Detecting DSM-5 somatic symptom disorder: criterion validity of the Patient Health Questionnaire-15 (PHQ-15) and the Somatic Symptom Scale-8 (SSS-8) in combination with the Somatic Symptom Disorder – B Criteria Scale (SSD-12)

Anne Toussaint, Paul Hüsing, Sebastian Kohlmann and Bernd Löwe

Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Introduction

The fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5; APA, 2013) changed the diagnostic category of somatoform and related disorders to somatic symptom and related disorders (SSD). This revision fundamentally shifted the way somatoform disorders are defined (Dimsdale et al., 2013). While medically unexplained symptoms were a key feature of somatoform and related disorders in DSM-IV (APA, 2000) and ICD-10 (WHO, 1992), SSD does not require the persistent symptoms to be medically unexplained. Regardless of their etiology, SSD is characterized by somatic symptoms that are either very distressing for the patients or result in significant disruption of daily functioning (A criterion). To be diagnosed with SSD, the individual must additionally experience excessive and disproportionate thoughts, feelings, and behaviors associated to the somatic symptoms (B criteria) which typically persist at least for 6 months (C criterion).

These new diagnostic criteria have been widely discussed. They are criticized as being too liberal, raising fears about mislabeling patients with comorbid medical illness as having a mental disorder (Frances and Chapman, 2013). In addition, the positive psychological criteria lack an empirical foundation (Rief and Martin, 2013). To be diagnosed with SSD, the individual must additionally experience excessive and disproportionate thoughts, feelings, and behaviors associated to the somatic symptoms (B criteria) which typically persist at least for 6 months (C criterion).

The fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5; APA, 2013) changed the diagnostic category of somatoform and related disorders to somatic symptom and related disorders (SSD). This revision fundamentally shifted the way somatoform disorders are defined (Dimsdale et al., 2013). While medically unexplained symptoms were a key feature of somatoform and related disorders in DSM-IV (APA, 2000) and ICD-10 (WHO, 1992), SSD does not require the persistent symptoms to be medically unexplained. Regardless of their etiology, SSD is characterized by somatic symptoms that are either very distressing for the patients or result in significant disruption of daily functioning (A criterion). To be diagnosed with SSD, the individual must additionally experience excessive and disproportionate thoughts, feelings, and behaviors associated to the somatic symptoms (B criteria) which typically persist at least for 6 months (C criterion).

These new diagnostic criteria have been widely discussed. They are criticized as being too liberal, raising fears about mislabeling patients with comorbid medical illness as having a mental disorder (Frances and Chapman, 2013). In addition, the positive psychological criteria lack an empirical foundation (Rief and Martin, 2013). To be diagnosed with SSD, the individual must additionally experience excessive and disproportionate thoughts, feelings, and behaviors associated to the somatic symptoms (B criteria) which typically persist at least for 6 months (C criterion).

The fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5; APA, 2013) changed the diagnostic category of somatoform and related disorders to somatic symptom and related disorders (SSD). This revision fundamentally shifted the way somatoform disorders are defined (Dimsdale et al., 2013). While medically unexplained symptoms were a key feature of somatoform and related disorders in DSM-IV (APA, 2000) and ICD-10 (WHO, 1992), SSD does not require the persistent symptoms to be medically unexplained. Regardless of their etiology, SSD is characterized by somatic symptoms that are either very distressing for the patients or result in significant disruption of daily functioning (A criterion). To be diagnosed with SSD, the individual must additionally experience excessive and disproportionate thoughts, feelings, and behaviors associated to the somatic symptoms (B criteria) which typically persist at least for 6 months (C criterion).

These new diagnostic criteria have been widely discussed. They are criticized as being too liberal, raising fears about mislabeling patients with comorbid medical illness as having a mental disorder (Frances and Chapman, 2013). In addition, the positive psychological criteria lack an empirical foundation (Rief and Martin, 2013). To be diagnosed with SSD, the individual must additionally experience excessive and disproportionate thoughts, feelings, and behaviors associated to the somatic symptoms (B criteria) which typically persist at least for 6 months (C criterion).

The fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5; APA, 2013) changed the diagnostic category of somatoform and related disorders to somatic symptom and related disorders (SSD). This revision fundamentally shifted the way somatoform disorders are defined (Dimsdale et al., 2013). While medically unexplained symptoms were a key feature of somatoform and related disorders in DSM-IV (APA, 2000) and ICD-10 (WHO, 1992), SSD does not require the persistent symptoms to be medically unexplained. Regardless of their etiology, SSD is characterized by somatic symptoms that are either very distressing for the patients or result in significant disruption of daily functioning (A criterion). To be diagnosed with SSD, the individual must additionally experience excessive and disproportionate thoughts, feelings, and behaviors associated to the somatic symptoms (B criteria) which typically persist at least for 6 months (C criterion).

These new diagnostic criteria have been widely discussed. They are criticized as being too liberal, raising fears about mislabeling patients with comorbid medical illness as having a mental disorder (Frances and Chapman, 2013). In addition, the positive psychological criteria lack an empirical foundation (Rief and Martin, 2013). To be diagnosed with SSD, the individual must additionally experience excessive and disproportionate thoughts, feelings, and behaviors associated to the somatic symptoms (B criteria) which typically persist at least for 6 months (C criterion).
costs (Jacobi et al., 2004). Recently, SSD is proposed as a percep-
tual disorder (Henningen et al., 2018) in which adverse events,
dysfunctional cognitions, expectations, negative affectivity, or
maladaptive behaviors influence the perception, perpetuation,
and deterioration of somatic symptoms (Löwe and Gerloff,
2018). Since their clinical relevance is high, strategies to improve
an early identification of patients with high symptom burden are
essential (Kohlmann et al., 2013). The correct diagnostic label
cannot only legitimize the patients’ concerns but also enable a tar-
gested treatment (Murray et al., 2016).

Standardized patient-reported outcome measures are generally
a valuable option to assess, quantify, and monitor patients’ per-
ceptions and experiences in medical care (Black, 2013). With
respect to the rather complex SSD criteria, well-established tools
like the Patient Health Questionnaire-15 (PHQ-15; Kroenke
et al., 2002) or the Somatic Symptom Scale-8 (SSS-8; Gierk
et al., 2014) can assist to assess the A criterion of distressing som-
atic symptoms. However, the changes in the diagnostic criteria
imply a need to review existing self-report questionnaires since
their feasibility and diagnostic accuracy toward the new diagnosis
cannot be presupposed (Klaus and Mewes, 2013). The Somatic
Symptom Disorder – B Criteria Scale (SSD-12) was developed
to assess the psychological B criteria of SSD (Toussaint et al.,
2016). Although these questionnaires alone are insufficient to
form the basis of a diagnosis and should always be used in con-
junction with a comprehensive clinical evaluation, they may
help to screen for SSD in clinical settings, or to guide discussions
about goals, expectations, and shared decisions for symptom
management.

The diagnostic accuracy of questionnaires for detecting the
SSD diagnosis is not well studied. It can be assumed that a com-
bination of measures assessing the A criterion (PHQ-15 or SSS-8)
together with a measure assessing the B criteria (SSD-12) should
altogether improve the detection rate. To the best of our know-
ledge, this is the first study to test the individual and combined cri-
terion validity of the PHQ-15, SSS-8, and SSD-12 within a clinical
sample.

Methods
Study participants and design

Data were collected within the German Research Foundation
(DFG) funded project ‘Somatic Symptom Disorder according to
DSM-5: Development and validation of a new self-report meas-
ure’ (Project number: TO 908/1-1), a study carried out at the
Department of Psychosomatic Medicine and Psychotherapy of
the University Medical Center Hamburg, Germany. Data were
collected between October 2015 and November 2016. All patients
presenting to the outpatient clinic, which is specialized in affect-
ive, anxiety, somatoform, and eating disorders, were invited to
participate in a study regarding bothersome somatic symptoms.
To be eligible to participate, patients needed to meet the following
criteria: be 18 or older, provide informed consent, and speak
German. Exclusion criteria were having a current psychotic dis-
order, indications of substance abuse, organic brain disease, or
active suicidality. The routine clinical consultation includes a
comprehensive assessment by a set of self-report questionnaires.
If patients agreed to participate and gave informed consent,
they were called within a period of 2 weeks after the consultation
to be interviewed about their symptoms. The study was approved
by the medical ethics board of the Hamburg Medical Chamber.

Study variables
Reference standard: assessment of SSD

Given that a German version of the Structured Clinical Interview
(SCID) for DSM-5 (First et al., 2015) was not available at the time,
we developed a research version of a semi-structured clinical inter-
view to assess the diagnostic criteria of SSD. The interview was
adapted to the English version of SCID-5, which is considered
the gold standard measure for DSM diagnoses. It was adminis-
tered by one of the seven trained assessors (M.D.; B.Sc. and
M.Sc. Psychology), who read questions aloud and recorded the
participant’s responses to the items. Calls took 20–30 min on
average. Based on the DSM-5 criteria, individuals were diagnosed
with an SSD once they fulfilled the A criterion of one or more
somatic symptoms that are distressing or result in significant dis-
ruption of daily life, as well as at least one of the three B criteria
either (1) disproportionate and persistent thoughts about the ser-
iousness of one’s symptoms, (2) a persistently high level of anxiety
about health or symptoms, or (3) excessive time and energy
devoted to these symptoms or health concerns. Although the
reported symptoms may not be continuously present, their state
of being symptomatic had to be persistent (typically more than
6 months). Patients were sub-classified as having a mild (only
one of the symptoms specified in Criterion B), moderate (two
or more of the symptoms specified in Criterion B), or severe
[two or more of the symptoms specified in Criterion B, and mul-
tiple somatic complaints (or one very severe symptom)] condition.

Self-report questionnaires assessing SSD Criterion A
Patient Health Questionnaire-15. The PHQ-15 is one of the most
frequently used instruments to identify people with elevated
symptom burden. It assesses the presence and severity of common
somatic symptoms in primary care, such as fatigue, gastrointes-
tinal, musculoskeletal, pain, and cardiopulmonary symptoms
within the last 4 weeks using 15 items. Each symptom can be
scored from 0 (‘not bothered at all’) to 2 (‘bothered a lot’). Sum
scores range from 0 to 30 and indicate the self-rated symptom
burden with higher scores indicating a higher burden (0–4 no
to minimal; 5–9 low; 10–14 medium; 15–30 high). Cronbach’s
α ranges around 0.80 (Kroenke et al., 2002). The PHQ-15 has
well-established psychometric properties and is recommended
for use in large-scale studies (Zijlema et al., 2013).

Somatic Symptom Scale-8. The SSS-8 is an abbreviated version
of the PHQ-15, which was developed within DSM-5 field trials
(Narrow et al., 2013). Items were selected on the basis of symptom
prevalence in primary care, association with measures of func-
tioning, and statistical commonalities with the items of the com-
plete PHQ-15 scale. A five-point response option (0–4) for each
item and a 7-day time frame are used. Cut-off scores indicate
whether a patient suffers from minimal (0–3 points), low (4–7),
medium (8–11), high (12–15), or very high (16–32) somatic
symptom burden. Previous studies demonstrated good item char-
acteristics and excellent reliability, a sound factor structure, and
significant associations with related constructs like depression,
anxiety, quality of life, and health care use. Gender- and age-
specific norms are available (Gierk et al., 2014), and sensitivity
to change has been demonstrated recently (Gierk et al., 2017).

Self-report questionnaire assessing SSD Criteria B
Somatic Symptom Disorder – B Criteria Scale 12. The Somatic
Symptom Disorder – B Criteria Scale 12 (SSD-12) is composed

https://doi.org/10.1017/S003329171900014X Published online by Cambridge University Press
of 12 items. Each of the three psychological sub-criteria is measured by four items with all item scores ranging between 0 and 4. The SSD-12 has good item characteristics, excellent reliability (Cronbach’s $\alpha = 0.95$), and a three-factorial structure, which reflects the three criteria. External and internal validity has been established (Toussaint et al., 2016, 2018). Norm values, which enable comparisons of SSD-12 scores with representative data, were derived from a large sample of the German general population (Toussaint et al., 2017a).

Statistical analysis

We computed means, standard deviations, and corrected item-total correlations to reflect the psychometric properties of the items. Cronbach’s $\alpha$ was determined as a measure of internal consistency.

Two multivariate stepwise logistic regression analyses with forced entry were calculated to evaluate the merit of combining each symptom questionnaire (PHQ-15 or SSS-8) with the SSD-12, entering each questionnaire at a different step (1. PHQ-15, 2. SSD-12 and 1. SSS-8, 2. SSD-12) and using SSD diagnosis ($0 = ‘not present’, 1 = ‘present’) as dependent variable.

For each single questionnaire, as well as for their combinations, we calculated sensitivity [the proportion of true positives correctly identified by the test as meeting a certain condition (e.g. SSD): a test with 100% sensitivity correctly identifies all patients with the condition], specificity [the proportion of true negatives correctly identified by the test as not meeting a certain condition: a test with 100% specificity correctly identifies all patients without the condition], positive (PPV) and negative predictive values (NPV) [the probability that individuals with a positive (negative) test result truly have (do not have) the condition], and efficiency [the total percentage of correct diagnosis, combining positive and negative diagnosis]. ROC curves were created for each instrument. The area under the curve (AUC) is a measure that provides an overall summary of the utility of the scale to correctly identify SSD. A measure that would have no ability to discriminate would be an AUC of 0.50. An AUC of 0.80 represents good accuracy, an AUC of 0.90–1.0 excellent accuracy. We used pROC, a package for R (Robin et al., 2011), to test for significant differences between the areas under the curve.

The level of statistical significance was $\alpha = 0.05$ (two-tailed) for all analyses. Analyses were conducted using IBM SPSS 23.

Results

Sample characteristics

A total of $n = 1149$ patients were eligible for participation in the study. Of these, 655 refused to participate due to various reasons, mainly acute psychological distress. Forty-three percent of the approached patients gave informed consent, 53 patients were subsequently excluded because they did not meet the inclusion criteria, moved abroad, or failed to respond to calls. The clinical telephone interview was conducted with 441 patients, whereas three of them were unable to complete the interview. Some of these participants did not provide viable data on the questionnaires during their clinical consultation, so that in the end data of $n = 372$ patients could be analyzed (dropout rate: 16%; see Fig. 1 for participant flow). Comparisons between participants and drop-outs yielded no significant differences regarding age, sex, nationality, education, or psychopathology (PHQ-15, SSS-8, and SSD-12 scores).

The average age of participants was 39.3 years (s.d. = 13.6). About two-thirds were female. The majority indicated that they had received school education for more than 10 years. Table 1 presents the demographic characteristics of the sample. In total, 86.8% of the participants fulfilled the A criterion of SSD, whereas a total of 56.2% met the full DSM-5 diagnostic criteria. Of those, 21.8% were classified as mild, 21.2% as moderate, and 13.2% as severe cases. Stratified analyses based on SSD status (yes or no) did not find any significant differences with respect to demographic factors. With respect to our self-report measures, those patients who fulfilled the SSD diagnosis reported significantly higher scores in all three questionnaires (Table 1).

Diagnostic accuracy

The combination of the PHQ-15 or SSS-8 with the SSD-12 in the regression analysis explained significantly more variance in predicting SSD than the PHQ-15 or SSS-8 alone ($R^2 = 0.15$ v. $R^2 = 0.29$, and $R^2 = 0.16$ v. $R^2 = 0.31$; Table 2).

Combination of screening instruments

The combination of the PHQ-15 or SSS-8 with the SSD-12 in the regression analysis explained significantly more variance in predicting SSD than the PHQ-15 or SSS-8 alone ($R^2 = 0.15$ v. $R^2 = 0.29$, and $R^2 = 0.16$ v. $R^2 = 0.31$; Table 2).

Table 3 shows sensitivities, specificities, predictive values for both positive and negative test results (PPV, NPV), and efficiency rates. These validation scores were calculated for the total scores of PHQ-15 (range: 0–30), SSS-8 (range: 0–32), and SSD-12 (range: 0–48) to establish optimal cutoff points. Table 3 shows relevant ranges only.

For the PHQ-15, efficiency was 66% at a cutpoint of $\geq 11$ and $\geq 15$, respectively. At a cutpoint of 11, sensitivity was higher (79%) than at a cutpoint of 15 (54%), whereas specificity was higher at a cutpoint of 15 (80%) than at a cutpoint of 11 (49%). NPVs were 64% and 58%, and PPVs were 62% and 77% for cutpoints 11 and 15. For the SSS-8, efficiency was highest with 66% at a cutpoint of $\geq 12$. Sensitivity was 72% and specificity 59%. NPV was 62% and PPV 69%. For the SSD-12, efficiency was highest with 69% at a cutpoint of $\geq 26$. Sensitivity was 70% and specificity 67%. NPV was 63% and PPV 73%.

When combining the instruments, the best efficiency values could be achieved by applying a cutoff of $\geq 9$ in the PHQ-15 or SSS-8, and $\geq 23$ in the SSD-12 (Table 4). Sensitivity and specificity were 69% and 70%, respectively. NPVs and PPVs were 64% and 74% for both combinations. Since previous studies reported severity thresholds of $\geq 10$ (medium somatic symptom burden) and $\geq 15$ (high somatic symptom burden) for both the PHQ-15 and
**Fig. 1.** Participant flow through recruitment and assessment process in a psychosomatic outpatient clinic from Hamburg, Germany (October 2015–November 2016).

**Table 1.** Baseline data of the study sample (n = 372)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Total (n = 372)</th>
<th>Somatic symptom disorder – no (n = 163)</th>
<th>Somatic symptom disorder – yes (n = 209)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N (%) or mean (s.d.)</td>
<td>N (%) or mean (s.d.)</td>
<td>N (%) or mean (s.d.)</td>
</tr>
<tr>
<td>Demographic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>39.3 (13.6)</td>
<td>38.2 (13.0)</td>
<td>40.2 (13.9)</td>
</tr>
<tr>
<td>Male sex</td>
<td>116 (31.2)</td>
<td>49 (30.1)</td>
<td>67 (32.1)</td>
</tr>
<tr>
<td>Married</td>
<td>80 (21.5)</td>
<td>28 (17.2)</td>
<td>52 (24.9)</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No qualifications</td>
<td>11 (3.0)</td>
<td>7 (4.3)</td>
<td>4 (1.9)</td>
</tr>
<tr>
<td>Less than 10 years</td>
<td>47 (12.6)</td>
<td>19 (11.7)</td>
<td>28 (13.4)</td>
</tr>
<tr>
<td>10 years of education</td>
<td>123 (33.1)</td>
<td>49 (30.0)</td>
<td>74 (35.4)</td>
</tr>
<tr>
<td>More than 10 years</td>
<td>183 (49.2)</td>
<td>84 (51.6)</td>
<td>99 (47.4)</td>
</tr>
<tr>
<td>Students</td>
<td>8 (2.2)</td>
<td>4 (2.5)</td>
<td>4 (1.9)</td>
</tr>
<tr>
<td>Scale scores</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Somatic symptom severity (PHQ-15)</td>
<td>range = 0–30</td>
<td>13.1 (5.2)</td>
<td>11.1 (4.7)*</td>
</tr>
<tr>
<td>Somatic symptom severity (SSS-8)</td>
<td>range = 0–32</td>
<td>13.3 (5.6)</td>
<td>11.1 (4.8)*</td>
</tr>
<tr>
<td>Psychological symptom severity (SSD-12)</td>
<td>range = 0–48</td>
<td>26.0 (10.6)</td>
<td>21.1 (10.0)*</td>
</tr>
</tbody>
</table>

*Denotes significance at p < 0.05.
and the SSS-8 (Toussaint et al., 2017b), and the corresponding thresholds for the SSD-12 can be determined at \(\geq 20\), and \(\geq 25\) (Toussaint et al., 2017a), cutpoints for these combinations are also reported in Table 4. All other possible cut-off combinations with the respective sensitivity, specificity, and PPVs and NPVs are available from the authors on request. Please note that sensitivity and specificity values are the same for combining either the PHQ-15 or the SSS-8 with the SSD-12. In both combinations, the rate of false negatives (sensitivity) raises when applying the ‘severe’ cutpoints (PHQ-15/SSS-8 \(\geq 15\); SSD-12 \(\geq 25\)).

ROC analyses

As shown in Fig. 2, the PHQ-15 (AUC = 0.70; \(p < 0.001\); 95% CI = 0.65–0.76), SSS-8 (AUC = 0.71; \(p < 0.001\); 95% CI = 0.66–0.77), and the SSD-12 (AUC = 0.74; \(p < 0.001\); 95% CI = 0.69–0.80) demonstrated moderate-to-good individual diagnostic accuracy. Differences between AUCs were not statistically significant (PHQ-15 v. SSD-12: \(p = 0.49\); PHQ-15 v. SSS-8: \(p = 0.41\); SSS-8 v. SSD-12: \(p = 0.70\)).

The combination of the PHQ-15 and SSS-8 with the SSD-12 slightly improved the accuracy and discriminatory power (PHQ-15 + SSD-12: AUC = 0.77; \(p < 0.001\); 95% CI = 0.72–0.82; SSS-8 + SSD-12: AUC = 0.79; \(p < 0.001\); 95% CI = 0.74–0.84).

\(95\%\) CI = 1.10–1.21.

\(95\%\) CI = 1.06–1.11.

\(95\%\) CI = 1.06–1.11.

\(95\%\) CI = 0.04.

\(95\%\) CI = 1.15–1.20.

\(95\%\) CI = 1.06–1.19.

\(95\%\) CI = 1.06–1.11.

\(95\%\) CI = 0.03.

Table 2. Stepwise logistic regression analysis evaluating the PHQ-15/SSS-8 and SSD-12 as predictors for SSD diagnosis (\(n = 372\))

<table>
<thead>
<tr>
<th>Variables</th>
<th>B</th>
<th>s.e.</th>
<th>(p)</th>
<th>OR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHQ-15</td>
<td>0.15</td>
<td>0.02</td>
<td>&lt;0.001</td>
<td>1.16</td>
<td>1.10–1.21</td>
</tr>
<tr>
<td>Constant</td>
<td>-1.63</td>
<td>0.33</td>
<td>&lt;0.001</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Step 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHQ-15</td>
<td>0.11</td>
<td>0.03</td>
<td>&lt;0.001</td>
<td>1.11</td>
<td>1.06–1.17</td>
</tr>
<tr>
<td>SSD-12</td>
<td>0.08</td>
<td>0.01</td>
<td>&lt;0.001</td>
<td>1.08</td>
<td>1.06–1.11</td>
</tr>
<tr>
<td>Constant</td>
<td>-3.16</td>
<td>0.45</td>
<td>&lt;0.001</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>Step 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSS-8</td>
<td>0.14</td>
<td>0.02</td>
<td>&lt;0.001</td>
<td>1.15</td>
<td>1.10–1.20</td>
</tr>
<tr>
<td>Constant</td>
<td>-1.57</td>
<td>0.31</td>
<td>&lt;0.001</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>Step 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSS-8</td>
<td>0.12</td>
<td>0.02</td>
<td>&lt;0.001</td>
<td>1.13</td>
<td>1.08–1.19</td>
</tr>
<tr>
<td>SSD-12</td>
<td>0.08</td>
<td>0.01</td>
<td>&lt;0.001</td>
<td>1.09</td>
<td>1.06–1.11</td>
</tr>
<tr>
<td>Constant</td>
<td>-3.41</td>
<td>0.46</td>
<td>&lt;0.001</td>
<td>0.03</td>
<td></td>
</tr>
</tbody>
</table>

Note:

Model summary statistics:

PHQ-15:

Step 1: \(\chi^2 (1) = 47.60, p < 0.001\); model: \(-2\log\text{likelihood} = 450.79\); Cox and Snell \(R^2 = 0.12\); Nagelkerke \(R^2 = 0.16\).

Step 2: \(\chi^2 (1) = 46.70, p < 0.001\); model: \(-2\log\text{likelihood} = 404.09\); Cox and Snell \(R^2 = 0.21\); Nagelkerke \(R^2 = 0.29\).

SSS-8:

Step 1: \(\chi^2 (1) = 47.47, p < 0.001\); model: \(-2\log\text{likelihood} = 431.03\); Cox and Snell \(R^2 = 0.15\); Nagelkerke \(R^2 = 0.29\).

Step 2: \(\chi^2 (1) = 48.30, p < 0.001\); model: \(-2\log\text{likelihood} = 388.93\); Cox and Snell \(R^2 = 0.29\); Nagelkerke \(R^2 = 0.31\).

PHQ-15: range 0–30 (higher scores indicate high somatic symptom burden); SSS-8: range 0–32 (higher scores indicate high somatic symptom burden); SSD-12: range 0–48 (higher scores indicate high psychological burden associated with somatic symptoms).

B, unstandardized regression coefficient; s.e., standard error; OR, odds ratio; CI, confidence interval.

Discussion

The present study evaluates and compares the diagnostic accuracy of the PHQ-15 and SSS-8 in combination with the SSD-12 for detecting DSM-5 somatic symptom disorder within a sample of psychosomatic outpatients. The results show that all three instruments are useful screening tools. With respect to the AUC analyses, all questionnaires showed a moderate ability to discriminate patients suffering from SSD v. patients without SSD. Combining the respective symptom questionnaires (to assess the A criterion) with the SSD-12 (to assess the B criterion) incrementally increased the diagnostic value (in the sense of explained variance) and slightly improved diagnostic accuracy. Altogether, the AUC values are lower than the operating characteristics reported for measures to detect depressive and anxiety disorders (Kroenke et al., 2003, 2007). This is probably due to the greater complexity and heterogeneity of the SSD construct.

Comparing the ROCs identified in our study with results from other studies is difficult as there are only few studies on the PHQ-15 available, whereas most of them do not focus on the rather inclusive definition of SSD, but on the former somatoform disorders. The diagnostic accuracy of the PHQ-15 for detecting patients at risk for somatoform disorders has been evaluated by Van Ravesteijn et al. (2009). In their primary care-based study, they reported a sensitivity of 78% and a specificity of 71% at a cut-off level \(\geq 3\) with an accuracy (AUC) of 0.76. De Vrooge et al. (2012) reported an AUC of 0.63, a sensitivity of 62%, and a specificity of 57% at a cut-off \(\geq 9\) for detecting somatoform disorders in sick-listed employees. A recent study, which examined the applicability of the PHQ-15 (A criterion), and the Whiteley-7 (WI-7; Pilowsky, 1967) (B criterion) for evaluating DSM-5 SSD via convenience sampling within psychiatric patients and healthy controls, reported a reasonable AUC of 0.73, a sensitivity of 85%, and a specificity of 49% for the PHQ-15 (cut-off 4/5), and a rather poor diagnostic accuracy for the Whiteley-7 (WI-7) (AUC = 0.66) (Liao et al., 2016). The study does not provide any information on a combined use of these instruments. Laferton et al. (2017) investigated the diagnostic accuracy of the PHQ-15 (A criterion) in combination with the WI-7 and the Scale for the Assessment of Illness Behavior (SAIB; Riel et al., 2003) (B criterion) in detecting somatic symptom disorders within a representative general population survey. They found moderate diagnostic accuracy for each individual questionnaire (PHQ-15: AUC = 0.79; WI-7: AUC = 0.76; SAIB: AUC = 0.77), whereas the combination of the instruments (PHQ-15 plus WI-7: AUC = 0.82; PHQ-15 plus WI-7 plus SAIB: AUC = 0.85) slightly improved diagnostic accuracy.

In our study, the combination of the PHQ-15 with the SSD-12 also only slightly improved diagnostic accuracy (AUC = 0.77), as did the combination with the SSS-8 (AUC = 0.77). Although the differences were not statistically significant, also not in comparison with the AUCs of each single questionnaire.
Table 3. Sensitivity, specificity, negative predictive values, positive predictive values, efficiency of PHQ-15, SSS-8, SSD-12 (n = 372)

<table>
<thead>
<tr>
<th></th>
<th>PHQ-15</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>≥6</td>
<td>≥7</td>
<td>≥8</td>
<td>≥9</td>
<td>≥10</td>
<td>≥11</td>
<td>≥12</td>
<td>≥13</td>
<td>≥14</td>
<td>≥15</td>
<td>≥16</td>
<td>≥17</td>
</tr>
<tr>
<td>(%)</td>
<td>Sensitivity</td>
<td>0.95</td>
<td>0.94</td>
<td>0.93</td>
<td>0.89</td>
<td>0.83</td>
<td>0.79</td>
<td>0.71</td>
<td>0.66</td>
<td>0.60</td>
<td>0.54</td>
<td>0.45</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>0.10</td>
<td>0.17</td>
<td>0.23</td>
<td>0.32</td>
<td>0.42</td>
<td>0.49</td>
<td>0.56</td>
<td>0.59</td>
<td>0.69</td>
<td>0.80</td>
<td>0.83</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>NPV</td>
<td>0.64</td>
<td>0.70</td>
<td>0.72</td>
<td>0.70</td>
<td>0.66</td>
<td>0.64</td>
<td>0.60</td>
<td>0.58</td>
<td>0.58</td>
<td>0.58</td>
<td>0.54</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>PPV</td>
<td>0.57</td>
<td>0.59</td>
<td>0.61</td>
<td>0.63</td>
<td>0.64</td>
<td>0.66</td>
<td>0.67</td>
<td>0.67</td>
<td>0.71</td>
<td>0.77</td>
<td>0.77</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>Efficiency</td>
<td>0.58</td>
<td>0.60</td>
<td>0.62</td>
<td>0.64</td>
<td>0.65</td>
<td>0.66</td>
<td>0.63</td>
<td>0.64</td>
<td>0.66</td>
<td>0.62</td>
<td>0.61</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(%)</td>
<td>Sensitivity</td>
<td>0.96</td>
<td>0.94</td>
<td>0.91</td>
<td>0.85</td>
<td>0.81</td>
<td>0.76</td>
<td>0.72</td>
<td>0.66</td>
<td>0.60</td>
<td>0.52</td>
<td>0.46</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>0.10</td>
<td>0.17</td>
<td>0.26</td>
<td>0.32</td>
<td>0.41</td>
<td>0.50</td>
<td>0.59</td>
<td>0.64</td>
<td>0.72</td>
<td>0.73</td>
<td>0.81</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>NPV</td>
<td>0.62</td>
<td>0.70</td>
<td>0.69</td>
<td>0.63</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.59</td>
<td>0.58</td>
<td>0.55</td>
<td>0.54</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>PPV</td>
<td>0.57</td>
<td>0.59</td>
<td>0.61</td>
<td>0.62</td>
<td>0.64</td>
<td>0.66</td>
<td>0.69</td>
<td>0.70</td>
<td>0.73</td>
<td>0.70</td>
<td>0.76</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>Efficiency</td>
<td>0.58</td>
<td>0.61</td>
<td>0.62</td>
<td>0.62</td>
<td>0.63</td>
<td>0.65</td>
<td>0.66</td>
<td>0.65</td>
<td>0.65</td>
<td>0.61</td>
<td>0.61</td>
<td>0.59</td>
</tr>
<tr>
<td>(%)</td>
<td>Sensitivity</td>
<td>0.91</td>
<td>0.89</td>
<td>0.86</td>
<td>0.85</td>
<td>0.84</td>
<td>0.83</td>
<td>0.80</td>
<td>0.78</td>
<td>0.76</td>
<td>0.73</td>
<td>0.70</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>0.32</td>
<td>0.37</td>
<td>0.40</td>
<td>0.44</td>
<td>0.46</td>
<td>0.48</td>
<td>0.52</td>
<td>0.56</td>
<td>0.58</td>
<td>0.63</td>
<td>0.67</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>NPV</td>
<td>0.74</td>
<td>0.73</td>
<td>0.69</td>
<td>0.69</td>
<td>0.68</td>
<td>0.69</td>
<td>0.67</td>
<td>0.66</td>
<td>0.65</td>
<td>0.64</td>
<td>0.63</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>PPV</td>
<td>0.64</td>
<td>0.65</td>
<td>0.65</td>
<td>0.66</td>
<td>0.67</td>
<td>0.68</td>
<td>0.69</td>
<td>0.70</td>
<td>0.70</td>
<td>0.72</td>
<td>0.73</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Efficiency</td>
<td>0.65</td>
<td>0.67</td>
<td>0.66</td>
<td>0.67</td>
<td>0.67</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
</tr>
</tbody>
</table>

NPV, negative predictive value; PPV, positive predictive value.
AUCs for all three questionnaires are comparable, the diagnostic efficiency, specificity, and sensitivity somewhat differ between the individual questionnaires. Whereas the SSD-12 showed the highest efficiency value of 69% (sensitivity: 70%, specificity: 67%) at a cutpoint of \(\geq 26\), the PHQ-15 and SSS-8 showed similar efficiency values of 66% at cutpoints of \(\geq 11\) and \(\geq 15\) for the PHQ-15 (sensitivity: 79%, specificity: 49% v. sensitivity: 54%, specificity: 80%) and \(\geq 12\) for the SSS-8 (sensitivity: 72%, specificity: 39%). Optimal

### Table 4. Combination of relevant cut-off scores of PHQ-15 and SSD-12/SSS-8 and SSD-12 (n = 372)

<table>
<thead>
<tr>
<th></th>
<th>PHQ-15 and SSD-12</th>
<th>Pragmatic cutpoints based on established severity scores (medium severity)</th>
<th>Pragmatic cutpoints based on established severity scores (high severity)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Optimal cutpoints determined in the current sample</td>
<td>PHQ-15 (\geq 10) and SSD-12 (\geq 20)</td>
<td>PHQ-15 (\geq 15) and SSD-12 (\geq 25)</td>
</tr>
<tr>
<td>(%)</td>
<td>PHQ-15 (\geq 9) and SSD-12 (\geq 23)</td>
<td>PHQ-15 (\geq 10) and SSD-12 (\geq 20)</td>
<td>PHQ-15 (\geq 15) and SSD-12 (\geq 25)</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.69</td>
<td>0.70</td>
<td>0.42</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.70</td>
<td>0.63</td>
<td>0.91</td>
</tr>
<tr>
<td>NPV</td>
<td>0.64</td>
<td>0.62</td>
<td>0.55</td>
</tr>
<tr>
<td>PPV</td>
<td>0.74</td>
<td>0.71</td>
<td>0.85</td>
</tr>
<tr>
<td>Efficiency</td>
<td>0.70</td>
<td>0.67</td>
<td>0.63</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SSS-8 and SSD-12</th>
<th>Pragmatic cutpoints based on established severity scores (medium severity)</th>
<th>Pragmatic cutpoints based on established severity scores (high severity)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSS-8 (\geq 9) and SSD-12 (\geq 23)</td>
<td>SSS-8 (\geq 10) and SSD-12 (\geq 20)</td>
<td>SSS-8 (\geq 15) and SSD-12 (\geq 25)</td>
</tr>
<tr>
<td>(%)</td>
<td>SSS-8 (\geq 9) and SSD-12 (\geq 23)</td>
<td>SSS-8 (\geq 10) and SSD-12 (\geq 20)</td>
<td>SSS-8 (\geq 15) and SSD-12 (\geq 25)</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.69</td>
<td>0.70</td>
<td>0.42</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.70</td>
<td>0.63</td>
<td>0.91</td>
</tr>
<tr>
<td>NPV</td>
<td>0.64</td>
<td>0.62</td>
<td>0.55</td>
</tr>
<tr>
<td>PPV</td>
<td>0.74</td>
<td>0.71</td>
<td>0.85</td>
</tr>
<tr>
<td>Efficiency</td>
<td>0.70</td>
<td>0.67</td>
<td>0.63</td>
</tr>
</tbody>
</table>

**Fig. 2.** Receiver operating characteristic (ROC) curves of PHQ-15, SSS-8, SSD-12 and their combinations in detecting the diagnosis of DSM-5 SSD (\(N=372\)).
defined by DSM-5. This rate is higher than the estimated prevalence-focused might generally be favored and recommended. False positives (e.g., referral to a diagnostic interview) is considered to the new criteria of SSD, which seem to include more patients.

In our sample, we found a prevalence of 56.2% of SSD as defined by DSM-5. This rate is higher than the estimated prevalence in the general population of 5–7%. It is also higher than the prevalence rates reported in comparable studies, which used the PHQ-15 and SSS-8 in settings with somatoform disorders. The differences regarding the cutpoints might therefore be related to the new criteria of SSD, which seem to include more patients. Patients in a psychosomatic setting probably suffer from a higher level of dysfunctioning than patients from primary care or from the general population, so that operating characteristics of our measures would probably also differ in general population or primary care samples, where the prevalence of SSD is lower. Some of the included patients suffered from severe mental symptoms, and there was a presumably high comorbidity rate between somatoform, depressive, and/or anxiety disorders. The high overlap between these three disorders, at least in the ICD-10 and DSM-IV conceptualization (Löwe et al., 2008; Kohlmann et al., 2016), supports the assumption that these disorders are not easily distinguished from each other.

The relation between the type of sample and the AUC score is also of importance: For example, Van Ravesteijn et al. (2009) excluded patients with a diagnosis of depression from their primary care sample, because these patients also often report somatic complaints. This may explain their somewhat higher AUC score of 0.76.

Compared with the study of Laferton et al. (2017), the SSD-12 performed similarly to the WI-7 and the SAIB in combination with the PHQ-15. Since the SSD-12 was developed to directly reflect the SSD criteria, we would have expected a superior performance. Again, one explanation might lie in our more severely impaired sample. Another explanation might be found in the rather imprecise psychological SSD criteria themselves. Previous studies with the SSD-12 (Toussaint et al., 2016) showed a high intercorrelation between the three subscales reflecting the content of the B criteria (cognitive, affective, and behavioral psychological features). Health anxiety, which is explicitly addressed by the items of the WI-7, might thereby be the most important underlying core concept of these psychological features.

**Implications**

We hope that this paper adds to the debate on the so far poorly tested SSD diagnosis. The provision of reliable and valid self-report questionnaires to operationalize the concept of SSD in research may facilitate the collection of empirical data on the often criticized A and B criteria. Our results suggest that the PHQ-15, SSS-8, and SSD-12 are suitable instruments to detect SSD in a psychosomatic outpatient sample. The PHQ-15 and SSS-8 are well-established self-report questionnaires to assess somatic symptom burden. Our results support the idea that combining these instruments with a questionnaire measuring psychological features improves the diagnostic accuracy. The SSD-12 was developed to specifically assess the psychological features of SSD. Its performance in combination with either the PHQ-15 or SSS-8 shows good results. Even though self-reported questionnaire results cannot replace a clinical evaluation, they could – in the sense of screening tools – be used to assist clinicians to rule out or confirm a suspected SSD diagnosis, especially in medical settings with a limited consultation time like primary care. We assume that the questionnaires could be easily integrated in routine care to monitor symptom courses, and to support an early identification of patients at risk for developing an SSD. Used in terms of continuous scores, they may have additional value reflecting severity of the SSD-D A and B criteria, both in research and practice.

When choosing or recommending the most suitable questionnaire, there are some considerations to apply, including the fit with the respective clinical population, response options, number of questions asked, time to complete the questionnaire, and symptom duration as suggested by the timeframes underlying the respective questionnaires. The choice of which questionnaire to use may best be decided by the needs of the respective clinical or research setting (Rief et al., 2017). If patients and clinicians have limited time only, the rather short SSS-8 should probably be favored in combination with the SSD-12, as a time- and cost-efficient screening strategy for routine care. It asks about common somatic symptoms within the last 7 days on a five-point Likert scale. However, if clinicians are interested in a broad spectrum of symptoms within the last 4 weeks, the PHQ-15 covers most of the common symptoms in primary care on a three-point Likert scale. All three questionnaires offer a quick and easy to understand scoring, and can be conducted and scored in <10 min. To assist clinicians with the interpretation of individual test scores on each measure, we evaluated sensitivity and specificity for all possible cut-off combinations of the individual questionnaires (available on request).

**Limitations**

Despite its strengths through the simultaneous investigation of the three questionnaires in a large sample, a few limitations of our study have to be noted: The data of this study are monocentric and rely on a psychosomatic outpatient sample. Participant dropout was fairly high throughout the recruitment process (only 1/3 of the eligible patients participated in the study). Although the final sample was still representative of a German psychosomatic sample in terms of socio-demographics, this fact might limit the generalizability of the results to some extent. In general, the German health care system is comparable to other Western health care systems, so that the results should be representative for similar patient groups. Furthermore, although the interview used as reference standard strictly applied DSM-5 criteria, comparability with other reference standards might be subject to some limitations due to the lack of a gold standard for diagnosing SSD. The interview was provided via telephone. However, previous studies could show that clinical interviews conducted via telephone provide reliable results (Cacciola et al., 1999; Fine et al., 2013). We did not control for comorbid mental or physical disorders. Since the concept of SSD may be applicable to a rather
heterogeneous patient group, it may be useful to review our findings on the diagnostic accuracy in different patient groups (e.g., patients with or without comorbid medical disease, with monomorphic or poly-symptomatic disorder).

Conclusions

To conclude, PHQ-15, SSS-8, and SSD-12 are amongst the first self-report screening instruments, which have been evaluated for diagnostic accuracy in detecting SSD. Although there are still possibilities for improvement, the combination of the PHQ-15 or SSS-8 with the SSD-12 provides an easy-to-use and time- and cost-efficient opportunity to identify patients at risk for SSD. If systematically applied, effective screening and subsequent diagnosis-appropriate treatment might be useful strategies to reduce disease burden and health care excess costs.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S003329171900014X

Acknowledgements. This study was funded by the German Research Foundation (DFG: TO 908/1-1). We would like to thank the participating clinicians and patients.

Conflict of interest. None.

References


