Abschlussbericht zum Vorhaben

Ein Beitrag zur Beschwerdenvalidierung in der Begutachtung psychischer Störungen – Prävalenz von Antwortverzerrungen und Validerung einer deutschen Version des Structured Interview of Reported Symptoms (SIRS-2).

Kennziffer: 412.02-FR223

Laufzeit
01.09.2014 – 31.08.2016

Bericht vom 09.11.2016

Autor:
Thomas Schmidt

BG Klinikum Bergmannstrost Halle
Abteilung Medizinische Psychologie
Merseburger Straße 165
D-06112 Halle (Saale)
Tel.: +49 345 132 74 79
E-Mail: Thomas.Schmidt@bergmannstrost.de

Inhaltsverzeichnis

Kurzfassung deutsch

In **Teilprojekt 1** zeigt die retrospektive Untersuchung eines Gutachtenpools (n = 1175) im Verlauf von 16 Jahren einen steigenden Aufwand zur Erstellung psychologischer Gutachten mit zunehmender Gutachtenlänge und -komplexität. Mit komplexer werdender Methodik werden häufiger Inkonsistenzen aufgeführt und eine höhere Rate verfälschter Beschwerdendarstellung festgestellt. Ein einheitliches und multimethodales Vorgehen gewährleistet, dass kein Bias-Effekt in der Beurteilung dieses Aspektes über verschiedene Gutachter hinweg entsteht. Entsprechend vorhandener Schätzungen in der Literatur finden sich im untersuchten Gutachtenpool häufig einzelne Inkonsistenzen (40.9%). Diese sind jedoch nicht gleichzusetzen mit Aggravation oder Simulation, es bedarf einer weiteren Gewichtung und Interpretation der Inkonsistenzen. Die hieraus eruierte Rate einer insgesamt eingeschränkten Beschwerdenvalidität von 15.8% entspricht aktuellen Reviews, die im Vorfeld deutlich höher eingeschätzte Raten kritisch hinterfragen. Es liegt damit erstmals eine Schätzung der Basisrate von Verfälschungen im Begutachtungskontext für die UV-Träger vor, die über einzelne Indikatoren oder Testwerte hinausgeht. Beschwerdenvalidierungsverfahren (BVT) werden als bedeutende Entscheidungsquelle von psychologischen Gutachtern als Indikator für Verfälschungen (in 79.6% der Fälle) genutzt. Für die Gesamtentscheidung werden jedoch zusätzlich weitere Inkonsistenzen berücksichtigt. Antwortverzerrungen in herkömmlichen BVT finden sich häufig (47.2%). Insgesamt besteht jedoch selten das Ausmaß einer bewussten, zielgerichteten Manipulation (8.2%).

In **Teilprojekt 2** wurde eine deutsche Version des SIRS-2 an einer Stichprobe im berufsgenossenschaftlichen Behandlungskontext (n = 92) validiert. Es wurde damit erstmals deutschsprachig ein BVT überprüft, das als standardisiertes vollstrukturiertes Interview konzipiert ist und eine spezifische Analyse des Antwortverhaltens erlaubt. Trotz der Besonderheiten der psychologischen Begutachtungspopulation für die UV-Träger (hohe Rate somatischer Komorbiditäten) zeigt das Verfahren im durchgeführten Simulationsdesign hohe Effektstärken zur Differenzierung authentischen und verfälschten Antwortverhaltens ($d_{gemittelt}$ = 2.36), eine hohe Interrater-Reliabilität ($r_{gemittelt}$ = .99) und Stabilität der Klassifikationen im Retest (mittlere Konkordanz 96.4%). Für die Praxis können eine Sensitivität von bis zu 93.2% und eine Spezifität von bis zu 97.9% angenommen werden. Das Verfahren ist in der untersuchten Stichprobe bisher eingesetzten, als spezifisch geltenden Tests (WMT) überlegen. Hervorzuheben ist die minimale falsch-positiv-Rate in der untersuchten Stichprobe bei der Beurteilung durch das SIRS-2 von 1%. Aufgrund seines Aufwandes ist die Kombination mit einem vorgeschalteten Screeningverfahren (z.B. SFSS) zu empfehlen. Ein nächster Schritt ist die Publikation in einem Testverlag, um das Verfahren für die Anwenderpraxis zur Verfügung zu stellen und an realen Begutachtungspopulationen überprüfen zu können. Die Diagnostikstandards werden hinsichtlich eines zentralen Aspektes der Begutachtung optimiert.

Kurzfassung englisch

In **sub-project 1**, the retrospective analysis of 1175 forensic psychological evaluations has shown a grown effort over the course of 16 years, together with the increasing length and complexity of the evaluations. With the use of increasingly complex methods, inconsistencies are reported more frequently and a higher rate of feigning and malingering is noted. A uniform multi-methods approach guarantees that there does not develop any bias in the assessment of this aspect across various evaluators. In agreement with existing estimations in the literature, the analyzed pool of evaluations shows many individual inconsistencies (40.9%). However, these should not be interpreted as malingering; further assessment and interpretation of these inconsistencies is necessary. A resulting rate of an overall rate of 15.8% of malingering matches recent reviews that call into question estimations which had yielded substantially higher rates. This represents the first estimation of a base rate of malingering in evaluation contexts for workers' compensation claim samples which goes beyond individual indicators or test scores. Used as an indicator of feigning (in 79.6% of the cases), symptom validity tests (SVT) serve as an important decision-making tool for psychological evaluators. For the overall assessment, however, other possible inconsistencies are taken into account. Among the common SVTs, negative response bias is especially frequent in screening measures (47.2%). Yet overall, they seldom reach the level of deliberate, purposeful manipulations (8.2%).

For **sub-project 2**, we were able to validate a German version of the SIRS-2 and thus, to assess for the first time in German-speaking contexts a SVT that was designed as a standardized fully structured interview and allows for a specific analysis of response styles. Despite the particularities of the evaluated population of workers' compensation claim samples (high rates of somatic co-morbidities), the measures yielded in the performed simulation design large effect sizes ($d_{average}$ = 2.36) for the discrimination between genuine and feigned response styles, high inter-rater reliability ($r_{average}$ = 0.99), and stable classifications in the retest (medium concordance of 96.4%). For application in practice, sensitivity values of up to 93.2% and specificity values of up to 97.9% can be assumed. With this, this measure is superior to tests classified as specific and used up till now (WMT). Notably, the number of false positives of the analyzed sample was minimal (1%). Due to its time-consuming and complex application, it is recommendable to precede this test with a screening measure (SFSS). A further step will be its publication with a test publisher in order to make the measure available to practitioners and for testing it with real-life populations. This project adds to recent developments and enhances the diagnostic standards for a central aspect of forensic psychological evaluation.

1. Problemstellung

Ein bedeutsamer Anteil von Arbeitsunfällen kann unmittelbare oder sich im Verlauf entwickelnde psychische Folgen nach sich ziehen. Nach klinischer Erfahrung in der antragstellenden Einrichtung sind etwa 20% der Unfallverletzten in störungsrelevantem Ausmaß betroffen (Schulz & Ullman, 2006). Gegenüber den häufig gut objektivierbaren körperlichen Gesundheitsstörungen erfordert die Begutachtung psychischer Erkrankungen die Beurteilung nicht direkt beobachtbarer innerer Vorgänge. Dabei kann die von außen sichtbare Beschwerdendarstellung durch viele Faktoren beeinflusst werden: Durch die Erkrankung selbst (z.B. durch Stimmungsbeeinträchtigungen, Ängstlichkeit, Scham), durch Interaktionsschwierigkeiten mit dem Gutachter (z.B. durch Misstrauen, negative Vorerfahrungen), aber auch bewusst, z.B. um eine finanzielle Absicherung zu erhalten. Führen solche Verzerrungen zielgerichtet zu einer verstärkten Darstellung oder zum Vortäuschen von Beschwerden (Aggravation, Simulation), können tatsächlich vorliegende Beeinträchtigungen nicht valide abgegrenzt werden. In der Regel führt dies in der Leistungsprüfung zu einer Ablehnung von Versicherungs- bzw. Rentenleistungen.

Aufgrund methodischer Probleme sind nur Schätzungen über die Häufigkeit bewusst vorgetäuschter Beschwerden verfügbar, sodass die Zahlen je nach angewandter Operationalisierung der Studien variieren. Wie häufig solche Verfälschungen im Begutachtungsprozess in Deutschland vorkommen, ist noch unklar (Schmidt et al., 2011).

Die Beantwortung der Frage, ob berichtete psychische Beschwerden authentisch sind, stellt Gutachter also regelmäßig vor Herausforderungen. Im psychiatrisch-nervenärztlichen bzw. psychologischen Gutachten kommt dabei dem psychischen Querschnittsbefund zentrale Bedeutung zu. Hier erfolgt eine sorgfältige Erhebung der berichteten Beschwerden unter kritischer Würdigung weiterer Datenquellen. In einem diagnostischen Prozess werden hieraus vom Gutachter störungsspezifische Charakteristika und die vorliegenden Beeinträchtigungen abgeleitet. Der gutachtliche Entscheidungsprozess ist aber bisher wenig untersucht. Ohne Berücksichtigung standardisierter Methoden werden in solchen komplexen Entscheidungsprozessen häufig nur allgemeine Kriterien angewandt, die nur begrenzt zu einem validen oder reliablen Urteil führen (Dreßing & Foerster, 2010). Die Beurteilung der Beschwerdenauthentizität allein nach dem klinischen Eindruck ist zudem kaum besser als der Zufall (Hall & Poirier, 2001). Nach allgemeinem Konsens sollte daher eine Plausibilitäts- und Konsistenzprüfung verschiedener erhobener Datenquellen erfolgen (z.B. Inkonsistenzen in der Beschwerdenschilderung, Verhaltensbeobachtung, Aktenlage und fremdanamnestischen Angaben, klinischer Befund, Wissen über die entsprechende Störung, Profil testpsychologischer Untersuchung, Gegenübertragungsphänomene).

Ein solcher Punkt der „Beschwerdenvalidierung“ wird auch von den verschiedenen Fachgesellschaften und in den aktuellen Richtlinien (z.B. AWMF-Leitlinie, 2012) als notwendig

erachtet. Vorgeschlagen wird ein multimethodales Vorgehen, um verschiedene Fehlerquellen zu minimieren. Demnach stellen auch testpsychologische Verfahren einen wichtigen Zugang bei der Begutachtung dar. Aufgeführt werden in der Leitlinie auch Beispiele für mögliche einzusetzende spezielle Beschwerdenvalidierungsverfahren (BVT).

Im deutschsprachigen Raum liegen bisher jedoch nur wenige spezifische Beschwerdenvalidierungsverfahren vor, die entweder Screeningcharakter haben oder für den neuropsychologischen Bereich entwickelt wurden. Hieraus entstehen jeweils methodische Limitierungen für den Einsatz bei psychischen Erkrankungen. Als Verfahren, das diese Kritikpunkte und Schwachpunkte anderer Tests berücksichtigt, ist das US-amerikanische Structured Interview of Reported Symptoms (SIRS-2, Rogers et al., 2010) zu nennen, das in der Literatur häufig als „Goldstandard" für die Fragestellung der Beschwerdenvalidierung genannt wird. Es wurde ursprünglich für psychiatrische Störungen (z.B. psychotische Erkrankungen) entwickelt. Im Verlauf hat es sich jedoch auch für den Einsatz „weicherer" psychischer Störungen (z.B. Anpassungsstörungen, depressive Erkrankungen, Traumafolgestörungen) bewährt (Rogers et al., 2009, 2010), die in der Begutachtung für die UV-Träger von Bedeutung sind. Eine einsetzbare deutsche Version existiert bisher nicht. Eine wissenschaftliche Überprüfung ist lohnenswert.

2. Forschungszweck/-ziel

In einem *ersten Teilprojekt* soll eine für die UV-Träger bereits erstellte Gutachtenpopulation nach der Häufigkeit eingeschätzter nichtauthentischer Beschwerden und deren zugrunde liegender Kriterien analysiert werden. Hierdurch kann erstmals eine aussagekräftige Schätzung der Basisrate möglicher Verfälschungen im Begutachtungskontext erfolgen, die über die Aussage einzelner Indikatoren oder Testwerte hinausgeht. Es wird eruiert, ob mit einer relevanten Häufigkeit zu rechnen ist und welche der wissenschaftlich empfohlenen Indikatoren am aussagekräftigsten sind.

Die in Deutschland bisher einsetzbaren BVT speziell für den Bereich psychischer Störungen entsprechen nicht den wissenschaftlichen Möglichkeiten und Anforderungen (insbes. hohe Spezifität), woraus sich ein substanzieller Anteil der geäußerten Kritik ergibt. Entwicklungen, die z.B. als Fremdbeurteilungsverfahren konstruiert wurden und weitere Kritikpunkte berücksichtigen, existieren bisher nur für den englischen Sprachgebrauch. Als Verfahren ist das US-amerikanische Structured Interview of Reported Symptoms (SIRS-2) hervorzuheben, das noch nicht für den deutschen Sprachraum validiert wurde.

In einem *zweiten Teilprojekt* soll daher das SIRS-2 in einer deutschen Version an einer Patientenpopulation mit verifizierten psychischen Erkrankungen (jedoch ohne laufende Begutachtungs- bzw. Berentungsverfahren) in einem „Simulationsdesign" validiert und auf seine

Praxistauglichkeit überprüft werden. Eine reale Begutachtungspopulation scheidet aus ethischen Gründen aus. Die Klassifikationsgüte des SIRS-2 soll mit der von herkömmlichen BVT verglichen werden, um eine differenziertere Einschätzung bzw. Gewichtung dieser Verfahren zu ermöglichen.

Das Forschungsfeld kann in das festgelegte Forschungsziel „(Akut)Trauma – psychosoziale Traumafolgen“ des Klinikverbundes der BG-Kliniken eingeordnet werden. Das Projekt dient der Optimierung eines zentralen Aspektes der Begutachtung psychischer Störungen (Beschwerdenvalidierung) und entsprechend auch der Leistungsprüfung. In der Anwendung soll die Beurteilung von Unfallfolgen methodisch ausgebaut und für alle Beteiligten (Antragsteller, UV-Träger, Sachverständige, Gerichte) transparenter gestaltet werden. Nach den Kriterien zur Anerkennung von Begutachtungen in der Rechtspraxis erweitert dies die wissenschaftlich-medizinische Sachkunde und Aktualität.

Die Kompetenzen der UV-Träger hinsichtlich der Analyse und Aufbereitung von Diagnostikverfahren in Sachverständigengutachten werden ausgebaut.

Hinsichtlich der Praxisverknüpfung sollen sowohl authentisch vorliegende Beeinträchtigungen, als auch die Einschätzung von Verfälschungen valider gesichert werden. Die Untersuchungen dienen als Pilotprojekt, um hieraus bei dann validiertem Untersuchungsinstrument ggf. ein multizentrisches Vorgehen zur Optimierung der Begutachtungsstandards für die UV-Träger abzuleiten.

3. Methodik

3.1 Arbeits- und Zeitabläufe

Anhang 1 zeigt schematisch den zur Antragstellung geplanten Verlauf sowie die Abweichungen, die bei der Umsetzung des Projektes entstanden sind. Die Abweichungen werden untenstehend erläutert:

Arbeitspaket 0:

A1 Sitzungen/Treffen mit Kooperationspartner:

Durch das Ausscheiden des Studienleiters im Kooperationszentrum Basel (Dr. med. Lanquillon) konnten die geplanten unabhängigen Stichprobenuntersuchungen dort nicht erfolgen. Die methodische Zusammenarbeit (Erstellung und Herausgabe des Testverfahrens, Datenauswertung und –interpretation) mit dem Department für Klinische Psychologie und Psychiatrie/ Psychodiagnostik der Universität Basel (Prof. Dr. R. D. Stieglitz) erfolgt weiterhin. Hierfür waren aber keine gesonderten Treffen erforderlich. Um die im Projektantrag aufgeführten Stichprobengrößen zu erzielen, wurde eine Kooperation mit der Klinik für Psychiatrie,

Psychotherapie und Psychosomatik des Universitätsklinikums Halle (PD Dr. Dr. St. Watzke, siehe auch Kooperationsanzeige im Anhang 2) vereinbart (Beginn 07/2016, Projektmonat 12).

Arbeitspaket 1 (Retrospektive Gutachtenauswertung):

A1 Vorbereitung der Untersuchung/ Einarbeitung Hilfskraft/ Erstellung Studienprotokoll/ Auswertungsmatrix:

Im BG-Klinikum Bergmannstrost erfolgte die Einstellung der studentischen Hilfskräfte verzögert ab Projektmonat 3 (verwaltungstechnische Gründe, Schwierigkeiten bei der Auswahl geeigneter Bewerber) und konnte bis zum Projektmonat 12 auch nicht im gewünschten Umfang erfolgen (Auslastung ca. 80%). Zur Kompensation erfolgte ein verstärkter Personaleinsatz. Durch zwischenzeitliches Ausscheiden studentischer Hilfskräfte aus dem Projekt waren mehrfach Einarbeitungsphasen erforderlich (Projektmonate 12, 15).

A2 Datenerhebung/ A3 Datenkonsolidierung und –analyse:

Durch die o.g. Verzögerungen konnte die im Projektantrag bis 08/2015 geplante Datenerhebung im Teilprojekt 1 (Gutachtenauswertung) nur zu ca. 70% erfolgen. Nach Ausweitung des Zeitraumes bis 05/2016 wurde der geplante Datenpool (Gutachten der Jahre 2000-2014) vollständig erfasst. Zusätzlich konnten durch die Erweiterung des Untersuchungszeitraumes die im Jahr 2015 fertiggestellten Gutachten (n=124) mit eingeschlossen und die zur Verfügung stehende Stichprobe entsprechend erweitert werden.

Arbeitspaket 2 (Patientenuntersuchung):

A1 Vorbereitung der Untersuchung/ Einarbeitung Hilfskraft/ Erstellung Studienprotokoll/ Auswertungsmatrix/ A2 Beginn der Patientenrekrutierung:

Durch die o.g. genannten Verzögerungen begann der Patienteneinschluss ebenso verzögert. Die Patientenuntersuchung wurde daher bis zum Ende des Förderungszeitraumes ausgeweitet. Das Rekrutierungsziel (n=100) wurde zu 92% erreicht.

A3 unabhängige Patientenuntersuchung (UKH)/ A5 Zusammenführung der Daten mit d. UKH:

Durch die neue Forschungskooperation für die ergänzenden Stichprobenuntersuchungen war der ursprüngliche Zeitplan für die Patientenuntersuchung durch den Kooperationspartner hinfällig. Der Beginn der unabhängigen Stichprobenuntersuchungen konnte hier erst ab Dezember 2015 erfolgen. Der Untersuchungszeitraum ist hier bis November 2016 vorgesehen. Zusammenführende Analysen der Stichproben können daher erst im Anschluss und außerhalb des Projektzeitraumes erfolgen.

3.2 Vorarbeiten

Vor Beginn des Förderungszeitraumes wurde eine deutsche Version des SIRS-2 zusammen mit einem früheren Kooperationspartner (Dr. med. Stefan Lanquillon, Schweiz) nach den notwendigen wissenschaftlichen Maßstäben (forward-backward-Translation) unter Einbezug billingualer Übersetzer erstellt. Eine Autorisierung dieser Version und Genehmigung zur Validierung an deutschen Stichproben wurden vom Erstautor der Originalversion (Prof. R. Rogers, USA) und den Rechteinhabern (PAR inc., USA/ Hogrefe Verlag, Schweiz) eingeholt. In einer Übersichtsarbeit (Schmidt et al., 2011) wurde nach entsprechenden Literaturreviews eine ausführliche und kritische IST-Analyse zu Möglichkeiten von Beschwerdenvalidierungsverfahren erarbeitet und der theoretische Rahmen für die geplante Untersuchung bereitet. Ein positives Votum der zuständigen Ethikkomission (Medizinische Fakultät Martin-Luther-Universität Halle-Wittenberg, 19.10.2012) zur geplanten Untersuchung wurde eingeholt.

3.3 Teilprojekt 1 (Gutachtenauswertung)

Im Studienzentrum BG-Klinikum Bergmannstrost (Halle) erfolgte i.R. des ersten Teilprojektes die retrospektive Analyse und elektronische Datenerfassung von Gutachten zur Zusammenhangsfrage bei psychischen Gesundheitsschäden (Gesamtstichprobe n=1175/ Jahre 2000-2015). Neuropsychologische Gutachten wurden aufgrund einer differierenden Struktur und Kausalitätsbeurteilung nicht berücksichtigt. Die Häufigkeit einer gutachterlich als eingeschränkt beurteilten Beschwerdenauthentizität wurde erfasst. Zudem, welche Datenquellen (u.a. Inkonsistenzen in der Beschwerdenschilderung, in der Verhaltensbeobachtung, in der Aktenlage und fremdanamnestische Angaben, klinischer Befund, Wissen über die entsprechende Störung, Profil genereller testpsychologischer Untersuchung, spezielle BVT, Gegenübertragungsphänomene) den Gutachter zu der Entscheidung gebracht haben. Hierfür wurde ein Kategoriensystem entwickelt. Ein Auszug der zu beurteilenden Variablen findet sich zur Illustration im Anhang 3 (Kategoriensystem zur Inhaltsanalyse des Gutachtenpools aus Teilprojekt 1). Unterschiede zwischen verschiedenen Gutachtern in Bezug auf die gutachterlichen Entscheidungen (Abhängigkeit von der Expertise) wurden eruiert. Zudem wurden Unterschiede bzgl. der gutachterlichen Entscheidungen in Abhängigkeit von Veränderungen der Gutachtengestaltung im Zeitverlauf eruiert. Das abgeleitete Kategoriensystem zur Datenreduktion und -erfassung zeigte sich als handhabbar und erlaubte eindeutige Zuordnungen. Der veranschlagte Zeitaufwand (ca. 2h/ Gutachtenakte) war ebenfalls realistisch.

Auf Basis der deskriptiven Daten erfolgten Häufigkeitsanalysen zur Basisrate nicht-authentisch eingeschätzter Beschwerden und zu Auftretenshäufigkeiten zugrunde liegender Kriterien in der gutachterlichen Beurteilung. Regressionsanalytisch sollten Gewichte der Indikatoren zur Beschwerdenvalidierung ermittelt werden. Mögliche Einflussgrößen auf die Häufigkeit nichtauthentisch-eingeschätzter Beschwerden wurden ermittelt (z.B. Expertise des Gutachters,

Veränderung der Gutachtenstruktur und Einführung erweiterter Messmethoden über die Zeit). Explorative Analysen dienten z.B. der Bewertung herkömmlicher BVT in der untersuchten Gutachtenpopulation (Inwieweit führt eine testpsychologisch eruierte Antwortverzerrung auch zur gutachterlichen Einschätzung einer generell eingeschränkten Beschwerdenauthentizität?).

Zusammenfassung Methodik Teilprojekt 1: Retrospektive deskriptive Auswertung von bereits fertiggestellten Gutachten in der antragstellenden Einrichtung 2000-2015.
Einschlusskriterien: Alle fertiggestellten Psychologischen Gutachten der antragstellenden Einrichtung bis 2015.
Ausschlusskriterien*:* Neuropsychologische Gutachten.
Primäre Zielkriterien: Häufigkeit als nicht-authentisch eingeschätzter Beschwerden in der Gutachtenpopulation.
Sekundäre Zielkriterien*:* Evaluierung von Zusammenhängen mit den zugrunde liegenden Kriterien und deren Gewichtung (Inkonsistenzen in der Beschwerdenschilderung, in der Verhaltensbeobachtung, in der Aktenlage und fremdanamnestischen Angaben, klinischer Befund, Wissen über die entsprechende Störung, Profil genereller testpsychologischer Untersuchung, spezielle BVT, Gegenübertragungsphänomene). Zusammenhänge zu Veränderungen der Gutachtenmethodik über die Zeit.

3.4 Teilprojekt 2 (Patientenuntersuchung)

In Teilprojekt 2 wurde eine experimentelle Untersuchung zur Validierung der erstellten deutschen Version des Beschwerdenvalidierungsverfahrens SIRS-2 geplant. In der Abteilung Medizinische Psychologie des BG-Klinikums Bergmannstrost Halle wurden hierfür Patienten mit bereits gesicherten psychoreaktiven Störungen rekrutiert, um der potentiellen Begutachtungspopulation für die der Einsatz des SIRS-2 bestimmt ist, möglichst nahe zu kommen (externe Validität). Bei Einwilligung in die Studienbedingungen erfolgte die Untersuchung durch zwei unabhängige Versuchsleiter. Der genaue Ablauf der Untersuchung sowie eine Beschreibung der verwendeten Testverfahren findet sich zur Illustration in Anhang 4.

Im Experiment wurden die Patienten dabei angehalten, ihre tatsächlich vorhandenen psychischen Beschwerden entweder authentisch (Kontrollgruppe) oder übertrieben (Experimentalgruppe) zu schildern. Es wurde dadurch eine bewusste Antwortverzerrung, nach dem DSM-5 also Simulation, induziert. Die Zuordnung zu den Untersuchungbedingungen erfolgte randomisiert (Losverfahren). Der Untersuchungsleiter, der das SIRS-2 und die weiteren BVT durchführt, war für die Untersuchungsbedingung verblindet. Dieses sog. „Simulationsdesign“ (interne Validität) findet in der Forschungslage für diese Fragestellungen am häufigsten Anwendung (Rogers et al., 2008) und wurde auch für die Originalversion des SIRS-2 und die spanische Adaptation (Correa et al., 2010) eingesetzt. Hierdurch wird zudem eine Vergleichbarkeit der Daten gewährleistet. Der Zeitaufwand für die Untersuchung eines Probanden betrug etwa 2-3 Zeitstunden. Die

Auswertungszeit der Testverfahren betrug insgesamt ca. 1 Zeitstunde. Dabei war die Testbatterie prinzipiell im vorgesehenen Zeitrahmen für jeden Patienten einsetzbar.

Zur Überprüfung der Retestreliabilität wurde das SIRS-2 mit einer Subgruppe der Patienten der authentischen Untersuchungsbedingung (n=25) nach einem Zeitintervall (ca. 1 Woche – 4 Wochen) erneut durchgeführt. Zur Überprüfung der Interraterreliabilität wurde bei einer Subgruppe (n=36) eine Videoaufnahme der Untersuchung von einem zweiten Interviewer ausgewertet, der die entsprechenden Antworten und das Verhalten der Probanden unabhängig vom ersten Interviewer protokollierte.

Es sollte eruiert werden, wie gut das SIRS-2 und die weiteren eingesetzten Testverfahren zwischen den beiden Untersuchungsbedingungen unterscheiden können. Im Einzelnen erfolgten Analysen:

- der Klassifikationsgüte (Sensitivität, Spezifität, falsch positiv-Rate, falsch negativ-Rate, Effektgrößen über die verschiedenen Skalen)
- der internen Konsistenz der Skalen des SIRS-2
- der Retestreliabilität bei wiederholter Vorgabe des SIRS-2
- der Interraterreliabilität bei unabhängiger Protokollierung des SIRS-2
- der Kriteriumsvalidität (Zusammenhänge zu anderen Beschwerdenvalidierungsverfahren)
- weiterführende explorative Analysen z.B. zu Zusammenhängen der Basisdiagnostik zur Untersuchungsbedingung und zum SIRS-2.

Zusammenfassung Methodik Teilprojekt 2: Experimentelle Datenerfassung bei Patienten mit psychischen Erkrankungen. Randomisierte Zuordnung der Untersuchungsbedingung (authentische vs. nicht-authentische Beschwerdenschilderung). Einfache Verblindung des Untersuchungsleiters für die Untersuchungsbedingung. Die Bedingung der authentischen Beschwerdenschilderung dient als Kontrollbedingung.
Einschlusskriterien: Alle volljährigen Patienten der antragstellenden Einrichtung bei denen eine ambulant behandelbare psychische Störungslage vorliegt. Einschluss der Patienten konsekutiv ab Studienbeginn.
Ausschlusskriterien: Schwere psychische Störung, die die Einwilligungsfähigkeit beeinträchtig und die Untersuchung nicht möglich macht (Störungen der Orientierung, des Bewusstseins, psychotische Störungen, schwere Antriebsstörungen); keine deutsche Muttersprache, kein Einverständnis zur Teilnahme, laufendes Begutachtungs- oder Berentungsverfahren.
Primäre Zielkriterien: Klassifikationsgüte des deutschen SIRS-2 zwischen authentisch und nicht-authentisch geschilderten psychischen Beschwerden zu unterscheiden.
Sekundäre Zielkriterien: Evaluierung von Zusammenhängen mit Ergebnissen weitere verwendeter Testverfahren (WMT/ SFSS), zur kognitiven Leistungsfähigkeit, allgemeiner Symptomatik, zur Schwere der tatsächlichen Beschwerden.

Hinsichtlich der Rekrutierung von Patienten war im Antrag zur Förderung ursprünglich ein Fokus auf Ambulanzpatienten vorgesehen. Neben „Akutfällen“ frühzeitig nach einem Unfall werden hier im bg-lichen Kontext aber vor allem chronifizierte und langanhaltende Unfallfolgen behandelt. Für das Projekt zeigten sich in diesem Zusammenhang Limitierungen durch einen unerwartet hohen Anteil von Patienten mit anstehendem oder laufendem Begutachtungsprozess zur Bewertung der Unfallfolgen (Ausschlusskriterium). Diesbezüglich erfolgte eine Ausweitung auf stationäre Patienten in den somatischen Fachkliniken der antragstellenden Einrichtung (mit zusätzlicher psychischer Störungslage). Diese Patienten befinden sich oft in einem früheren Status des Heilverfahrens oder im Rehabilitationsprozess, in dem eine Begutachtung oft noch keine Rolle spielt. Hierdurch konnte im Verlauf ein flüssigerer Patienteneinschluss erreicht werden.

Insgesamt erfüllten 130 Patienten die Einschlusskriterien für eine Studienteilnahme. Hiervon willigten 31 Patienten (23,8%) nicht in eine Studienteilnahme ein. Neben logistischen Schwierigkeiten (Anreise, terminliche Einteilung) wurde als Grund vor allem die mögliche Teilnahme in der „nichtauthentischen“ Gruppe angegeben. Die entsprechenden Patienten äußerten ein Unbehagen, bzw. subjektiv das Unvermögen, bewusst Beschwerden zu übertreiben oder zu simulieren (kognitive Dissonanz). Von den verbliebenen 99 Studienteilnehmern (76,2%) zeigten weitere 7 (5,4%) Patienten eine ungenügende Adhärenz für die Untersuchungsinstruktionen (Unvermögen, über den gesamten Untersuchungszeitraum nicht-authentisch bzw. übertrieben zu antworten). Diese Datensätze wurden für die weiteren Analysen ausgeschlossen.

Durch den ursprünglichen Kooperationspartner der Universität Basel erfolgten keine Stichprobenuntersuchungen. Im Rahmen einer neuen Forschungskooperation (Universitätsklinikum Halle) wurde nach der Untersuchungsplanung, logistischer Vorbereitung und Einholung eines positiven Ethikvotums mit der Rekrutierung von Patienten der psychiatrischen Institutsambulanz begonnen (12/2015). Der Abschluss der Patientenuntersuchungen ist für 11/2016 vorgesehen. Der Kooperationspartner liefert unabhängig Daten, die über den Projektabschluss hinaus eine größere Datenbasis (über DGUV-relevante Bereiche hinaus) und die allgemeine Anwendung des Testinstrumentes SIRS-2 in Fachkreisen sichern.

4. Ergebnisse des Gesamtvorhabens

Als Vorarbeiten zum Projekt bzw. im Zeitraum der Antragsstellung wurden folgende projektassoziierte Publikationen und Kongressbeiträge erarbeitet:

Schmidt, T., Lanquillon, S. & Ullmann, U. (2011). Kontroverse zu Beschwerdenvalidierungsverfahren bei der Begutachtung psychischer Störungen. Forensische Psychiatrie, Psychotherapie, Kriminologie, 5, 177-183.

Lanquillon, S. & Schmidt, T. (2012). Beschwerdevalidierung qualitativer psychischer Symptome bei zivil- und strafrechtlicher Begutachtung. In: N. Saimeh. (Hrsg.), Respekt – Kritik – Entwicklung: Therapeutische Aspekte im Maßregelvollzug. Forensik 2012. Bonn: Psychiatrie Verlag. S 137-150.

Lanquillon, S. (2013) - German-language validation of the Structured Interview of Reported Symptoms (SIRS) – Vortrag - Third European Symposium of Symptom Validation Würzburg, 2013.

Lanquillon, S. & Schmidt, T. (2013) - Reported mental health problems - a practical approach to assess their symptom validity – Workshop - Third European Symposium of Symptom Validation Würzburg, 2013.

Aus dem Berichtszeitraum 09/2014-08/2016 liegen bisher keine projektbezogenen Publikationen vor. Diese werden erst nach Abnahme des Abschlussberichtes auf Basis der unten stehenden Ergebnisse (Punkte 4.1; 4.2) erstellt. Vorgesehen sind Publikationen in Fachzeitschriften im Peer-Review-Verfahren sowie die Publikation des SIRS-2 beim deutschsprachigen Rechteinhaber (Hogrefe Verlag, Bern, Schweiz).

4.1 Teilprojekt 1 (Gutachtenauswertung)

In den ausgewerteten Gutachtenpool des Studienzentrums BG-Klinikum Bergmannstrost der Jahre 2000-2015 flossen insgesamt 1175 Gutachten zur Zusammenhangsfrage bei psychischen Gesundheitsstörungen ein.

In Anhang 5 findet sich eine Übersicht über die deskriptiven Kennwerte. Tabelle 1 zeigt den Anstieg erstellter Gutachten im erfassten Zeitraum. Auch die Gutachtenlänge und -komplexität, gemessen anhand der Seitenanzahl und der beantworteten Fragestellungen nehmen über die Zeit signifikant zu (ANOVA: $F_{\text{Seitenzahl}}(15, 1159) = 33.36$, $p<.001$; $F_{\text{Fragestellungen}}(15, 1159) = 9.48$, $p<.001$). Post hoc-Einzelvergleiche (Games-Howell-Tests) zeigen für die Seitenzahl einen signifikanten Anstieg der Seitenzahlen ab dem Jahr 2003 ($p = .002$), ausgehend vom Jahr 2003 ab 2009 ($p = .001$) und ausgehend vom Jahr 2009 ab 2012 ($p = .003$). Hinsichtlich der

Fragestellungen zeigen sich entsprechend signifikante Anstiege ab 2005 (p = .026) und 2013 (p < .001). Eine Zunahme der Gutachtenkomplexität bildet sich auch methodisch-inhaltlich ab. So erfolgt ab dem Jahr 2006 durchgängig eine explizite Argumentation zur Kausalitätsprüfung, ab dem Jahr 2007 findet sich zusätzlich durchgängig eine Beurteilung der Beschwerdenvalidität (jedoch ohne explizite Struktur oder Messverfahren), ab dem Jahr 2009 werden beginnend Beschwerdenvalidierungsverfahren eingesetzt, ab dem Jahr 2010 finden sich diese durchgängig.

Tabelle 1: Gutachtenanzahl, Seitenzahl, Fragestellungen 2000-2015

Jahr	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*
Anzahl, n	17	22	16	33	51	61	68	75
Seitenzahl,	19,0	20,3	24,0	25,9	26,3	26,2	23,5	26,0
M (SD)	(4,8)	(5,4)	(2,8)	(4,6)	(6,2)	(6,2)	(5,5)	(7,5)
Fragenzahl,	4,9	4,5	6,1	5,9	5,8	8,9	8,3	9,2
M (SD)	(3,4)	(1,2)	(2,9)	(2,7)	(2,5)	(4,5)	(6,0)	(6,4)
Jahr	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*
Anzahl, n	85	68	81	104	110	134	126	124
Seitenzahl,	28,5	30,9	33,3	34,4	35,5	34,6	35,6	35,7
M (SD)	(7,6)	(6,1)	(8,7)	(7,4)	(7,8)	(7,5)	(7,0)	(8,5)
Fragenzahl,	9,6	9,2	10,8	10,4	10,2	14,2	14,4	14,8
M (SD)	(5,3)	(5,7)	(8,6)	(8,6)	(8,8)	(10,8)	(11,5)	(11,6)

M = Mittelwert; n = Anzahl; SD = Standardabweichung

Prävalenz eingeschränkter Beschwerdenvalidität und Zusammenhang zur verwendeten Methodik

Hinsichtlich des primären Zielkriteriums wurde insgesamt in 145 Gutachten (12.4%) eine eingeschränkte Beschwerdenvalidität (BV) im Sinne einer negativen Verfälschung (Aggravation/ Simulation) eingeschätzt. Über den Untersuchungszeitraum zeigt sich dabei folgende Verteilung (Tabelle 2):

Tabelle 2: Einschätzung einer eingeschränkten Beschwerdenvalidität 2000-2015

Jahr	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*
Anzahl, n (%)	0	3	2	3	4	2*	4	4
	(0.0)	(13.6)	(12.5)	(9.4)	(7.8)	(3.3)	(5.9)	(5.3)
Jahr	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*
Anzahl, n (%)	7	13	14	8	12	23	28*	18
	(8.2)	(19.1)	(17.3)	(7.7)	(10.9)	(17.2)	(22.2)	(14.5)

n = Anzahl; * = statistisch signifikante Unterschiede zwischen beobachteten und erwarteten Häufigkeiten bei Betrachtung der standardisierten Residuen

Bei stark unterschiedlichen Stichprobengrößen (siehe Tabelle 1) zu Beginn des Untersuchungszeitraumes im Vergleich zu späteren Jahren (insbes. ab 2011) zeigen lediglich die Jahre 2005 (geringere Anzahl eingeschränkter BV als erwartet) und 2014 (höhere Anzahl eingeschränkter BV als erwartet) statistisch signifikante Unterschiede in den eruierten Häufigkeiten (χ^2 (15, N = 1174) = 37.40, p = .001; Tabelle 3). Bei Betrachtung der veränderten Methodik über die Zeit zeigen sich deutliche Unterschiede in den eruierten Häufigkeiten (Tabelle 3):

Tabelle 3: Einschätzung einer eingeschränkten Beschwerdenvalidität bzgl. der Methode

Methode	*keine KP, BV, BVT (n = 304)*	*nur KP (n = 68)*	*KP + BV (n = 92)*	*KP + BV + BVT (n = 710)*
Anteil eingeschränkt beurteilter BV, n (%)	20* (6.6)	2* (2.9)	11 (12)	112* (15.8)

BV = Beschwerdenvalidität; Keine KP, BV, BVT = im Gutachten findet sich keine besondere Methodik zur Beurteilung der Beschwerdenvalidität (Zeitraum bis 2005); nur KP = im Gutachten findet sich nur eine Kausalitätsprüfung (ab 2006); KP + BV = im Gutachten findet sich neben dem Punkt der Kausalitätsprüfung ein Punkt zur Einschätzung der Beschwerdenvalidität (ab 2007); KP+BV+BVT = zusätzlich zur Kausalitätsprüfung und Einschätzung der Beschwerdenvalidität werden gezielt Beschwerdenvalidierungstests eingesetzt (ab 2009, durchgängig ab 2010).* = statistisch signifikante Unterschiede zwischen beobachteten und erwarteten Häufigkeiten bei Betrachtung der standardisierten Residuen; n = Anzahl

Ohne besondere Methodik (keine KP, BV, BVT) und bei einfacher Kausalitätsbeurteilung werden weniger Fälle als erwartet hinsichtlich der BV als eingeschränkt beurteilt (6.6% bzw. 2.9%). Die zusätzliche Betrachtung der Beschwerdenvalidität (ohne spezifische Struktur oder gesonderte Messverfahren) zeigt keinen bedeutsamen Unterschied. Mit zusätzlicher Einführung von Beschwerdenvalidierungstests (KP+BV+BVT) steigt der Anteil signifikant auf 15.8% an (χ^2 (3, N = 1174) = 22.62, p < .001). Diese Häufigkeit entspricht den Angaben bei Young (2015), der in einem kritischen Review der bisherigen Studienlage ebensolche Basisraten (15%) vorschlägt, die wesentlich niedriger sind, als bisher angenommen (bis zu 40%; siehe auch Punkt 5).

Unterschiede zwischen Gutachtern

Ein Gutachterbias i.S. statistisch signifikanter Unterschiede in der eingeschätzten Häufigkeit einer eingeschränkten BV zwischen den verschiedenen Gutachtern, entsprechend auch der Expertise ergibt sich nicht (χ^2 (12, N = 1175) = 19.35, p = .08).

Beitrag von eruierten Inkonsistenzen zur Einschätzung der Authentizität

Jedes Gutachten wurde nach folgenden möglichen vom Gutachter aufgeführten Inkonsistenzen untersucht: Inkonsistenzen in den Vorbefunden, in den explorierten Angaben, zwischen

Schilderung und Verhaltensbeobachtung, zwischen aktuellen Daten und Vorbefunden, zwischen Exploration und Expertenwissen, sekundären Motive, Gegenübertragungsphänomene, Inkonsistenzen im Testverhalten sowie in Testergebnissen inkl. Beschwerdenvaldierungsverfahren. Eine Übersicht über die jeweiligen Häufigkeiten findet sich in Anhang 5. Am häufigsten werden inkonsistente Testergebnisse (inbes. auffällige BVT) in 79.6% der Fälle, sekundäre Motive (z.B. Rentenbegehren) in 23.8% der Fälle und Inkonsistenzen zwischen Beschwerdenschilderung und Verhaltensbeobachtung in 23.2% der Fälle aufgeführt. Gegenübertragungsphänomene (z.B. Gefühl des Unechten) werden hingegen kaum genannt (4.1% der Fälle). In 40.9% der Gutachten (n = 480) findet sich zumindest eine Inkonsistenz (M = 1.81, SD = 1.4, Range 1 – 7). In jedem Gutachten mit negativ eingeschränkt beurteilter BV finden sich Inkonsistenzen (M = 2.96, SD = 1.38, Range 1 – 7). In Gutachten ohne eingeschränkt beurteilte BV finden sich in 32.5 % der Fälle Inkonsistenzen (M = .43, SD = .74, Range 0 – 4). Die mittlere Anzahl der eruierten Inkonsistenzen unterscheidet sich dabei signifikant zwischen negativ eingeschränkt und authentisch beurteilten Gutachten (Mann-Whitney-U-Test, $U(1029,145) = -19.74$, $p < .001$).

Auch mit Veränderung der Methodik über die Zeit nimmt die Anzahl eruierter Inkonsistenzen entsprechend der Befunde zur letztlichen Einschätzung einer eingeschränkten Beschwerdenvalidität (s.o.) signifikant zwischen dem Zeiträumen ohne besondere Methodik bzw. lediglich mit Kausalitätsprüfung und ab 2009/ 2010 (Kausalitätsprüfung, Beschwerdenvalidierung + Beschwerdenvalidierungstests) zu (Kruskall-Wallis-Rangtest, χ^2 (3, N = 1175) = 49.26, p < .001); Tabelle 4:

Tabelle 4: Anzahl eruierter Inkonsistenzen bzgl. der Methode

Methode	*keine KP, BV, BVT (n=304)*	*nur KP (n=68)*	*KP + BV (n=92)*	*KP + BV + BVT (n=710)*
Anzahl Inkonsistenzen n (SD), Range	.49 (1.03) 0-6	.43 (.82) 0-4	.74 (1.22) 0-5	.89 (1.26) 0-7

Keine KP, BV, BVT = im Gutachten findet sich keine besondere Methodik zur Beurteilung der Beschwerdenvalidität (Zeitraum bis 2005); nur KP = im Gutachten findet sich nur eine Kausalitätsprüfung (ab 2006); KP + BV = im Gutachten findet sich neben dem Punkt der Kausalitätsprüfung ein Punkt zur Einschätzung der Beschwerdenvalidität (ab 2007); KP+BV+BVT = zusätzlich zur Kausalitätsprüfung und Einschätzung der Beschwerdenvalidität werden gezielt Beschwerdenvalidierungstests eingesetzt (ab 2009, durchgängig ab 2010); n = Anzahl; SD = Standardabweichung

Hinsichtlich der Gewichtung der aufgeführten Inkonsistenzen für die gutachterliche Entscheidung einer eingeschränkten BV zeigt eine logistische Regressionsanalyse sowohl für das Gesamtmodell (χ^2 (9, N = 1174) = 478.91, p <. 001) als auch die meisten eruierten Inkonsistenzen als Koeffizienten (Tabelle 5) ein signifikantes Ergebnis. Inkonsistenzen innerhalb der Vorbefunde,

Inkonsistenzen zwischen aktueller Datenlage und Vorbefunden sowie Gegenübertragungsphänomene zeigen als Prädiktoren keinen signifikanten Einfluss. Diese Inkonsistenzen werden hinsichtlich ihrer Häufigkeit am wenigsten in Gutachten aufgeführt (Inkonsistenz innerhalb der Vorbefunde 7.7%, zwischen aktuellen Daten und Vorbefunden 6.0%, Gegenübertragung 4.1%). Den größten Beitrag zur Einschätzung der BV leistet die Testdiagnostik (Inkonsistenz Testverhalten *OR* = 9.50, Inkonsistenz Testergebnisse *OR* = 12.01). Insgesamt wurden 92.9% der Gutachten durch das Modell entsprechend ihrer tatsächlichen Einschätzung klassifiziert. Dies ist aber durch den großen Stichprobenunterschied zwischen dem Anteil authentisch und als eingeschränkt beurteilten Gutachten bedingt. Von den als authentisch beurteilten Gutachten (n = 1029) werden 98.1% (n = 1009) korrekt vorhergesagt, von den als eingeschränkt beurteilten Gutachten (n = 124) jedoch nur 56.6% (n = 82). Insgesamt erklärt das Gesamtmodell dementsprechend lediglich 63.6% der auftretenden Varianz (R^2), was einen komplexeren „inneren“ Entscheidungsalgorithmus der Gutachter nahelegt, als durch die aufgeführten Inkonsistenzen erklärt werden kann.

Tabelle 5: Beitrag der einzelnen Inkonsistenzen als Prädiktoren zur Einschätzung einer eingeschränkten Beschwerdenvalidität (binär-logistische Regressionsanalyse)

Inkonsistenz	*B*	*Standardfehler*	*p*	*OR*	*CI*
Innerhalb der Vorbefunde	.81	.54	ns	2.24	(.77 -6.49)
In den explorierten Angaben	2.07	.43	<.001	7.92	(3.38 – 18.52)
Zwischen Schilderung und Verhaltensbeobachtung	1.62	.32	<.001	5.05	(2.68 – 9.53)
Zwischen aktuellen Daten und Vorbefunden	.61	.64	ns	1.83	(.52 – 6.49)
Zwischen Exploration und Expertenwissen	.98	.49	.044	2.67	(1.03 – 6.94)
Sekundäre Motive	2.28	.32	<.001	9.74	(5.20 – 18.27)
Spezifische Gegenübertragung	-.14	.66	ns	.87	(.24 – 3.17)
Im Testverhalten	2.51	.43	<.001	9.50	(4.08 – 22.10)
Testergebnisse inkl. BVT	2.49	.32	<.001	12.01	(6.44 – 22.41)

B = Regressionskoeffizient; CI = 95%-Konfidenzintervall; ns = nicht signifikant; OR = Odds Ratio; p = Signifikanzniveau

Ergebnisse in Beschwerdenvalidierungstests (BVT) und die Einschätzung einer eingeschränkten Beschwerdenvalidität

Aufgrund der eruierten Bedeutung der Testdiagnostik (insbes. BVT) für die gutachterliche Entscheidung erfolgte eine Analyse der Zusammenhänge der Ergebnisse bisher eingesetzter Beschwerdenvalidierungstests (Strukturierter Fragebogen simulierter Symptome (SFSS), Word Memory Test (WMT)) und der Einschätzung einer insgesamt als eingeschränkt beurteilten Beschwerdenauthentizität im Gutachten. In 710 Gutachten (60.4%) wurden BVT eingesetzt.

Lediglich in 52.8% (*n* = 375) der Fälle zeigen sich dabei keine Hinweise auf eine Antwortverzerrung. In 33.7% (*n* = 239) der Fälle zeigt sich in einem Verfahren eine Antwortverzerrung, in weiteren 13.5% (*n* = 96) sind in beiden Verfahren Antwortverzerrungen zu verzeichnen. Dies entspricht bisherigen Befunden zu Basisraten von Antwortverzerrungen auf Testebene in deutschen Begutachtungspopulationen (z.B. Merten et al., 2010: 48%). Entsprechend der dortigen Interpretation ist eine testpsychologische Antwortverzerrung jedoch nicht mit genereller Aggravation oder Simulation gleichzusetzen und bedarf weiteren Einbezugs verschiedener Datenquellen.

Die Häufigkeiten testpsychologischer Antwortverzerrungen in Abhängigkeit von der gutachterlichen Entscheidung der Beschwerdenauthentizität zeigt Tabelle 6.

Tabelle 6: Antwortverzerrungen in BVT in Abhängigkeit der gutachterlichen Einschätzung

	kein BVT auffällig (n=375)	*1 BVT auffällig (n=239)*	*2 BVT auffällig (n=96)*
BV nicht eingeschränkt n (%)	368 (98.1)*	204 (85.4)	26 (27.1)*
BV eingeschränkt n (%)	7 (1.9)*	35 (14.6)	70 (72.9)*

* = statistisch signifikante Unterschiede zwischen beobachteten und erwarteten Häufigkeiten bei Betrachtung der standardisierten Residuen (p < 0.01); n = Anzahl

In der Auswertung (χ^2 (2, N = 710) = 290.75, p < .001) zeigen sich dabei erwartungsgemäß signifkant höhere Raten eines als authentisch eingeschätzten Gutachtenprofils sowie signifikant niedrigere Raten einer eingeschränkt beurteilten Beschwerdenauthentizität, wenn keine testpsychologische Antwortverzerrung vorliegt. Wenn in zwei Verfahren Antwortverzerrungen vorliegen, zeigt sich demgegenüber ein entgegengesetztes Entscheidungsverhalten auf Seiten der Gutachter. Auch bei zwei auffälligen BVT gelangen die Gutachter jedoch noch in 27.1% der Fälle zu der Einschätzung, dass keine generelle Aggravation oder Simulation vorliegt.

Dabei unterscheiden sich Gutachten mit auffälligen BVT und trotzdem als ausreichend eingeschätzter Beschwerdenauthentiziät von Gutachten mit auffälligen BVT und eingeschränkter Beschwerdenvalidität insbesondere durch eine niedrigere Anzahl der aufgeführten Inkonsistenzen ($BVT_{auffällig + authentisch}$: n = 230, M = .87, SD = .82, Range 1 - 7; $BVT_{auffällig + BV eingeschränkt}$: n = 105, M = 2.85, SD = 1.43, Range 1 - 7; U (105,230) = -12.17, p < .001).

Neben dem bloßen Vorliegen von Antwortverzerrungen ist insbesondere das Ausmaß der Verzerrungen zu berücksichtigen. In beiden BVT zeigen sich in allen Kennwerten dabei durchweg weniger „auffällige“ Ausprägungen in der Gruppe, die trotz auffälliger BVT insgesamt als authentisch eingeschätzt wurde. Im WMT liegen die Ausprägungen der als authentisch

eingeschätzten Gruppe im Mittel nicht unterhalb der vorgegebenen cut-off-Werte für eine Antwortverzerrung (IR, DR, CNS ≤ 82,5; MC ≤ 70, PA ≤ 50), in der als nicht-authentisch eingeschätzten Gruppe bis auf den Kennwert WMT-PA jeweils darunter (Tabelle 7).

Tabelle 7: Ergebnisse der BVT bzgl. der Gruppen „BVT auffällig – nicht authentisch eingeschätzt“ vs. „BVT auffällig – authentisch eingeschätzt“

Kennwert			*M*	*SD*	*Range*	*p*
SFSS-Summenwert	*BVT-auffällig + nicht-auth.*	n=102	27.75	9.27	3-55	< .001
	BVT-auffällig + authentisch	n=227	22.60	6.23	1-47	
WMT-IR	*BVT-auffällig + nicht-auth.*	n=82	73.18	15.16	35-100	< .001
	BVT-auffällig + authentisch	n=102	90.65	11.94	45-100	
WMT-DR	*BVT-auffällig + nicht-auth.*	n=81	70.11	15.78	40-100	< .001
	BVT-auffällig + authentisch	n=96	98.60	11.98	40-100	
WMT-CNS	*BVT-auffällig + nicht-auth.*	n=81	67.74	13.18	40-100	< .001
	BVT-auffällig + authentisch	n=96	87.39	12.50	45-100	
WMT-MC	*BVT-auffällig + nicht-auth.*	n=81	54.40	16.23	15-100	< .001
	BVT-auffällig + authentisch	n=96	77.19	16.32	20-100	
WMT-PA	*BVT-auffällig + nicht-auth.*	n=66	51.53	15.75	10-95	< .001
	BVT-auffällig + authentisch	n=83	74.49	16.76	22.5-100	

M = Mittelwert; n = Anzahl; SD = Standardabweichung; p = Signifikanzniveau (U-Tests); SFSS = Strukturierter Fragebogen Simulierter Symptome; WMT = Word Memory Test; IR = Immediate Recognition; DR = Delay Recognition; CNS = Consistency Response; MC = Multiple Choice; PA = Paired Associate

Als extreme Antwortverzerrung kann bzgl. des WMT, der als Alternativwahlverfahren konzipiert ist, zudem das Unterschreiten der Ratewahrscheinlichkeit in den ersten drei Kennwerten (Maße für Anstrengungsbereitschaft) betrachtet werden. In diesem Fall ist von einer bewussten Manipulation beim Bearbeiten des Tests auszugehen. Insgesamt findet sich in 22 (8.2%) Testprotokollen mindestens eine Skala unterhalb der Ratewahrscheinlichkeit mit höherer Wahrscheinlichkeit (23.2%, n = 19) in der nicht-authentischen Gruppe (χ^2 (1, N = 178) = 16.4, p<.001, Tabelle 8).

Alle Probanden dieser Gruppe gaben eine adäquate Anstrengungsbereitschaft bei der Durchführung an, sodass ein zielgerichtetes aggravierendes Verhalten (entsprechend auch der als nicht-authentisch eingeschätzten Beschwerdenschilderung) deutlich wird.

In der dennoch als authentisch eingeschätzten Gruppe zeigten sich 3 Testprofile (3.1%) mit Kennwerten im Bereich der Ratewahrscheinlichkeit. In zwei Fällen wurde dabei eine fehlende Zugänglichkeit für die Sinnhaftigkeit der Durchführung des Testverfahrens auf Seiten der Probanden dokumentiert und die fehlende Kooperationsbereitschaft nicht im Sinne einer Aggravation begründet. In einem weiteren Fall wurde eine massive Übererregung und Dissoziationsneigung i.R. einer durch das Gutachtengespräch akut getriggerten Traumafolgestörung bei der Testdurchführung angegeben. Der Proband wurde vom Gutachter als anstrengungsbereit, jedoch kaum in der Lage, den Instruktionen zu folgen, beschrieben.

Tabelle 8: Antwortverzerrungen unterhalb der Ratewahrscheinlichkeit im WMT bzgl. der Gruppen „BVT auffällig – nicht authentisch eingeschätzt“ vs. „BVT auffällig – authentisch eingeschätzt“

WMT	*BVT-auffällig + nicht-authentisch (n=82)*	*BVT-auffällig + authentisch (n=96)*
keine Kennwerte unterhalb der Ratewahrscheinlichkeit, n (%)	63 (76.8)	93 (96.9)
mind. ein Kennwert unterhalb der Ratewahrscheinlichkeit, n (%)	19 (23.2)*	3 (3.1)*

* = statistisch signifikante Unterschiede zwischen beobachteten und erwarteten Häufigkeiten bei Betrachtung der standardisierten Residuen

4.2 Teilprojekt 2 (Patientenuntersuchung)

Insgesamt gingen 92 ($n_{\text{authentisch}}$ = 48; $n_{\text{nicht-authentisch}}$ = 44) Datensätze in die weiteren Analysen ein. Basierend auf dem MINI zeigten sich als hauptsächliche Störungsgruppen (Mehrfachnennungen möglich) bei 42.5% der Patienten Anpassungsstörungen, bei 39.1% eine Schmerzstörung, bei 38.1% depressive Störungen und bei 23.9% eine Traumafolgestörung (PTBS oder PTBS-Teilsymptomatik). Hier spiegelt sich das generelle Profil des Patientenklientels mit bg-lichem Schwerpunkt wieder. Andere Angststörungen zeigten sich in geringerem Ausmaß (Spezifische Phobie 6.5%, Panikstörung 2.2%, soziale Phobie 2.2%). Psychotische Störungen und Suchterkrankungen waren nicht vertreten. Insbesondere das Fehlen psychotischer Erkrankungen ist als Unterschied zu den bisherigen Validierungen des SIRS-2 zu sehen. Eine detaillierte Darstellung der soziodemografischen Angaben und der Begleitdiagnostik (außerhalb der BVT) findet sich in Anhang 6.

Gruppenunterschiede in soziodemografischen Maßen und der Begleitdiagnostik

Die beiden Gruppen unterscheiden sich dabei nicht hinsichtlich des Alters, der Geschlechterverteilung, des Bildungsniveaus, des beruflichen bzw. Rentenstatus, der geschätzten fluiden und kristallinen Intelligenzmaße (LPS, WST) sowie bzgl. der Symptombelastung, weder auf Subskalenebene, der Bandbreite an berichteten Beschwerden, noch hinsichtlich der Gesamtbelastung (BSCL). Auch bzgl. der Verteilung der zugrunde liegenden psychischen Störungen zeigen sich kein Unterschiede (Anhang 6). Die beiden Gruppen können in diesem Zusammenhang als ausreichend parallelisiert gelten.

Gruppenunterschiede in Beschwerdenvalidierungsmaßen

Tabelle 9 zeigt die Kennwerte der durchgeführten BVT. Bis auf die Zusatzskala *Direct Appraisal of Honesty - DA* des SIRS-2 zeigen sich in allen vorgegebenen Kennwerten signifikante Unterschiede zwischen beiden Gruppen. Bei ansonsten parallelen Stichproben führt die Simulationsinstruktion also zu entsprechenden Verfälschungen, die durch die Testverfahren auch abgebildet werden. Die *SIRS-DA* Skala weist bei einem Score > 4 auf eine reduzierte Offenheit im Kontakt zu Behandlern hin, die in der untersuchten Stichprobe im Mittel auch bei der nicht-authentischen Gruppe nicht vorliegt. Die Skala gehört dabei zu den Zusatzskalen, die weitere Informationen zur Interpretation des jeweiligen Antwortverhaltens geben. In die Klassifikation des SIRS-Profils (authentisch vs. verfälscht) fließen nur die Hauptskalen ein, so dass die letztliche Einordnung hierdurch nicht beeinflusst wird.

Tabelle 9: Kennwerte der Beschwerdenvalidierungsverfahren und Gruppenunterschiede

Kennwert, M (SD), Range	*authentisch (n = 48)*	*nicht-authentisch (n = 44)*	*Gruppenvergleich*
SIRS-2			
Hauptskalen:			
Rare Symptoms (RS)	1.06 (1.79) 0-6	7.57 (3.6) 0-16	*U* (48,44) = -7.46, *p* <.001
Symptom Combinations (SC)	.92 (1.53) 0-6	5.86 (3.85) 0-18	*U* (48,44) = -6.99, *p* <.001
Improbable or Absurd Symptoms (IA)	.77 (1.13) 0-4	4.95 (2.94) 0-12	*U* (48,44) = -7.14, *p* <.001
Blatant Symptoms (BL)	3.44 (3.35) 0-15	16.5 (4.1) 9-29	*U* (48,44) = -8.1, *p* <.001
Subtle Symptoms (SU)	9.85 (6.67) 1-25	21.68 (5.71) 11-31	*t* (90) = 9.09, *p* < *.001*
Selectivity of Symptoms (SEL)	8.94 (4.64) 1-18	21.05 (3.6) 13-30	*t* (90) = 13.89, *p* < .001
Severity of Symptoms (SEV)	4.35 (5.05) 0-16	17.14 (4.89) 7-27	*U* (48,44) = -7.52, *p* <.001
Reported vs. Observed Symptoms (RO)	1.38 (1.25) 0-5	5.95 (2.83) 0-12	*U* (48,44) = -7.16, *p* <.001
Zusatzskalen:			
Direct Appraisal of Honesty (DA)	2.4 (1.5) 0-8	3.5 (2.9) 0-12	*U* (48,44) =-1.67 , *p* = .095
Defensive Symptoms (DS)	22.54 (5.48) 12-34	30.7 (6.54) 12-38	*U* (48,44) = -5.62, *p* <.001
Improbable Failure (IF)	1.37 (2.49) 0-10	4.84 (4.98) 0-18	*U* (48,44) = -4.33, *p* <.001
Overly Specified Symptoms (OS)	1.04 (1.22) 0-10	3.52 (3.27) 0-12	*U* (48,44) = -4.65, *p* <.001
Inconsistency of Symtoms (INC)	2.77 (1.99) 0-7	5.14 (4.1) 0-18	*U* (48,44) = -2.89, *p* =.004
Einzelne Kennwerte:			
Modified Total Index (MT Index)	6.19 (5.72) 0-23	34.89 (11.06) 14-71	*U* (48,44) = -8.16, *p* <.001
Rare Symptoms Total (RS-Total)	1.65 (2.52) 0-10	9.68 (8.37) 0-34	*U* (48,44) = -6.53, *p* <.001
Supplementary Scale Index (SS Index)	27.35 (5.46) 13-40	42.57 (10.82) 17-66	*t* (90) = 8.62, *p* < *.001*
SFSS			
Summenwert	12.73 (8.26) 1-38	46.25 (11.31) 24-72	*U* (48,44) = -8.07, *p* <.001
WMT			
Immediate Recognition (IR)	96.98 (5.92) 75-100	61.54 (23.46) 12.5-100	*U* (48,44) = -7.45, *p* <.001
Delay Recognition (DR)	96.27 (6.66) 70-100	56.7 (25.74) 3-100	*U* (48,43) = -7.35, *p* <.001
Consistency Response (CNS)	94.71 (7.42) 72.5-100	65.43 (15.91) 37.5-97.5	*U* (48,43) = -7.42, *p* <.001
Multiple Choice (MC)	87.92 (12.58) 55-100	43.26 (22.04) 37.5-97.5	*U* (48,43) = -7.45, *p* <.001
Paired Associate (PA)	82.5 (17.17) 40-100	43.57 (22.53) 0-95	*U* (48,42) = -6.72, *p* <.001
Free Recall (FR)	56.2 (18.07) 23-95	26.44 (13.54) 0-60	*t* (88) = -8.74, *p* < *.001*

M = Mittelwert; n = Anzahl; SD = Standardabweichung, SFSS = Strukturierter Fragebogen Simulierter Symptome, SIRS-2 = Structured Interview of Reported Symptoms, WMT = Word Memory Test; p = Signifikanzniveau (bei Normalverteiltheit des Merkmals erfolgten t-Tests für unabhängige Stichproben, bei Verletzung der Normalverteiltheit U-Tests)

Validität des SIRS-2

Effektstärken

Für die Hauptskalen des SIRS wurden Effektstärken berechnet und den Angaben aus der Originalversion und der spanischen Adaptation gegenübergestellt (Tabelle 10). Nach den Vorschlägen von Rogers et al. (2008) für Beschwerdenvalidierungsmaße (hohe Effektstärke, d ≥ 1.25; sehr hohe Effektstärke, d ≥ 1.5) zeigen sich durchgängig sehr starke Effekte (*M* = 2.36, Range 1.71 - 3.50). Dabei zeigen Skalen, die eine *umfangsbasierte* Strategie zur Erfassung einer eingeschränkten BV nutzen (BL, SU, SEL, SEV) größere Effekte, als Skalen mit einer *wahrscheinlichkeitsbasierten* Strategie (RS, SC, IA, RO) (mittlere Effektstärke 2.71 vs. 2.02).

Tabelle 10: Effektstärken der Hauptskalen des SIRS (eigene Untersuchung vs. Originalversion vs. Spanische Adaptation)

Kennwert	*d*	*Original-SIRS-2*[a] *d*	*Spanisches SIRS-2*[b] *d*
SIRS-2			
Hauptskalen:			
Rare Symptoms (RS)	2.32	2.04	1.92
Symptom Combinations (SC)	1.71	1.75	2.07
Improbable or Absurd Symptoms (IA)	1.91	1.80	1.84
Blatant Symptoms (BL)	3.50	2.49	2.47
Subtle Symptoms (SU)	1.90	2.12	1.87
Selectivity of Symptoms (SEL)	2.90	2.37	2.25
Severity of Symptoms (SEV)	2.57	1.95	2.18
Reported vs. Observed Symptoms (RO)	2.12	2.12	1.38
Mittlere Effektstärke	2.36	2.08	2.00

[a] = Daten aus Rogers, Sewell & Gillard (2010), [b] = Daten aus Correa, Rogers & Hoersting (2011), d = Effektstärke (Cohens *d*), SIRS-2 = Structured Interview of Reported Symptoms

Wahrscheinlichkeitsbasierte Nachweisstrategien erkennen negative Antwortverzerrungen daran, dass unwahrscheinliche Symptome und falsche Beschwerden berichtet werden. *Umfangbasierte Strategien* dagegen erkennen ein nicht-authentisches Antwortverhalten daran, dass generell nachvollziehbare Beschwerden in übertriebener Menge angegeben werden.

Insgesamt zeigt sich durch die Effektstärken das hohe Potential der SIRS-Hauptskalen zwischen authentischem und verfälschtem Antwortverhalten zu unterscheiden.

Interne Konsistenz

Hinsichtlich der internen Konsistenz wurden *Cronbach's α* für sechs der Hauptskalen und die Klassifikationsskala berechnet. Ausgenommen wurden die Skalen *SEL* und *SEV*. Da diese über mehrere Beschwerdenbereiche hinweg erfasst werden, ist nicht von Unidimensionalität auszugehen.

Tabelle11: Interne Konsistenzen, Interrater-Reliabilität und Retest-Reliabilität des SIRS-2.

Kennwert	*Cronbach's α* (n = 92)	*Interrater r (Intraklassen-korrelationen)* (n = 36)	*Retest r (Pearson-Korrelationen)* (n = 25)	*Konkordanzen der Klassifikationen (%)* (n = 25)
SIRS-2				
Hauptskalen:				
Rare Symptoms - RS (8 Items)	.78	1.0	.56**	96.0
Symptom Combinations - SC (10 Items)	.75	1.0	.52**	100.0
Improbable or Absurd Symptoms - IA (7 Items)	.68	1.0	.29*	100.0
Blatant Symptoms - BL (15 Items)	.87	1.0	.74**	96.0
Subtle Symptoms - SU (17 Items)	.89	1.0	.86**	96.0
Selectivity of Symptoms - SEL (32 Items)	-	1.0	.87**	92.0
Severity of Symptoms - SEV (32 Items)	-	1.0	.78**	92.0
Reported vs. Observed Symptoms - RO (12 Items)	.80	.98	.64**	100.0
Klassifikationsskala:				
RS-Total (15 Items)	.84	1.0	.70**	96.0
Gemittelter Wert	.80	.99	.66	96.4

SIRS-2 = Structured Interview of Reported Symptoms; *$p \leq .05$; **$p \leq .01$

Wie Tabelle 11 zu entnehmen ist, zeigen die Skalen *BL*, *SU*, *RO* und *RS-Total* jeweils eine gute interne Konsistenz ($\alpha \geq .80$). Für die Skala *IA* ($\alpha < .70$) wird nur eine marginale interne Konsistenz deutlich. Hier ist ggf. die geringe Anzahl der Items der Skala ($n = 7$) als limitierend zu berücksichtigen. Die Skalen *RS* und *SC* erreichen zumindest akzeptable Reliabilitätskoeffizienten. Auch hier ist im Vergleich zu den anderen Skalen die geringere Itemzahl zu berücksichtigen. Bei Betrachtung der korrigierten Item-Skalen-Korrelationen dieser beiden Skalen fallen bei zwei Items nur geringe bzw. negative Korrelationen zur Skala auf, die entsprechend jeweils den Reliabilitätskoeffizienten vermindern (RS: Item-155 $r = .19$; SC: Item-149 $r = -.009$). Diese beiden Items erfragen körper- bzw. schmerzbezogene Einschränkungen.

Hier zeigt sich im Vergleich zu den Original-Untersuchungen eine Besonderheit der hier untersuchten Stichprobe. Im Behandlungskontext einer Unfallklinik haben Patienten mit psychischen Erkrankungen häufig mehr und schwerere somatische Komorbiditäten (in vorliegender Stichprobe z.B. 39.1% Schmerzerkrankungen), als in anderen Einrichtungen (z.B. Psychiatrische Kliniken). In der aktuellen Stichprobe werden in diesem Zusammenhang Items, die körperliche Einschränkungen erfragen, möglicherweise unabhängig von Items mit anderen Inhalten der zugehörigen Skala beantwortet. Abzuwarten bleibt die Auswirkung dieses Effektes bei Zusammenführung mit den Daten des Kooperationspartners (Patienten einer psychiatrischen Institutsambulanz).

Insgesamt ist beim SIRS-2 zu beachten, dass die Skalen nicht zur Messung stabiler Konstrukte erstellt wurden, sondern dazu dienen, zustandsabhängige authentische Beschwerdenbilder mit

eher begrenztem Skalen-Bereich von verfälschten Beschwerden mit hoher Variabilität in den Skalenwerten zu trennen. Unabhängig vom jeweiligen Inhalt (Content) unterscheiden sich die SIRS-Skalen daher vor allem bzgl. der oben bereits aufgeführten zugrunde liegenden Nachweisstrategie (wahrscheinlichkeitsbasiert vs. umfangsbasiert). Wie Rogers et al. (2010) bereits anmerken, kann sich dies ebenfalls auf die rechnerisch erhobene interne Konsistenz der Skalen auswirken. Wenn z.B. eine Person keine psychotischen Beschwerden vortäuscht, wird sie entsprechende Items einer Skala negieren (z.B. RS-Skala: 4 aus 8 Items). Obwohl die Nachweisstrategie hiervon nicht beeinflusst wird, verringert dies jedoch die α-Schätzungen, indem aufgrund des Inhalts auf bestimmte Skalenitems selektiv geantwortet wird.

Trotz dieser Problematik zeigt sich im Mittel eine zufriedenstellende bis gute interne Reliabilität (mittleres α = .80), vergleichbar mit dem Original-SIRS (mittleres α = .86).

Interrater- und Retestreliabilität

Angesichts der Konsequenzen des Einsatzes des Verfahrens (z.B. gutachterliche Entscheidung) ist eine möglichst hohe Übereinstimmung unabhängig durchgeführter Messungen entscheidend (Interrater-Reliabilität).

Entsprechend den Anforderungen an ein vollstrukturiertes standardisiertes Interview gewährleistet das SIRS-2 nach entsprechender Schulung dabei fast ausschließlich eindeutige Zuordnungen, die in den meisten Skalen zu vollständigen Beurteilerübereinstimmungen führen (Interrater-Reliabilitäten in Tabelle 11). Lediglich für die Subskala *RO* ergeben sich minimale Abweichungen. Auf dieser Skala muss vom Interviewer entgegen den anderen Skalen keine einfach zu bewertende Antwort auf ein Item, sondern beobachtetes Verhalten eingeschätzt werden. Die Interrater-Reliabilität kann insgesamt als gegeben gelten (mittleres *r* = .99).

Um die Stabilität des SIRS über die Zeit einzuschätzen, wurde das Interview mit einem Teil der authentischen Gruppe (*n* = 25) erneut durchgeführt, wobei das Zeitintervall allerdings recht breit zwischen 7 und 35 Tagen variierte. Die Retest-Reliabilitäten für die einzelnen Skalen finden sich ebenfalls in Tabelle 11. Dabei zeigen sich überwiegend moderate bis hohe Korrelationen zwischen den Testdurchführungen mit vergleichbaren Ergebnissen (mittleres *r* = .66) zur Original-Validierung (mittleres *r* = .71). Wie im Original zeigt sich lediglich für die Skala *IA* (Original *r* = .24) ein geringer Zusammenhang zwischen den Skalenausprägungen zu den Testzeitpunkten.

Da der Fokus des SIRS-2 auf der Differenzierung nicht-authentischen und authentischen Antwortverhaltens liegt, ist aber weniger die Stabilität des absoluten Skalenwertes über die Zeit, sondern die Stabilität der Klassifikation (authentisch vs. verfälscht) entscheidend. Daher wurden zusätzlich die Klassifikationen der einzelnen Skalen anhand ihrer cut-off-Werte berechnet (authentisch = Skala liegt im unauffälligen oder unbestimmten Bereich; verfälscht = Skala liegt im

wahrscheinlichen oder definitiven Bereich) und die Konkordanzen bezogen auf die beiden Messzeitpunkte bestimmt. Die Ergebnisse finden sich in Tabelle 11. Hier zeigen sich über alle Skalen hinweg hohe Übereinstimmungen, die Skala *IA* zeigt trotz geringer Korrelation der Skalenwerte zwischen den Testzeitpunkten eine absolute Übereinstimmung hinsichtlich ihrer Klassifikation. Alle Testprotokolle wurden in der Gesamteinschätzung erwartungsgemäß auch im Retest als authentisch klassifiziert.

Konstruktvalidität

Als Maße der Konstruktvalidität wurden Pearson-Korrelationen zwischen den Hauptskalen bzw. der Klassifikationsskala des SIRS und den Ausprägungen der beiden anderen BVT (SFSS, WMT) sowie einem Maß der generellen Symptombelastung berechnet (Tabelle 12).

Tabelle 12: Konvergente Validität der SIRS-2 Primärskalen und der Klassifikationsskala: Korrelationen mit SFSS-, WMT- und BSCL-Ausprägungen

SIRS-Skalen:	*RS*	*SC*	*IA*	*BL*	*SU*	*SEL*	*SEV*	*RO*	*RS-Total*
SFSS-Summenwert	.777**	.789**	.839**	.867**	.770**	.868**	.829**	.764**	.722**
WMT-IR	-.641**	-.624**	-.657**	-.756**	-.628**	-.722**	-.717**	-.651**	-.662**
WMT-DR	-.634**	-.643	-.696**	-.759**	-.665**	-.732**	-.750**	-.711**	-.656**
WMT-CNS	-.637**	-.614**	-.676**	-.770**	-.689**	-.760**	-.760**	-.708**	-.650**
WMT-MC	-.633**	-.650**	-.701**	-.767**	-.649**	-.732**	-.745**	-.661**	-.660**
WMT-PA	-.555**	-.609**	-.635	-.690**	-.596**	-.652**	-.680**	-.589**	-.631**
BSCL-GSI	.165	.163	.217*	.246*	.365**	.281**	.312**	.247*	.053

SIRS = Structured Interview of Reported Symptoms; RS = Rare Symptoms; SC = Symptom Combinations; IA = Improbable or Absurd Symptoms; BL = Blatant Symptoms; SU = Subtle Symptoms; SEL = Selectivity of Symptoms; SEV = Severity of Symptoms; RO = Reported vs. Observed Symptoms; SFSS = Strukturierter Fragebogen Simulierter Symptome; WMT = Word Memory Test; IR = Immediate Recognition; DR = Delay Recognition; CNS = Consistency Response; MC = Multiple Choice; PA = Paired Associate; BSCL-GSI = Brief Symptom Checklist Global Severity Index; *$p \leq .05$; **$p \leq .01$

Dabei zeigen sich erwartungsgemäß hohe Zusammenhänge zwischen den SIRS-Skalen und den Ausprägungen im SFSS (mittleres r = .802) sowie im WMT (mittleres r = -.719). Der negative Zusammenhang zwischen SIRS und WMT ist durch die Richtung der Items bedingt (SIRS: hohe Ausprägung als Hinweis auf Verfälschung; WMT als Leistungstest: geringe Ausprägung als Hinweis auf Verfälschung). Die Korrelationen sind bei allen SIRS-Primärskalen vergleichbar für wahrscheinlichkeitsbasierte und umfangbasierte Nachweisstrategien, was einen grundlegenden Beleg für konvergente Validität darstellt. Für das tatsächliche Belastungserleben (BSCL-GSI) wurden nur geringe Zusammenhänge zu den SIRS-Skalen erwartet, da das SIRS zwar nachvollziehbare Beschwerden mit erfasst, für seine Messintention letztlich aber extreme oder

unglaubwürdige Muster hiervon ableitet. Erwartungsgemäß zeigen sich hier die höchsten Korrelationen bzgl. der Skala *SIRS-SU*, die generelle Beschwerden in der Allgemeinbevölkerung erfasst und der Skala *SIRS-SEV*, die die angegebene Intensität der Beschwerden misst. Generell bleiben die Zusammenhänge jedoch niedrig (mittleres *r* = .228), was bestätigt, dass die Ausprägungen im SIRS vor allem von Verfälschungen (SFSS, WMT) und nicht vom tatsächlichen Belastungserleben beeinflusst werden.

Klassifikationsgüte der Beschwerdenvalidierungsverfahren

Klassifikation des Structured Interview of Reported Symptoms (SIRS-2)

Die rechnerische Bestimmung der Klassifikationsgüte verlangt eine dichotome Zuordnung zum interessierenden Kriterium, die das SIRS-2 aber nicht bietet. Neben eindeutig positivem (*Feigning*) oder negativem Testprofil (*Genuine*) finden sich nach einem Zuordnungsalgorithmus Zwischenkategorien (*Indeterminate-General* und *Indeterminate-Evaluate*). Bzgl. der Kategorie *Indeterminate-Evaluate* geben die Autoren eine erhöhte Wahrscheinlichkeit für Verfälschungen (> 50%) an. Eine Zuordnung allein auf Basis des SIRS-2 sei in diesem Fall jedoch nicht zulässig. Es sollten zusätzliche Informationsquellen genutzt werden. In der Kategorie *Indeterminate-General* geben die Autoren eine deutlich geringere Wahrscheinlichkeit für Verfälschungen (geschätzte Prävalenz 34.3%) an, als Klassifizierung gebe ein solches Ergebnis keine weiteren Auskünfte.

Die Zuordnung der in Teilprojekt 2 untersuchten Stichprobe auf den verschiedenen Ebenen des vorliegenden Algorithmus findet sich in Anhang 7.

In der Original-Validierung (Rogers et al., 2010) werden für das SIRS-2 eine Sensitivität von 80% und eine Spezifität von 97.5% angegeben. Die Autoren haben dabei aber lediglich die eindeutig positiven und eindeutig negativen Testprofile berücksichtigt, was den ursprünglichen Datensatz um 23% reduziert. Entsprechend finden sich Kritiken (z.B. Rubenzer, 2010, DeClue, 2011; Green et al., 2013), die von einer Überschätzung der Sensitivität ausgehen, wenn nur die offensichtlichsten authentischen und verfälschten Fälle einbezogen werden. Green et al. (2013) schlagen in diesem Zusammenhang zwei weitere Varianten zur Bestimmung der Klassifikationsgüte vor. In Variante 2 werden die Kategorie *Indeterminate-Evaluate* aufgrund der erhöhten Wahrscheinlichkeit für Verfälschungen zusammen mit der Kategorie *Feigning* und die Kategorie *Indeterminate-General* (kein Informationszugewinn bzgl. einer Verfälschung) zusammen mit der Kategorie *Genuine* kombiniert. In Variante 3 wird die Klassifikationsgüte einer eindeutigen Verfälschung (*Feigning*) in Abhängigkeit von anderen Antwortstilen (alle weiteren Kategorien) bestimmt.

Alle drei Varianten wurden auch für die untersuchte Stichprobe berechnet und finden sich zusammen mit der Klassifikation des SFSS und des WMT in Tabelle 13.

Tabelle 13: Klassifikationsgüte der einzelnen Beschwerdenvalidierungsverfahren

Testverfahren		*tatsächliche Gruppenzugehörigkeit* *nicht-authentisch*	*authentisch*
SIRS-2			
Klassifikation[a] *nach Rogers et al. (2010)*			
	n = 72		
	nicht-authentisch	29 (richtig positiv)	0 (falsch positiv)
	authentisch	0 (falsch negativ)	43 (richtig negativ)
Sensitivität		100%	CI: 88.06% - 100%
Spezifität		100%	CI: 91.78% - 100%
Positiver Vorhersagewert		100%	CI: 88.06% - 100%
Negativer Vorhersagewert		100%	CI: 91.78% - 100%
Klassifikation[b],	*n = 92*		
	nicht-authentisch	41 (richtig positiv)	1 (falsch positiv)
	authentisch	3 (falsch negativ)	47 (richtig negativ)
Sensitivität		93.18%	CI: 81.34% - 98.57%
Spezifität		97.92%	CI: 88.93% - 99.95%
Positiver Vorhersagewert		97.62%	CI: 87.43% - 99.94%
Negativer Vorhersagewert		94.00%	CI: 83.45% - 98.75%
Klassifikation[c]	*n = 92*		
	nicht-authentisch	29 (richtig positiv)	0 (falsch positiv)
	authentisch	15 (falsch negativ)	48 (richtig negativ)
Sensitivität		65.91%	CI: 50.08% - 79.51%
Spezifität		100%	CI: 92.60% - 100%
Positiver Vorhersagewert		100%	CI: 88.06% - 100%
Negativer Vorhersagewert		76.19%	CI: 63.79% - 86.02%
SFSS *(Score >16)*	*n = 92*		
Klassifikation	*nicht-authentisch*	44 (richtig positiv)	12 (falsch positiv)
	authentisch	0 (falsch negativ)	36 (richtig negativ)
Sensitivität		100.00%	CI: 91.96% - 100%
Spezifität		75.00%	CI: 60.4% - 86.36%
Positiver Vorhersagewert		78.57%	CI: 65.56% - 88.41%
Negativer Vorhersagewert		100%	CI: 90.26% - 100%
WMT *(IR, DR od. CNS ≤ 82.5)*			
	n = 92		
Klassifikation, n (%)	*nicht- authentisch*	37 (richtig positiv)	6 (falsch positiv)
	authentisch	7 (falsch negativ)	42 (richtig negativ)
Sensitivität		84.09%	CI: 69.93% -93.36%
Spezifität		87.50%	CI: 74.75% - 95.27%
Positiver Vorhersagewert		86.05%	CI: 72.07% - 94.70%
Negativer Vorhersagewert		85.71%	CI: 72.76% - 94.06%

Klassifikation[a] = definitiv als verfälscht vs. definitiv als authentisch klassifizierte Protokolle (unbestimmte Protokolle werden nicht berücksichtigt); Klassifikation[b] = erhöhte Wahrscheinlichkeit von Verfälschung (definitiv verfälscht + indeterminate evaluate) vs. keine erhöhte Wahrscheinlichkeit von Verfälschung (definitiv authentisch + indeterminate general); Klassifikation[c] = definitiv verfälscht vs. andere Antwortmuster (definitiv authentisch + indeterminate evaluate + indeterminate general); CI = 95%-Konfidenzintervall; n = Anzahl; SFSS = Strukturierter Fragebogen Simulierter Symptome, SIRS-2 = Structured Interview of Reported Symptoms, WMT = Word Memory Test

Bei gleicher Berechnung wie im Original (Tabelle 13: Klassifikation[a]) zeigt das SIRS-2 in der vorliegenden Stichprobe eine korrekte Zuordnung jedes Testprofiles (Sensitivität 100%; Spezifität: 100%), allerdings werden aus der Berechnung 20 Testprofile (21.7%) ausgeschlossen, die nicht eindeutig klassifiziert werden. Hier zeigt sich eine ähnliche Rate von Testprofilen in den Zwischenkategorien, wie in der Originalversion.

Da das SIRS-2 in der Regel nicht standardmäßig durchgeführt wird, sondern aufgrund seines Aufwandes nur bei Verdacht auf eine Verfälschung (z.B. anhand eines zuvor durchgeführten Screenings od. anderer Inkonsistenzen im Untersuchungsprozess), bietet sich für die Anwendung in der Praxis insbesondere die Schätzung der Klassifikationsgüte anhand der zweiten aufgeführten Klassifikation an (Klassifikation[b]).

Über alle Klassifikations-Varianten hinweg zeigt sich aber eine sehr hohe Spezifität des Verfahrens (97.9% - 100%; CI: 88.93% - 100%). Das Ziel der Testautoren, ein Verfahren bereitzustellen, das falsch-positive Zuordnungen minimiert, kann also auch für die deutsche Version als erfüllt gelten (falsch-positive Zuordnung max. 1%). Klassifikation[a] bietet hingegen die optimistischste Einschätzung der Sensitivität (100%: eindeutige Verfälschung vs. eindeutig authentisches Antwortverhalten) und Klassifikation[c] die vorsichtigste Einschätzung der Sensitivität (65.9%: eindeutige Verfälschung vs. andere Antwortstile).

Für klarere Interpretationsrichtlinien sind zusätzliche Befunde zur Differenzierung von authentischen und verfälschten Testprofilen erforderlich, die nach dem SIRS-2 in eine der Zwischenkategorien klassifiziert werden. Für die vorliegende Stichprobe wurden hierfür die Gruppen (authentisch vs. nicht-authentisch) statistisch verglichen, die in eine der beiden Zwischenkategorien eingeordnet wurden. Die Ergebnisse sind jedoch nur explorativ zu interpretieren, da die Stichprobengrößen gering und stark unterschiedlich sind:

Die authentische Gruppe mit SIRS-Klassifikation *Indeterminate* (*n* = 5) zeigt dabei hinsichtlich der Diagnoseverteilungen einen größeren Anteil depressiver Störungen als die nichtauthentische Gruppe (n = 15) mit dieser Klassifikation (100% vs. 33.3%). Hinsichtlich des Alters, der Geschlechterverteilung und der Intelligenzmaße (LPS, WMT) zeigen sich keine Unterschiede. In der weiteren Begleitdiagnostik zeigen sich jedoch höhere Ausprägungen der authentischen Gruppe hinsichtlich verschiedener Beschwerdenbereiche in der BSCL (Aggressivität, Paranoides Denken, Somatisierung, Zwanghaftigkeit) und dem Globalkennwert Positiv-Symptom-Distress-Index (Tabelle 14). Hinsichtlich der BVT zeigen sich bis auf die Skala WMT-DR erwartungsgemäß „auffälligere“ Ausprägungen in der nicht-authentischen Gruppe. Trotzdem liegt in der authentischen Gruppe im Mittel eine deutlich über dem cut-off-Wert (> 16) liegende SFSS-Ausprägung vor (*M* = 28.2, *SD* = 7.67), die WMT-Werte liegen im Durchschnitt nicht im auffälligen Bereich. Insgesamt zeigen also insbesondere authentische Patienten mit manifester depressiver Erkrankung und deutlich überdurchschnittlicher Bandbreite an erlebten Beschwerdenbereichen ein ähnliches

Antwortverhalten im SIRS, wie weniger beeinträchtigte Patienten mit Täuschungsabsicht. Dies kann zu uneindeutigen Klassifikationen beitragen. Im SFSS besteht in diesem Fall die Gefahr falsch-positiver Einschätzungen. In der Praxis sollte daher insbesondere bei der Beurteilung der Beschwerdenvalidität bei dieser Patientengruppe auf umfassendere Informationen zusätzlich zur Testpsychologie zurückgegriffen werden.

Tabelle 14: Gruppenunterschiede authentischer und nicht-authentischer Probanden mit SIRS-Klassifikation im Indeterminate-Bereich

Kennwert, M (SD)	authentisch (n = 5)	nicht-authentisch (n = 15)	Gruppenvergleich
BSCL-Aggressivität	74.00 (8.25)	64.6 (8.51)	*U* (15,5) = -2.02, *p* =.025
BSCL-Ängstlichkeit	76.00 (5.52)	69.20 (8.69)	*U* (15,5)) = -1.63, *p* = .119
BSCL-Depressivität	72.40 (5.34)	66.40 (14.23)	*U* (15,5)) = -.67, *p* = .553
BSCL-Paranoides Denken	72.6 (6.54)	66.40 (7.33)	*U* (15,5) = -2.04, *p* =.042
BSCL-Phobische Angst	78.00 (4.47)	70.33 (11.59)	*U* (15,5)) = -1.48, *p* = .197
BSCL-Psychotizismus	71.20 (12.38)	64.07 (12.13)	*U* (15,5)) = -1.07, *p* = .306
BSCL-Somatisierung	79.80 (.45)	63.47 (10.05)	*U* (15,5)) = -3.01, *p* ≤ .001
BSCL-Soziale Unsicherheit	74.40 (10.36)	68.33 (9.89)	*U* (15,5)) = -1.58, *p* = .119
BSCL-Zwanghaftigkeit	76.20 (5.22)	65.00 (10.02)	*U* (15,5)) = -2.33, *p* = .019
BSCL-Global Severity index	77.60 (3.29)	71.60 (9.51)	*U* (15,5)) = -1.24, *p* = .215
BSCL-Positiv Symptom Total	75.40 (6.31)	70.20 (9.09)	*U* (15,5)) = -1.29, *p* = .230
BSCL-Positiv Symptom Distress Index	77.20 (3.42)	65.47 (9.79)	*U* (15,5)) = -2.68, *p* = .005
SFSS-Summenwert	28.20 (7.66)	38.07 (8.55)	*U* (15,5)) = -2.19, *p* = .025
WMT-IR	94.50 (8.37)	71.36 (20.34)	*U* (15,5)) = -2.45, *p* = .011
WMT-DR	88.70 (7.84)	72.87 (18.47)	*U* (15,5)) = -1.88, *p* = .066
WMT-CNS	85.20 (7.46)	71.70 (12.89)	*U* (15,5)) = -2.01, *p* = .042

M = Mittelwert; n = Anzahl; SD = Standardabweichung, BSCL = Brief Symptom Checklist; SFSS = Strukturierter Fragebogen Simulierter Symptome; WMT = Word Memory Test; IR = Immediate Recognition; DR = Delay Recognition; CNS = Consistency Response

Klassifikation des Strukturierten Fragebogens Simulierter Symptome (SFFS)

Der SFSS klassifiziert als Screening 87% der Testprofile korrekt (36 richtig-negative/ 44 richtig-positive). Es zeigen sich aber gehäuft falsch-positive Einschätzungen (12 Testprofile (13%)). Hieraus ergibt sich für die untersuchte Stichprobe eine außerordentliche Sensitivität (100%), bei eingeschränkter Spezifität (75%) (Tabelle 13).

Klassifikation des Word Memory Test (WMT)

Der WMT gilt als Verfahren mit höherer Spezifität. Er entstammt dem neuropsychologischen Bereich zur Validierung geltend gemachter kognitiver Leistungseinbußen, z.B. nach Schädel-Hirn-Traumata und objektiviert die gezeigte Leistungsbereitschaft. Eine Kritik beim Einsatz für psychische Störungen (ohne SHT) beruht auf Befunden, die nahelegen, dass Verfälschungen auf verschiedenen Domänen (psychisch, kognitiv, körperlich-somatisch) unabhängig voneinander auftreten können. Zudem bestehen Unklarheiten, inwieweit der WMT auch bei authentischen psychischen Störungen zu falsch-positiven Einschätzungen führen kann (z.B. Schmidt et al., 2011). In der vorliegenden Stichprobe klassifiziert der WMT 86% der Testprofile korrekt (42 richtig-negative/ 37 richtig-positive). Es zeigen sich 6 (6.5%) falsch-positive Einschätzungen und 7 falsch-negative Einschätzungen (7.6%). Hieraus ergeben sich für die untersuchte Stichprobe eine Sensitivität von 84.1% und eine Spezifität von 87.5% (Tabelle 13).

Kombination von Testverfahren

In der Begutachtungspraxis wird in der antragstellenden Einrichtung herkömmlich der SFSS als Screening eingesetzt. Wenn dieser auffällig ist, erfolgt zusätzlich der Einsatz des WMT. In der vorliegenden Stichprobe klassifiziert der SFSS 36 Testprofile als authentisch. Die verbleibenden 56 Testprofile werden als verfälscht klassifiziert und werden durch den WMT spezifiziert. Hier schätzt der WMT 9 weitere Profile als richtig-negativ (authentisch) ein, 7 Testprofile werden falsch-negativ, 3 Testprofile falsch-positiv und 37 Testprofile als richtig-positiv klassifiziert. Hieraus ergibt sich eine zusammenfassende Sensitivität der beiden Verfahren von 84.09% und eine Spezifität von 93.75%.

In der Kombination von SFSS und SIRS-2, klassifiziert letzteres die verbleibenden 56 Testprofile wie folgt: 11 Testprofile richtig-negativ, 3 Profile falsch-negativ, 1 Profil falsch-positiv, 41 Profile richtig-positiv. Es ergibt sich eine zusammenfassende Sensitivität von 93.18% und eine Spezifität 97.87%.

Der Einsatz des SIRS-2 anstatt des WMT zeigt sich in diesem Zusammenhang als überlegen. Angesichts der möglichen Konsequenzen einer gutachterlichen Entscheidung ist insbesondere die hierdurch mögliche Reduktion der falsch-positiven Einschätzungen zu berücksichtigen, ohne dass sich ein Anstieg der falsch-negativen Einschätzungen ergibt.

Zusammenfassung

Der *SFSS* zeigt als Screeningverfahren in der untersuchten Stichprobe insbesondere eine beachtliche Sensitivität, jedoch eine relevante falsch-positiv-Rate (13%). Das *SIRS-2* zeigt sowohl eine hohe Sensitivität als auch Spezifität. Für den Einsatz bei psychischen Störungen ohne neuropsychologischen Schwerpunkt (kein SHT) ist das SIRS-2 nach diesen Ergebnissen dem herkömmlich eingesetzten *WMT* hinsichtlich der Spezifität und je nach Berücksichtigung der Klassifikations-Variante auch hinsichtlich der Sensitivität überlegen. Für die Praxis bietet sich zusammenfassend ein Screening durch den SFSS und bei Hinweis auf mögliche Verfälschungen eine Spezifizierung mit dem SIRS-2 an.

5. Auflistung der für das Vorhaben relevanten Veröffentlichungen, Schutzrechtsanmeldungen und erteilten Schutzrechte von nicht am Vorhaben beteiligten Forschungsstellen

Ergänzend zum Zwischenbericht (09/2015) erfolgte eine erneute Literaturrecherche (Medline bis 08/2016) zu den im Erstantrag genannten Themenbereichen. Dabei finden sich im Untersuchungszeitraum insbesondere Arbeiten zur Evaluation der Anwendungsbreite vorhandener Testverfahren sowie zur Interpretation von Testprofilen unter Berücksichtigung individueller Krankheitsumstände und der Testsituation.

Die bisher größte Umfrage unter europäischen Neuropsychologen (Dandachi-Fitz Gerald, Ponds & Merten, 2013) stellte heraus, dass es trotz ausreichendem Fachwissen weitverbreitete überholte Überzeugungen zum Thema und insbes. Unklarheiten zur Interpretation von „Testversagen“ in Beschwerdenvalidierungstests gibt. Eine weitere Kritik betrifft die häufige Schätzung der Basisrate von verfälschten Beschwerden anhand der Untersuchung studentischer Gruppen (Silk-Eglit, Stenclik & Gavett, 2014).

Auf diese Aspekte geht Teilprojekt 1 insbesondere ein. Durch die Untersuchung einer realen Begutachtungspopulation unter Berücksichtigung verschiedener möglicher Indikatoren für verfälschte Beschwerden leistet das Projekt eine aussagekräftigere Schätzung und einen weiteren Beitrag, wie eine Beschwerdenvalidierung insgesamt objektiver erfolgen kann.

Das bisher umfassendste Review zu empirischen Befunden zur Häufigkeit von verfälschten Beschwerden (Young, 2015) übt Kritik an den bisher z.T. vorgeschlagenen hohen Wahrscheinlichkeitsschätzungen von bis zu 40 ± 10% (Mittenberg et al. 2002, Larabee et al., 2009). Kritikpunkte sind vor allem eine Mittelwertbildung von Häufigkeitseinschätzungen über verschiedene Anwender mit unterschiedlichen Stichprobengrößen und eine stark uneinheitliche Methodik zur Definition und Erfassung von verzerrtem Antwortverhalten. Nach erneuter Durchsicht

der zitierten Studien ist nach dem Autor von einer wesentlich niedrigeren Rate definitiver Verfälschung von Antwortverhalten im Untersuchungskontext von 15 ± 15% auszugehen.

Dies entspricht auch den Ergebnissen aus dem in Teilprojekt 1 ausgewerteten Gutachtenpool mit einer Rate von 15.8% als verfälscht eingeschätztem Antwortverhalten bei der Nutzung multipler Datenquellen.

Bezogen auf das Teilprojekt 2 ist eine Metaanalyse zum international am häufigsten eingesetzten Screeninginstrument (SIMS, dt. SFSS) zur Beurteilung der Beschwerdenvalidität (van Impelen, Merckelbach & Jelicic, 2014) herauszuheben. Hierin wird eine hohe Sensitivität bestätigt. Die Spezifität ist jedoch eingeschränkt, so dass eine Kombination mit einem weiteren Testverfahren, das diesbezüglich aussagekräftiger ist, empfohlen wird.

Das in Teilprojekt 2 validierte Verfahren (SIRS-2) sollte in der Konsequenz genau diese Spezifität leisten können und zeigt eine optimalere Klassifikationsgüte, insbesondere mit minimaler falsch-positiv-Rate, als bisher eingesetzte Testverfahren (Word Memory Test).

Aus der Kritik am SFSS wurden inzwischen auch zwei originäre deutschsprachige Screeningverfahren entwickelt, zu denen seit kurzem erste Befunde vorliegen (BEVA-Beschwerdenvalidierungstest, Walter et al., 20016/ SRSI-Self Report Symptom Inventory, Merten et al., 2016). Beide Verfahren benötigen weitere Validierung, erscheinen aber vielversprechend und spezifischer für den Einsatz im sozial- und zivilrechtlichen Bereich, als der SFSS, der insbesondere für den forensischen Bereich validiert wurde.

Eine direkte Konkurrenz zum entwickelten SIRS-2 wird nicht gesehen, da Screeningverfahren und spezifischere aber auch aufwendigere Verfahren, wie das SIRS-2, in Zusammenschau der Befunde in Kombination eingesetzt werden sollten. Von Interesse sind Untersuchungen, inwieweit eine Kombination von BEVA bzw. SRSI und SIRS-2 zu einer besseren Klassifikationsgüte als SFSS und SIRS-2 führt. Entsprechende Kooperationsanfragen mit den Autoren der beiden neuen Testverfahren laufen bereits.

6. Bewertung der Ergebnisse hinsichtlich des Forschungszwecks/-ziels, Schlussfolgerungen

Bisher besteht eine Kontroverse zur Auftretenswahrscheinlichkeit von Verfälschungen und deren Bedingungen im Begutachtungskontext. Ziel war daher eine im Vergleich zu bisherigen Untersuchungen aussagekräftigere Schätzung, die verschiedene von Gutachtern genutzte Informationsquellen und Veränderungen der Methodik über die Zeit berücksichtigt.

Über die Zeit zeigt sich dabei mit komplexer werdender Methodik ein steigender Aufwand zur Gutachtenerstellung, aber auch ein höheres Bewusstsein der Gutachter für mögliche

Beschwerdenverfälschungen. Es bestätigt sich, dass psychologische Gutachter testpsychologischen Befunden einen hohen Stellenwert beimessen. Dabei zeigt sich eine beachtliche Replikation bisheriger Einschätzungen zur Häufigkeit von einzelnen Verzerrungen in Testverfahren in der berufsgenossenschaftlichen Begutachtung (Merten et al., 2010: 48%). Im untersuchten Gutachtenpool zeigen sich in diesem Zusammenhang in über 47% der Fälle, in denen Beschwerdenvalidierungstests (BVT) eingesetzt wurden, Antwortverzerrungen. Die tatsächliche Einschätzung der Gutachter nach Berücksichtigung verschiedener anderer Informationsquellen ist aber geringer, als angesichts einzelner Indikatoren. Die Gesamteinschätzung einer eingeschränkten Beschwerdenauthentizität liegt nach Gewichtung durch die Gutachter bei aktueller Methodik bei 15.8%. Auch mit diesem Ergebnis können aktuelle internationale Literaturreviews (Young, 2015; geschätzte Rate nicht authentischer Beschwerdenschilderung: 15%) repliziert und für den deutschen Sprachraum übertragen werden. Die Ergebnisse bestätigen noch einmal, dass gutachterliche Aussagen i.S. einer Prüfung der vorliegenden Informationen jeden Einzelfalls multimethodal erfolgen sollten. Der steigende Aufwand zur Erstellung von Gutachten im Zeitverlauf bzgl. des Aspektes der Beschwerdenvalidität ist gerechtfertigt, um sowohl authentisch vorliegende Beeinträchtigungen, als auch die Einschätzung von Verfälschungen valider sichern zu können. Auch ein Bias-Effekt des einzelnen Gutachters kann durch ein methodisch ausgereiftes und strukturiertes Vorgehen vermieden werden.

Die Kritik zu bisher eingesetzten BVT (z.B. lediglich Screening-Charakter, Gefahr falsch-positiver Einschätzungen, eingeschränkter konzeptueller Rahmen) wurde aufgegriffen und ein spezifisches Verfahren für psychische Störungen in einer deutschen Version validiert (SIRS-2). Das SIRS-2 zeigt sich dabei in der Untersuchung herkömmlichen anderen BVT überlegen. Hervorzuheben ist die stichprobenbezogene geringe falsch-positiv Rate (1%) im Vergleich zu den anderen Verfahren (SFSS: 13%, WMT: 6.5%). Erstmals steht somit ein BVT zur Verfügung, das nicht als reines Selbstbeurteilungsinstrument oder Leistungstest konzipiert ist und durch seinen komplexen Aufbau eine Vielzahl zusätzlicher Informationen zur Interpretation des Antwortverhaltens bietet. Erstmals wurde solch ein Verfahren auch spezifisch an Patienten- bzw. Begutachtungsklientel der UV-Träger (hohe somatische Komorbiditäten bei psychischen Erkrankungen) überprüft und zeigt sich entsprechend geeignet.

Für die Routineanwendung außerhalb von Forschungsfragestellungen ist aber zuerst urheberrechtlich eine Bereitstellung als für Fachkreise zugängliches Testverfahren in einem Fachverlag erforderlich. Diese erfolgt nach Zusammenführung mit den Daten des Kooperationspartners beim Rechteinhaber. Hiernach muss noch eine Erprobung an einer realen Begutachtungspopulation erfolgen, um die experimentell erhobenen Gütekriterien für die Praxisverknüpfung zu überprüfen (siehe auch Punkt 7 – Umsetzungs- und Verwertungsplan).

Das Projekt optimiert insgesamt einen zentralen Aspekt der Begutachtung psychischer Störungen (Beschwerdenvalidierung) und entsprechend auch die Leistungsprüfung. In der Anwendung kann die Beurteilung von Unfallfolgen methodisch ausgebaut und für alle Beteiligten (Antragsteller, UV-Träger, Sachverständige, Gerichte) transparenter gestaltet werden. Nach den Kriterien zur Anerkennung von Begutachtungen in der Rechtspraxis erweitern die Ergebnisse der beiden Teilprojekte die wissenschaftlich-medizinische Sachkunde und Aktualität. Die Kompetenzen der UV-Träger hinsichtlich der Analyse und Aufbereitung von Diagnostikverfahren in Sachverständigengutachten werden erweitert.

7. Aktueller Umsetzungs- und Verwertungsplan

Die Ergebnisse des *Teilprojektes 1 (Gutachtenauswertung)* werden nach Abnahme des Abschlussberichtes themenbezogen aufbereitet und sollen ab 01/2017 in entsprechenden Fachzeitschriften im „Peer-Review-Verfahren“ publiziert werden. Hierdurch wird die Zugänglichkeit und Diskussion in Fachkreisen gesichert.

Nach Abschluss der zusätzlichen Stichprobenuntersuchungen des Kooperationspartners (UKH Halle, anvisiert 11/2016) erfolgen bezüglich des *Teilprojektes 2 (Sitchprobenuntersuchung/ Validierung des SIRS-2)* zusammenfassende Analysen, eine Übertragung des englischsprachigen Manuals und Einreichung zur Publikation der deutschen Version des SIRS-2 beim Rechteinhaber (Hogrefe AG, Bern anvisiert 01-02/2017).

Anwenderempfehlungen bzgl. des Routine-Einsatzes des SIRS-2 sind erst nach Überprüfung der experimentell gewonnenen Daten im realen Begutachtungskontext möglich. Wenn das Verfahren als offiziell erhältliches Testverfahren zur Verfügung steht (voraussichtlich 3. Quartal 2017) ist in der antragstellenden Einrichtung daher ein einjähriger Versuch als Standardeinsatz im Gutachtenprozess im known-groups-Design vorgesehen. Schulungen der entsprechenden Mitarbeiter in der Durchführung des komplexen Verfahrens erfolgen ab dem 2. Quartal 2017.

Zudem ist eine Zusammenführung der Befunde des SIRS-2 mit den beiden 2016 vorgestellten originär deutschsprachigen Beschwerdenvalidierungsverfahren BEVA und SRSI vorgesehen. Hier ist angesichts noch notwendiger Absprachen mit den Autoren und Rechteinhabern (die Verfahren sind für die Anwendung noch nicht publiziert) noch keine genauere Zeitplanung möglich.

8. Literatur

AWMF-Leitlinie zur Begutachtung psychischer und psychosomatischer Erkrankungen. AWMF – Registernr. 051/029 (2012).

Correa, A.A., Rogers, R. & Hoersting, R. (2010). Validation of the Spanish SIRS with Monolingual Hispanic Outpatients. Journal of Personality Assessment. 92(5): 458-464.

Dandachi-Fitz Gerald, B., Ponds, R.W. & Merten, T. (2013). Symptom validity and neuropsychological assessment: a survey of practices and beliefs of neuropsychologists in six European countries. Archives of clinical Neuropsychology. 28(8):771-783.

DeClue, G. (2011). Harry Potter and the Structured Interview of Reported Symptoms? Journal of Forensic Psychology. 3: 1-18.

Dressing, H. & Foerster, K. (2010) Forensik: Begutachtung in der Sozial- und Versicherungsmedizin. Psychiatrie und Psychotherapie up2date. 4:109–122.

Green, D., Rosenfeld, H. & Belfi, B. (2013). New and Improved? A Comparison of the Original and Revised Versions of the Structured Interview of Reported Symptoms. Assessment. 20(2): 210-218.

Hall, H.V. & Poirier, J.G. (2001) Detecting Malingering and Deception. CRC Press: Boca Raton.

Larabee, G.J., Millis, S.R. & Meyers, J.E: (2009). 40 plus or minus 10, a new magical number: Reply to Russell. The Clinical Neuropsychologist. 23:841-849.

Merten T. (2014). Beschwerdenvalidierung. Göttingen: Hogrefe.

Merten, T., Krahl, G., Krahl, Ch. & Freytag, H.W. (2010). Prävalenz von negativen Antwortverterrungen in der berufsgenossenschaftlichen Begutachtung. Versicherungsmedizin. 62: 126-131.

Merten, T., Merckelbach, H., Giger, P. & Stevens, A. (2016). The Self-Report Symptom Inventory (SRSI): a New Instrument fort he Assessment of Distorted Symptom Endorsement. Psychological Injury and Law. 9(2): 102-111.

Mittenberg, W., Patton, C., Canyock, E.M. & Condit, D.C. (2002). Base Rates of malingering and symptom exaggeration. Journal of Clinical and Experimental Neuropsychology. 24:1094-1102.

Rogers, R. (2008). Detection strategies for malingering and defensiveness. In: Rogers, R. (Hrsg.) Clinical assessment of malingering and deception. Guilford Press: New York.

Rogers, R., Payne, J.W., Berry, D.T. & Granacher, R.P. (2009). Use of the SIRS in compensation cases: an examination of its validity and generalizability. Law and Hum Behavior. 33:213-224.

Rogers, R., Sewell, W. & Gillard, N.D. (2010). Structured Interview of Reported Symptoms 2nd Edition (SIRS-2) and professional manual. Lutz: Psychological Assessment Resources.

Rubenzer, S. (2010). Review of the Structured Interview of Reported Symptoms-2 (SIRS-2). Journal of Forensic Psychology. 2: 273-286.

Schmidt, T., Lanquillon, S. & Ullmann, U. (2011). Kontroverse zu Beschwerdenvalidierungsverfahren bei der Begutachtung psychischer Störungen. Forensische Psychiatrie Psychologie Kriminologie. 5:177-183.

Schulz, B. & Ullmann, U. (2006). Psychotraumatologische Versorgung im bg-lichen Heilverfahren. Trauma und Berufskrankheit, 9:109-112.

Silk-Eglit, G.M., Stenclik, J.H., Gavett, B.E. et al. (2014). Base rate of performance invalidity among non-clinical undergraduate research participants. Archives of Clinical Neuropsychology. 29(5):415-421.

van Impelen, A., Merckelbach, H. & Jelicic, M. et al. (2014). The Structured Inventory of Malingered Symptomatology (SIMS): a systematic review and meta-analysis. Clinical Neuropsychology. 28(8):1336-65.

Walter, F., Petermann, F: & Kobelt, A. (2016). Erfassung von negative Antwortverzerrungen – Entwicklung und Validierung des Beschwerdenvalidierungstests BEVA. Die Rehabilitation. 55:182-190.

Young, G. (2015). Malingering in Forensic Disability-Related Assessments: Prevalence 15±15%. Psychological Injury and Law. 8:188-199.

9. Anhänge

- Anhang 1: Vorgesehener Arbeitsplan zur Antragstellung und bei der Umsetzung aufgetretene Veränderungen

- Anhang 2: Kooperationsanzeige

- Anhang 3: Kategoriensystem zur Inhaltsanalyse des Gutachtenpools aus Teilprojekt 1

- Anhang 4: Ablauf der Untersuchung und verwendete testpsychologische Verfahren in Teilprojekt 2

- Anhang 5: deskriptive Kennwerte der Gutachtenauswertung (Teilprojekt 1)

- Anhang 6: deskriptive Beschreibung der Stichprobenuntersuchung – soziodemografische Kennwerte und Begleitdiagnostik (Teilprojekt 2)

- Anhang 7: Entscheidungsalgorithmus des SIRS-2 bzgl. der untersuchten Stichprobe (Teilprojekt 2)

- Anhang 8: Unterschriftenseite Kooperationsprojekte