Ergänzende Stellungnahme zum STEP.De Evaluationsbericht[1,2]

(Zusatz gemäß Nr. 14.1 ANBest-IF)

Konsortialführung:	BKK·VBU
Förderkennzeichen:	01NVF17050
Akronym:	STEP.De
Projekttitel:	Sporttherapie bei Depression
Autoren:	Prof. Dr. Dr. Michael Rapp (Projektleitung, Projektevaluation), Universität Potsdam: Erstellung Stellungnahme und Erratum, statistische Analysen Prof. Dr. Stephan Heinzel (Projektleitung, Prozessevaluation), Technische Universität Dortmund: Erstellung Stellungnahme und Erratum, statistische Analysen Dr. Andreas Heißel (Gesamtprojektleitung, Sporttherapie und Schulungen), Universität Potsdam: Erstellung Stellungnahme und Erratum Marlen Du Bois (Konsortialführung), BKK VBU: Revision Stellungnahme und Erratum

Vorwort

Zum Projekt STEP.De liegen uns mittlerweile Ergebnisse der Daten der Nachbeobachtung im Langzeitverlauf (12 Monate nach Ende der Intervention) vor. Diese Daten wurden von uns außerhalb des mit dem G-BA abgestimmten Evaluationskonzepts erhoben und nun außerhalb der Förderung analysiert. Anhand dieser Daten und im Rahmen der Vorbereitung weiterer Publikationen sind neue Erkenntnisse aufgekommen, die auch für die Interpretation der Ergebnisse des Projektes weiterführende Erkenntnisse bringen. Dabei zeigt die Sporttherapie im Langzeitverlauf im 12-monatigen Zeitraum nach Ende der Intervention weiterhin eine klinisch relevante Verringerung der Depressionssymptome. Konkret sehen wir im Langzeitverlauf jedoch eine zunehmende statistische Unsicherheit in Bezug auf die Nicht-Unterlegenheit der Sporttherapie und können im 12-monatigen Zeitraum nach Ende der Intervention die Nicht-Unterlegenheit der Sporttherapie nicht mehr

[1] Aus Gründen der besseren Lesbarkeit wird auf die gleichzeitige Verwendung männlicher, weiblicher und diverser Sprachformen verzichtet. Sämtliche Personenbezeichnungen gelten gleichermaßen für alle Geschlechter.

[2] Die Autoren danken Darlene Heinen, Christiane Stielow und Frederike Schneider für die Unterstützung in der weiteren Datenaggregation und der Beschreibung der Stichprobe; Dr. Fabian Arntz für erweiterte Sensitivitätsanalysen und Dr. Gregor Wilbertz für Hinweise zu notwendigen Korrekturen und für erweiterte Datenaggregation und -analysen im Langzeitverlauf.

durchgängig nachweisen. Die uns jetzt vorliegenden Ergebnisse legen nahe, dass dieser Befund angesichts stabiler Effekte der Sporttherapie auch im Langzeitverlauf insbesondere durch eine weitere Verbesserung in der Psychotherapiegruppe entstehen könnten. Wir haben deshalb eine ergänzende Stellungnahme vorbereitet. Zudem möchten wir im Evaluationsbericht unklare Formulierungen auflösen sowie spezifische Ergänzungen und Korrekturen vornehmen. Diese betreffen die Zuteilung der Patienten und die Depressionsdiagnostik in der Studie, sowie die geschätzte Power der primären Datenanalyse und die Analyse im Langzeitverlauf.

Diese Ergänzungen und Korrekturen ändern nicht die Einordnung der Studienergebnisse und die von den wissenschaftlichen Projektleitern gegebene Empfehlung zur Implementierung der Sporttherapie in die Regelversorgung im Ergebnis- und Evaluationsbericht. Diese neuen Ergebnisse sollten aber in der Implementierung der Sporttherapie in die Regelversorgung berücksichtigt werden.

1. Stabilität der Ergebnisse im Verlauf nach Ende der Intervention

In unseren gegenwärtigen Analysen des Langzeitverlaufes (mITT Stichprobe) zeigt die Sporttherapie auch 12 Monate (T5) nach Beendigung der Therapie eine klinisch relevante Verringerung der Depressionssymptome (-9.37 Punkte; 95% Konfidenzintervall, KI [-11.60 bis -7.14]) im BDI-II (primäres Outcome, Selbsterhebung durch Fragebogen), ist jedoch zu T5 um -1.55 Punkte schlechter im Vergleich zur Psychotherapie und das 95% KI [KI -4.86 bis 1.28] unterschreitet klar die Non-Inferiority-Grenze von -3, so dass sich eine Nichtunterlegenheit nach 12 Monaten nicht mehr nachweisen lässt.

In Bezug auf die Verringerung der Depressionssymptome finden wir auch zum Zeitpunkt T4 (vgl. Tabelle S2) sowohl für die Sporttherapiegruppe (-8.60 Punkte; 95% KI [-10.83 bis -6.37]) als auch für die Psychotherapiegruppe (-9.40 Punkte; 95% KI [-12.37 bis -6.43]) eine klinisch relevante Verringerung der Symptome. Wir haben daraufhin auch die Nichtunterlegenheitsgrenzen im Gruppenvergleich zum Zeitpunkt T4 (nach 6 Monaten) weiterführend analysiert.

Numerisch (vgl. Tabelle S2 im Bericht, d. h. ohne Imputation und ohne Modellierung) ist der Unterschied zwischen beiden Gruppen zum Zeitpunkt T4 Null (0.00), unterschreitet aber die Non-Inferiority-Grenze von -3 im 95% Konfidenzintervall [-3.06 bis 3.06] knapp. Für den Messzeitpunkt T4 (6 Monate nach Intervention) zeigt im clusteradjustierten GEE Modell (vgl. korrigierte Tabelle S6 des Evaluationsberichtes) der Schätzer für den Interaktionseffekt (Gruppe und Zeit von T0 bis T4) eine um 1.17 Punkte im BDI-II geringere Symptomverbesserung durch die Sporttherapie im Vergleich zur Psychotherapie (-1.17 [-3.34 bis 1.00]) und das 95% Konfidenzintervall unterschreitet die Non-Inferiority-Grenze von -3 knapp. Zum Follow-up Zeitpunkt T4 (6 Monate nach Ende der Intervention)

lässt sich somit anhand der relevanten Konfidenzintervalle (auf dem Niveau des 95% Konfidenzintervalls) keine Non-Inferiority im primären Outcome nachweisen. Weiterführende Analysen haben hier ergeben, dass das Non-Inferiority-Kriterium im primären Outcome BDI-II zu T4 hingegen auf dem 90% Konfidenzintervall, das jedoch nicht das übliche und vorab festgelegte Zielkriterium der Analyse war (vgl. Heissel et al., 2020), gegeben ist (-1.17; 90% KI -2.99 bis 0.65). Wir möchten darauf hinweisen, dass im Nicht-Unterlegenheitsdesign neben dem Kriterium der Non-Inferiority (Nichtunterlegenheit; welches erfüllt ist, wenn das 95% Konfidenzintervall des Unterschiedes die Non-Inferiority-Grenze nicht unterschreitet) auch das Kriterium der Inferiority (Unterlegenheit) zu berücksichtigen ist (Schumi und Wittes, 2011). Inferiority liegt vor, wenn der Unterschied numerisch größer ist als die Non-Inferiority-Grenze *und* das 95% Konfidenzintervall null (0) nicht umschließt. Da das Inferiority-Kriterium der Sporttherapie zu T4 (nach 6 Monaten) nicht erfüllt ist, lässt sich somit auch keine Unterlegenheit der Sporttherapie zeigen.

Für das sekundäre Outcome HAM-D (Depressionssymptome erhoben durch verblindete Assessoren) zeigt im GLMM der Interaktionseffekt (Gruppe und Zeit von T0 bis T4) eine numerische Verbesserung der Sporttherapiegruppe gegenüber der Psychotherapiegruppe (-0.66 [-1.88 bis 3.19]), deren 95% Konfidenzintervall die Non-Inferiority-Grenze von -2.16 nicht unterschreitet (vgl. Tabelle S8 im Evaluationsbericht). Zum Zeitpunkt T4 (vgl. Tabelle S2) ist der Unterschied 0.20 (0.20 [-1.81 bis 2.21]), und das 95% Konfidenzintervall unterschreitet die Non-Inferiority-Grenze von -2.16 damit ebenfalls nicht. Für das sekundäre Zielkriterium lässt sich die Nichtunterlegenheit somit auf dem Niveau des 95% Konfidenzintervalls also auch für T4 zeigen.

Wir führen derzeit im Rahmen der erweiterten Publikation der Ergebnisse zusätzlich zum Evaluationsbericht Sensitivitätsanalysen zur Stabilität der Effekte für das primäre Outcome BDI-II im Langzeitverlauf durch, bei der wir für 16 Monate nach Interventionsbeginn alle Messzeitpunkte von T0 bis T5 (einschließlich der intermediären Messzeitpunkte T1 (2 Monate nach Interventionsbeginn) und T3 (2 Monate nach Ende der Intervention)) für die ITT Analyse berücksichtigen. Dabei können wir auch die zeitliche Zuordnung zu den Messzeitpunkten (u.a. durch Verschiebungen durch pandemiebedingte Schließungen der Therapieeinrichtungen) genauer berücksichtigen.

In etablierten Analysemodellen (gemischte lineare Modelle mit linearen Kontrasten) finden wir ebenfalls erste Hinweise, dass die Unterschiede zwischen den Gruppen bis T5 (12 Monate nach Ende der Intervention) größer werden (vgl. Abbildung 1). Bis zum Ende der Intervention findet sich kein signifikanter Unterschied, im Langzeitverlauf wird dieser aber zugunsten der Psychotherapie (TAU) größer und in einigen Modellen auch statistisch signifikant. Grundsätzlich ist dabei allerdings zu berücksichtigen, dass im STEP.De Projekt ein größerer Anteil an Patienten der Psychotherapie-Intervention (76.8%) im Vergleich zur Sporttherapie (21.9%) im Anschluss an die Intervention (nach

T2) eine weiterführende Psychotherapie erhalten haben. Somit könnte auch ein Dosis-Wirkungs-Zusammenhang den Unterschied im Langzeitverlauf erklären.

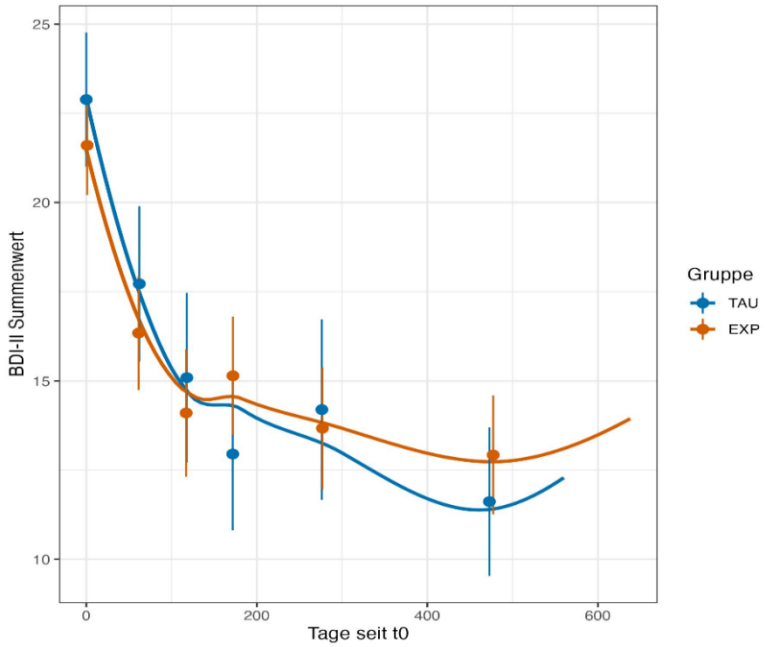

Abbildung 1. Langzeitverlauf nach Interventionsbeginn (Darstellung der Mittelwerte zu den Zeitpunkten T0 bis T5 +/- 2 Wochen (Punkte mit 95% KI) und aller verfügbaren Daten mittels lokal gewichteter Regressionsfunktion (Linien)).

2. Power der primären Datenanalyse

Die Studienteilnehmer wurden nach Regelpsychotherapieverfahren und Praxisgröße stratifiziert randomisierten Clustern zugeordnet. Bei den Fallzahlschätzungen wurde diese Clusteradjustierung mittlerweile berücksichtigt (Heissel et al., 2020). Anders als in den Limitationen beschrieben, ist die Fallzahl bei Interventionsende (T2) damit für die adjustierten Analysen nicht zu 83%, sondern nur zu 67% erreicht. Für die Fallzahlschätzung gingen wir für die adjustierten Analysen von einer Intraclusterkorrelation (ICC) von 0.01 aus. Die tatsächliche ICC bezüglich des primären Outcomes BDI-II beträgt für die Messzeitpunkte T0-T2-T4 0.023. Wir gehen nach den vorliegenden Daten davon aus, dass eine Power von 80% für die clusteradjustierte Analyse mit hoher Wahrscheinlichkeit nicht erreicht wurde.

Für nichtadjustierte Analysen (s. S. 16 und S. 25 Evaluationsbericht) wurde zum Zeitpunkt T2 eine Fallzahl von 312 geplant, wovon n = 259, also 83% der geplanten Stichprobe erreicht wurden. In der last observation carried forward (LOCF) Analyse (die ursprünglich mit Clusteradjustierung geplant war) konnte aber eine Fallzahl von n = 334 berücksichtigt werden, die die ursprünglich für nichtadjustierte

Analysen geschätzte Fallzahl überschreitet (vgl. Overall et al., 2009). Es wurden deshalb zusätzlich nichtadjustierte Analysen in der geplanten (vgl. Heissel et al., 2020) last observation carried forward (LOCF) Analyse vorgenommen. Hier zeigt der Interaktionseffekt (Gruppe und Zeit) für das primäre Outcome BDI-II von T0 bis T2 im GEE ohne Clusteradjustierung eine numerisch geringere Verbesserung durch die Sporttherapie im Vergleich zur Psychotherapie (-1.19 [-2.73 bis 0.35]), deren 95% Konfidenzintervall die Non-Inferiority-Grenze von -3 jedoch nicht unterschreitet. Für den Messzeitpunkt T4 (6 Monate nach Intervention) zeigt der Schätzer für den Interaktionseffekt (Gruppe und Zeit von T4 bis T0) eine um 1.15 Punkte im BDI-II geringere Symptomverbesserung durch die Sporttherapie im Vergleich zur Psychotherapie (-1.15 [-3.32 bis 1.03]) und das 95% Konfidenzintervall unterschreitet die Non-Inferiority-Grenze von -3 (s. Tabelle S6E in der Ergänzung). Somit können wir für das primäre Outcome für den Zeitpunkt T2 (direkt nach der Intervention) mit den nichtadjustierten Analysen die Nichtunterlegenheit der Sporttherapie zeigen, für den Zeitpunkt T4 (6 Monate nach der Intervention) hingegen nicht mehr. Das Kriterium der Inferiority ist aber weder zu T2 noch zu T4 gegeben, so dass unsere Daten eine Unterlegenheit der Sporttherapie in den nichtadjustierten LOCF Analysen mit besserer Power auch zu T4 nicht belegen können.

Die Auswertung des Evaluationsoutcomes erfolgte (auch im Zeitraum nach der Förderung) von einer Mitarbeiterin der Universität Potsdam unter Supervision durch Profs. Rapp (Uni Potsdam) und Heinzel (FU Berlin). Aufgrund der 2:1 Randomisierung und den damit zusammenhängenden unterschiedlichen Gruppengrößen konnte im weiteren Projektumfeld die Datenanalyse hinsichtlich der Gruppenzugehörigkeit nicht verblindet werden.

3. Studieneinschluss, Zuteilung der Patienten und Diagnostik der Depression

Die Studienteilnehmer wurden über Arbeitsunfähigkeitsdiagnosen oder Versorgungsdiagnosen oder einem gestellten Antrag auf Psychotherapie durch die beteiligten Krankenkassen rekrutiert. Weiterhin konnten sich Betroffene, die auf das Projekt aufmerksam wurden, direkt bei den Fallmanagern der beteiligten Krankenkassen informieren und eine potenzielle Eignung und Studienteilnahme abklären. Die Fallmanager erfragten eine vorläufige Einwilligung, führten ein erstes Screening der Ein- und Ausschlusskriterien durch und wiesen die Studieninteressierten anhand einer Randomisierungsliste einer der beiden Interventionsgruppen zu. Da es sich hierbei formal um eine Randomisierung mit offener Zuteilung handelte, ist ein möglicher Selektionsbias nicht sicher auszuschließen. Die Fallmanager waren aber streng instruiert, die Randomisierungsliste erst nach Ende des Screeninggesprächs zu öffnen und keine Zeile zu überspringen. Zudem hatten zeitgleich alle Fallmanager (n = 22) Zugriff auf die Liste, so dass nach Eintragung für nachfolgende Rekrutierungsgespräche nicht klar ersichtlich war, welche Zuordnungen zukünftig verfügbar sein

werden. Eine Nachuntersuchung der Zeitmarken der 560 Zeilen in der Zuteilungsliste ergab, dass 8 Zeilen falsch zugeordnet und 5 übersprungen wurden (also insgesamt 2,32% der Zuteilungen).

Der eigentliche Studieneinschluss, formale Studienaufklärung und Einwilligung sowie die Zulassung zu einer der Interventionen erfolgte später in einem diagnostischen Interventionsgespräch bei an der Studie beteiligten Psychotherapeuten. Einschlusskriterien (vgl. S. 22 Ergebnisbericht) waren sechs Diagnosen aus dem Depressionsspektrum (leichte bis mittelschwere depressive Episode (F32.0, F32.1), rezidivierende depressive Störung mit einer aktuellen leichten bis mittelschweren Episode (F33.0, F33.1), Dysthymie (F34.1) oder gemischte Angst- und depressive Störungen (F41.2)) und vier weitere verwandte psychiatrische Diagnosen (Anpassungsstörungen (F43.2), Neurasthenie (F48.0), und Reaktionen auf schweren Stress (F43.8, F43.9).

Die Indikation für eine therapeutische Behandlung wurde von den beteiligten Psychotherapeuten geprüft. Sie waren angehalten, für die klinische Diagnostik und Indikationsstellung das strukturierte klinische Interview für DSM-IV (SKID-I) durchzuführen und wurden dafür in der Durchführung des SKID-I (Sektionen A, E und I) im Rahmen des Projekts geschult. Die Entscheidung, wie die Diagnostik durchgeführt wurde, oblag, analog zur Regelversorgung, letztendlich den Psychotherapeuten. Es muss davon ausgegangen werden, dass die Absicherung der Diagnose nicht regelhaft durch den SKID erfolgt ist; dies konnte auch extern nicht überprüft werden.

Aus Gründen des Datenschutzes konnte die finale Diagnose der Psychotherapeuten nicht direkt an die Studienmitarbeiter übermittelt werden. Im Rahmen der erweiterten Ergebnispublikation haben wir nunmehr die Versorgungsdiagnosen aufgenommen. Diese sind die Diagnosen aus den Abrechnungsdaten der Krankenkassen im Rahmen der gesundheitsökonomischen Analyse, die aus der Abrechnung der besonderen Versorgungsform gemäß § 295 SGB V hervorgehen oder im zeitlichen Zusammenhang (<=30 Tage zu T0; bis zu 7 Tage nach dem Erstgespräch beim Psychotherapeuten) in den ambulanten Abrechnungsdaten erfasst wurden. Demnach lagen in der mITT Stichprobe (n=344) bei 65.99% (n=227) der Studienteilnehmer eine Versorgungsdiagnose aus dem Depressionsspektrum (leichte bis mittelschwere depressive Episode (F32.0, F32.1), rezidivierende depressive Störung mit einer aktuellen leichten bis mittelschweren Episode (F33.0, F33.1), Dysthymie (F34.1) oder gemischte Angst- und depressive Störungen (F41.2)) und bei 26.17% (n=90) der Studienteilnehmer eine verwandte psychiatrische Einschlussdiagnose (Anpassungsstörungen (F43.2), Neurasthenie (F48.0), und Reaktionen auf schweren Stress (F43.8, F43.9) vor. Von den Studienteilnehmern, bei denen auf Basis der Versorgungsdiagnosen keine definierte Einschlussdiagnose vorlag, war bei 2.61% (n=9) eine Depressionsdiagnose (F32.9; depressive Episode, nicht näher bezeichnet) diagnostiziert. Von den verbleibenden Studienteilnehmern haben 2.03% (n=7) keine Einschlussdiagnose und bei 3.20% (n=11)

der Studienteilnehmer lagen keine Versorgungsdiagnosen vor (vor allem, weil diese die Einwilligung zur Datenverarbeitung zu Evaluationszwecken nicht gaben oder diese widerrufen haben).

Zusammenfassung und Schlussfolgerungen

Wir beobachten im Langzeitverlauf eine zunehmende statistische Unsicherheit und können im 12-monatigen Zeitraum nach Ende der Intervention die Nicht-Unterlegenheit der Sporttherapie nicht mehr durchgängig nachweisen. Zudem legen erste Sensitivitätsanalysen nahe, dass dieser Befund angesichts stabiler Effekte der Sporttherapie auch im Langzeitverlauf insbesondere durch eine weitere Verbesserung in der Psychotherapiegruppe entstehen könnte. Dass die Fortsetzungsraten der Psychotherapie nach Abschluss der Intervention in beiden Gruppen signifikant unterschiedlich waren (in der Sporttherapiegruppe 21.9% Psychotherapien, in der Psychotherapiegruppe 76.8% fortgeführte Psychotherapien), könnte eine Erklärung für diese Ergebnisse im Langzeitverlauf sein. Dies könnte aber auch der verminderten Power in clusteradjustierten Analysen und im Langzeitverlauf geschuldet sein.

Für nichtadjustierte Analysen mit besserer Power (s. Tabelle S6E) bleibt aber für das primäre Zielkriterium BDI-II der Effekt der Nichtunterlegenheit für den Interventionszeitraum (T0-T2) erhalten. Für das sekundäre Zielkriterium Hamilton Depressionsskala bleibt der Nichtunterlegenheitseffekt auch für den Zeitraum bis T4 (6 Monate nach der Intervention) erhalten. Zudem findet sich für unsere Prüfhypothese 1 für das primäre Outcome BDI-II („Nach 16 Wochen Intervention und bei der Nachbeobachtung nach 6 Monaten zeigt die Gesamtstichprobe eine signifikante Verbesserung der depressiven Symptomatik“) auch im 6-Monats Follow-Up in clusteradjustierten Analysen ein Effekt von 7.97 Punkten mit reliablem 95% Konfidenzintervall (6.17 bis 9.77) ohne signifikanten Gruppenunterschied (p = .81; vgl. Tabelle S6), für den die genannten Einschränkungen der statistischen Power nicht zutreffen. Eine rezente Vergleichsstudie zwischen Sporttherapie und medikamentöser Therapie bei Depression aus den Niederlanden folgert allein aus einem diesem Ergebnis ähnlichen Befund eine Vergleichbarkeit der Therapieverfahren (Verhoeven et al., 2023).

Durch die offene Gruppenzuteilung sowie die auch im Versorgungskontext zu erwartende Unsicherheit bezüglich der Diagnosen besteht eine gewisse Unsicherheit hinsichtlich der Generalisierbarkeit der Ergebnisse, die aber in Teilen bereits diskutiert wurde (vgl. S 25 Evaluationsbericht). Die im Evaluationsbericht gemachte Empfehlung für leichte und mittelschwere Depression ist hinsichtlich der beschriebenen Versorgungsdiagnosen aus dem Depressionsspektrum und verwandte psychiatrische Diagnosen (vgl. Ergebnisbericht S. 22) zu konkretisieren. Letztlich beruht die gegebene Empfehlung hinsichtlich ihrer Generalisierbarkeit nicht allein auf den Ergebnissen der STEP.De Intervention, sondern ist in umfangreiche metaanalytische Evidenz (Cooney

et al., 2013; Heissel et al., 2023; Krogh et al., 2017; Schuch et al., 2016) und Leitlinien (z.B. S3-Leitlinie unipolare Depression) zur Sporttherapie bei depressiven Erkrankungen und Symptomen eingebettet. **Aus diesen Erwägungen heraus bleibt unsere im Evaluationsbericht gegebene Empfehlung für Patienten aus dem Depressionsspektrum bestehen.**

Es bleibt allerdings festzuhalten, dass im Rahmen der Analyse des Follow-Up-Zeitraums bis T4 und T5 (6 und 12 Monaten nach Abschluss der Intervention) im primären Outcome keine Nichtunterlegenheit nach gängigen Beurteilungsmaßstäben vorliegt. Zudem finden wir in ersten Sensitivitätsanalysen des Langzeitverlaufs, die den Interventionseffekt während der Sportintervention vom Langzeitverlauf nach Ende der Intervention abgrenzen, erste Hinweise für einen signifikanten Gruppenunterschied zugunsten der Psychotherapie. Dies kann einerseits mit der geringeren statistischen Power bzw. den relativ großen Unsicherheiten im Konfidenzintervall zu tun haben. Dass die Fortsetzungsraten der Psychotherapie nach Abschluss der Intervention in beiden Gruppen signifikant unterschiedlich waren (in der Sporttherapiegruppe 21.9% Psychotherapien, in der Psychotherapiegruppe 76.8% fortgeführte Psychotherapien), könnte aber auch eine Erklärung für diese Ergebnisse im Langzeitverlauf sein. In zukünftigen Studien sollten deshalb die Dosis-Wirkungszusammenhänge noch umfangreicher geprüft werden.

Das Design der Sportintervention des Projektes in der von uns gegebenen Evaluationsempfehlung sah bereits die Begleitung durch Psychotherapeuten zu Beginn und Ende der Sportintervention vor, um weiteren psychotherapeutischen Behandlungsbedarf zu identifizieren und bei Bedarf ein Behandlungsangebot zu machen.

Die Ergänzung der Empfehlung ist deshalb, eine Verlaufskontrolle zur Klärung eines weiteren möglichen Therapiebedarfs nicht nur nach Abschluss der Sporttherapie, sondern auch sechs bis zwölf Monate später (analog zu unseren Zeitpunkten T4 und T5) anzubieten und in der Umsetzung in die Regelversorgung zu berücksichtigen.

Literatur

Cooney, G. M., Dwan, K., Greig, C. A., Lawlor, D. A., Rimer, J., Waugh, F. R., McMurdo, M., & Mead, G. E. (2013). Exercise for depression. *The Cochrane Database of Systematic Reviews*, 9, CD004366. https://doi.org/10.1002/14651858.CD004366.pub6

Heissel, A., Pietrek, A., Schwefel, M., Abula, K., Wilbertz, G., Heinzel, S., & Rapp, M. (2020). STEP. De study—a multicentre cluster-randomised effectiveness trial of exercise therapy for patients with depressive symptoms in healthcare services: study protocol. *BMJ open*, 10(4), e036287.

Heissel, A., Heinen, D., Brokmeier, L. L., Skarabis, N., Kangas, M., Vancampfort, D., Stubbs, B., Firth, J., Ward, P. B., Rosenbaum, S., Hallgren, M., & Schuch, F. (2023). Exercise as medicine for depressive

symptoms? A systematic review and meta-analysis with meta-regression. *British Journal of Sports Medicine*, bjsports-2022-106282. https://doi.org/10.1136/bjsports-2022-106282

Krogh, J., Hjorthøj, C., Speyer, H., Gluud, C., & Nordentoft, M. (2017). Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential analysis. *BMJ Open*, *7*(9), e014820.

Overall, J. E., Tonidandel, S., & Starbuck, R. R. (2009). Last-observation-carried-forward (LOCF) and tests for difference in mean rates of change in controlled repeated measurements designs with dropouts. Social Science Research, 38(2), 492-503.

Rhodes S., Richards, D.A., Ekers, D., et al. (2014). Cost and outcome of behavioural activation versus cognitive behaviour therapy for depression (cobra): study protocol for a randomised controlled trial. Trials 15:29.doi:10.1186/1745-6215-15-29.

Richards, D.A., Ekers, D., McMillan, D., et al. (2016). Cost and outcome of behavioural activation versus cognitive behavioural therapy for depression (cobra): a randomised, controlled, non-inferiority trial. *Lancet*,388:871–80.doi:10.1016/S0140-6736(16)31140-0.

Schuch, Vancampfort, D., Richards, J., Rosenbaum, S., Ward, P. B., & Stubbs, B. (2016). Exercise as a treatment for depression: A meta-analysis adjusting for publication bias. *Journal of Psychiatric Research*, *77*, 42–51. https://doi.org/10.1016/j.jpsychires.2016.02.023

Schumi, J., & Wittes, J. T. (2011). Through the looking glass: understanding non-inferiority. *Trials*, 12, 1-12.

Verhoeven, J. E., Han, L. K., Lever-van Milligen, B. A., Hu, M. X., Révész, D., Hoogendoorn, A. W., ... & Penninx, B. W. (2023). Antidepressants or running therapy: Comparing effects on mental and physical health in patients with depression and anxiety disorders. *Journal of Affective Disorders*, 329, 19-29.

Erratum und Ergänzungen im Evaluationsbericht an den G-BA

2.1 Übersicht

Bisher (S. 8, 2. Absatz):

Die Fallmanager waren für die Gruppenzugehörigkeit der Psychotherapeuten verblindet.

Erratum:

Die Fallmanager erfragten eine vorläufige Einwilligung, führten ein erstes Screening der Ein- und Ausschlusskriterien durch und wiesen die Studieninteressierten anhand einer Randomisierungsliste einer der beiden Interventionsgruppen zu. Da es sich hierbei formal um eine Randomisierung mit offener Zuteilung handelte, ist ein möglicher Selektionsbias nicht sicher auszuschließen. Die Fallmanager waren aber streng instruiert, die Randomisierungsliste erst nach Ende des Screeninggesprächs zu öffnen und keine Zeile zu überspringen. Zudem hatten zeitgleich alle Fallmanager (n = 22) Zugriff auf die Liste, so dass nach Eintragung für nachfolgende Rekrutierungsgespräche nicht klar ersichtlich war, welche Zuordnungen zukünftig verfügbar sein werden. Eine Nachuntersuchung der Zeitmarken der 560 Zeilen in der Zuteilungsliste ergab, dass 8 Zeilen falsch zugeordnet und 5 übersprungen wurden (also insgesamt 2,32% der Zuteilungen).

Bisher (S.8, vorletzter Absatz):

Hierfür wurden alle Psychotherapeuten darin geschult, das Strukturelle Klinische Interview I für das Diagnostische und Statistische Manual Psychischer Störungen 4 (SKID-I, Structured Clinical Interview for DSM IV, SCID-I), Achse 1, Abschnitt A, E und I anzuwenden.

Erratum und Ergänzung:

Die Indikation für eine therapeutische Behandlung wurde von den beteiligten Psychotherapeuten geprüft. Sie waren angehalten, für die klinische Diagnostik und Indikationsstellung das strukturierte klinische Interview für DSM-IV (SKID-I) durchzuführen und wurden dafür in der Durchführung des SKID-I (Sektionen A, E und I) im Rahmen des Projekts geschult. Die Entscheidung, wie die Diagnostik durchgeführt wurde, oblag, analog zur Regelversorgung, letztendlich den Psychotherapeuten. Es muss davon ausgegangen werden, dass die Absicherung der Diagnose nicht regelhaft durch den SKID erfolgt ist; dies konnte auch extern nicht überprüft werden.

2.6 Fallzahlenberechnung

Bisher (S.16., 3. Absatz):

Die Verwendung des R-Pakets "SampleSize4ClinicalTrials" mit einem wahren Effekt von Null, einer α-Fehlerrate von 5 % und einer Power von 80 % sowie einer Nichtunterlegenheitsgrenze von 3 und einer SD von 10 ergibt eine Gesamtzahl von 312 Teilnehmern, um die Nichtunterlegenheit zwischen den Gruppen zum Zeitpunkt der Nachbehandlung zu testen. Unter Berücksichtigung zusätzlicher 25 % Ausfälle führt dies zu einer geschätzten Stichprobengröße von 390 Teilnehmern.

Ergänzung:

Für die geplante Clusteradjustierung wurden noch einmal ca. 23% mehr Probanden eingeplant, nämlich insgesamt N=384 für die finale Stichprobe (Heissel et al., 2020).

2.7 Statistische Analysen

Bisher (S. 16, 4. Absatz):

Die Auswertung des Evaluationsoutcomes erfolgte durch die FU Berlin/ Prof. Stephan Heinzel. Diese erhielten die anonymisierten und den Gruppen nicht zuzuordnenden Datensatz. Damit sollte die Unabhängigkeit der primären Evaluation gewährleistet werden.

Erratum:

Die Auswertung des Evaluationsoutcomes erfolgte (auch im Zeitraum nach der Förderung) von einer Mitarbeiterin der Universität Potsdam unter Supervision durch Profs. Rapp (Uni Potsdam) und Heinzel (FU Berlin). Aufgrund der 2:1 Randomisierung und den damit zusammenhängenden unterschiedlichen Gruppengrößen konnte im weiteren Projektumfeld die Datenanalyse hinsichtlich der Gruppenzugehörigkeit nicht verblindet werden.

3.1.1 Probandencharakteristika

Bisher (S. 17, 4. Absatz):

Es wurden 393 Patienten mit leichter bis mittelschwerer Depression (Mittelwert des BDI-II = 22.7, SD 9.9) in die STEP.De-Studie aufgenommen.

Erratum und Ergänzung:

Es wurden 393 Patienten in die STEP.De-Studie aufgenommen (Mittelwert des BDI-II = 22.7, SD 9.9). Die Studienteilnehmer wurden über Arbeitsunfähigkeitsdiagnosen oder Versorgungsdiagnosen oder einem gestellten Antrag auf Psychotherapie durch die beteiligten Krankenkassen rekrutiert. Weiterhin konnten sich Betroffene, die auf das Projekt aufmerksam wurden, direkt bei den Fallmanagern der beteiligten Krankenkassen informieren und eine potenzielle Eignung und Studienteilnahme abklären.

Der eigentliche Studieneinschluss, formale Studienaufklärung und Einwilligung sowie die Zulassung zu einer der Interventionen erfolgte später in einem diagnostischen Interventionsgespräch bei an der Studie beteiligten Psychotherapeuten. Einschlusskriterien (vgl. S. 22 Ergebnisbericht) waren sechs Diagnosen aus dem Depressionsspektrum (leichte bis mittelschwere depressive Episode (F32.0, F32.1), rezidivierende depressive Störung mit einer aktuellen leichten bis mittelschweren Episode (F33.0, F33.1), Dysthymie (F34.1) oder gemischte Angst- und depressive Störungen (F41.2)) und vier weitere verwandte psychiatrische Diagnosen (Anpassungsstörungen (F43.2), Neurasthenie (F48.0), und Reaktionen auf schweren Stress (F43.8, F43.9). Die im Evaluationsbericht gemachte Empfehlung für leichte und mittelschwere Depression ist somit hinsichtlich der Versorgungsdiagnosen aus dem Depressionsspektrum und verwandte psychiatrische Diagnosen (vgl. Ergebnisbericht S. 22) zu konkretisieren.

Aus Gründen des Datenschutzes konnte die finale Diagnose der Psychotherapeuten nicht direkt an die Studienmitarbeiter übermittelt werden. Im Rahmen der erweiterten Ergebnispublikation haben wir nunmehr die Versorgungsdiagnosen aufgenommen. Diese sind die Diagnosen aus den Abrechnungsdaten der Krankenkassen im Rahmen der gesundheitsökonomischen Analyse, die aus der Abrechnung der besonderen Versorgungsform gemäß § 295 SGB V hervorgehen oder im zeitlichen Zusammenhang (<=30 Tage zu T0; bis zu 7 Tage nach dem Erstgespräch beim Psychotherapeuten) in den ambulanten Abrechnungsdaten erfasst wurden. Demnach lagen in der mITT Stichprobe (n=344) bei 65.99% (n=227) der Studienteilnehmer eine Versorgungsdiagnose aus dem Depressionsspektrum (leichte bis mittelschwere depressive Episode (F32.0, F32.1), rezidivierende depressive Störung mit einer aktuellen leichten bis mittelschweren Episode (F33.0, F33.1), Dysthymie (F34.1) oder gemischte Angst- und depressive Störungen (F41.2)) und bei 26.17% (n=90) der Studienteilnehmer eine verwandte psychiatrische Einschlussdiagnose (Anpassungsstörungen (F43.2), Neurasthenie (F48.0), und Reaktionen auf schweren Stress (F43.8, F43.9) vor. Von den Studienteilnehmern, bei denen auf Basis der Versorgungsdiagnosen keine definierte Einschlussdiagnose vorlag, war bei 2.61% (n=9) eine Depressionsdiagnose (F32.9; depressive Episode, nicht näher bezeichnet) diagnostiziert. Von den verbleibenden Studienteilnehmern haben 2.03% (n=7) keine Einschlussdiagnose und bei 3.20% (n=11) der Studienteilnehmer lagen keine Versorgungsdiagnosen vor (vor allem, weil diese die Einwilligung zur Datenverarbeitung zu Evaluationszwecken nicht gaben oder diese widerrufen haben).

Bisher (S. 17, 4. Absatz):

Demnach konnten 344 Patienten in die mITT-Analyse einbezogen werden (219 EXP und 125 TAU) (Anhang B, Abbildung S1 für weitere Informationen).

Ergänzung:

Von diesen wiesen 25 Patienten (19 EXP und 6 TAU), also 7,27% der mITT-Stichprobe, einen BDI-II Wert von 8 und weniger auf, was die Abwesenheit auch minimaler depressiver Symptomatik nahelegt. Von diesen lag für zwei Studienteilnehmer keine Einschlussdiagnose vor.

Bisher (S.18, 3. Absatz):

Im Rahmen der Psychotherapiegruppe kamen zwei Psychotherapieverfahren zum Einsatz: Die Verhaltenstherapie und die tiefenpsychologisch fundierte Therapie. In ersterem Therapieverfahren waren 26 Psychotherapeutencluster ausgebildet. In zweiterem waren 2 Psychotherapeutencluster ausgebildet. Die Psychotherapieverfahren entsprachen der in der Versorgung vorgefundenen Verteilung in den eingebundenen Psychotherapiezentren. Durch die geringe Anzahl an Psychotherapeuten, welche in tiefenpsychologisch fundierter Therapie ausgebildet sind, wurden keine Subgruppenanalysen in Bezug auf die Psychotherapieverfahren durchgeführt.

Ergänzung:

Die Bildung der Psychotherapeutencluster erfolgte randomisiert. Zwei wissenschaftliche Mitarbeiterinnen der Professur für Sozial- und Präventivmedizin der Universität Potsdam nahmen die Randomisierung vor. Zuerst wurden Psychotherapeutenpaare nach Ort (Psychotherapiepraxis) und Psychotherapieverfahren gebildet. Person A entschied dann, welche Interventionsgruppe zuerst gezogen wird, und notierte dies schriftlich ohne Wissen von Person B. Weiter ordnete Person A Kopf und Zahl einer Münze jeweils einem Psychotherapeuten zu und notierte dies schriftlich ohne Wissen von Person B. Anschließend erfolgte der Münzwurf durch Person B. Das Ergebnis wurde anschließend in der Liste der Psychotherapeutencluster dokumentiert und den Psychotherapeuten abschließend mittgeteilt, ob sie der Vergleichsgruppe-Psychotherapie oder der Sporttherapiegruppe zugeordnet wurden.

3.1.2 Analyse des primären Outcome-Parameters

Bisher (S. 18, letzter Absatz):

In einem zweiten Modell (Anhang C, Tabelle S6, Modell 2) wurde die 6-monatige Nachbeobachtung (t4) miteinbezogen, um zu untersuchen, ob die Auswirkungen der Intervention 6 Monate nach Ende der Behandlung erhalten blieben. Es zeigte sich ein signifikanter Haupteffekt der Zeit (Tabelle S6), der auf eine Verbesserung des BDI-II Scores in beiden Gruppen von t0 zu t4 (7.97 [6.17 bis 9.77]; $p<0.001$)

und von t2 zu t4 (2.16 [0.90 bis 3.42]; p<0.001) hinweist. Es wurde weder ein Haupteffekt der Gruppe noch eine Interaktion zwischen Gruppe und Zeit festgestellt.

Ergänzung:

Numerisch (vgl. Tabelle S2 im Bericht, d. h. ohne Imputation und ohne Modellierung) ist der Unterschied zwischen beiden Gruppen zum Zeitpunkt T4 Null (0.00), unterschreitet aber die Non-Inferiority-Grenze von -3 im 95% Konfidenzintervall [-3.06 bis 3.06] knapp. Für den Messzeitpunkt T4 (6 Monate nach Intervention) zeigt im clusteradjustierten GEE Modell (vgl. korrigierte Tabelle S6 des Evaluationsberichtes) der Schätzer für den Interaktionseffekt (Gruppe und Zeit von T0 bis T4) eine um 1.17 Punkte im BDI-II geringere Symptomverbesserung durch die Sporttherapie im Vergleich zur Psychotherapie (-1.17 [-3.34 bis 1.00]) und das 95% Konfidenzintervall unterschreitet die Non-Inferiority-Grenze von -3. Zum Follow-up Zeitpunkt T4 (6 Monate nach Ende der Intervention) lässt sich somit anhand der relevanten Konfidenzintervalle (auf dem Niveau des 95% Konfidenzintervalls) keine Non-Inferiority im primären Outcome nachweisen. Weiterführende Analysen haben hier ergeben, dass das Non-Inferiority-Kriterium im primären Outcome BDI-II zu T4 hingegen auf dem 90% Konfidenzintervall, das jedoch nicht das übliche und vorab festgelegte Zielkriterium der Analyse war (vgl. Heissel et al., 2020), gegeben ist (-1.17; 90% KI -2.99 bis 0.65). Da das Inferiority-Kriterium der Sporttherapie zu T4 (nach 6 Monaten) ebenfalls nicht erfüllt ist (das Konfidenzintervall schließt die Null mit ein), lässt sich allerdings auch keine Unterlegenheit der Sporttherapie zeigen.

Für nichtadjustierte Analysen (s. S. 16 und S. 25 Evaluationsbericht) wurde zum Zeitpunkt T2 eine Fallzahl von 312 geplant, wovon n = 259, also 83% der geplanten Stichprobe erreicht wurden. In der last observation carried forward (LOCF) Analyse (die ursprünglich mit Clusteradjustierung geplant war) konnte aber eine Fallzahl von n = 334 berücksichtigt werden, die die ursprünglich für nichtadjustierte Analysen geschätzte Fallzahl überschreitet (vgl. Overall et al. 2009). Es wurden deshalb zusätzlich nichtadjustierte Analysen in der geplanten (vgl. Heissel et al., 2020) last observation carried forward (LOCF) Analyse vorgenommen. Hier zeigt der Interaktionseffekt (Gruppe und Zeit) für das primäre Outcome BDI-II von T0 bis T2 im GEE ohne Clusteradjustierung eine numerisch geringere Verbesserung durch die Sporttherapie im Vergleich zur Psychotherapie (-1.19 [-2.73 bis 0.35]), deren 95% Konfidenzintervall die Non-Inferiority-Grenze von -3 jedoch nicht unterschreitet. Für den Messzeitpunkt T4 (6 Monate nach Intervention) zeigt der Schätzer für den Interaktionseffekt (Gruppe und Zeit von T4 bis T0) eine um 1.15 Punkte im BDI-II geringere Symptomverbesserung durch die Sporttherapie im Vergleich zur Psychotherapie (-1.15 [-3.32 bis 1.03]) und das 95% Konfidenzintervall unterschreitet die Non-Inferiority-Grenze von -3 (s. Tabelle S6E in der Ergänzung). Somit können wir für das primäre Outcome für den Zeitpunkt T2 (direkt nach der Intervention) mit den nichtadjustierten Analysen die Nichtunterlegenheit der Sporttherapie zeigen, für den Zeitpunkt T4 (6 Monate nach der

Intervention) hingegen nicht mehr. Das Kriterium der Inferiority ist weder zu T2 noch zu T4 gegeben, so dass unsere Daten eine Unterlegenheit der Sporttherapie in den nichtadjustierten LOCF Analysen mit besserer Power auch zu T4 nicht belegen können. Grundsätzlich ist zu berücksichtigen, dass im STEP.De Projekt ein größerer Anteil an Patienten der Psychotherapie-Intervention (76.8%) im Vergleich zur Sporttherapie (21.9%) im Anschluss an die Intervention (nach T2) eine weiterführende Psychotherapie erhalten haben. Somit besteht auch die Möglichkeit, dass ein Dosis-Wirkungs-Zusammenhang diese Ergebnisse erklärt.

3.1.3 Analyse der sekundären Outcome-Parameter

Bisher (S. 19., 3. Absatz):

Für depressive Symptome, die mit dem HAM-D gemessen wurden, zeigten die mITT-Analysen einen signifikanten Effekt der Zeit vor vs. nach der Behandlung (3.95 [1.99 bis 5.91]; p<0.001) (Anhang C, Tabelle S7, Modell 3), was auf einen Rückgang der HAM-D Gesamtscores nach 16 Wochen Intervention für beide Gruppen hinweist. Bei Einbeziehung der 6-monatigen Nachbeobachtung in das Modell (t4) (Anhang C, Tabelle S8, Modell 4) wurde ebenfalls ein signifikanter Haupteffekt der Zeit beobachtet (5.96 [3.94 bis 7.99]; p<0.001), der eine Abnahme der depressiven Symptome vom Interventionsbeginn bis zur 6-monatigen Nachbeobachtung anzeigt.

Ergänzung:

Für das sekundäre Outcome HAM-D (Depressionssymptome erhoben durch verblindete Assessoren) zeigt im GLMM der Interaktionseffekt (Gruppe und Zeit von T0 bis T4) eine numerische Verbesserung der Sporttherapiegruppe gegenüber der Psychotherapiegruppe (-0.66 [-1.88 bis 3.19]), deren 95% Konfidenzintervall die Non-Inferiority-Grenze von -2.16 nicht unterschreitet (vgl. Tabelle S8 im Evaluationsbericht). Zum Zeitpunkt T4 (vgl. Tabelle S2) ist der Unterschied 0.20 (0.20 [-1.86 bis 2.25]), und das 95% Konfidenzintervall unterschreitet die Non-Inferiority-Grenze von -2.16 damit ebenfalls nicht. Für das sekundäre Zielkriterium lässt sich die Nichtunterlegenheit somit auf dem Niveau des 95% Konfidenzintervalls also auch für T4 zeigen.

4.1. Schlussfolgerungen des Evaluators

Bisher (S.24, 5. Absatz):

Darüber hinaus konnte mit dem Nicht-Unterlegenheits-Design gezeigt werden, dass die STEP.De Sporttherapie bei Depression der leitliniengerechten Erstlinientherapie (Psychotherapie) sowohl nach dem 16-wöchigen Interventionszeitraum als auch im 6-monatigen Nachbeobachtungszeitraum nicht

unterlegen war und damit eine gleichwertige Alternative in der Behandlung leichter bis mittlerer Depression zur Psychotherapie darstellen kann.

Erratum und Ergänzung:

Darüber hinaus konnte mit dem Nicht-Unterlegenheits-Design gezeigt werden, dass die STEP.De Sporttherapie bei Depression der leitliniengerechten Erstlinientherapie (Psychotherapie) im 16-wöchigen Interventionszeitraum nicht unterlegen war und damit eine gleichwertige Alternative in der Behandlung von Patienten aus dem Depressionsspektrum in der untersuchten Personengruppe zur Psychotherapie darstellen kann.

Im Nachbeobachtungszeitraum überschreitet das 95% Konfidenzinterval die Non-Inferiority-Grenze. Somit lässt sich im Nachbeobachtungszeitraum die Nichtunterlegenheit der Sporttherapie bei der a priori festgelegten Unsicherheitstoleranz für das primäre Zielkriterium (BDI) nicht nachweisen. Für das sekundäre Outcome (Hamilton Depressionsskala) ist die Nichtunterlegenheit aber auch nach 6 Monaten gegeben.

Bisher (S.25, 1. Absatz):

Insgesamt konnten so trotz des Erreichens der Einschlusszahlen zu Beginn der Studie zum Zeitpunkt nach der Intervention für die mITT Analyse nur 259 Probanden erhoben werden. Dies entspricht 83% der ursprünglich angestrebten Fallzahl von 312. Die für die Überprüfung der Interventionseffekte auf das primäre Zielkriterium verwendeten verallgemeinerten lineare Schätzungsgleichungen (GEE) gelten aber für längsschnittlichen Dropout als robust (Lin & Rodriguez, 2015).

Erratum und Ergänzung:

Zum Zeitpunkt T2 nach der Intervention standen für die mITT Analyse BDI-Daten von 259 Probanden zur Verfügung. Dies entspricht 67% der ursprünglich angestrebten Fallzahl von 384, in der eine Intraclusterkorrelation von 0.01 berücksichtigt wurde. Im Nachhinein erwies sich diese Annahme in der Fallzahlschätzung als zu gering, da die ICC für die Messzeitpunkte T0-T2-T4 mit 0.023 berechnet wurde. Die für die Überprüfung der Interventionseffekte auf das primäre Zielkriterium verwendeten verallgemeinerten lineare Schätzungsgleichungen (GEE) gelten zwar für längsschnittlichen Dropout als robust (Lin & Rodriguez, 2015). Angesichts des Anteils von Studienteilnehmern ohne gesicherte (2.03%) bzw. vorliegende Einschlussdiagnose (3.20%) sowie der korrigierten Fallzahlerreichung und der tatsächlichen Intraclustercorrelation ist die Wahrscheinlichkeit hoch, dass die angestrebte Power von 80% nicht erreicht wurde. Darüber hinaus könnte das gewählte Design mit mITT-Analyse und die geplanten LOCF Analysen Non-Inferiority begünstigt haben. In internationalen Studien wurden andererseits aber auch liberalere Non-Inferiority Kriterien als die von uns gewählten 0.3 Standardabweichungen berichtet (0.35*SD in Rhodes et al.; 0.39*SD in Richards et al.).

Bisher (S.25)

Des Weiteren sicherte der Psychotherapeut die Diagnose zwar entsprechend der Einschlussdiagnosen ab, war instruiert diese anhand des affektiven Moduls des SKID-I abzusichern und bestätigte therapeutischen Behandlungsbedarf. Trotzdem ergibt sich zu Beginn der Interventionen, dass ca. 20% der Teilnehmer einen BDI Wert von unter 14 Punkten aufweisen, was auf eine minimale bis keine Depression hinweist und damit auch eine weitere Reduktion der Symptomatik limitiert.

Ergänzung:

Zusätzlich zeigten zu T0 25 Teilnehmer (7.62%) einen BDI-Wert unter 9, der für die Abwesenheit depressiver Symptome spricht. Von diesen 25 hatten 2 keine Einschlussdiagnosen.

Bisher (S.27, 1. Absatz):

Insgesamt konnte in der Evaluation gezeigt werden, dass die neue Versorgungsform STEP.De zur Verbesserung der Behandlung von Patienten mit leichter bis mittlerer Depression in einer Cluster-randomisierten kontrollierten Nicht-Unterlegenheits-Studie beiträgt, vergleichbar mit der Standardbehandlung.

Erratum und Ergänzung:

Insgesamt konnte in der Evaluation gezeigt werden, dass die neue Versorgungsform STEP.De zur Verbesserung der Behandlung von Patienten mit Erkrankungen aus dem Depressionsspektrum in einer Cluster-randomisierten kontrollierten Nicht-Unterlegenheits-Studie beiträgt, vergleichbar mit der Standardbehandlung.

Diese Empfehlung der Umsetzung der Sporttherapieintervention bedarf aufgrund der geringen Power der adjustierten Analysen, der Unsicherheit in den Versorgungsdiagnosen und des genannten möglichen Selektionsbias einer Relativierung hinsichtlich dieser methodischen Limitationen.

Für nichtadjustierte Analysen mit besserer Power (s. Tabelle S6E) bleibt aber für das primäre Zielkriterium BDI-II der Effekt der Nichtunterlegenheit für den Interventionszeitraum (T0-T2) erhalten. Für das sekundäre Zielkriterium Hamilton Depressionsskala bleibt der Nichtunterlegenheitseffekt auch für den Zeitraum bis T4 (6 Monate nach der Intervention) erhalten.

Zudem findet sich für unsere Prüfhypothese 1 für das primäre Outcome BDI-II („nach 16 Wochen Intervention und bei der Nachbeobachtung nach 6 Monaten zeigt die Gesamtstichprobe eine signifikante Verbesserung der depressiven Symptomatik“) auch im 6-Monats Follow-Up in clusteradjustierten Analysen ein Effekt von 7.97 Punkten mit reliablem 95% Konfidenzintervall (6.17 - 9.77) ohne signifikanten Gruppenunterschied (p = .81; vgl. Tabelle S6), für den die genannten Einschränkungen der statistischen Power nicht zutreffen. Eine rezente Vergleichsstudie zwischen Sporttherapie und medikamentöser Therapie bei Depression aus den Niederlanden folgert allein aus

einem diesem Ergebnis ähnlichen Befund eine Vergleichbarkeit der Therapieverfahren (Verhoeven et al., 2023).

Durch die offene Gruppenzuteilung sowie die auch im Versorgungskontext zu erwartende Unsicherheit bezüglich der Diagnosen besteht eine gewisse Unsicherheit hinsichtlich der Generalisierbarkeit der Ergebnisse, die aber in Teilen bereits diskutiert wurde (vgl. S 25 Evaluationsbericht). Die im Evaluationsbericht gemachte Empfehlung für leichte und mittelschwere Depression ist hinsichtlich der beschriebenen Versorgungsdiagnosen aus dem Depressionsspektrum und verwandte psychiatrische Diagnosen (vgl. Ergebnisbericht S. 22) zu konkretisieren. Letztlich beruht die gegebene Empfehlung hinsichtlich ihrer Generalisierbarkeit nicht allein auf den Ergebnissen der STEP.De Intervention, sondern ist in umfangreiche metaanalytische Evidenz (Cooney et al., 2013; Heissel et al., 2023; Krogh et al., 2017; Schuch et al., 2016) und Leitlinien (z.B. S3-Leitlinie unipolare Depression) zur Sporttherapie bei depressiven Erkrankungen und Symptomen eingebettet. Aus diesen Erwägungen heraus bleibt unsere im Evaluationsbericht gegebene Empfehlung für Patienten aus dem Depressionsspektrum bestehen.

Es ist allerdings zu ergänzen, dass im Rahmen einer erweiterten Analyse des Follow-Up-Zeitraums im primären Outcome keine Nichtunterlegenheit nach gängigen Beurteilungsmaßstäben mehr vorliegt. Dies kann einerseits mit der geringeren statistischen Power bzw. den relativ großen Unsicherheiten im Konfidenzintervall zu tun haben. Dass die Fortsetzungsraten der Psychotherapie nach Abschluss der Intervention in beiden Gruppen signifikant unterschiedlich waren (in der Sporttherapiegruppe 21.9% Psychotherapien, in der Psychotherapiegruppe 76.8% fortgeführte Psychotherapien), könnte aber auch eine Erklärung für diese Ergebnisse im Langzeitverlauf sein. In zukünftigen Studien sollten deshalb die Dosis-Wirkungszusammenhänge noch umfangreicher geprüft werden.

Das Design der Sportintervention des Projektes in der von uns gegebene Evaluationsempfehlung sah bereits die Begleitung durch Psychotherapeuten zu Beginn und Ende der Sportintervention vor, um weiteren psychotherapeutischen Behandlungsbedarf zu identifizieren und bei Bedarf ein Behandlungsangebot zu machen.

Die Ergänzung der Empfehlung ist deshalb, eine Verlaufskontrolle zur Klärung eines weiteren möglichen Therapiebedarfs nicht nur nach Abschluss der Sporttherapie, sondern auch sechs bis zwölf Monate später anzubieten und in der Umsetzung in die Regelversorgung zu berücksichtigen.

Anhang C:

Erratum:

Die Modellschätzer in Zeile 12 der Tabelle S6 waren fälschlicherweise in Zeile 11 angeben. Zudem waren hier nicht clusteradjustierte Schätzer berichtet worden. Dies wurde wie folgt korrigiert:

Tabelle S6. GEE-Analysis mit BDI-II als abhängige Variable (Modell 2) in der mITT-Population (mit Clusteradjustierung)

	Koeffizient (B)	SE	Wald X^2	p-Wert
Konstante	17·92 (12·69 to 23 14)	2·67	45·162	**<0·001**
Zeit				
T0	7·99 (6·20 to 9·78)	0·92	76·177	**<0·001**
T2	2·16 (0·90 to 3·42)	0·64	11·322	**<0·001**
T4	Ref.			
Intervention				
Sporttherapie	0·02 (-2·63 to 2·66)	1·35	0·000	0·990
Psychotherapie	Ref.			
Intervention*Zeit				
Sporttherapie*T0	-1·17 (-3·34 to 1·00)	1·11	1·115	0·291
Psychotherapie*T0	Ref.	Ref.	Ref.	Ref.
Sporttherapie*T2	-1·19 (-2·72 to 0·35)	0·78	2·283	0·131
Psychotherapie*T2	Ref.			
Sporttherapie*T4	Ref.			
Psychotherapie*T4	Ref.			
Alter	-0·001 (-0·10 to 0·10)	0·05	0·001	0·980

BDI-II=Beck Depression Inventory II. mITT=modified intention to treat.
Quasi Likelihood Under Independence Model Criterion (QIC): 114217.739
Corrected QIC (QICC): 114211.1292

Ergänzung:

Zusätzlich haben wir die ursprünglich dargestellte Tabelle ohne Clusteradjustierung korrigiert und als Ergänzung (Tabelle S6E) aufgenommen.

Tabelle S6E. GEE-Analysis mit BDI-II als abhängige Variable (Modell 2) in der mITT-Population (ohne Clusteradjustierung)

	Koeffizient (B)	SE	Wald X^2	p-Wert
Konstante	15·21 (10·39 to 20 04)	2·48	38·145	**<0·001**
Zeit				
T0	7·97 (6·17 to 9·77)	0·92	75·276	**<0·001**
T2	2·16 (0·90 to 3·42)	0·64	11·322	**<0·001**
T4	Ref.			
Intervention				
Sporttherapie	0·32 (-2·32 to 2·96)	1·35	0·058	0·810
Psychotherapie	Ref.			
Intervention*Zeit				
Sporttherapie*T0	-1·15 (-3·2 to 1·03)	1·11	1·072	0·301
Psychotherapie*T0	Ref.	Ref.	Ref.	Ref.
Sporttherapie*T2	-1·19 (-2·72 to 0·35)	0·78	2·283	0·131
Psychotherapie*T2	Ref.			
Sporttherapie*T4	Ref.			
Psychotherapie*T4	Ref.			

BDI-II=Beck Depression Inventory II. mITT=modified intention to treat.
Quasi Likelihood Under Independence Model Criterion (QIC): 115707.781
Corrected QIC (QICC): 115704.300

5. Literaturverzeichnis

Ergänzung (S. 30)

Heissel, A., Heinen, D., Brokmeier, L. L., Skarabis, N., Kangas, M., Vancampfort, D., Stubbs, B., Firth, J., Ward, P. B., Rosenbaum, S., Hallgren, M., & Schuch, F. (2023). Exercise as medicine for depressive symptoms? A systematic review and meta-analysis with meta-regression. *British Journal of Sports Medicine*, bjsports-2022-106282. https://doi.org/10.1136/bjsports-2022-106282

Krogh, J., Hjorthøj, C., Speyer, H., Gluud, C., & Nordentoft, M. (2017). Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential analysis. *BMJ Open*, *7*(9), e014820.

Ergänzung (S. 32)

Overall, J. E., Tonidandel, S., & Starbuck, R. R. (2009). Last-observation-carried-forward (LOCF) and tests for difference in mean rates of change in controlled repeated measurements designs with dropouts. Social Science Research, 38(2), 492-503.

Schumi, J., & Wittes, J. T. (2011). Through the looking glass: understanding non-inferiority. Trials, 12, 1-12.

Ergänzung (S. 32)

Verhoeven, J. E., Han, L. K., Lever-van Milligen, B. A., Hu, M. X., Révész, D., Hoogendoorn, A. W., ... & Penninx, B. W. (2023). Antidepressants or running therapy: Comparing effects on mental and physical health in patients with depression and anxiety disorders. *Journal of Affective Disorders*, *329*, 19-29.