pessimismus zu beobachten, die Patienten übertrafen die Erwartungen etwas.

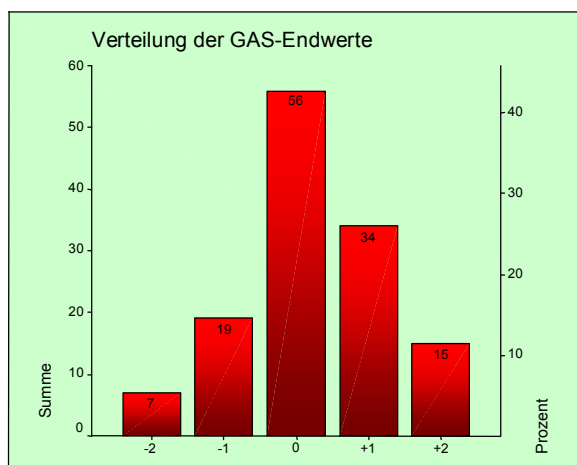


Abb. 22: GAS: Abweichung vom erwarteten Ergebnis / prognostische Genauigkeit

Diese Ergebnisse zeigen eine gute Prognosefähigkeit des GAS. Im Vergleich zu den Voruntersuchungen von Bolm [11, 1996] konnte diese deutlich verbessert werden: In einer Auswertung von 160 GAS-Skalen von 36 Patienten (66 % davon mit einer Psychose aus dem schizophrenen Formenkreis) der Abteilung für Sozialpsychiatrie der Freien Universität Berlin in den Jahren 1984-86 trafen lediglich in ca. 32 % der Skalen Prognose und Therapiezielerreichung

überein, 28 % der Endwerte auf der GAS-Skala wichen erheblich von den Erwartungen ab, wobei sich in der Schiefe der Verteilung eine deutliche Tendenz zum therapeutischen Pessimismus erkennen ließ.

Eine weitere Untersuchung Bolms an 339 zwischen 1986-89 in der DRK-Klinik Phönix aufgenommenen Patienten (80 % davon mit einer Psychose aus dem schizophrenen Formenkreis) ergab in der Auswertung von insgesamt 1133 GAS-Skalen in ca. 36 % eine Übereinstimmung von erwartetem und erreichtem Wert auf der GAS-Skala, ca. 26 % der Einschätzungen wichen erheblich vom wahrscheinlich erreichbaren Ergebnis ab. In ca. 57 % aller Therapieziele hatten die Rater die Ressourcen des Patienten und therapeutischen Möglichkeiten überschätzt, in ca. 25 % war es dabei zu erheblichen Fehleinschätzungen gekommen.

5.6.3. Interpretation und Fazit

Die Genauigkeit der Prognose und damit die Güte des GAS [vgl. 11, 55, 117] konnte in dieser Studie deutlich verbessert werden, systematische Verzerrungen im Sinne eines übermäßigen therapeutischen Pessimismus oder Optimismus wie in den Untersuchungen an der Abteilung für Sozialpsychiatrie und der Klinik Phönix traten zumindest in der zweiten Projekthälfte nicht mehr auf. Dabei dürften die folgenden Maßnahmen erfolgreich gewesen sein (siehe Abschnitte 1.2. und 4.1.1.4):

Durch GAS-Training und externe Überprüfung der Skalenentwürfe konnten formale Fehlerquellen minimiert werden (siehe Abschnitte 2. und 3.2.1.). Lediglich 2 GAS-Skalen (1,5 %) mussten wegen einer Änderung der Therapieziele im Behandlungsverlauf von der Auswertung ausgeschlossen werden. In der oben erwähnten Voruntersuchung Bolms an der Abteilung für Sozialpsychiatrie der FU Berlin [9] waren 6 % der Ziele ersatzlos gestrichen worden, andere Änderungen betrafen weitere 44 % der GAS-Skalen.

Durch ein systematisches GAS-Training, Überprüfung von Relevanz und Realismus der Therapieziele durch externe Besucher sowie Patientenbeteiligung und Zielsetzung im Team scheint es offensichtlich gelungen, inhaltliche Fehlerquellen bei Skalenkonstruktion zu reduzieren und die realistische Einschätzung dessen, was ein Patient in einer Behandlung erreichen kann, in für das therapeutische Team und Patient wichtigen Zielbereichen deutlich zu verbessern.

5.7. Korrelationsanalyse – Validitätsaspekte der eingesetzten Instrumente

5.7.1. Einfache Korrelationen

Die Ergebnisse zum Schweregrad der Erkrankung bzw. zur Zustandsänderung (CGI), zum globalen Funktionsniveau (GAF), der Summenwert der Prognoseskala nach Strauss und Carpenter sowie die für jeden Patienten der Experimentalgruppe gemittelten Ausgangs- und Endwerte auf der Goal Attainment Scale wurden zur Überprüfung auf statistische Zusammenhänge einer Korrelationsanalyse unterworfen (siehe Tab. 23).

		Korrelationen - Experimentalgruppe								
		cgi 1	cgi 2a	cgi 2b	gaf 1	gaf 2	S&C	mGASaus	mGASend	
cgi 1	r	1,000								
	Sig.	,								
	N	35								
cgi 2a	r	,412*	1,000							
	Sig.	,014	,							
	N	35	35							
cgi 2b	r	,084	,729**	1,000						
	Sig.	,630	,000	,						
	N	35	35	35						
gaf 1	r	-,434**	-,355*	-,078	1,000					
	Sig.	,009	,037	,654	,					
	N	35	35	35	35					
gaf 2	r	-,278	-,831**	-,662**	,466**	1,000				
	Sig.	,106	,000	,000	,005	,				
	N	35	35	35	35	35				
S&C	r	-,238	-,340*	-,033	,608**	,380*	1,000			
	Sig.	,169	,045	,850	,000	,024	,			
	N	35	35	35	35	35	35			
mGASaus	r	,151	,022	-,058	-,082	,022	,139	1,000		
	Sig.	,387	,900	,741	,639	,902	,426	,		
	N	35	35	35	35	35	35	35		
mGASend	r	-,121	-,630**	-,574**	-,065	,499**	,094	,274	1,000	
	Sig.	,490	,000	,000	,712	,002	,592	,111	,	
	N	35	35	35	35	35	35	35	35	

Tab. 23: Nichtparametrische Korrelationen nach Spearman-Rho

r: Korrelationskoeffizient; Sig.: 2-seitige Signifikanz; N: Anzahl der gewerteten Fälle.

* Korrelation ist auf dem Niveau von 0,05 signifikant (2-seitig)

** Korrelation ist auf dem Niveau von 0,01 signifikant (2-seitig)

cgi 1 : Schweregrad der Erkrankung zum Zeitpunkt U1

cgi 2a : Schweregrad der Erkrankung zum Zeitpunkt U2

cgi 2b : Zustandsänderung des Krankheitsbildes seit Aufnahme in die Untersuchung, eingeschätzt zum Zeitpunkt U2

gaf 1 : Globales Funktionsniveau zum Zeitpunkt U1

gaf 2 : Globales Funktionsniveau zum Zeitpunkt U2

S&C : Summenwert auf der Skala nach Strauss & Carpenter

mGASaus : Arithmetisches Mittel der Ausgangswerte auf der GAS je Fall

mGASend : Arithmetisches Mittel der Endwerte auf der GAS je Fall

5.7.2. CGI und GAF – Interpretation von Zusammenhängen

Es findet sich eine mit $r = -.43^8$ mäßige und auf dem Niveau von $p = 0,01$ signifikante Korrelation zwischen den zum Zeitpunkt U1 erhobenen Einschätzungen auf der CGI- und GAF-Skala.

Schweregrad und Zustandsänderung zum Zeitpunkt U2, gemessen mit CGI 2a und CGI 2b zeigen untereinander mit einem Spearman-Koeffizienten von $r = .73$ als auch zum zeitgleich eingeschätzten globalen Funktionsniveau (GAF 2) mit $r = -.83^8$ respektive $-.66^8$ mäßig (bis enge) und ebenfalls signifikante Zusammenhänge ($p < 0,001$).

Vergleichbare Ergebnisse reproduzieren sich auch in anderen Studien und deuten darauf hin, dass beide Skalen – obwohl ihnen unterschiedliche Konstrukte zugrunde liegen – auch Ähnliches beschreiben.

In einer US-amerikanischen (unkontrollierten) Längsschnitt-Untersuchung an 102 akut psychotischen, stationär in eine Universitätsklinik aufgenommenen Patienten mit einer Diagnose aus dem schizophrenen Formenkreis hinterfragten Goldman et al. [47, 1999] in welchem Ausmaß CGI- und GAF-Skala die Schwere und die Änderung in der Ausprägung von Positiv- und Negativsymptomatik, von Depression und Agitation abbilden. Besonders enge und signifikante Zusammenhänge fanden sich zwischen CGI-Skala und Positivsymptomatik, gefolgt von Negativsymptomatik und Agitation, wohingegen zur depressiven Symptomatik keine signifikanten Korrelationen bestanden. Bei der mitgeführten GAF-Skala lagen die Verhältnisse ähnlich, die Zusammenhänge waren jedoch durchweg etwas weniger eng, was nach den Autoren auf die Mitbewertung des psychosozialen Funktionsniveaus mit der GAF-Skala zurückzuführen sein dürfte.

Da in anderen Studien zu medikamentösen Behandlungseffekten affektiver Störungen sehr wohl Zusammenhänge zwischen Globalen Messinstrumenten und depressiver Symptomatik gefunden worden waren [Wernicke et al., 1987, referiert in 47], schien es den Autoren nahe liegend, dass Globale Messinstrumente die Kernsymptome der jeweiligen Störung abbilden. Alternativ wird diskutiert, dass die Rater bei schizophrenen Patienten depressive Symptome weniger stark in der Bewertung auf den Globalskalen berücksichtig bzw. gewichtet haben, da es sich bei den depressiven Symptomen nicht um Kernelemente dieser Störung handelt.

⁸ Das negative Vorzeichen erklärt sich aus der entgegen gerichteten Polarität der Skalen, es können daher die Beträge der dargestellten Koeffizienten als Maß eines gleichsinnigen Zusammenhangs gelten.

Die beschriebene ungleich geringere Abbildung depressiver Symptomatik in der CGI-Skala bestätigte sich in den Ergebnissen einer Untersuchung an 114 schizophrenen Patienten im Rahmen der SOHO-Studien [Haro et al. 2003, 53, siehe auch 4.1.2.]. Für den Zusammenhang zwischen dem Schweregrad auf der CGI-Skala und dem globalen Funktionsniveau (GAF) wurde ein Korrelationskoeffizienten nach Pearson von $r = -.67^9$ ($p < 0,001$) mitgeteilt.

Für beide Globalmaße (Schweregrad der CGI und GAF) besteht ein negativer Zusammenhang zwischen Ausgangswert und erreichter Veränderung (berechnet mit Hilfe der konstruierten Parameter $\text{diff cgi } 2a-1$ und $\text{diff gaf } 2-1$, siehe 5.4.). Die Korrelationskoeffizienten nach Spearman-Rho sind auf dem Niveau von $p = 0,05$ signifikant von Null verschieden und liegen bei $r = -.37$ für das Variablenpaar $\text{cgi } 1 / \text{diff cgi } 2a-1$ respektive $-.39$ für $\text{gaf } 1 / \text{diff gaf } 2-1$ (Daten nicht in Tabellenform gezeigt).

Damit bestätigt sich im Konsens mit anderen Studien [z.B. 31, 38, 43, 118, siehe Abschnitt 5.4.], dass besonders schwer kranke und / oder im Funktionsniveau eingeschränkte Patienten im Behandlungszeitraum größere Fortschritte erreichen.

Dabei könnte es – unterstützt durch die Fokussierung auf individuelle Therapieziele in der Experimentalgruppe – besonders in den Bereichen zu einer Verbesserung im Verlauf gekommen sein, in denen die einzelnen Patienten die größten Defizite aufwiesen. So ließe sich die erhebliche Zunahme der Enge des Zusammenhangs zwischen den Globalmaßen (Schweregrad und Funktionsniveau) in der Experimentalgruppe von $r = -.43^9$ zu U1 auf $-.83^9$ zu U2 erklären.

Gestützt wird diese These durch den Blick auf die entsprechenden Korrelationskoeffizienten der Kontrollgruppe: Zum Zeitpunkt U1 korrelieren Schweregrad und Funktionsniveau mit $r = -.56^9$ ($p < 0,01$), zu U2 mit $r = -.80^9$ ($p < 0,01$). Die Zunahme des Zusammenhangs fällt hier deutlich milder aus (Daten nicht in Tabellenform gezeigt).

⁹ Das negative Vorzeichen erklärt sich aus der entgegen gerichteten Polarität der Skalen, es können daher die Beträge der dargestellten Koeffizienten als Maß eines gleichsinnigen Zusammenhangs gelten.

5.7.3. Prognostischer Wert der Skala nach Strauss und Carpenter – Konsequenzen einer Analyse

Der Summenwert der Prognoseskala korreliert zwar auf dem Niveau von $p = 0,05$ signifikant mit Schweregrad ($r = -.34^{10}$) und globalem Funktionsniveau ($r = .38$) zum Ende des Beobachtungszeitraums, bezüglich des Funktionsniveaus bestand jedoch zum Eintritt in die Untersuchung mit einem Spearman-Koeffizienten von $r = .61$ ein weitaus engerer Zusammenhang auf dem Signifikanzniveau von $p = 0,01$. Folgende Schlüsse sind hier zulässig: Während ein geringer Zusammenhang zwischen Prognose und Schweregrad der Erkrankung als Ergebniskriterium nachweisbar ist, besteht kein statistisch relevanter Zusammenhang zur Zustandsänderung ($r = -.03^{10}$). Hinsichtlich des globalen Funktionsniveaus kehren sich die Verhältnisse sogar um: In der Prognoseskala ist das Funktionsniveau zum Aufnahmezeitpunkt gut abgebildet, dieser Zusammenhang verliert sich jedoch zum Zeitpunkt U2, was im Sinne einer fehlenden prognostischen Aussagefähigkeit der Skala von Strauss und Carpenter hinsichtlich des globalen Funktionsniveaus in dieser Studie interpretiert werden muss. Es wurde daher darauf verzichtet, weitere Daten zur Prognoseskala nach Strauss und Carpenter zu zeigen.

5.7.4. Goal Attainment Scaling

Vor jeder Überprüfung auf statistische Zusammenhänge steht – nach den Regeln guter methodologischer Praxis und um falschen Schlussfolgerungen vorzubeugen – eine Theorie über die möglichen und erwarteten Ergebnisse.

Gerade in diesem Punkt fanden sich jedoch bei der Durchsicht entsprechender Literatur uneinheitliche Konzeptionen. Die Verhältnisse beim GAS sind im Vergleich zu den klassischen Messinstrumenten ungleich komplexer.

In diesem Abschnitt sollen daher zunächst und in Ergänzung zu den Abschnitten 4.1.1.3., 4.1.1.4. und 5.6.2.3. wesentliche Positionen zur Problematik von Validität und psychometrischer Theorie des GAS dargestellt werden. Besondere Berücksichtigung findet dabei das Modell der „idiosynkratischen Messung“ von Heavlin et al. [55]. Basierend auf diesem Konzept werden die Ergebnisse der Korrelationsanalyse interpretiert und methodische Strategien zur Bestimmung der „concurrent validity“ des GAS aufgezeigt.

¹⁰ Das negative Vorzeichen erklärt sich aus der entgegen gerichteten Polarität der Skalen, es können daher die Beträge der dargestellten Koeffizienten als Maß eines gleichsinnigen Zusammenhangs gelten.

5.7.4.1. Was misst das Goal Attainment Scaling?

Mit der Erstbeschreibung des GAS von Kiresuk & Sherman 1968 [64] war eine Methode zur Bestimmung des Ausmaßes vorgestellt worden, in dem vorher spezifizierte Therapieziele erreicht werden. Es war nicht die Rede davon, ob der GAS-Score als Maß von Veränderung bei einer Reihe von Zielbereichen anzusehen ist oder einer Zustandsbeschreibung einer solchen Gruppe von Zielbereichen nach der Behandlung entspricht. Dies sollte weitreichende Konsequenzen haben:

Mit der Verbreitung des Instrumentes folgten die zahlreichen Anwender nicht einheitlich einer dieser Richtungen, hinsichtlich der Interpretation des GAS zeigten sich weder Konsistenz noch Übereinstimmung [107] und oft genug war die Frage, was das GAS misst, gar nicht reflektiert worden.

Einige Untersuchungen [z.B. Austin et al., 2] errechneten aus den Ausgangs- und Endwerten auf der GAS-Skala einen Change-Score, was eine Interpretation des GAS-Score im Sinne einer Zustandsmessung nach der Behandlung impliziert. Das GAS rückt so in die Nähe anderer – klassischer – psychiatrischer Evaluationsmethoden, welche den Zustand am Ende einer Behandlung beschreiben [vgl. 107].

Smith und Cardillo kritisierten die Anwendung eines Change-Scores [referiert in Roecken, 95] nicht nur als – aufgrund einer unzulässigen Vermischung von Rohwerten – statistisch unsauber (unter Berufung auf Cronbach und Furby 1970 [23]) sondern auch als inkongruent mit der „real meaning“ des GAS-Scores. Smith [107] stellte in diesem Zusammenhang fest:

„... the Goal Attainment Score is not a measurement of posttreatment standing but is itself a summary index of the progress or change that has taken place by the end of treatment.”

Kiresuk & Sherman argumentierten noch 1977 [referiert in Heavlin et al., 55], dass die Konstruktvalidität offenbar sei und sich aus dem Grundsätzlichen des GAS ableite und dass das zugrunde liegende Konstrukt nur durch „content validity assessments“ evaluiert werden könne. Oder in Kurzform mit den Worten Sherman's 1974 [106, siehe auch 66]:

„The Goal Attainment-Score is as valid as it is reliable...”

Die häufig genannten Argumente zur Augenschein- und inhaltlichen Validität überzeugten die Kritiker nicht [25, 107]. Es mangelte ihnen an präzisen Definitionen, empirischen Beweisen und Datenanalysen, mit der die behauptete Konstruktvalidität begründet werden konnte [vgl. 55, 101].

Dieser Mangel an einer substanziellen, definierten psychometrischen Theorie war nicht nur dafür verantwortlich, dass sich sehr unterschiedliche Auffassungen über die Bedeutung des GAS-Scores herausbilden konnten [siehe oben, 95], er kann auch als Ursache vielgestaltiger konzeptueller [z.B. Seaburg & Gillespie 1977, 101] und in Konsequenz praktischer Schwierigkeiten [z.B. Cytrynbaum et al. 1979, 25] gelten [55].

Aus dieser wachsenden Kritik der späten 70er Jahre heraus entwickelte sich der folgende, bemerkenswerte Ansatz.

Heavlin, Lee-Merrow und Lewis [55] stellten 1982 ein mathematisches Model vor, welches die psychometrischen Eigenschaften des GAS definiert und diese zum einen mit der Theorie normativer Messungen zum anderen mit der Theorie der Generalisierbarkeit verbindet.

Zwei Eigenschaften des GAS wurden dabei als essentiell hervorgehoben: seine „within-subjects standardization“ und seine Abhängigkeit von Wahrnehmung und Erfahrungshintergrund der Rater. In einem nächsten Schritt wurde die formale Definition des GAS als „idiosynkratische“ Messung eingeführt:

„A measure is called idiosyncratic if for each subject the expected level at follow up, conditional upon information assessible at baseline, is constant. ... (By this definition) a measure is idiosyncratic if its conditional expectation is *not* a function of factors assessible at baseline.“

Die Definition erklärt die Prognose – im Fall des GAS – als in die Skalierung des erwarteten Ergebnisses eingegangen. Eine „idiosynkratische Messung“ zeigt somit nicht nur eine Veränderung im „level of functioning“ an, sondern versucht implizit jegliche Korrelation mit einem Prognoseindikator zu unterdrücken. Letzteres kann auch zur Erklärung für eine zusätzliche Fehlerquelle dienen, der normative Messungen für gewöhnlich nicht unterliegen: Die Gefahr, sich bei der Skalenkonstruktion im „erwarteten Ergebnis“ zu verschätzen (siehe Ergebnisse 5.6.2.2.-3.: Schiefe von Verteilungen).

In ihren Ausführungen zur Konstruktvalidität nehmen die Autoren Bezug auf die u. a. von Seaberg und Gillespie [101] beanstandete Vagheit hinsichtlich der Definitionen, was GAS-Skalen messen und gelangen damit in die Nähe zur „theory of generalizability“ von Cronbach [weiterführende Literatur: 24, 104]:

„We suggest that such vagueness is intrinsic to attainment scaling, not necessarily bad, and even useful in understanding the aspect of outcome that GAS attempting to measure. Outcome itself is a vague and global notion: what is good for one subject is not always the most

desirable result for another; clinicians proposing different courses of treatment are the first to recognize this. Further, clinicians are trained to judge when their clients are making adequate progress. GAS also measures vaguely and globally; it seems a small step to assert that GAS and clinical judgment may measure the same construct, amorphous though it be. In other words, GAS may be thought as attempting to measure what would be an intelligent consensus of clinicians' judgement of outcome." [55]

5.7.4.2. Übereinstimmungsvalidität – die Suche nach geeigneten Außenkriterien

Im vorangegangenen Abschnitt sind wesentliche und sehr unterschiedliche Positionen zum Konstrukt des GAS dargestellt worden. Wenn aber schon kein Konsens über das zugrunde liegende Konstrukt besteht, nach welchem Kriterium soll dann ein geeignetes Referenzkriterium ausgewählt werden, um die Übereinstimmungs- oder Kriterienvalidität festzustellen?

Im Anschluss an seine Ausführungen zu GAS- und Change-Score empfiehlt Smith 1981, je nachdem, ob die Ergebnisse des GAS als Zustandsmessung nach einer Behandlung oder als Index für erreichte Veränderung betrachtet werden, geeignete Referenzkriterien auszuwählen:

„... one's view of what it is that Goal Attainment Scaling is measuring will determine, in a major way, how one thinks its concurrent validity should be tested." [107]

Die kompromißhafte Natur dieses Vorschlags kann in testtheoretischer Hinsicht kaum akzeptiert werden, da sie mehr als eine zugrunde liegende Theorie zulässt. Die Absicht von Smith jedoch, durch den expliziten Einbezug eines Konstruktes das Design von Kreuzvalidierungsstudien annehmbarer und ihre Ergebnisse glaubwürdiger zu machen, kann als eine wichtige Grundlage weiterer Evaluationsbemühungen der psychometrischen Eigenschaften des GAS gelten.

Ungeachtet aller kontrovers geführten Diskussionen zur Güte des GAS scheinen sich Kritiker und Befürworter der Methode doch in einem Punkt einig zu sein:

Der hohe Grad der Individualisierung des GAS-Scores (Auswahl der Ziele, Konstruktion der Skalenniveaus) erschwert die Vergleichbarkeit mit Behandlungsergebnissen anderer Methoden erheblich [vgl. z.B. 25].

Kiresuk und Sherman selbst erwarteten nur geringe bis mäßige Korrelationen zwischen dem GAS und normativen Messungen und begründeten dies mit eben dieser individualspezifischen Natur des Instruments [66, siehe auch 25]. Garwick äußerte sich 1974 in diesem Zusammenhang [referiert in 25] wie folgt:

„... there is something wrong if GAS which is an individualized measure, correlates with something which applies the same standards to everyone.“

Folgt man der Argumentationsweise Sherman's und betrachtet das GAS als „adjustable to an individual“ [Sherman 1974, referiert in 25] wären lediglich in sehr homogenen Patientenpopulationen und bei speziell auf die Bedürfnisse dieser Klientel abgestimmter Meßmethoden höhere Korrelationen zwischen GAS-Score und normativen Messungen zu erwarten [vgl. 95].

Nicht nur, dass man es in der Anwendung des GAS überwiegend mit weitaus heterogeneren Kollektiven zu tun hat [95], auch empirische Studien zur Kriterienvalidität ergaben, dass am ehesten Globalschätzungen mit den GAS-Ergebnissen korrelieren [25, 95]. In seiner Übersichtsarbeit teilt Roecken diesbezüglich Korrelationen zwischen $r = .12$ und $r = .85$ mit und resümiert,

„... dass das GAS einem Global-Rating zur Klient-Besserung näher kommt als ursprünglich angenommen wurde...“.

Bolm [9] berichtet in der schon erwähnten Untersuchung an 36 Patienten der Abteilung für Sozialpsychiatrie (siehe 5.6.2.3.) über Korrelationen nach Spearman von $r = .40$ zwischen GAS-Score und Globalmaßen (BPRS-Differenzwerte und CGI - Zustandsänderung).

Die in der vorliegenden Studie ermittelten Korrelationen zwischen GAS-Score und den zu U2 erreichten Werten in den Globalmaßen CGI 2a und 2b sowie GAF 2 (Tab. 25) liegen inmitten der Spanne der berichteten Ergebnisse. Die Beträge¹¹ der Spearman-Koeffizienten reichen dabei von $r = .50$ für das globale Funktionsniveau bis $r = .63$ für den Schweregrad der Erkrankung und erweisen sich als durchweg auf dem Niveau $p = 0,01$ signifikant.

¹¹ Aufgrund der entgegen gerichteten Polarität der CGI-Skala („Schweregrad“) gilt der Betrag als Maß eines gleichsinnigen Zusammenhangs.

Wenden wir uns an dieser Stelle noch einmal den Überlegungen von Heavlin et al. zu: Wie oben dargestellt und in der Definition als idiosynkratischer Messung erklärt, sollten GAS-Scores mit keinem zur Ausgangsuntersuchung bestimmten Prädiktor oder Prognoseindikator korrelieren.

Dies grenzt das GAS entschieden von normativen Messungen ab, bei denen oft ein möglichst großer Zusammenhang zu Prognoseindikatoren in der Ausgangsmessung erwünscht ist.

Aufgrund des Fehlens dieser Varianzkomponente wird das Ausmaß des Zusammenhangs, den eine idiosynkratische mit einer normativen Messung teilen kann, auf natürliche Weise begrenzt [55].

Um diese Behauptungen anhand der vorliegenden Ergebnisse zu überprüfen, war es notwendig, den Fokus nicht nur auf markante Zusammenhänge zu richten, sondern auch fehlende Zusammenhänge zu hinterfragen und mögliche Unterschiede zwischen den in diesem Projekt eingesetzten Instrumenten aufzudecken.

Neben den bereits beschriebenen Zusammenhängen können aus der Korrelationsanalyse (Tab. 25) folgende Befunde abgeleitet werden:

- Zwischen den jeweiligen Ausgangs- und Endwerten der Globalmaße GAF und CGI ließen sich mit $r = .47$ respektive $r = .41$ erhebliche Zusammenhänge nachweisen. Diese besagen nichts anderes als, dass Patienten mit einem hohen Funktionsniveau zum Eintritt in die Untersuchung auch im Behandlungsergebnis ein hohes Funktionsniveau erreichen und zu U1 als besonders krank beurteilte Patienten auch zu U2 als eher schwerer krank eingeschätzt werden. Bei einer statistischen Irrtumswahrscheinlichkeit von 0,5 respektive 1,4 % scheint den Einschätzungen zu U1 plausibel eine prädiktive Potenz innezuwohnen, die sich als Teil der Varianz der Ergebnisse wiederfindet.
- Nur schwache oder sehr lose einfache Zusammenhänge (siehe Tab. 25) fanden sich dagegen zwischen GAS-Score (mGASend) und den Ausgangswerten auf der GAS-Skala ($r = .27$) sowie zwischen GAS-Score und allen anderen zu U1 durchgeführten Fremdbeurteilungen (Summenwert der Prognoseskala: $r = .09$; cgi 1: $r = -.12$; gaf 1: $r = -.07$).

Damit erfüllen sich in dieser Studie die Erwartungen von Heavlin et al., das Konzept des GAS als einer idiosynkratischen Messung wird dadurch empirisch gestützt.

Um den GAS-Score mit den zum Ende einer Behandlung erreichten Skalenwerten normativer Messungen angemessen vergleichen und so die Kriterienvalidität des GAS bestimmen zu können, schlagen Heavlin et al. [55] methodisch die Anwendung partieller Korrelationen vor.

5.7.4.3. Partielle Korrelationen zwischen GAS-Score und globalen Ergebnismaßen

Partielle Korrelationen beschreiben die Beziehung zwischen zwei Variablen unter Berücksichtigung weiterer möglicher Einflussfaktoren. Wenn vermutet wird, dass weitere Variablen – im folgenden Kontrollvariablen genannt – die Werte der zu vergleichenden Variablen beeinflussen und dadurch den einfachen Korrelationskoeffizienten verzerren, kann der lineare Einfluss der Kontrollvariablen durch das Verfahren der partiellen Korrelation mittels Regressionsschätzungen ausgeschaltet werden. So lässt sich mit der partiellen Korrelation die wichtige experimentelle Technik des Konstanthaltens simulieren. [Online-Handbuch für SPSS sowie 12, 108].

Bezüglich der Anwendung im vorliegenden Kontext ist kritisch zu bemerken, dass in dieses Verfahren die Korrelationskoeffizienten nach Pearson, einem Maß für lineare Zusammenhänge, eingehen und damit eine bivariate Normalverteilung für jedes Variablenpaar angenommen wird, eine Tatsache die in dieser Studie nicht überprüft worden ist.

Um den Effekt einer Bereinigung mittels partieller Korrelationen nicht verzerrt darzustellen, sind in Tab. 24 auch die entsprechenden einfachen Korrelationen nach Pearson angegeben.

Die Auswertung erfolgte mit dem Statistikprogramm SPSS.

Ziel der Analyse partieller Korrelationen hinsichtlich der Validitätsbestimmung des GAS ist es, Varianzkomponenten (oder Kontrollvariablen) zu eliminieren, die den „inhaltlichen“ Zusammenhang zwischen GAS-Score und normativen Erfolgsmaßen aufgrund ihrer unterschiedlichen psychometrischen Eigenschaften verzerren können.

Die Analyse partieller Korrelationen zwischen GAS-Score (mGASend) und den Ergebnisparametern CGI und GAF wurde auf drei unterschiedlichen Ebenen durchgeführt (Tab. 24):

In einem ersten Schritt (Ebene-1) wurden die Korrelationskoeffizienten zwischen GAS-Score und globalen Ergebnismaßen (z.B. gaf 2) um den Einfluss der Messung zu U1 in dem entsprechenden Globalmaß (z.B. gaf 1) bereinigt. Wie angenommen, erhöhte sich die Stärke der Interkorrelationen im Vergleich zu den einfachen Korrelationen nach Pearson deutlich (Tab. 24), wenn der Einfluss eines Prädiktors für die globalen Erfolgsmaße unterdrückt wurde.

Eine zweite Analyse (Ebene-2) bezog zusätzlich die Ausgangswerte auf der GAS-Skala als Kontrollvariable mit ein. Wie aus den bisherigen Ausführungen hervorgeht, waren dadurch keine großen Effekte auf die Enge der Korrelation erwartet worden, dies konnte in den Ergebnissen bestätigt werden: Die Bereinigung um die Ausgangswerte der GAS-Skala führte lediglich zu einer diskreten Zunahme der partiellen Interkorrelationen (im Vergleich zu Ebene-1).

		Partielle Korrelationen: GAS-Score – Outcome mit CGI, GAF			
		mGASend			nicht kontrolliert (Pearson)
kontrolliert für:		Ebene-1	Ebene-2	Ebene-3	
		cgi 1 bzw. gaf 1	cgi 1 bzw. gaf 1 und mGASaus	cgi 1, gaf 1, S&C und mGASaus	
cgi 2a	r	-,706**	-,730**	-,732**	-,650**
	Sig.	,000	,000	,000	,000
	N	32	31	29	35
gaf 2	r	,704**	,709**	,685**	,608**
	Sig.	,000	,000	,000	,000
	N	32	31	29	35

Tab. 24: Partielle Korrelationen

r: Korrelationskoeffizient; Sig.: 2-seitige Signifikanz; N: Anzahl der gewerteten Fälle.

** Korrelation ist auf dem Niveau von 0,01 signifikant (2-seitig)

cgi 1 : Schweregrad der Erkrankung zum Zeitpunkt U1

cgi 2a : Schweregrad der Erkrankung zum Zeitpunkt U2

gaf 1 : Globales Funktionsniveau zum Zeitpunkt U1

gaf 2 : Globales Funktionsniveau zum Zeitpunkt U2

S&C : Summenwert auf der Skala nach Strauss & Carpenter

mGASaus : Arithmetisches Mittel der Ausgangswerte auf der GAS je Fall

mGASend : Arithmetisches Mittel der Endwerte auf der GAS je Fall

Schließlich (Ebene-3) wurden die Effekte aller in der Untersuchung zum Zeitpunkt U1 eingesetzten Fremdbeurteilungsverfahren auf die Korrelationskoeffizienten der Beziehungen zwischen den globalen Erfolgsmaßen (cgi 2 und gaf 2) und dem GAS-Score (mGASend) überprüft. In den Ergebnissen scheinen sich (in Analogie zu den einfachen Korrelationen, Tab

23) die Variablen gaf 1 und der Summenwert der Prognose als Faktoren mit prädiktivem Wert für die Einschätzung des Schweregrads zu U2 abzubilden. Bezüglich der Höhe der partiellen Interkorrelation zwischen GAS-Score und dem Ergebnis im globalen Funktionsniveau zeigt sich hier im Vergleich zur Ebene-2 eine leichte Reduktion. Dieser Effekt lässt sich durch die bereits in der Auswertung für einfache Korrelationen begründete fehlende prognostische Aussagekraft der Skala von Strauss und Carpenter hinsichtlich des globalen Funktionsniveaus hinreichend erklären.

5.7.4.4. Zusammenfassung und Fazit

Die dargestellten Ergebnisse weisen darauf hin, dass Goal Attainment Scaling und Globalmaße wie die CGI oder das GAF trotz ihrer Verschiedenheit hinsichtlich theoretischer Hintergründe, psychometrischen Eigenschaften und Handhabung Ähnliches messen.

Darüber hinaus finden sich empirische Belege für die Richtigkeit der Theorie von Heavlin et al. (Konstruktvalidität).

Ihr Konzept der „Idiosyncraticity“ des GAS löst Aspekte des Referenzkriterienproblems auf elegante Weise per Definitionem. In Konsequenz können durch die Technik der partiellen Korrelation um (für globale Ergebnismaße) prädiktive Faktoren bereinigte Zusammenhänge als Indices für „concurrent-“ oder Kriteriengültigkeit gelten. Um bei diesem Verfahren Interpretationsfehler zu vermeiden, sollten die Befunde theoretisch erwartbar sein.

Die partiellen Zusammenhänge zwischen GAS-Score und dem Behandlungsergebnis gemessen mit CGI (Schweregrad) bzw. GAF erwiesen sich mit einer Irrtumswahrscheinlichkeit unter 0,1 % als hochsignifikant und sind als moderat (bis hoch) einzuschätzen.

Die Nähe der Beurteilungen mittels GAS zu Einschätzungen auf normativen Globalskalen mag auf den ersten Blick überraschen, sie wird jedoch verständlich legt man die weiter oben zitierten Überlegungen von Heavlin et al. zugrunde, wonach der GAS-Score so etwas wie eine klinische „common sense – Schätzung“ [95] zu sein scheint.

Letztendlich scheinen andererseits auch Globalmaße nicht so „global“ wie dieser Begriff suggeriert. Allein die Formulierung der Anleitung zur CGI: „ziehen Sie Ihren gesamten Erfahrungsschatz an *dieser* Art von Kranken in Betracht...“ lädt den Beurteiler zum Differenzieren ein (vgl. Abschnitt 4.1.2.); die in Abschnitt 5.7.2. zitierten Studien von Goldman et al. [47] und Haro et al. [53] sowie das im Abschnitt 5.4.3. Gesagte liefern weitere Argumente.

5.8. Befragung der besuchten Teams

5.8.1. Methode – Projektfragebogen für besuchte Teams

Wie den Besuchern, wurde auch jedem an der peer review teilnehmenden Kollegen des besuchten Teams ein Projektfragebogen ausgehändigt, auf dem nach der Teamsitzung mit Besuchern, eine Einschätzung über diese kollegiale Beratung abgegeben werden sollte. Neben 8 Fragen, die durch Setzen einer Markierung im Intervall einer metrischen Skala von 1 bis 10 zu beantworten waren, ließ der Fragebogen Platz für Kommentare.

Die Fragen im Einzelnen:

- "Wie nützlich war diese kollegiale Beratung über Therapieziele für die weitere Behandlung?"
- "Wie nützlich waren die Anregungen des Besuchers, die über das Thema 'Therapieziele' hinausgingen?"

Der Wert 1 steht bei diesen beiden Fragen für "völlig unnützlich", der Wert 10 für "sehr nützlich".

- "Hinsichtlich der Therapieziele fühle ich mich als Behandler nach dieser Beratung..."

Der Wert 1 steht hier für "...gar nicht sicherer.", der Wert 10 für "...ja, auf jeden Fall sicherer".

- "Nach dieser Beratung verstehe ich als Behandler Angehörige oder Patienten besser..."
- "Durch die Beratung ist für mich klarer, wer im menschlichen Umfeld des Patienten welches Interesse an der Behandlung hat..."
- "Nach dieser Beratung können wir in der Behandlung des Patienten besser an einem Strang ziehen..."
- "Nach dieser Beratung fließt die Sichtweise anderer Berufsgruppen besser als vorher in die Therapieziele ein..."
- "Nach dieser Beratung sollte die geplante Aufenthaltsdauer des Patienten verkürzt werden..."

Der Wert 1 steht bei den letzten fünf Items für "...nein, gar nicht.", der Wert 10 für "...ja, auf jeden Fall."

5.8.2. Ergebnisse

Insgesamt gingen 252 Fragebögen in die Auswertung ein. Auf einen der 35 untersuchten Fälle heruntergerechnet sind das 7,2 Fragebögen pro Patient, die Spannweite liegt hier zwischen 2 und 11. Die Einschätzungen der verschiedenen Beurteiler zu einer Fragestellung wurden, ohne dabei zwischen den einzelnen Berufsgruppen zu unterscheiden, zu einem Mittelwert für jeden Fall zusammengefasst und anschließend wurde über die 35 Fälle wiederum das arithmetische Mittel gebildet.

Aufgrund der Dateneigenschaften wurde auf eine weiterführende statistische Auswertung verzichtet. Die Ergebnisse können daher nur allgemein, im Sinne einer Umfrage interpretiert werden (siehe Abschnitt 4.2.).

		mKOLL	mANREG	mFÜHL	mVERST	mINT	mSTRAN	mSICHT	mAUFEN
Befragung der besuchten Teams									
1. Projekthälfte	M	7,18	6,94	6,70	6,58	6,17	7,18	6,39	2,94
	N	18	18	18	18	18	18	18	18
	s	1,24	1,03	1,16	1,28	1,07	1,01	1,15	,81
2. Projekthälfte	M	7,30	6,84	7,24	6,32	5,41	7,52	6,48	2,83
	N	17	17	17	17	17	17	17	17
	s	1,18	1,22	1,24	,90	1,30	1,12	1,09	1,25
Gesamt	M	7,24	6,89	6,96	6,46	5,80	7,35	6,43	2,89
	N	35	35	35	35	35	35	35	35
	s	1,20	1,11	1,21	1,10	1,23	1,06	1,10	1,03

Tab. 25: Projektfragebogen für die besuchten Teams - Mittelwerte

M: Mittelwerte; N: Anzahl der gewerteten Fälle; s: Standardabweichung.

mKOLL: Auswertung der Frage (1) „Wie nützlich war diese kollegiale Beratung über Therapieziele für die weitere Behandlung?“

mANREG: Auswertung der Frage (2) „Wie nützlich waren die Anregungen des Besuchers, die über das Thema ‘Therapieziele’ hinausgingen?“

mFÜHL: Auswertung der Frage (3) „Fühle ich mich als Behandler nach dieser Beratung hinsichtlich der Therapieziele sicherer?“

mVERST: Auswertung der Frage (4) „Nach dieser Beratung verstehe ich als Behandler Angehörige oder Patienten besser...“

mINT: Auswertung der Frage (5) „Durch die Beratung ist für mich klarer, wer im menschlichen Umfeld des Patienten welches Interesse an der Behandlung hat...“

mSTRAN: Auswertung der Frage (6) „Nach dieser Beratung können wir in der Behandlung des Patienten besser an einem Strang ziehen...“

mSICHT: Auswertung der Frage (7) „Nach dieser Beratung fließt die Sichtweise anderer Berufsgruppen besser als vorher in die Therapieziele ein...“

mAUFEN: Auswertung der Frage (8) „Nach dieser Beratung sollte die geplante Aufenthaltsdauer des Patienten verkürzt werden...“

Die Mittelwerte für das gesamte Projekt liegen für die Fragen 1 bis 4, 6 und 7 zwischen 6,4 und 7,3 und damit entweder im oberen Abschnitt des mittleren oder unteren Abschnitt des oberen Drittels der Skalen (siehe Tab. 25). Im Vergleich der Projekthälften zeigt sich eine relative Stabilität der Mittelwerte. Eine Änderung von über 0,5 ist allein hinsichtlich der Einschätzungen zur Fragestellung, ob sich die Beurteiler hinsichtlich der Therapieziele nach der Beratung als Therapeuten sicherer fühlen (3), zugunsten der zweiten Projekthälfte zu verzeichnen. Hier könnte es sich um einen generellen Qualifizierungs- und Fortbildungseffekt durch das Programm handeln (vgl. Abschnitt 1.2.).

Zur Frage, ob durch die Beratung klarer wurde, wer im Umfeld des Patienten welches Interesse an der Behandlung hat (5), liegt der Gesamt-Mittelwert mit 5,8 nahe der Skalenmitte. Im Vergleich 1. zu 2. Projekthälfte sinkt dieser Mittelwert um 0,8 auf 5,4.

Interessant ist auch die Beantwortung der Frage, ob nach der Beratung die geplante Aufenthaltsdauer des Patienten verkürzt werden sollte: Mit einem Wert von 2,9 liegt die gemittelte Einschätzung im unteren Skalendrittel; sie ist über die Projekthälften relativ stabil. (siehe Abschnitt 5.3.)

Ähnlich der Frage wie sie den besuchten Teams vorgelegt wurde, sollten auch die Besucher mit den Projektfragebögen zur Frage Stellung beziehen, ob nach der Beratung die Sichtweise anderer Berufsgruppen besser in die Therapieziele einfließen kann. Die Einschätzungen waren wieder auf einer Skala im Intervall zwischen 1 und 10 zu markieren. Der Wert 1 steht für "nein, überhaupt nicht", der Wert 10 für "ja, vollständig".

Hatten zwei oder drei Besucher einen Patienten besucht, wurde über die einzelnen Einschätzungen eines Falles das arithmetische Mittel errechnet und diese Daten dann weiter im Mittelwertvergleich über die angegebenen Fallgruppen zusammengefasst.

Antworten der Besucher auf die Frage „ob nach der Beratung (peer review) die Sichtweise anderer Berufsgruppen besser in die Therapieziele einfließen kann“			
	Mittelwert	N	s
1. Projekthälfte	6,96	16	1,57
2. Projekthälfte	6,71	17	1,89
Gesamt	6,83	33*	1,72

* entspricht 94,3% der Fälle der Gesamtstichprobe, zu zwei Fällen der 1. Projekthälfte lagen keine verwertbaren Daten vor.

Tab. 26: Einfließen von Sichtweisen anderer Berufsgruppen - Mittelwerte

N: Anzahl der gewerteten Fälle, s: Standardabweichung

In der Auswertung (siehe Tab. 26) liegt der Mittelwert bei 6,8, also – wie bei der Beurteilung dieser Fragestellung durch die Mitglieder der besuchten Teams (siehe oben) – im oberen Abschnitt des mittleren Skalendrittels und bleibt im Projekthälftenvergleich gleichfalls relativ stabil.

5.8.3. Interpretation und Fazit

Die Daten lassen folgende Interpretationen in Übereinstimmung mit freien Kommentaren der Projektfragebögen, den Protokollen zur peer review und persönlichen Gesprächen mit einzelnen Projektteilnehmern zu:

Die kollegiale Beratung

- wird von den besuchten Teams als nützlich empfunden und bewertet,
- fördert das Verständnis für die Situation und Interessen des Patienten und seines Umfeldes,
- lässt die Sichtweisen anderer Berufsgruppen besser einfließen und fördert eine zielgerichtete Zusammenarbeit und
- lässt sich die besuchten Kollegen als Therapeuten sicherer fühlen.

Die ablehnend-indifferenten Einschätzungen zu einer Verkürzung der Behandlungsdauer spiegeln im Kontext der Auseinandersetzung mit den Kostenträgern sicher zum Teil auch diesbezügliche Ängste wider, das Projekt könne einer weiteren Verschlechterung der Patientenversorgung Vorschub leisten.

Auch die Frage, wer welches Interesse an der Behandlung hat, könnte von einigen in diese Richtung interpretiert worden sein. Zur Diskussion der Dynamik dieses Prozesses siehe auch Abschnitt 5.5..

Teams wie Besucher schätzen übereinstimmend die besseren Möglichkeiten des Einfließens von Sichtweisen anderer Berufsgruppen in die Therapieziele nach und durch den externen Besuch (siehe Abschnitt 1.2.).