

Anhang II zum Abschlussbericht

Wissenschaftliche Begleitung des
Modellprojekts Anthroposophische Medizin
der Innungskrankenkasse Hamburg

Zusatzevaluation LWS-Syndrom

Multivariater Vergleich:
Patienten der Basisevaluation vs. LWS-Kontrollgruppe

Testbeschreibung und Ergebnisse

Prof. Dr. Karl Wegscheider

Vorbemerkung

Im Folgenden wird der biometrische Bericht einer von Prof. Dr. Karl Wegscheider durchgeführten Untersuchung abgedruckt: Die multivariaten Vergleiche der Patienten der Basisevaluation, Zusatzevaluation LWS-Syndrom (im Abschlussbericht BE-LWS abgekürzt) mit der schulmedizinisch behandelten LWS-Kontrollgruppe. Diese Vergleiche erfolgten im Kontext weiterer wissenschaftlicher Untersuchungen im Rahmen anderer Teilbereiche des Modellprojekts Naturheilverfahrens der Innungskrankenkasse Hamburg. Aus diesem Grund weicht die Nomenklatur z. T. von der im vorliegenden Abschlussbericht ab (Tabelle 1).

Abschlussbericht	Biometrischer Bericht Prof. Wegscheider
Zusatzevaluation LWS-Syndrom	Studie Chronische Erkrankungen: Diagnosegruppe D5: Erwachsene, LWS-Syndrom
BE-LWS	Anthroposophie
LWS-Kontrollgruppe	Schulmedizin
Symptomscore	D5-B3: Durchschnittliche Beschwerde
Krankheitsscore*	D5-B4: Schwere*

Tabelle 1 Abweichende Nomenklatur im Abschlussbericht und im Biometrischen Bericht Prof. Wegscheider. *Die vergleichende Analyse des Krankheitsscores wurde in den Abschlussbericht nicht aufgenommen, aufgrund abweichender Follow-up-Zeiträume in den zwei Gruppen.

Studie Chronische Krankheiten

Beschwerden, Schweregrad (nach Arzteinschätzung) und Lebensqualität nach Diagnosegruppen:

Graphische Darstellung und inferenzstatistische Analyse der Verläufe

Statistische Methoden

Im folgenden werden

- für jede Altersgruppe und innerhalb der Altersgruppen für jede Diagnosegruppe,
- für jede der ausgewählten allgemeinen bzw. spezifischen Zielgrößen und
- für jeden der beiden Zwei-Gruppenvergleiche Homöopathie bzw. Anthroposophie vs. Schulmedizin

eine Verlaufskurvengraphik präsentiert und ein zugehöriger Testplan absolviert, der es erlaubt, einzuschätzen, welche Aspekte der Verlaufskurvengraphik als systematisch interpretiert werden können und welche möglicherweise zufällig zustande gekommen sind.

Die Reihenfolge der Graphiken entspricht der Reihenfolge in der zugehörigen Übersichtstafel, in der die Testergebnisse zusammengefasst sind. Diagnosegruppen und Variable sind entsprechend mit korrespondierenden Kürzeln belegt. „D“ bedeutet „Diagnosegruppe“, „B“ bezeichnet die für alle Diagnosegruppen identischen Beschwerde- bzw. Schweregradskalen, „S“ die diagnosegruppenspezifischen Beschwerdeskalen, „L“ die Lebensqualitätsskalen des SF36.

Verlaufskurven-Darstellungen:

Die Verläufe innerhalb einer Behandlungsgruppe werden durch boxplot-gebündelte individuelle Polygonzüge dargestellt. Dabei werden die Gruppen farblich unterschieden.

Die Boxen umfassen dabei zeitpunktweise die jeweils mittleren 50% der Patienten. Der Gürtelstrich markiert den Median. Die Barthaare reichen bis zum jeweiligen Minimum bzw. Maximum, es sei denn, die Extremwerte liegen weiter als zwei Boxbreiten vom Median entfernt; in diesem Fall werden die dann sog. Ausreißer einzeln markiert (die Ziffer gibt die Vielfachheit des Ausreißers an).

Die fetten Linien verbinden die arithmetischen Mittel. Die zugehörigen Zahlenwerte und die optisch nicht umgesetzten Standardabweichungen sind jeweils unterhalb der Abszisse zu finden.

Direkt unterhalb der Abszisse sind die Stichprobenumfänge vermerkt. Es sind nur Patienten mit vollständigen Verläufen vermerkt. Die Ausgangswerte von Patienten, die unvollständige Verläufe vorzuweisen haben, sind in Boxplotform abgetrennt auf der linken Seite dargestellt.

Für die Lebensqualitäts-Teilscores des SF-36 sind zwei zusätzliche Graphiken beigelegt, die die Ausgangswerte ins Verhältnis zur amerikanischen Norm-Stichprobe (schwarze Linie) setzen. Die erste Graphik enthält die originalen Skalenwerte. Die p-Werte unter der Abszisse vergleichen die Ausgangswerte zwischen den Behandlungsgruppen für die Gesamtpopulation (inkl. Patienten mit unvollständigen Verläufen). Die zweite Graphik enthält die bzgl. der amerikanischen Norm-Stichprobe standardisierten Werte. Die p-Werte geben hier wieder, inwieweit sich die einzelnen Behandlungsgruppen-Populationen signifikant von der US-Norm-Stichprobe unterscheiden (zweiseitige t-Tests). Die jeweils dargestellten beiden Summen-Scores (KSK und PSK) sind mit US-Gewichten berechnet und in der US-Norm-Stichprobe unkorreliert. In einer deutschen Population bzw. einer Patienten-Population ist aufgrund der Konstruktion der Summenscores nicht unbedingt zu erwarten, dass die Unkorreliertheit erhalten bleibt. Diese Einschränkung ist bei der Interpretation zu berücksichtigen.

Testplan:

Der Testplan enthält sowohl Testelemente zum Gruppenvergleich als auch Elemente zur zeitlichen Veränderung innerhalb der Gruppen. Ferner werden mögliche Störeffekte (Selektion bzw. Selbstselektion, Confounding) im Testplan berücksichtigt und gegebenenfalls mituntersucht bzw., wo möglich, statistisch korrigiert.

Im einzelnen:

Schritt 0: (nur SF-36-Teilscores, für andere Scores mangels Normwerten nicht möglich)

In je einem zweiseitigen t-Test wird untersucht, ob die Werte in den jeweiligen Vergleichsgruppen signifikant niedriger sind als die Werte in der Normalbevölkerung. Liegt keine Signifikanz vor, so besteht Unsicherheit, inwieweit überhaupt eine behandlungswürdige Beeinträchtigung vorliegt. Die zugehörigen p-Werte sind in den Boxplot-Verlaufskurven-Darstellungen der standardisierten SF-36-Teilscores enthalten (s.o.). Alle weiteren maßgeblichen Test-Ergebnisse sind in den Überblickstabellen zu finden.

Schritt 1: Vorgeschalteter univariater Test zur Patientenselektion durch unvollständiges Follow-up

Nicht für alle Patienten liegen vollständige Verläufe vor. Es ist denkbar, dass das Nicht-wieder-Erscheinen beim Arzt mit dem Behandlungsverlauf korreliert. Dieser Effekt könnte verfälschend wirken. Als Sicherheitsmaßnahme wird deshalb für jede Gruppe mit Hilfe eines Zwei-Stichproben-t-Testes untersucht, ob die Patienten ohne vollständigen Verlauf andere (höhere oder tiefere) Ausgangswerte hatten als die Patienten mit vollständigem Verlauf. Dieser Test vergleicht den jeweiligen linken separaten Boxplot mit dem 0M-Boxplot in der Verlaufskurvendarstellung. Falls ein signifikanter ($p < 0.05$) Unterschied besteht, wird der p-Wert in der Tabelle zur Warnung schwarz hinterlegt. (Man beachte, dass dieser Test nicht geeignet ist, *alle* Arten von Selektionsverfälschung zu erkennen.)

Schritt 2 – Schritt 4: Anpassung eines Allgemeinen Linearen Modells (Varianzanalyse mit Messwiederholungen, gegebenenfalls Einschluss von Kovariaten) an die Daten der Patienten mit vollständigen Verläufen.

Es werden zwei Modelle angepasst: das unadjustierte Modell (Faktoren: Zeit, Gruppe, Zeit*Gruppe) und das adjustierte Modell (Faktoren: Zeit, Gruppe, Geschlecht, Abitur ja/nein, Kovariaten: Alter (in Jahren), Geschlecht*Alter, Log(Symptombdauer), sowie die jeweiligen Wechselwirkungen mit der Zeit). Die Symptombdauer wurde bei 999 Tagen winsorisiert und logarithmiert, um den Hebelwirkung (leverage) isolierter Ausreißer zu begrenzen und die Modellvoraussetzungen des Varianzanalysemodells besser zu treffen.

Schritt 2: Test auf (Ausgangs-)Niveau-Unterschiede zwischen den Gruppen

Z.B. durch Selbstselektion kann es vorkommen, dass sich die Patienten der beiden Gruppen in den Ausgangsbedingungen unterscheiden. In diesem Fall wäre ein Vergleich der Verläufe in den beiden Gruppen problematisch, da man zwischen selektionsbedingten und therapiebedingten Effekten nicht unterscheiden könnte. Ungleiche Ausgangsbedingungen lägen insbesondere dann vor, wenn sich die Ausgangswerte systematisch unterscheiden würden. Aus diesem Grund wird in Schritt 2 getestet, ob es Lageunterschiede zwischen den Aufnahmewerten der Patienten mit vollständigen Verläufen gibt. Das verwendete Testverfahren hängt in diesem Fall davon ab, ob der (unter Schritt 3 behandelte) Interaktionstest, der die Parallelität der Verläufe testet, zur Ablehnung der Nullhypothese führt. Wird die Parallelität der Verläufe verneint, so wird in diesem Schritt ein Zwei-Stichproben-t-Test der Ausgangswerte (0 Monate) auf Gruppenunterschiede gerechnet. Kann die Verlaufparallelität nicht abgelehnt werden, so wird der Test auf Niveau-Unterschiede aus der Messwiederholungs-Varianzanalyse verwendet (dieser Test hat die größere Power). In jedem Fall werden signifikante p-Werte zur Warnung schwarz hinterlegt.

Schritt 3: Test auf Parallelität der Verläufe

Bestehen unterschiedliche Erfolgchancen zwischen den Gruppen, so sind die Verlaufskurven nicht parallel. Die Hypothese „Parallele Verlaufskurven“ kann mit dem Interaktions-Test Zeit*Gruppe untersucht werden. Signifikante p-Werte deuten auf Gruppen-Unterschiede in den Behandlungsverläufen hin und werden grau hinterlegt. Man beachte, dass diese Effekte dennoch mit äußerster Zurückhaltung zu interpretieren sind, wenn bei der jeweiligen Zielgröße schwarz hinterlegte p-Werte zu verzeichnen sind.

Schritt 4: Test auf zeitliche Veränderungen

Die bisher durchgeführten Tests lassen offen, inwieweit überhaupt über das jeweilige Kollektiv konsistente Veränderungen unter den Therapien stattgefunden haben. Hierzu kann der Test auf Zeiteffekte (Messwiederholungseffekte) innerhalb des Varianzanalysemodells herangezogen werden. Sind die Verläufe nicht parallel (signifikanter Interaktionstest in Schritt 4), so müssen die Zeiteffekte behandlungsgruppenweise getestet werden.

Adjustierte Analysen:

Da es sich bei der vorliegenden Studie nicht um eine Interventionsstudie handelt und somit nicht von einer randomisierten Zuteilung der Patienten auf die Ärzte ausgegangen werden kann, ist ein unadjustierter Vergleich nicht ausreichend. Die in Schritt 2 bis 5 aufgedeckten Gruppen-Unterschiede könnten sehr wohl andere Ursachen haben, die von der Selbstselektion der Patienten herrühren. Es könnte z.B. sein, dass Männer und Frauen unterschiedlich schnelle Heilungsverläufe aufweisen. Würden dann z.B. die Frauen bevorzugt zum Homöopathen gehen, so würde möglicherweise in den vorstehenden unadjustierten Analysen ein scheinbarer Gruppenunterschied konstatiert werden, der jedoch in Wirklichkeit ein verkappter Geschlechtsunterschied ist. Um solchen Fehlinterpretationen vorzubeugen, wurde in den adjustierten Analysen eine statistische Kontrolle einer Reihe möglicher weiterer Determinanten des Therapieerfolges versucht.

Folgende Abweichungen zwischen adjustierter und unadjustierter Analyse sind denkbar:

- 1.) Die adjustierte Analyse ergibt einen signifikanten Effekt an einer Stelle, an der die unadjustierte Analyse keinen Effekt ergab. In diesem Fall ist davon auszugehen, dass ein tatsächlich vorhandener Gruppenunterschied durch Confounder maskiert wurde. Der adjustierte p-Wert wird grau hinterlegt.
- 2.) Die unadjustierte Analyse ergibt einen signifikanten Effekt an einer Stelle, an der die adjustierte Analyse keinen Effekt ergab. In diesem Fall ist denkbar, dass ein Gruppenunterschied durch Confounder vorgetäuscht wurde. Der adjustierte p-Wert wird in diesem Fall schwarz hinterlegt, da er das unadjustierte Ergebnis in Frage stellt. Zu beachten ist jedoch, dass insbesondere bei kleinen Stichprobenumfängen auch tatsächlich vorhandene Effekte häufiger nach Adjustierung nicht mehr signifikant sind, da die Power der Teilfragestellung durch die Adjustierung zu klein wird.

Von weiter gehendem Interesse dürfte sein, welche Confounder jeweils besonders einflussreich bzgl. der jeweiligen Zielgröße war. Die Aufarbeitung dieser Frage sprengt jedoch den Rahmen des hier Präsentierbaren. Auf den Einfluß spezieller Confounder wird deshalb nur gelegentlich im Diskussionsteil an Stellen eingegangen, an denen sie zum Verständnis des Behandlungsgruppenvergleiches beitragen.

Zusammenfassend ist festzuhalten: nur grau hinterlegte Signifikanzen bei Zielgrößen, bei denen keiner der p-Werte schwarz hinterlegt wurde, können mit gewisser Berechtigung als mögliche systematische Gruppenunterschiede diskutiert werden.

Studie Chronische Erkrankungen

Diagnosegruppe D5: Erwachsene, LWS-Syndrom

Vergleich: Anthroposophie vs. Schulmedizin

Wegscheider Biometrie und Statistik GmbH

10. März 2002

Diagnosegruppe D5: Erwachsene, LWS-Syndrom

Vergleich: Anthroposophie vs. Schulmedizin

<u>Beschwerden</u>		Vorgeschalteter univariater Test	Varianzanalyse mit Messwiederholungen 0/6/12 Monate			
		Systematische Patientenselektion durch unvollständiges Follow-Up?	Test der Nullhypothese:			
Zielgrößen	Analyse		p	Differenz A-S	p	Verlaufsparellität zwischen den Gruppen
					p	p
D5-B1: Beschwerde 1	unadj.	A: 0.868 S: 0.122	-0.678	0.179	0.627	<0.001
	adj.		-1.324	0.063	0.502	0.054
D5-B2: Maximale Beschwerde	unadj.	A: 0.583 S: 0.156	-0.421	0.410	0.432	<0.001
	adj.		-1.005	0.166	0.939	0.093
D5-B3: Durchschnittliche Beschwerde	unadj.	A: 0.660 S: 0.134	-1.049	0.024	0.100	<0.001
	adj.		-1.972	0.003	0.425	0.030
D5-S1: FFbH-R	unadj.	A: 0.759 S: 0.821	7.797	0.126	0.055	<0.001
	adj.		13.700	0.055	0.079	0.151
D5-S2: LBPRS	unadj.	A: 0.771 S: 0.435	-13.546	0.004	0.499	<0.001
	adj.		-22.946	0.001	0.080	0.751

Diagnosegruppe D5: Erwachsene, LWS-Syndrom
Vergleich: Anthroposophie vs. Schulmedizin

<u>Schweregrad Arzt</u>		Vorgeschalteter univariater Test	Varianzanalyse mit Messwiederholungen Anfang/Ende			
		Systematische Patientenselektion durch unvollständiges Follow-Up?	Test der Nullhypothese:			
			Gleiche Ausgangsbedingungen	Verlaufsparellität zwischen den Gruppen	Keine zeitlichen Veränderungen innerhalb der Gruppen	
Zielgrößen	Analyse	p	Differenz A-S	p	p	p
D5-B4: Schwere	unadj.	A: 0.530 S: 0.846	0.418	0.245	0.118	<0.001
	adj.		0.306	0.515	0.415	0.017

Diagnosegruppe D5: Erwachsene, LWS-Syndrom

Vergleich: Anthroposophie vs. Schulmedizin

SF-36 Körper		Vorgeschalteter univariater Test	Varianzanalyse mit Messwiederholungen 0/6/12 Monate			
		Systematische Patientenselektion durch unvollständiges Follow-Up?	Test der Nullhypothese:			
Zielgrößen	Analyse	p	Differenz A-S	p	Verlaufsparallelität zwischen den Gruppen	Keine zeitlichen Veränderungen innerhalb der Gruppen
					p	p
D5-L1: Körperliche Funktionsfähigkeit	unadj.	A: 0.664 S: 0.679	0.520	0.039	0.415	0.029
	adj.		0.974	0.004	0.579	0.025
D5-L2: Körperliche Rollenfunktion	unadj.	A: 0.586 S: 0.568	-0.270	0.241	0.036	A: 0.003 S: 0.912
	adj.		0.546	0.141	0.068	0.014
D5-L3: Körperliche Schmerzen	unadj.	A: 0.073 S: 0.511	0.348	0.058	0.232	<0.001
	adj.		0.898	<0.001	0.099	0.003
D5-L4: Allgemeine Gesundheitswahrnehmung	unadj.	A: 0.461 S: 0.112	0.250	0.172	0.019	A: 0.043 S: 0.310
	adj.		0.290	0.260	0.006	A: 0.280 S: 0.576
D5-L5: KSK	unadj.	A: 0.817 S: 0.665	5.771	0.009	0.162	<0.001
	adj.		10.172	0.001	0.192	0.004

Diagnosegruppe D5: Erwachsene, LWS-Syndrom

Vergleich: Anthroposophie vs. Schulmedizin

SF-36 Psyche		Vorgeschalteter univariater Test	Varianzanalyse mit Messwiederholungen 0/6/12 Monate			
		Systematische Patientenselektion durch unvollständiges Follow-Up?	Test der Nullhypothese:			
			Gleiche Ausgangsbedingungen	Verlaufsparallelität zwischen den Gruppen		Keine zeitlichen Veränderungen innerhalb der Gruppen
Zielgrößen	Analyse	p	Differenz H-S	p	p	p
D5-L6: Vitalität	unadj.	A: 0.914 S: 0.199	-0.061	0.711	0.080	0.004
	adj.		-0.188	0.407	0.005	A: 0.032 S: 0.102
D5-L7: Soziale Funktions- fähigkeit	unadj.	A: 0.104 S: 0.455	-0.184	0.555	0.185	0.473
	adj.		0.437	0.208	0.512	0.109
D5-L8: Emotionale Rollenfunktion	unadj.	A: 0.220 S: 0.195	-0.261	0.356	0.056	0.670
	adj.		-0.031	0.941	0.076	0.977
D5-L9: Psychisches Wohlbefinden	unadj.	A: 0.354 S: 0.212	-0.484	0.036	0.012	A: 0.015 S: 0.400
	adj.		-0.206	0.519	0.045	A: 0.128 S: 0.069
D5-L10: PSK	unadj.	A: 0.155 S: 0.193	-4.492	0.084	0.061	0.961
	adj.		-0.417	0.905	0.093	0.729

Diskussion (Testergebnisse und zugehörige Grafiken)

1.) Durch fehlende Follow-ups bedingte Selektionseffekte: Die Verlaufgrafiken suggerieren, dass in beiden Gruppen bevorzugt solche Patienten nicht zur Nachuntersuchung erschienen sind, die im SF-36-Summscore-Psyche und drei der psychischen Teilscores niedrige Werte aufwiesen. Ein solcher Effekt wäre plausibel. Die Selektionstests zeigen jedoch, dass die beobachteten Tendenzen auch sehr gut zufällig sein können. Der nachfolgende Gruppenvergleich wird somit durch psychisch bedingte Selektionseffekte nicht in Frage gestellt.

2.) Selbstselektion der Patienten (unterschiedliche Ausgangsbedingungen): Die Anthroposophie LWS-Patienten unterscheiden sich von den Schulmedizin-LWS-Patienten in folgenden Belangen signifikant:

- weniger durchschnittliche Beschwerden (mit ähnlicher nicht-signifikanter Tendenz bei Beschwerde1/maximalen Beschwerden),
- niedrigere LBPRS-Werte (mit nicht-signifikanter Tendenz zu höheren FFbHR-Werten),
- höhere SF36-Werte im körperlichen Summscore und körperlicher Funktionsfähigkeit/körperlichen Schmerzen,
- niedrigere SF36-Werte im psychischen Wohlbefinden.

Nur die Unterschiede im psychischen Wohlbefinden lassen sich ausreichend auf die ins Modell aufgenommenen Confounder-Variablen (hier insbesondere: die Symptombdauer: anthroposophische Patienten haben längere Symptombdauern, die mit geringerem psychischem Wohlbefinden verschwistert sind) zurückführen. In allen anderen Fällen sind die Bedingungsfaktoren offen.

Keine signifikanten Differenzen gibt es beim Schweregrad bei Arzteinschätzung.

In der Summe sind damit die Anthroposophie-Patienten bei den Beschwerden sowie bei den spezifischen und unspezifischen körperbezogenen Skalen signifikant weniger beeinträchtigt als die Schulmedizin-Patienten, während sie in Bezug auf Schwere und Psyche in etwa vergleichbar sind. Die Vergleichbarkeit der LWS-Kollektive steht damit generell in Frage.

3.) Beeinträchtigung der Lebensqualität: Im Vergleich zur US-Normbevölkerung weisen die LWS-Patienten in allen SF-36-Teilskalen signifikant erniedrigte Werte auf. Ausnahme: der psychische Summenscore bei den schulmedizinischen Patienten.

4.) Differentielle Verläufe:

Signifikant unterschiedliche Verläufe zwischen den Behandlungsgruppen finden sich ausschließlich in einzelnen Lebensqualitäts-Teilskalen des SF-36.

Bei der körperlichen Rollenfunktion und der allgemeinen Gesundheitswahrnehmung sind bei den Anthroposophie-Patienten bei vergleichbaren Ausgangsbedingungen signifikante Besserungen im Verlauf zu beobachten, bei den schulmedizinischen Patienten hingegen nicht. Auch nach Adjustierung ergibt sich kein grundsätzlich anderes Bild, auch wenn einzelne p-Werte über die Signifikanzgrenze rutschen.

Im psychischen Wohlbefinden wird der initiale Nachteil der Anthroposophischen Patienten im Verlauf ausgeglichen. Nach Adjustierung für die Symptombdauer nivellieren sich indes die Gruppen-Unterschiede, Veränderungen mit der Zeit können nicht mehr statistisch gesichert werden.

Unterschiede zwischen den Gruppen finden sich auch bei der Vitalität: nach Adjustierung (hier insbesondere für Geschlecht und Alter) lässt sich ein positiver Trend bei den Anthroposophie-Patienten, nicht aber bei den Schulmedizin-Patienten statistisch sichern.

5.) Veränderungen im Therapieverlauf:

Bei allen Patienten bessern sich die berichteten Beschwerden, der Schweregrad nach Arzteinschätzung sowie die körperbezogene Lebensqualität mit der Zeit.

Die Veränderungen erfolgen fast ausschließlich in den ersten sechs Monaten.

Nach Adjustierung für potentielle Confounder lassen sich einige die Veränderungen mit der Zeit allerdings nicht mehr statistisch sichern.

Zusammenfassung:

Die Beschwerden der LWS-Patienten sowie die körperbezogene Lebensqualität bessern sich sowohl in der Anthroposophie-Gruppe wie in der initial stärker beeinträchtigten Schulmedizin-Gruppe.

Die anthroposophischen Patienten zeigen bei initialer Vergleichbarkeit stärkere Besserungen in der Lebensqualität bezogen auf die körperliche Rollenfunktion und die allgemeine Gesundheitswahrnehmung.

Durch längere Symptombdauer bedingte Reduktionen des psychischen Wohlbefindens sind bevorzugt in der Anthroposophie-Gruppe zu finden und nivellieren sich im Verlauf.