Was heißt signifikant – geheilt oder nur etwas besser? Ein Vergleich verschiedener Berechnungsmethoden zur Bewertung von Veränderungen der depressiven Symptomatik bei ambulanten Rehabilitanden

Birgit Senft · Daniela Fischer-Hansal · Alexandra Schosser

Eingegangen: 29. November 2019 / Angenommen: 13. Februar 2020 / Online publiziert: 11. März 2020
 © Der/die Autor(en) 2020

Zusammenfassung

Hintergrund Zur Bewertung des Behandlungserfolgs bei depressiven Symptomen stehen unterschiedliche statistische Methoden zur Verfügung. Effektgrößen, prozentuale Verbesserung, signifikante Verbesserung und klinisch signifikante Verbesserung sind gängige Methoden zur Bewertung von Behandlungsergebnissen basierend auf „Patient Reported Outcomes“.

Fragestellung Wie unterscheiden sich die Behandlungsergebnisse in Abhängigkeit von der verwendeten Methode?

Methodik Für $N=3018$ Patienten einer ambulanten psychiatrischen Rehabilitationsklinik wurden Verbesserungen anhand von Effektgrößen, prozentuellen Veränderungen, signifikanten und klinisch signifikanten Veränderungen im Beck Depressions-Inventar (BDI-II) berechnet.

Ergebnisse Für die Gesamtgruppe zeigt sich ein gut mittlerer Effekt von $d=0,62$, der nach Ausschluss von zum Zeitpunkt der Aufnahme nicht depressiven (euthymen) Patienten als groß bewertet werden kann ($d=0,79$). Die stärkste Aussagekraft für die Symptomreduktion der Depression hat die Methode der klinischen Signifikanz, bei einem Cut-off von 13 Punkten wurden 23 % als klinisch signifikant und weitere 19,3 % als signifikant gebessert klassifiziert. Bei der Methode der prozentuellen Verbesserung erreichten 30,4 % der Patienten einen Wert von mindestens 50 %.

Schlussfolgerungen Effektgrößen geben einen groben Überblick über den Behandlungserfolg eines Patientenkollektivs ohne Berücksichtigung des Endzustands und der Variabilität innerhalb des Kollektivs. Die prozentuelle Veränderung scheint eher für die Kontrolle des Ansprechens auf eine Behandlung geeignet zu sein. Ergebnisse zur klinischen Signifikanz geben am transparentesten Auskunft über signifikante Verbesserungen und den Wechsel vom dysfunktionalen in den funktionalen Bereich.

Schlüsselwörter Depression · Veränderungsmessung · Klinische Signifikanz · Patient reported outcomes · Rehabilitation

What is the meaning of significance – healed or only slightly improved? A comparison of different evaluation methods to assess depressive symptoms in ambulant rehabilitants

Summary

Background To assess treatment success of depressive symptoms, different statistical methods are available. Effect sizes, percentage improvement, significant improvement, and clinical significant improvement are established methods to assess treatment effects based on “patient reported outcomes”.

B. Senft · D. Fischer-Hansal ·
 Prim. Priv.-Doz. Dr. A. Schosser, PhD MBA (✉)
 BBRZ MED Zentren für seelische Gesundheit Wien,
 Schererstraße 30, 1210 Wien, Österreich
alexandra.schosser@meduniwien.ac.at

B. Senft
birgit.senft@bbrz.at

D. Fischer-Hansal · Prim. Priv.-Doz. Dr. A. Schosser, PhD
 MBA
 Universitätsklinik für Psychiatrie und Psychotherapie,
 Medizinische Universität Wien, Währinger Gürtel
 18–20, 1090 Wien, Österreich

Prim. Priv.-Doz. Dr. A. Schosser, PhD MBA
 Arbeitsgemeinschaft für Verhaltensmodifikation,
 Paris-Lodron-Straße 32, 5020 Salzburg, Österreich

B. Senft
 School of Social Sciences, Department of Psychology,
 University of Nicosia, Nicosia, Zypern

Aim of the study How do treatment effects differ according to the used method?

Methods For $N=3018$ patients treated in an ambulant psychiatric rehabilitation clinic, improvements were calculated based on effect sizes, percentage improvements, significant and clinical significant changes in Beck Depression-Inventory (BDI-II).

Results For the total sample, a sound medium effect of $d=0.62$ was found, after exclusion of patients that were not depressed (euthymic) at time of admission, a large effect ($d=0.79$) was calculated. The strongest informative value for symptom reduction of depression was found for the method of clinical significance, applying a cutoff score of 13 a total of 23% were classified as clinical/clinically significant and additional 19.3% as significantly improved. The method of percentage improvement resulted in 30.4% of patients that achieved a value of at least 50%.

Conclusions Effect sizes give a rough overview over treatment success of a collective of patients, without consideration of the final state and the variability within the collective. The percentage improvement seems to be better suitable for examination of treatment response. With regard to clinical significance, results provide the most transparent information about significant improvements and change from dysfunctional to functional values.

Keywords Depression · Measurement of change · Clinical significance · Patient reported outcomes · Rehabilitation

Hintergrund

Ambulante medizinisch-psychiatrische Phase-3 Rehabilitation (laut WHO schließt die Phase 2 entweder an die Phase 1, d.h. die Akutversorgung, im Sinne eines „Anschlussheilverfahrens“ an den Aufenthalt im Akutkrankenhaus oder an eine akute Krankenbehandlung im extramuralen Bereich an) wird in Österreich ambulant und stationär in einem sechswöchigen Setting angeboten. In den Zentren für seelische Gesundheit BBRZ-Med Wien werden im Rahmen eines verhaltenstherapeutischen Schwerpunkts in etwa 25 Therapieeinheiten pro Woche (142 Therapieeinheiten in 6 Wochen) in störungsspezifischen und störungsübergreifenden Einzel- und Gruppensettings angeboten. Ziele der Rehabilitation sind bestmögliche Wiederherstellung der Gesundheit im Sinne des bio-psycho-sozialen Modells sowie eine bestmögliche Re-Integration in das berufliche und soziale Umfeld. Das beinhaltet eine Reduktion der Symptomatik, aber auch ein erweitertes Verhaltensrepertoire und eine verbesserte Teilhabe im Sinne der International Classification of Functioning ICF [1].

Methoden zur Messung des Behandlungserfolgs

Nachfolgende methodische Ansätze haben gemeinsam, dass die Datenbasis für Berechnung und Bewertung auf Selbstauskünften von Rehabilitanden bei Beginn und Ende der Rehabilitation beruhen. Es handelt sich um sogenannte „Patient Reported Outcomes“ (PROs), also Fragebögen, die generisch oder indikationsübergreifend gestaltet sein können, und deren Gütekriterien in zahlreichen Validierungsstudien geprüft wurden. Wichtigste Informationsquelle sind somit die Rehabilitanden mit der Selbsteinschätzung ihrer Symptomatik oder ihres Wohlbefindens zu zwei oder mehr Beobachtungszeitpunkten. Das wird als indirekte Veränderungsmessung bezeichnet, da von der Differenz zweier Messwerte auf das Ausmaß der Veränderung geschlossen wird. Eine Aufstellung von Instrumenten für die unterschiedlichen Indikationen der Rehabilitation findet sich in einem Beitrag zu Diagnostik und Assessments in der Rehabilitation [2]. In verschiedenen medizinischen Disziplinen werden PROs als transparente Outcomeparameter für Behandlungserfolg, gemessen am gesundheitlichen Status der Patienten, sowohl für Behandler wie auch für Entscheidungsträger gesehen [3].

Eine Standardmethode zur Bewertung von Behandlungseffekten (im Optimalfall mit Kontrollgruppe) sind statistische Tests mit Messwiederholungsfaktor. Die Interpretation des p -Werts allein gibt allerdings noch nicht Aufschluss über die Bedeutsamkeit oder Größe des Effekts, da bei hinreichend großer Stichprobe schon kleine Effekte signifikante Ergebnisse liefern können. Eine Minimalanforderung für die Evaluation ist die Angabe von Effektgrößen, auch wenn diese noch keinen Aufschluss über die klinische Bedeutsamkeit der Ergebnisse liefern [4].

Nachfolgend wird auf praktikable und gängige Methoden zur Bewertung von Veränderungen näher eingegangen.

Berechnung von Effektgrößen

Eine breit eingesetzte Methode zur Bewertung von Behandlungseffekten sind Effektgrößen nach Cohen (1988). Dazu wird im Zuge von Gruppenauswertungen die Mittelwertsdifferenz von zwei Messzeitpunkten gebildet und diese durch die gepoolte Varianz geteilt (Gl. 1). Dadurch erhält man einen standardisierten Wert für das Ausmaß der Veränderung, der unabhängig vom jeweiligen Messverfahren interpretiert und mit anderen Effektgrößen verglichen werden kann. Zur Bewertung der Größe der Differenz empfiehlt Cohen folgende Interpretation: $d=0,2$ =kleiner Effekt, $d=0,50$ =mittlerer Effekt und $d=0,80$ =großer Effekt [5].

- Formel 1, Berechnung für Cohen's d (Cohen, 1988, S. 20):

$$d = \frac{|m_A - m_B|}{\sigma} \quad (1)$$

In der Literatur werden verschiedene Vorgehensweisen zur Berechnung der Effektgrößen beschrieben, da sich die Wahl des Maßes für die Streuung (σ) auf das Ergebnis auswirken kann. In der Praxis wird häufig die gepoolte Varianz zur Berechnung von Cohen's d angewendet, insbesondere dann, wenn starke Boden- oder Deckeneffekte zu einer systematischen Einschränkung der prä-Streuung führen. Effektgrößen sind einerseits abhängig von der Stichprobengröße, weil mit großen Stichproben die Streuung kleiner wird. Andererseits ist die prä- und post-Streuung abhängig von der Ausprägung des Merkmals. Wenn z.B. viele Mitglieder einer Stichprobe bei Behandlungsbeginn einen (ähnlich) hohen Wert aufweisen, kann von einem Deckeneffekt ausgegangen werden, was aufgrund der geringen Streuung zu einem größeren Effekt führt [6]. Bei der post-Erhebung kann es zu einer größeren Streuung kommen, weil sich nicht alle Mitglieder einer Gruppe gleichmäßig verändern. Bei der Angabe von Effektgrößen sollte immer das Konfidenzintervall (KI) berichtet werden. Nur wenn die 0 im Intervall nicht enthalten ist, kann von einem gegen den Zufall abgesicherten Ergebnis ausgegangen werden.

Steffanowski et al. (2007) haben im Rahmen einer Meta-Analyse die Effekte der stationären psychosomatischen Rehabilitation in Deutschland untersucht. In 56 Studien mit unterschiedlichen Ergebnismaßen wurde zwischen Aufnahme und Entlassung ein mittlerer Effekt von $d = 0,51$ [KI: $0,49 < d < 0,53$] errechnet. Zum Katamnesezeitpunkt zeigte sich zwar ein etwas geringerer Effekt von $d = 0,41$ [$0,39 < d < 0,43$], aber von einer nachhaltigen Wirkung kann ausgegangen werden [6].

Diese Ergebnisse dienen durchaus als Vergleichswert für österreichische Studien in der psychiatrischen Rehabilitation, auch wenn die Klientel in Österreich im Schnitt etwas älter und deutlich seltener berufstätig ist. Für die Symptomverringerung wurde eine mittlere Effektgröße von $0,53$ [KI: $0,45 - 0,60$] ermittelt [7].

Berechnung von statistisch signifikanter und klinisch signifikanter intraindividueller Veränderung

Bei dieser Vorgehensweise liegt der Fokus auf dem Einzelfall [8]. In der Praxis kommt die kritische Differenz zwischen zwei Testwerten, welche auf dem Ansatz des Reliable Change (RC) beruht, häufig zur Anwendung. Pro Fragebogen wird ein Wert errechnet, um den sich zwei Messungen einer Person mindestens unterscheiden müssen, um eine statistisch signifikante Veränderung zu postulieren. In Gl. 2 [9] wurde der z-Wert für eine einseitige Fragestellung eingefügt,

da nach Behandlungen von Veränderungen in die positive Richtung ausgegangen wird. Valide und reliable psychometrische Testverfahren geben in ihren Manu-
alen die kritische Differenz für Veränderungen an, was Anwendern eine komfortable Beurteilung intraindividueller Veränderungen ermöglicht.

- Formel 2, Berechnung der kritischen Differenz zwischen zwei Testpunktwerten (Lienert & Raatz, 1998, S. 370):

$$d_{\text{krit.5\%}} = z \cdot s_{e(i-k)} = 1,65 \cdot s_t \cdot \sqrt{2 \cdot (1 - r_{tt})} \quad (2)$$

Eine weitere Möglichkeit stellen diagnostische Interviews dar. Wenn die Kriterien einer Diagnose nicht mehr erfüllt sind, ist keine Trennung von der „Normalbevölkerung“ mehr gegeben. Stärker statistisch orientierte Möglichkeiten sind die die Nutzung von Fragebögen mit Normwerten und einem Cut-off-Wert für Recovery [4].

Jacobson und Truax (1991) haben mit dem Reliable Change Index (RC) einen Ansatz vorgestellt, der neben der Verringerung in einem Mindest-Ausmaß zusätzlich den Wechsel von einer dysfunktionalen in eine funktionale Kategorie beinhaltet. Dahinter steht die Annahme, dass Patienten als Teil einer dysfunktionalen Population in Therapie kommen und nach Abschluss der Behandlung nicht mehr länger dieser Gruppe angehören sollten. Die Berechnung des RC erfolgt aus der Differenz der Messwerte einer Person prä und post, die ins Verhältnis zur eigenen Standardabweichung gesetzt wird, welche die Reliabilität des Messverfahrens beinhaltet (Gl. 2). Der standardisierte Wert für das Mindestmaß einer Veränderung beträgt $1,96$. Bei einem RC-Wert $> 1,96$ ist es unwahrscheinlich, dass keine echte Veränderung vorliegt, bzw. diese auf Schwankungen eines unreliablen Fragebogens zurückzuführen ist [8].

- Formel 3, Berechnung den Reliable Change Index (RC) nach Jacobson & Truax (1991, S. 14):

$$RC = \frac{X_2 - X_1}{S_{\text{diff}}} \quad (3)$$

Der Ansatz der klinischen Signifikanz ist also strenger, weil die Veränderung nicht nur gegen den Zufall abgesichert ist, sondern zusätzlich der Wechsel vom dysfunktionalen in den funktionalen Bereich gefordert wird. Dadurch kann von „Recovery“, „Remission“ oder „Heilung“ gesprochen werden, üblich ist auch die Formulierung „klinisch relevant gebesserten Patienten“ oder „klinisch bedeutsame Veränderung“ [10]. Personen die zu Beginn der Behandlung nicht im auffälligen Bereich lagen, werden gesondert ausgewertet, da sie keine klinische Signifikanz erzielen können.

Nichtsdestotrotz können (klinisch) signifikante Veränderungen keine Garantie dafür sein, dass Patien-

ten oder Angehörige die Veränderung als bedeutsam wahrnehmen. Bei Patienten mit klinisch signifikanter Verbesserung zeigte sich aber eine höhere Zufriedenheit mit der Therapie als bei Patienten ohne klinisch signifikanter Veränderung, das wurde von den Autoren als Hinweis auf die Validität der Berechnungsmethode gesehen [4].

Berechnung prozentueller Verbesserung für intraindividuelle Veränderungen

Eine weitere Möglichkeit ist die Methode der prozentuellen Verbesserung. Im Beispiel einer Intervention gegen Kopfschmerz wurden aus den Daten eines Kopfschmerztagebuchs die Differenz von zwei Beobachtungszeiträumen in das Verhältnis zum Ausgangswert gesetzt (Gl. 4) definiert [4].

- Formel 4, Beispiel für die Berechnung einer prozentuellen Verbesserung (Ogles, Lunnen & Bonesteel, 2001, S. 431):

$$\begin{aligned} & \text{Percent Improvement} \\ & \frac{\text{Headache Index Pretreatment} - \text{Headache Index Posttreatment}}{\text{Headache Index Pretreatment}} \end{aligned} \quad (4)$$

In den S3-Leitlinien für Unipolare Depression [11] wird als Hauptkriterium für die Wirksamkeit einer Behandlung auch der Grad der Symptomreduktion bzw. eine Remission gesehen. Als Response oder Ansprechen auf die Therapie wird die Reduktion in einer Fragebogenskala um 50 % des Ausgangswerts gesehen. Unter 20 % liegt kein Effekt vor, 20–50 % weisen auf eine minimale oder geringe Wirkung hin und ab 50 % wird von einer Teilremission gesprochen. Eine vollständige Remission liegt bei 100 % Symptomreduktion oder bei Unterschreiten eines Cut-off-Werts vor. Für den BDI-II wird als Cut-off ein Wert von 13 Punkten angegeben, Ausprägungen darunter gelten als klinisch unauffällig oder remittiert [11].

Zusammenfassung zu den vier vorgestellten Methoden zur Veränderungsmessung

Effektgrößenberechnung

Basis für die Berechnung sind Mittelwerte von Gruppen vor und nach einer Intervention, aus der Größe der Differenzwerte und der statistischen Signifikanz wird auf den Erfolg einer Behandlung oder die Effizienz einer Einrichtung geschlossen. Vorteile sind die einfache Berechnung und Interpretierbarkeit, Nachteile können in der Abhängigkeit von der Stichprobengröße und der Streuung des Merkmals in der untersuchten Gruppe gesehen werden.

Signifikante Veränderung auf intraindividueller Ebene

Im Fokus steht die Frage, ob eine Veränderung auf intraindividueller Ebene bedeutsam genug (groß genug) ist, um über den Zufall erhaben zu sein. Fra-

gebögen mit einer hohen Reliabilität (Genauigkeit/Zuverlässigkeit) sind hier im Vorteil, weil durch die Berücksichtigung der Reliabilität in der Formel kleinere Differenzen für eine signifikante Veränderung ausreichen. Gerade im klinischen Bereich stehen viele Verfahren mit zufriedenstellenden Gütekriterien zur Verfügung. Im Unterschied zur Effektgröße wird das Ergebnis in einem leicht verständlichen Prozentsatz gebesserter Patienten ausgedrückt.

Klinisch signifikante Veränderung auf intraindividueller Ebene

Zum Kriterium der signifikanten Veränderung kommt bei diesem Ansatz die Frage hinzu, ob ein Patient am Ende einer Intervention der Normalpopulation zuzuordnen ist und damit als „gesund“ klassifiziert werden kann. Klinische Signifikanz ist also zu erzielen, wenn Patienten bei Beginn der Behandlung im auffälligen Bereich lagen und am Ende der Behandlung einen definierten Cut-off unterschreiten.

Prozentuelle Verbesserung auf intraindividueller Ebene

Bei dieser Berechnungsmethode werden die Ausprägungen nach einer Intervention in das Verhältnis zur Ausprägung vor einer Intervention gesetzt. Der reine Prozentwert sagt aber noch nichts über den Endzustand des Patienten aus.

Studiendesign und Untersuchungsmethoden

Die Daten stammen aus der prospektiven Evaluationsstudie der Zentren für seelische Gesundheit BBRZ-Med Wien-Leopoldau. Einbezogen wurden alle regulär beendeten Heilverfahren mit Fragebögen zu beiden Messzeitpunkten. Die Diagnosen wurden klinisch vom ärztlichen Personal gestellt. Das Standardassessment setzt sich aus Angaben zur Person und zu krankheitsbezogenen Daten sowie den standardisierten Fragebögen World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) [12], Brief Symptom Inventory 18 (BSI-18) [13], Fragebogen für die Selbstbeurteilung von Aktivitäten und Teilhabe bei psychischen Störungen (ICF 3F AT) [14] und Beck Depressions-Inventar (BDI-II) [15] zusammen. Ziel der Studie ist die Bewertung des Behandlungserfolgs anhand der unterschiedlichen Berechnungsmethoden.

Beschreibung der Stichprobe

Von $N=3018$ Rehabilitanden aus dem Zeitraum Jänner 2014 bis April 2019 der ambulanten Rehabilitation lagen zwei Messzeitpunkte im BDI-II vor. Soziodemographische und krankheitsbezogene Angaben finden sich in Tab. 1. Im Durchschnitt waren die Rehabilitanden 44,1 Jahre alt und zu knapp 2/3 weiblich. Gut 2/3 wiesen als Hauptdiagnose eine affektive Störung auf, 59 % wiesen aber mindestens eine weitere psychiatrische Diagnose auf. Von einer affektiven Störung waren

Tab. 1 Stichprobencharakteristika

Variable	Ausprägung	M	SD
Alter	Alter in Jahren	44,08	9,87
Variable	Ausprägung	N	%
Geschlecht	Männlich	1094	36,3
	Weiblich	1924	63,7
Hauptdiagnose nach ICD-10	F3 Affektive Störungen	2030	67,3
	F4 Neurol.-/Belast.-/somat. Störungen	670	22,2
	F6 Persönlichk.-/Verhaltensstörungen	182	6,0
	Sonstige	136	4,5
Affektive Störungen unter Einbeziehung aller psychiatrischen Diagnosen	Nicht-affektive Störung	684	22,7
	Nicht-chronische Depression	752	24,9
	Chronische Depression	1442	47,8
	Nicht-depressive affektive Störung	140	4,6
Variable	Ausprägung	M	SD
Depressive Symptomatik bei Aufnahme (BDI-II Score)	Nicht-affektive Störung	21,26	10,94
	Nicht-chronische Depression	23,81	10,65
	Chronische Depression	25,89	10,77
	Nicht-depressive affektive Störung	23,68	11,97
	Total	24,22	10,99
Variable	Ausprägung	N	%
Beruflicher Status	Berufstätig	886	29,4
	Arbeitslos	1774	58,8
	Reha-Geld	245	8,1
	Sonstige	35	1,2
	Keine Information	78	2,5

77,3 % aller Rehabilitanden betroffen, 4,6 % davon litten aber unter einer nicht-depressiven affektiven Störung. Zusammengefasst waren 47,8 % der Rehabilitanden von einer chronischen (rezidivierende depressive Störungen und Dysthymie) und 24,9 % von einer nicht-chronischen Depression betroffen. Die depressive Symptomatik war bei Rehabilitanden mit einer chronischen depressiven Störung signifikant höher als bei allen anderen Gruppen. Rehabilitanden mit einer nicht-affektiven Störung wiesen eine signifikant geringere depressive Symptomatik auf als die anderen Gruppen, allerdings lag auch diese Gruppe psychometrisch im Mittel im Bereich einer mittelschweren Depression.

Weniger als 1/3 der Rehabilitanden waren zum Zeitpunkt des Reha-Beginns berufstätig, 59 % waren arbeitslos, Notstandshilfebezieher oder befanden sich nach Kündigung im Krankenstand. Der Anteil der Reha-Geld-Bezieher war mit 8 % eher gering und 4 % fallen in sonstige Kategorien.

Beck Depressions-Inventar (BDI-II)

Der Fokus der Studie liegt auf der Veränderung depressiver Symptome zwischen Beginn und Ende der Rehabilitation im BDI-II. Die Originalversion des Fra-

gebogens BDI stammt aus dem Jahr 1994 von Beck und Mitarbeitern und liegt als BDI-II reliabel und valide in deutscher Version vor (Hautzinger, Keller F., & Kühner, 2006). Damit kann der Schweregrad einer Depression ab einem Alter von 13 Jahren beurteilt werden. Es enthält 21 Aussagen zu Depressionssymptomen, die den neun DSM-IV-Kriterien zugeordnet wurden. Als Maß für den Schweregrad einer Depression wird der Summenwert der Items herangezogen: 0–8 Punkte keine Depression, 9–13 Punkte minimale Depression, 14–19 Punkte leichte Depression, 20–28 Punkte mittelschwere Depression und 29–62 Punkte schwere Depression [15].

Für die deutsche Version wurde für eine bedeutsame und reliable intraindividuelle Veränderung als Richtlinie eine kritische Differenz von 8 Punkten angegeben. Zur Definition einer klinisch signifikanten Veränderung wurde als Wechsel vom dysfunktionalen in den funktionalen Bereich ein Cut-off-Wert von 19 Punkten angegeben. Dieser Wert entspricht der oberen Grenze einer „leichten Depression“.

In der S3-Leitlinie [11] wird der Cut-off-Wert für den BDI-II mit 13 Punkten festgelegt. Unter 13 Punkten gelten Patienten als klinisch unauffällig oder remittiert bzw. liegt keine Depression vor. Die übrigen Schwellenwerte stimmen mit jenem aus dem Manual des BDI-II überein.

In der vorliegenden Arbeit werden für die Bewertung von Symptomveränderungen folgende Begrifflichkeiten verwendet:

signifikant (statistisch)	auf Gruppen- oder Einrichtungs-Ebene als Ergebnis eines Signifikanztests mit Angabe von <i>p</i> -Werten
signifikant	auf Einzelfall-Ebene: Reduktion der Symptomatik mindestens im Ausmaß der kritischen Differenz
klinisch signifikant/ klinisch bedeutsam	auf Einzelfall-Ebene: Reduktion der Symptomatik mindestens im Ausmaß der kritischen Differenz und zusätzlich Wechsel vom dysfunktionalen in den funktionalen (gesunden) Bereich

Ergebnisse zur Symptomreduktion anhand verschiedener Bewertungsmethoden

Gut 1/3 der Rehabilitanden wies zu Beginn der Behandlung einen Wert im Bereich einer schweren Depression auf, dieser Anteil hat sich bis zum Ende der Rehabilitation halbiert. Umgekehrt wiesen nur 18 % bei Beginn der Rehabilitation eine Ausprägung im Bereich keiner oder einer minimalen Depression auf, dieser Anteil hat sich am Ende der Behandlung mit 44 % mehr als verdoppelt. Der Anteil von Rehabilitanden mit einer mittelschweren Depression sank von 29,6 auf 19,6 %, der Anteil von Rehabilitanden mit einer leichten Depression ist in etwa gleichgeblieben (siehe Abb. 1).

Abb. 1 Ausprägung der depressiven Symptomatik im BDI-II bei Beginn und Ende der Rehabilitation ($N = 3018$)

Bewertung der Symptomreduktion im BDI-II anhand der Effektgrößen für Cohen's d

Für die Gesamtgruppe der Rehabilitanden zeigte sich für die Veränderung zwischen Beginn und Ende der Rehabilitation eine Effektgröße von $d = 0,62$. Das entspricht nach Cohen (1988) einem guten mittleren Effekt [KI: 0,57–0,67]. Die Effektgröße im BDI-II wurde in Abhängigkeit von der Ausgangsbelastung für vier Subgruppen separat berechnet. Die beiden bei Aufnahme kleinsten Kategorien (0–8 Punkte keine Depression und 9–13 Punkte minimale Depression) wurden zusammengefasst und als „unauffällig“ bezeichnet.

In Tab. 2 wird deutlich, dass die Gruppe mit der geringsten Symptomatik auch eine erwartungsgemäß geringere Reduktion der Werte aufweist. Die Streuung ist zu beiden Messzeitpunkten gering, die Effektgröße kann bei einer mittleren Reduktion von 2,3 Punkten dennoch als mittel bewertet werden [$d = 0,48$; KI: 0,36–0,60].

Die Gruppe mit leichter Depression bei Reha-Beginn verbessert sich im Schnitt um 5,11 Punkte und weist eine große Effektgröße von 1,09 [KI: 0,96–1,21] auf. Der größte Effekt zeigt sich in der Gruppe mit mittelschwerer depressiver Symptomatik bei Reha-Beginn. Cohen's d ist groß und beträgt 1,31 [KI:

Tab. 2 Mittelwerte und Standardabweichungen im BDI-II zu t_0 und zu t_1 sowie Cohen's d als Maß für die Veränderung der Symptomatik

Gruppe nach Schweregrad der depressiven Symptomatik bei Beginn der Rehabilitation	M_{t_0}	SD_{t_0}	N	M_{t_1}	SD_{t_1}	d	KI unten	KI oben
Unauffällig (0–13 Punkte)	8,94	3,31	536	6,67	5,78	0,48	0,36	0,60
Leichte Depression (14–19 Punkte)	16,64	1,73	546	11,53	6,41	1,09	0,96	1,21
Mittelschwere Depression (20–28 Punkte)	23,83	2,56	893	16,07	7,95	1,31	1,21	1,42
Schwere Depression (29–63 Punkte)	36,37	6,22	1043	26,64	11,09	1,08	0,99	1,17
Gesamt	24,22	10,99	3018	17,23	11,46	0,62	0,57	0,67
Gesamt ohne unauffällige Gruppe	27,52	9,12	2482	19,51	11,09	0,79	0,73	0,85

M Mittelwert, SD Standardabweichung, N Anzahl Rehabilitanden, d standardised effect size bias corrected (Hedges), KI unten Konfidenzintervall für d untere Grenze, KI oben Konfidenzintervall für d obere Grenze, t_0 Reha-Beginn, t_1 Reha-Ende

Abb. 2 Mittelwerte und 95 %-Konfidenzintervall für die Werte im BDI-II zu t_0 und t_1

1,21–1,42]. Die mittlere Reduktion im BDI liegt bei 7,8 Punkten (etwa die kritische Differenz).

In der Gruppe mit schwer ausgeprägter depressiver Symptomatik bei Reha-Beginn kommt es zu einer mittleren Reduktion von 9,7 Punkten. Das ist die deutlichste Reduktion in der größten Gruppe der Studie, aber aufgrund der höheren Streuung fällt der Effekt etwas geringer aus ($d = 1,08$; KI: 0,99–1,17). Betrachtet man alle Rehabilitanden mit einer relevanten depressiven Symptomatik, zeigt sich ebenfalls ein großer Effekt von $d = 0,79$ [KI: 0,73–0,85].

In Abb. 2 wird deutlich, dass alle drei Gruppen mit krankheitsrelevanter depressiver Symptomatik eine deutliche Verbesserung erzielen können, aber dass die Werte bei Reha-Ende jeweils über jenen der Gruppen mit niedrigerer Ausgangsbelastung liegen.

Bewertung der Symptomreduktion im BDI-II auf Einzelfall-Ebene anhand der kritischen Differenz und der klinischen Signifikanz

Beim Erfolgskriterium einer Verminderung der depressiven Symptomatik im BDI-II um 8 Punkte (kritische Differenz) können 45 % der Rehabilitanden als signifikant gebessert eingestuft werden. In 51,5 % der Fälle lagen die Veränderungen jedoch in Bereich von bis zu 7 Punkten, dabei kann von keiner bedeutsamen Veränderung gesprochen werden. 3,5 % der Rehabilitanden haben sich um mindestens 8 Punkte im BDI-II verschlechtert.

Im Gegensatz zur signifikanten Veränderung stellt die klinische Signifikanz ein strengeres Kriterium für die Bewertung des Behandlungserfolgs dar. Zwei verschiedene Cut-off-Werte wurden als Erfolgskriterien analysiert.

Im Manual des BDI-II wird als Abgrenzung des funktionalen vom dysfunktionalen Bereich der Wert 19 angegeben. Damit werden Personen mit der Ober-

grenze einer „leichten Depression“ dem funktionalen Bereich zugeordnet. Bei dieser Analyse liegt 1/3 aller Patienten zu beiden Messzeitpunkten im unauffälligen Bereich, 1/4 der Patienten erzielt eine klinisch signifikante und 10 % eine signifikante Reduktion der depressiven Symptomatik, während 28 % keine signifikante Veränderung aufweisen auf und sich 3,5 % signifikant verschlechtern (siehe Abb. 3).

Legt man einen strengeren Maßstab für den funktionalen Bereich an und wählt als Cut-off den Wert 13 (Obergrenze einer minimalen Depression), liegen nur 15,9 % der Patienten zu beiden Messzeitpunkten im unauffälligen Bereich. Der Anteil nicht signifikant veränderter Patienten wird mit 38,6 % deutlich größer. Der Anteil erfolgreich behandelter Patienten ist mit insgesamt 42,1 % ebenfalls höher (19,3 % signifikant und 22,8 % klinisch signifikant gebessert).

In Tab. 3 finden sich in einer Kreuztabelle die übereinstimmenden und unterschiedlichen Bewertungen. Da bei einem Cut-off von 19 Punkten der unauffällige Bereich sehr groß ist ($N = 1006$), fallen von diesen Patienten beim niedrigeren Cut-off von 13 Punkten 316 Patienten in den Bereich „keine signifikante Veränderung“. Weil sie nun zum Entlassungszeitpunkt im auffälligen Bereich lagen, wechseln weitere 212 Patienten sogar in den Bereich einer klinisch signifikanten Verbesserung.

Umgekehrt galten 269 Patienten bei einem Cut-off von 19 Punkten als signifikant gebessert, bei einem Cut-off von 13 Punkten liegen sie jedoch immer noch im auffälligen Bereich, womit das zweite Kriterium der klinischen Signifikanz nicht erfüllt ist und nur eine signifikante Verbesserung vorliegt.

Insgesamt ergeben sich bei 798 Patienten (26,4 %) unterschiedliche Ergebnisse in der Bewertung des Behandlungserfolgs, aber in der Mehrheit überlappende Ergebnisse.

Bewertung der Symptomreduktion anhand der prozentuellen Verringerung

Im Schnitt zeigt sich eine Verbesserung der Symptomatik um 30,5 % ($SD = 45,51$), das zeugt von einer breiten Streuung innerhalb der Studiengruppe. Eine Reduktion der Symptomatik um 50 % oder mehr konnte von 30,4 % der Patienten erzielt werden. In dieser Gruppe finden sich mehr als die Hälfte jener Patienten, die im gesamten Beobachtungszeitraum im unauffälligen Bereich waren aber nur 16 % der Patienten mit einer signifikanten Reduktion der Symptomatik im Ausmaß der kritischen Differenz. Aus der Gruppe der klinisch signifikant gebesserten Patienten finden sich 88 % der Patienten in der Gruppe mit mindestens 50-%iger Reduktion der Symptomatik.

Diskussion

Die Bewertung von Behandlungseffekten wird von Kostenträgern gewünscht bzw. gefordert und ist we-

Veränderungen im BDI-II zwischen Beginn und Ende (gemessen an der kritischen Differenz von 8 Punkten)

klinische Signifikanz: Verringerung um kritische Differenz (8 P.) und Wechsel vom dysfunktionalen in den funktionalen Bereich mit Cut-off von 19 Punkten

klinische Signifikanz: Verringerung um kritische Differenz (8 P.) und Wechsel vom dysfunktionalen in den funktionalen Bereich mit Cut-off von 13 Punkten

Abb. 3 Signifikante und klinisch signifikante Veränderungen im BDI-II mit zwei verschiedenen Cut-off-Werten

Tab. 3 Gegenüberstellung der Ergebnisse in Abhängigkeit vom gewählten Cut-off-Wert im BDI-II

C13= Cut-off 13 Punkte C19= Cut-off 19 Punkte	C13 immer im unauffälligen Bereich	C13 signifikante Verschlechterung	C13 keine signifikante Veränderung	C13 signifikante Verbesserung	C13 klinisch signifikante Verbesserung	Summe und % bei Cut-off 19 Punkte
C19 immer im unauffälligen Bereich	478 15,8%	–	316 10,5%	–	212 7,0%	1006 33,3%
C19 signifikante Verschlechterung	1 0,1%	104 3,4%	–	–	–	105 3,5%
C19 keine sig. Veränderung	–	–	850 28,2%	–	–	850 28,2%
C19 signifikante Verbesserung	–	–	–	313 10,4%	–	313 10,4%
C19 klinisch sig. Verbesserung	–	–	–	269 8,9%	475 15,7%	744 24,6%
Summe und % bei Cut-off von 13 Punkten	479 15,9%	104 3,4%	1166 38,6%	582 19,3%	687 22,8%	3018 100%

sentlicher Bestandteil eines Qualitätsmanagements, eine transparente Vorgehensweise mit vergleichbaren Methoden ist eine wichtige Voraussetzung dazu.

Die verschiedenen Methoden zur Bewertung des Rehabilitationserfolgs zeigen differierende Ergebnisse. Auf Ebene der gesamten Stichprobe ($N=3018$) zeigte sich ein guter mittlerer Effekt von $d=0,62$. Bei dieser Berechnungsmethode kommt es aber zu einer Unterschätzung des Effekts, weil die konsekutive Stichprobe auch Rehabilitanden ohne bzw. mit minimaler depressiver Symptomatik bei Reha-Beginn beinhaltet. Bei der Analyse nach Subgruppen zeigen sich deshalb deutliche Unterschiede in Abhängigkeit von den Ausgangswerten im BDI-II, bei allen belasteten Gruppen zeigt sich ein großer Effekt von mindestens $d=0,8$. Wie in der Literatur berichtet, zeigt sich auch bei den vorliegenden Daten eine größere Variabilität beim zweiten Messzeitpunkt, weil sich viele Patienten deutlich verbesserten, ein kleiner Teil hat sich aber verschlechtert. Ogles et al. (2001) sprechen von einer Verwässerung des Effekts durch die große Varianz.

Bei der Analyse nach der Methode der kritischen Differenz zeigen sich mit 45 % die meisten Verbesserungen. Nimmt man als Kriterium die klinische Signifikanz mit einer reliablen Reduktion und einem Wechsel vom dysfunktionalen in den funktionalen Bereich, zeigen sich bei beiden Cut-off-Werten (13 und 19) jeweils ein knappes Viertel klinisch bedeutsam gebesserter Patienten. Die Wahl des Cut-off-Werts hat wesentlichen Einfluss auf das Ergebnis. Der Cut-off-Wert von 19 Punkten aus dem Manual klassifiziert 1/3 aller Rehabilitanden schon bei Beginn der Behandlung als unauffällig. Bei einem Cut-off-Wert von 13 Punkten liegt dieser Anteil nur bei 12,4 %, aber der Anteil nicht signifikant veränderter Patienten wird deutlich höher. Diese Vorgehensweise entspricht auch den Empfehlungen aus den S3-Leitlinien für unipolare Depression, wo ein Wert unter 13 als klinisch unauffällig oder remittiert angegeben wird.

Naturgemäß gibt es auch Kritik am Ansatz der klinischen Signifikanz. Ogles et al. (2001) meinen, Selbstbeurteilungsinstrumente könnten „zu reaktiv“, da keine Fremdbeurteilungen durch Angehörige oder Behandler im Sinne der sozialen Validität einfließen können (rater bias). Klinische Signifikanz sollte auch objektive und beobachtbare Kriterien beinhalten. Als Kritik wird auch angeführt, dass deutlich gebesserte Personen ohne Unterschreitung des Cut-off-Werts nicht als klinisch signifikant gebessert gelten können. Damit könne die Methode zu konservativ sein, da z. B. Patienten mit chronischer Depression, die in eine mildere Form der Depression wechseln, niemals als klinisch signifikant gebessert klassifiziert werden könnten. Ein Ausweg dazu wäre die Definition von verschiedenen Verteilungen, sodass beispielsweise Wechsel von einer Verteilung milder Beeinträchtigung in eine Verteilung mit leichter Beeinträchtigung als klinisch signifikant bewertet werden könnte.

Die Methode der klinischen Signifikanz mit Anwendung des strengerer Cut-off-Werts von 13 Punkten erwies sich als zielführend, weil damit nur jene Patienten als klinisch signifikant gebessert klassifiziert werden, die am Ende der Behandlung nur mehr eine Ausprägung im Bereich einer minimalen Depression aufweisen. Zwar zeigt sich damit ein höherer Anteil nicht signifikant gebesserter Patienten als beim höheren Cut-off von 19 Punkten, aber damit werden Patienten mit einer leichten Depression nicht dem unauffälligen Bereich zugeordnet.

Die Berechnung von Effektgrößen sollte deshalb nicht vernachlässigt werden, insbesondere dann, wenn auf Ebene von Einrichtungen oder Therapieverfahren Vergleiche durchgeführt werden. Der Fokus liegt dabei auf der mittleren Verbesserung einer Gruppe, dabei sollten aber Konfidenzintervalle für Mittelwerte und Mittelwertsdifferenzen angegeben werden [16]. Bei einer konsekutiven Studie mit heterogener Klientel und 60 % komorbiden Störungen erwies es sich jedoch als zielführend, Analysen nicht nur gesamt, sondern auch auf Ebene von Subgruppen in Abhängigkeit von der Ausgangsbelastung durchzuführen. Andernfalls könnten Effekte unterschätzt werden. Effektgrößen haben den Vorteil, dass die Berechnung auch möglich ist, wenn kein Schwellenwert für den funktionalen Bereich definiert ist oder wenn im Manual eines Fragebogens keine einheitliche kritische Differenz angegeben ist.

Schlussfolgerung/Fazit für die Praxis

Zur Bewertung des Behandlungserfolgs von depressiven Symptomen sollten neben Effektgrößen auch klinisch signifikante Veränderungen auf intraindividuel-ler Ebene herangezogen werden. Beim BDI-II hat sich ein Schwellenwert von 13 Punkten für eine Remission (S3-Leitlinie unipolare Depression) bewährt. Der prozentuelle Anteil der Symptomreduktion (S3-Leitlinie) eignet sich eher zur Prüfung des Ansprechens auf eine Behandlung, der Wechsel vom dysfunktionalen in den funktionalen Bereich geht aus dieser Berechnung nicht hervor. Die Angabe von signifikanten Verbesserungen im Mindestausmaß der kritischen Differenz sowie die Angabe klinisch signifikanter Verbesserungen – diese wechseln zusätzlich vom dysfunktionalen in den funktionalen Bereich – erweist sich als transparente und aussagekräftige Methode. Nach sechswöchiger ambulanter Rehabilitation in den Zentren für seelische Gesundheit BBRZ-Med Wien-Leopoldau, haben sich 19,3 % signifikant und weitere 22,8 % klinisch signifikant gebessert, knapp 1/4 kann damit als remittiert oder „geheilt“ und insgesamt 42 % als signifikant gebessert klassifiziert werden. Katamnestiche Studien weisen auf einen nachhaltigen Effekt der Symptomreduktion hin.

Funding Open access funding provided by Medical University of Vienna.

Einhaltung ethischer Richtlinien

Interessenkonflikt B. Senft, D. Fischer-Hansal und A. Schoser geben an, dass kein Interessenkonflikt besteht.

Ethische Standards Alle Patienten haben ihr schriftliches Einverständnis gegeben. Die Ethikkommission der Stadt Wien (Magistratsabteilung 15, Gesundheitsdienst der Stadt Wien) hat festgehalten, dass es sich hierbei nicht um eine angewandte medizinische Forschung am Menschen oder um eine medizinische Fragestellung handelt. (EK 13-063-VK_NZ).

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.

Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf <http://creativecommons.org/licenses/by/4.0/deed.de>.

Literatur

- World Health Organization. ICF: Internationale Klassifikation der Funktionsfähigkeit, Behinderung und Gesundheit. 2005. <http://www.dimdi.de/dynamic/de/klassi/downloadcenter/icf/endfassung/>. Zugriffen: 22. Nov. 2019.
- Biefang S, Schuntermann ME. Diagnostik und Assessment in der Rehabilitation. In: Bengel IJ, Koch U, Hrsg. Grundlagen der Rehabilitationswissenschaften: Themen, Strategien und Methoden der Rehabilitationsforschung. Berlin: Springer; 2000. S. 103–20.
- Calvert M, Brundage M, Jacobsen PB, Schünemann HJ, Effcace F. The CONSORT Patient-Reported Outcome (PRO) extension: Implications for clinical trials and practice. *Health Qual Life Out.* 2013;11:184. <https://doi.org/10.1186/1477-7525-11-184>.
- Ogles BM, Lunnen KM, Bonesteel K. Clinical significance: history, application, and current practice. *Clin Psychol Rev.* 2001;21(3):421–46.
- Cohen J. Statistical power analysis for the behavioral sciences (2.). Hillsdale: Lawrence Erlbaum; 1988.
- Steffanowski A, Löschmann C, Schmidt J, Wittmann W, Nübling R. Meta-Analyse der Effekte stationärer psychosomatischer Rehabilitation: Mesta-Studie (1st ed.). Arbeiten zur Theorie und Praxis der Rehabilitation in Medizin, Psychologie und Sonderpädagogik, Bd. 48. Bern: Huber; 2007.
- Sprung M, Münch HM, Kaiser E, Streibl L, Riffer F. Meta-Analyse der Evaluationsergebnisse psychiatrischer-psychosomatischer Rehabilitation in Österreich [Meta-analysis of evaluation results of psychiatric-psychosomatic rehabilitation in Austria]. *Neuropsychiatrie.* 2019;33(1):8–24. <https://doi.org/10.1007/s40211-018-0290-1>.
- Jacobson NS, Clinical Significance TP. A statistical approach to defining meaningful change in psychotherapy research. *Clin Psychol.* 1991;59(1):12–9.
- Lienert GA, Raatz U. Testaufbau und Testanalyse. 6. Aufl. Weinheim: Psychologie Verlags Union; 1998.
- Schmidt-Atzert L, Amelang M, Fydrich T, Moosbrugger H. Psychologische Diagnostik: Mit 82 Tabellen. 5. Aufl. Berlin: Springer; 2012. <https://doi.org/10.1007/978-3-642-17001-0>.
- DGPPN, BÄK, KBV, AMWE S3-Leitlinie/Nationale VersorgungsLeitlinie Unipolare Depression – Langfassung. 2015. www.depression.versorgungsleitlinien.de. Zugriffen: 25. Nov. 2019.
- Üstün TB, Kostanjsek N, Chatterji S, Rehm J. Measuring health and disability: Manual for WHO disability assessment schedule. WHODAS 2.0; Geneva. 2010.
- Nosper M. ICF AT-50: Entwicklung eines ICF-konformen Fragebogens für die Selbstbeurteilung von Aktivitäten und Teilhabe bei psychischen Störungen. In: Deutsche Rentenversicherung Bund (DRV Bund) (Chair), Hrsg. 17. Rehabilitationswissenschaftliches Kolloquium. Symposium conducted at the meeting of Deutsche Rentenversicherung Bund (DRV Bund) Bremen. 2008.
- Franke GH, Jäger S, Morfeld M, Salewski C, Reimer J, Rensing A, et al. Eignet sich das BSI-18 zur Erfassung der psychischen Belastung von nierentransplantierten Patienten? *Z Med Psychol.* 2010;19:30–7.
- Hautzinger M, Keller F, Kühner C. BDI-II Beck Depressions-Inventar Revision: Manual. Frankfurt: Harcourt Test Services; 2006.
- Höder J, Hüppe A. Zur Frage der klinischen Signifikanz in deutschen rehabilitationswissenschaftlichen Interventionsstudien – eine Bestandsaufnahme der gängigen Praxis. *Rehabilitation.* 2019;58(6):405–12. <https://doi.org/10.1055/a-0674-6360>.

Hinweis des Verlags Der Verlag bleibt in Hinblick auf geografische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutsadressen neutral.