Depression is more than the sum score of its parts: individual DSM symptoms have different risk factors

E. I. Fried¹², R. M. Nesse³⁴, K. Zivin⁴⁵⁶, C. Guille⁷ and S. Sen⁴*

¹Cluster of Excellence ‘Languages of Emotion’, Freie Universität Berlin, Germany
²Department of Education and Psychology, Freie Universität Berlin, Germany
³Department of Psychology, University of Michigan, Ann Arbor, MI, USA
⁴Department of Psychiatry, University of Michigan Medical School, Ann Arbor, MI, USA
⁵Department of Veterans Affairs, National Serious Mental Illness Treatment Resource and Evaluation Center, Ann Arbor, MI, USA
⁶Institute for Social Research, University of Michigan, Ann Arbor, MI, USA
⁷Medical University of South Carolina, Charleston, SC, USA

Background. For diagnostic purposes, the nine symptoms that compose the DSM-5 criteria for major depressive disorder (MDD) are assumed to be interchangeable indicators of one underlying disorder, implying that they should all have similar risk factors. The present study investigates this hypothesis, using a population cohort that shifts from low to elevated depression levels.

Method. We assessed the nine DSM-5 MDD criterion symptoms (using the Patient Health Questionnaire; PHQ-9) and seven depression risk factors (personal and family MDD history, sex, childhood stress, neuroticism, work hours, and stressful life events) in a longitudinal study of medical interns prior to and throughout internship (n=1289). We tested whether risk factors varied across symptoms, and whether a latent disease model could account for heterogeneity between symptoms.

Results. All MDD symptoms increased significantly during residency training. Four risk factors predicted increases in unique subsets of PHQ-9 symptoms over time (depression history, childhood stress, sex, and stressful life events), whereas neuroticism and work hours predicted increases in all symptoms, albeit to varying magnitudes. MDD family history did not predict increases in any symptom. The strong heterogeneity of associations persisted after controlling for a latent depression factor.

Conclusions. The influence of risk factors varies substantially across DSM depression criterion symptoms. As symptoms are etiologically heterogeneous, considering individual symptoms in addition to depression diagnosis might offer important insights obfuscated by symptom sum scores.

Received 1 March 2013; Revised 12 September 2013; Accepted 2 November 2013

Key words: Depression, depressive symptoms, DSM, heterogeneity, psychiatric diagnosis, residency.

Introduction

Major depressive disorder (MDD) is common, burdensome, recurrent and expensive (Solomon et al. 2000; Kessler et al. 2003, 2005; Lopez et al. 2006). The DSM-5 (APA, 2013) criteria for MDD are met when an individual exhibits at least five out of nine psychiatric symptoms during the same 2-week period, with at least one symptom being depressed mood or loss of interest. Four of the nine symptoms comprise different or opposite symptoms (insomnia or hypersomnia, psychomotor agitation or retardation, weight/appetite loss or weight gain, feelings of worthlessness or inappropriate guilt). This leads to 1497 potential unique symptom profiles that all qualify for the same diagnosis (Ostergaard et al. 2011), including profiles that do not have a single symptom in common.

Although variability in depression symptoms has been documented both across (Katschnig et al. 1986; Baumeister & Parker, 2012) and within individuals (Coryell et al. 1994; Oquendo et al. 2004), the DSM and depression screening instruments such as the Beck Depression Inventory (BDI; Beck et al. 1988) and the Hamilton Rating Scale for Depression (HAMD; Hamilton, 1960) build sum scores of depressive symptoms that are used to distinguish healthy and depressed individuals. Sum scores rely tacitly on the DSM-inherent assumption of symptom equivalence.
(Lux & Kendler, 2010): as criterion symptoms contribute equally to a diagnosis, they are considered to be equivalent to each other and thus interchangeable. Symptom equivalence, in turn, is the result of conceptualizing depression within the framework of reflective latent variable modeling (Bollen, 1989; Schmittmann et al. 2013): variation in the latent variable (depression) is understood to be the cause of variation of its observable indicators (DSM depressive symptoms), and as long as sufficient symptoms are present, a diagnosis can be given reliably.

Pronounced symptomatic variability of depression has led to numerous proposed subtypes (e.g. psychotic depression, neurotic depression, anxious depression and melancholic depression). However, no agreement has yet been reached as to the number or validity of these subtypes (Lichtenberg & Belmaker, 2010; Baumeister & Parker, 2012). Furthermore, although symptomatic heterogeneity has been explored by extracting principal components or latent factors, factor solutions differ markedly across and within clinical screening instruments for depression (Furukawa et al. 2005; Shafer, 2006), seem to be arbitrary (Wood et al. 2010), are not invariant across demographic variables such as ethnicity (Crockett et al. 2005), sex (Baas et al. 2011) or age (Williams et al. 2007), and vary depending on the extraction method (Widaman, 1993). Because of these potential limitations of methods that reduce symptom space, in the current study we examined depressive symptoms individually.

It has been suggested that individual symptoms are separate entities that have autonomous causal relevance (Staner, 2010; Borsboom et al. 2011; Schmittmann et al. 2013). This idea is supported by evidence in two small genetic reports that depressive symptoms are associated with specific genes (Jang et al. 2004; Myung et al. 2012), and by a larger study of 7500 twins concluding that the DSM symptomatic criteria for depression do not reflect a single underlying genetic factor (Kendler et al. 2013). It has also been suggested that different symptoms might have different risk factors reflecting different etiologies (Hasler et al. 2004; Cramer et al. 2010; Hasler & Northoff, 2011), but much remains to be done to test the validity of the assumption of symptom equivalence.

Identified risk factors for depression include demographic variables such as age and sex (Piccinelli & Wilkinson, 2000; Kendler et al. 2005), personality traits such as neuroticism (Angst & Clayton, 1986; Kendler et al. 2004), early life adversity (Gilmer & McKinney, 2003; Gutman & Nemeroff, 2003), previous episodes of depression (Beekman et al. 1995; Colman et al. 2011), family history of depression (Nierenberg et al. 2007) and stressful life events (Mazure, 1998; Paykel, 2003). If all depressive symptoms are caused by an underlying disorder, symptoms should have similar risk factors because risk factors only influence the liability to develop depression, and depression in turn causes the symptoms. If, however, risk factors differ for different depression symptoms, this suggests possible substantial benefits from analyses of individual symptoms.

A comprehensive study by Lux & Kendler (2010) investigated cross-sectional associations of all nine DSM criterion symptoms with 25 variables, including demographic information, personality traits, life events, history of depression and lifetime comorbidities in a sample of 10,15 individuals. The results revealed a complex association pattern. For instance, four symptoms were associated with years of education (sleep changes and fatigue with more education, psychomotor problems and suicidal ideation with less education), two symptoms were associated with family income (depressed mood with lower income, concentration problems with higher income), and two symptoms (depressed mood and psychomotor problems) had a significant positive correlation with current age. The authors concluded that the surprising degree of covert heterogeneity is difficult to reconcile with the assumption of symptom equivalence.

Longitudinal investigation of a population that shifts from low to elevated depression levels allows a more detailed examination of the influence of risk factors on specific symptoms. In the current study we used medical residency as a prospective stress model. Medical residency is known to be an intense stressor (Butterfield, 1988; Duffy, 2005), with residents facing long work hours, sleep deprivation, loss of autonomy, and also extreme emotional situations (Shanafelt & Habermann, 2002). In a previous longitudinal study of medical residents, depression levels increased from 3.9% at baseline to 25.7% during residency (Sen et al. 2010). Medical residency thus offers the rare opportunity to assess depressive symptoms and risk factors shortly before the onset of a severe and chronic stressor that drastically increases depression symptoms, allowing us to test the null hypothesis that increases in DSM depression criterion symptoms are predicted by similar risk factors.

Method

Sample

A total of 4005 interns entering residency programs in the USA during the 2009–2011 academic years were invited to participate in the study; 58% (2455/4005) accepted the invitation. The institutional review boards at participating hospitals approved the study. Participating subjects provided electronic informed consent, and were given US$50 in gift certificates.
Symptom had bothered them for the DSM-IV (APA, 2000) criterion symptoms of depression (DSM-IV and DSM-5 MDD symptoms are identical). For each of the nine symptoms, subjects indicated whether, during the previous 2 weeks, the symptom had bothered them ‘not at all’, ‘several days’, ‘more than half the days’ or ‘nearly every day’. Each item yields a score of 0, 1, 2 or 3. The nine symptoms assessed by the PHQ-9 are: little interest or pleasure in doing things (interest), feeling depressed or hopeless (depressed), sleep problems (sleep), feeling tired (fatigue), appetite problems (appetite), feeling bad about yourself/that you are a failure (self-blame), trouble concentrating on things (concentration), moving or speaking slowly/being fidgety or restless (psychomotor), and suicidal ideation (suicide).

Subjects completed a baseline survey 1–2 months prior to commencing internship that assessed general demographic factors (age, sex), personal factors (baseline PHQ-9 depressive symptoms, self-reported history of depression, self-reported family history of depression, childhood stress (Risky Families Questionnaire; Taylor et al. 2006) and neuroticism (Neuroticism–Extroversion–Openness Five-Factor Inventory, NEO-FFI; Costa & McCrae, 1997). Participants were contacted by email at months 3, 6, 9 and 12 of their internship year and asked to complete the PHQ-9 again. They were also queried regarding work hours and number of stressful events) were constructed analogous to symptoms. Average scores for the two risk factors that were assessed after baseline (work hours and number of stressful life events) were constructed analogous to symptoms.

Three analyses were conducted. First, we used a series of paired t tests comparing baseline levels of each symptom to the average symptom level during residency to determine which symptoms increased significantly.

Second, we investigated whether risk factors differentially predicted symptom change over time. We estimated a structural equation model (SEM) encompassing nine linear regressions, with one regression for each symptom. In each regression, the severity of a given symptom during residency was predicted by five baseline risk factors (sex, history of depression, family history of depression, neuroticism, and childhood stress) and two within-internship risk factors (work hours and number of stressful events), controlling for baseline severity of the symptom. To avoid overfitting, we used identical predictors in all nine regressions, and did not drop insignificant predictors to reach the best possible fit for each symptom. We did, however, decide to exclude predictors from the analysis in case they were not related to any of the nine symptoms. We then compared two models, the homogeneity model and the heterogeneity model. In the homogeneity model, regression weights were constrained to be equal for each risk factor across all symptoms. Specifically, a regression coefficient was estimated for each risk factor separately, but this coefficient was constrained to have the same effect on each symptom. This model represents the hypothesis that depression is the common cause for all symptoms, with each risk factor affecting all symptoms equally.

In the heterogeneity model, all regression coefficients were estimated freely. The models were compared using a χ² test, and model fit was assessed using the root mean square error of approximation (RMSEA; ≤0.06 indicating a good fit) and the comparative fit index (CFI; ≥0.95 indicating a good fit) (Hu & Bentler, 1999). If the constrained homogeneity model fit worse than the heterogeneity model, this would indicate that risk factors vary in their effects across symptoms. Both models were estimated using the maximum likelihood estimator (MLE), and the results of the model with better fit were visualized using the R-package qgraph (Epskamp et al. 2012).

Although our second analysis provides information about potential heterogeneity on the basis of individual depressive symptoms have different risk factors through repeated measurement. When averaging the within-internship symptom scores, 1166 of the 2455 subjects (47.5%) were dropped by listwise deletion because they had missing data on two or more time points, leaving 1289 subjects in the analytic sample. Average scores for the two risk factors that were assessed after baseline (work hours and number of stressful life events) were constructed analogous to symptoms.
observable symptoms, it does not rule out the possibility that heterogeneity would disappear in a latent disease model. Thus, we investigated whether controlling for a latent depression variable eliminates heterogeneity, using a longitudinal multiple indicators multiple causes (MIMIC) model. MIMIC models (Jöreskog & Goldberger, 1975) encompass one or more latent variables that have both observable indicators (in our case symptoms) and variables that influence the latent variables (in our case risk factors). Following the approach described by Jones (2006), we compared two versions of a MIMIC model. The first (model I) allows for risk factors to influence symptoms directly whereas the second (model II) represents the hypothesis that risk factors only affect the latent depression variable, which, in turn, influences individual depressive symptoms (Fig. 1). Evidence that these direct paths improve model fit significantly would indicate that symptom level effects of risk factors exist and are not mediated by D2.

Table 1. Demographic characteristics of study participants

<table>
<thead>
<tr>
<th>Variable</th>
<th>n (% )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>617 (47.9)</td>
</tr>
<tr>
<td>Female</td>
<td>672 (52.1)</td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
</tr>
<tr>
<td>≤ 25</td>
<td>221 (17.1)</td>
</tr>
<tr>
<td>26–30</td>
<td>867 (67.3)</td>
</tr>
<tr>
<td>31–35</td>
<td>111 (8.6)</td>
</tr>
<tr>
<td>&gt;35</td>
<td>28 (2.2)</td>
</tr>
<tr>
<td>History of depression</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>568 (44.1)</td>
</tr>
<tr>
<td>No</td>
<td>721 (55.9)</td>
</tr>
<tr>
<td>Family history of depression</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>682 (52.9)</td>
</tr>
<tr>
<td>No</td>
<td>607 (47.1)</td>
</tr>
<tr>
<td>Specialty</td>
<td></td>
</tr>
<tr>
<td>Internal medicine</td>
<td>481 (37.3)</td>
</tr>
<tr>
<td>Other</td>
<td>158 (12.3)</td>
</tr>
<tr>
<td>Pediatrics</td>
<td>150 (11.6)</td>
</tr>
<tr>
<td>General surgery</td>
<td>113 (8.8)</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>93 (7.2)</td>
</tr>
<tr>
<td>Emergency medicine</td>
<td>81 (6.3)</td>
</tr>
<tr>
<td>Family medicine</td>
<td>62 (4.8)</td>
</tr>
<tr>
<td>Obstetrics/gynecology</td>
<td>47 (3.6)</td>
</tr>
<tr>
<td>Internal medicine/pediatrics</td>
<td>35 (2.7)</td>
</tr>
<tr>
<td>Neurology</td>
<td>32 (2.5)</td>
</tr>
<tr>
<td>Transitional</td>
<td>27 (2.1)</td>
</tr>
<tr>
<td>Missing</td>
<td>10 (0.8)</td>
</tr>
</tbody>
</table>

Results

In total, 1289 interns were included in the final analyses. The mean age of the study participants was 27.6 years (s.d.=3.02) and 47.9% were male (Table 1).
Participants who had to be dropped due to missing values did not differ significantly from the retained participants in sex, neuroticism, childhood stress, personal and family history of depression, work hours, and number of stressful life events (all \( p > 0.05 \)).

**Effects of residency on depressive symptoms**

All depression symptoms increased significantly during residency training (\( t_{1288} \) ranging from 6.77 to 40.79, all \( p < 0.001 \)). On average, symptoms increased by 169% of their baseline means, with a range of 48% (sleep problems) to 270% (psychomotor problems).

<table>
<thead>
<tr>
<th>Table 2. Goodness-of-fit statistics and ( \chi^2 ) difference tests for the two model comparisons (heterogeneity v. homogeneity, MIMIC models I v. II)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Test for heterogeneity</td>
</tr>
<tr>
<td>Model I(^a)</td>
</tr>
<tr>
<td>Model II(^b)</td>
</tr>
<tr>
<td>MIMIC models</td>
</tr>
<tr>
<td>Model I(^c)</td>
</tr>
<tr>
<td>Model II(^d)</td>
</tr>
</tbody>
</table>

MIMIC, Multiple indicators multiple causes; df, degrees of freedom; RMSEA, root mean square error of approximation; CFI, comparative fit index; \( \chi^2_{\text{diff}} \), \( \chi^2 \) statistic of the \( \chi^2 \) difference test; df_{\text{diff}}, degrees of freedom of the \( \chi^2 \) difference test; \( p \), \( p \) value of the \( \chi^2 \) difference test.

\(^a\) Heterogeneity model.

\(^b\) Homogeneity model.

\(^c\) Nineteen significant direct paths from risk factors to symptoms.

\(^d\) No direct paths from risk factors to symptoms.

<table>
<thead>
<tr>
<th>Table 3. PHQ-9 symptoms predicted by risk factors (heterogeneity model)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Female sex</strong></td>
</tr>
<tr>
<td>( \beta )</td>
</tr>
<tr>
<td>Interest</td>
</tr>
<tr>
<td>Depressed</td>
</tr>
<tr>
<td>Sleep</td>
</tr>
<tr>
<td>Fatigue</td>
</tr>
<tr>
<td>Appetite</td>
</tr>
<tr>
<td>Self-blame</td>
</tr>
<tr>
<td>Concentration</td>
</tr>
<tr>
<td>Psychomotor</td>
</tr>
<tr>
<td>Suicide</td>
</tr>
</tbody>
</table>

PHQ-9, Nine-item depression scale of the Patient Health Questionnaire; \( \beta \), standardized regression coefficients of risk factors on symptoms; \( R^2 \), adjusted \( R \)-squared values.

Family history of depression is not displayed because none of the effects were significant.

\(* p < 0.05, ** p < 0.01, *** p < 0.001.\)

Impact of risk factors on symptoms

Family history of depression was the only risk factor that was not related to changes of any of the nine symptoms and was excluded from subsequent analyses. The homogeneity model, in which each risk factor was constrained to have equal impact on all symptoms, fit the data significantly worse than the heterogeneity model, in which the effects of risk factors on symptoms were estimated freely (\( p < 0.001 \)). The heterogeneity model fit the data well, and the highly significant \( \chi^2 \) difference test indicated a strong differential impact of risk factors on symptoms (Table 2).

The results for the heterogeneity model are summarized in Table 3 and visualized in Fig. 2. Of the various
Findings, three are particularly noteworthy. First, four risk factors (prior history of depression, higher number of stressful life events, childhood stress, and sex) predicted worsening of unique subsets of three to seven of the symptoms. Second, two predictors (neuroticism and work hours) had a significant impact on all nine PHQ-9 symptoms over time. Third, female residents showed increases in sleep problems, appetite problems and fatigue during residency whereas male residents reported increased suicidal ideation.

To assess whether the two uniform predictors neuroticism and work hours varied in their predictive strength across symptoms, we compared 95% confidence intervals (CIs) of the regression coefficients across symptoms. Of 36 symptom pairs, 14 CIs (38.9%) did not overlap for neuroticism and 14 CIs (38.9%) did not overlap for work hours (see online Supplementary Table S1). This implies considerable variability in the predictive strength of neuroticism and work hours across symptoms.

**Latent disease model**

In this analysis, we compared the fit of two longitudinal MIMIC models. As preliminary analyses revealed that factor loadings were not invariant across time, we allowed factor loadings to vary across measurement points in both MIMIC models. Because of high modification indices, residuals of the following symptoms were allowed to be correlated in both models: symptoms sleep and fatigue at baseline, and symptoms depressed and appetite, depressed and self-blame, fatigue and appetite, along with concentration and psychomotor within internship. As a result of the iterative fitting process described in the Method section, model I retained 19 significant direct effects of risk factors on symptoms after controlling for the latent depression factor.

The results of fitting the two models to the data are presented in Table 2. The \( \chi^2 \) difference test indicated that the 19 direct paths from risk factors to symptoms in model I improved the fit significantly (\( p<0.001 \)). This means that heterogeneity of risk factors and symptoms cannot be explained by a latent depression variable.

**Discussion**

In this study, we tested the hypothesis that the nine DSM-5 depression symptoms have similar risk factors. The evidence from a longitudinal population cohort of 1289 medical interns is inconsistent with this
Individual depressive symptoms have different risk factors

hypothesis. Instead, the results support the alternative hypothesis that risk factors have differential impact on depression symptoms. Four risk factors predicted increases of unique subsets of symptoms. Neuroticism and work hours did predict increases of all symptoms, albeit to varying magnitudes. We also tested whether the pronounced heterogeneity could be explained by a latent depression factor; this was not the case.

These results support the notion that depressive symptoms have different characteristics and properties. Studies of depression typically use sum scores to describe severity, a strategy that may obfuscate crucial information about the nature of depression symptoms and causes. For instance, in this report, women were more likely to report worsening of sleep and appetite problems in addition to fatigue during internship, whereas men reported increased suicidal ideation, information that would not have been available from assessing depression sum scores alone. Furthermore, the use of sum scores groups individuals with dramatic inter-individual symptom differences together in the same diagnostic group. This covert heterogeneity might help to explain recent studies reporting ‘disappointing’ results such as questionable reliability of depression diagnosis (Regier et al., 2012), and the striking lack of success of genome-wide association studies (GWAS) and other genetic studies in identifying loci definitively associated with depression and treatment response (Tansey et al., 2012; Hek et al., 2013; see also Power et al., 2012). Our results are consistent with the growing chorus of voices suggesting that covert heterogeneity of DSM diagnostic criteria in research may be inhibiting progress in elucidating the biological roots of mental illness (Lux & Kendler, 2010; Insel, 2013; Kendler et al., 2013).

Assessing depression symptoms individually promises to reveal new findings. Special attention is needed for the four compound symptoms (insomnia or hypersomnia, psychomotor agitation or retardation, weight loss or weight gain, feelings of worthlessness or inappropriate guilt). For example, chronic stress is more strongly associated with weight gain than weight loss (Keller et al., 2007), and it is plausible that these opposite symptoms might arise from different etiologies. In addition, the symptoms chosen as criteria for DSM depression are a small subset of possible depression symptoms (McGlinchey et al., 2006; Zimmerman et al., 2006a) and were determined largely by clinical consensus instead of empirical evidence (Zimmerman et al., 2006b; Kendler & Zachar, 2008; Lux & Kendler, 2010).

For instance, helplessness and hopelessness as compound symptoms have been shown to perform more strongly than some of the DSM-IV criterion symptoms in distinguishing depressed from healthy individuals (McGlinchey et al., 2006).

Furthermore, prior work shows that different life stressors provoke different symptoms in patterns that are somewhat consistent (Keller & Nesse, 2005, 2006; Keller et al., 2007; Cramer et al., 2011), supporting the need for close attention to heterogeneity of depression symptoms.

Future research informed about differences between individual symptoms will facilitate the search for etiologically and symptomatically homogeneous groups, and thus have the potential to increase treatment efficacy.

Limitations

These findings need to be interpreted in the light of four limitations. First, medical residents are relatively homogeneous in terms of age, education and socio-economic status, so our results may not generalize to other populations.

Second, the PHQ-9 does not contain detailed information about bidirectional symptoms. It is thus unclear whether, for instance, the increasingly severe sleep problems in female residents manifested themselves as insomnia or hypersomnia, or whether the increases in psychomotor problems reported by interns who experienced a higher number of stressful life events were due to psychomotor agitation or psychomotor retardation.

Third, in preliminary analyses conducted prior to MIMIC modeling, a one-factor solution emerged for the first measurement point whereas a two-factor solution fit the second measurement point better. As the MIMIC models were not interpretable from a substantive point of view with different numbers of factors across time, we estimated one factor for each measurement point but allowed factor loadings to be freely estimated. This approach penalizes the fit of both MIMIC models equally and should not favor one over the other.

Fourth, because multivariate models that contained nine dependent variables, six risk factors and five measurement points, and controlled for a latent depression factor, were not interpretable, our analyses are limited to two time points. As a result, if the relationship between specific variables changed during internship, these changes may have been masked. However, correlations of symptoms with time-varying risk factors at each individual time point were fairly stable across the four post-stress onset measurements, and comparable to the averaged time point used in our analyses. Our approach therefore does not seem to substantially distort the relationships of these variables.

Supplementary material

For supplementary material accompanying this paper, please visit http://dx.doi.org/10.1017/S0033291713002900.
Acknowledgments
We thank M. Schultze and Dr K. Shedden for their valuable statistical input. E. Fried is supported by fellowships from the Cluster of Excellence ‘Languages of Emotion’ (grant no. EXC302), the German Research Foundation, and the Dahlem Research School Berlin. Funding was also provided by the National Institute of Mental Health (R01 MH101459, K23 MH095109).

Declaration of interest
None.

References


reliability of selected categorical diagnoses. *American Journal of Psychiatry* 170, 59–70.


