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

E. I. Fried[1,2], R. M. Nesse[3,4], K. Zivin[4,5,6], C. Guille[7] and S. Sen[4]*

[1] *Cluster of Excellence 'Languages of Emotion', Freie Universität Berlin, Germany*

[2] *Department of Education and Psychology, Freie Universität Berlin, Germany*

[3] *Department of Psychology, University of Michigan, Ann Arbor, MI, USA*

[4] *Department of Psychiatry, University of Michigan Medical School, Ann Arbor, MI, USA*

[5] *Department of Veterans Affairs, National Serious Mental Illness Treatment Resource and Evaluation Center, Ann Arbor, MI, USA*

[6] *Institute for Social Research, University of Michigan, Ann Arbor, MI, USA*

[7] *Medical 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.

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

---

* Address for correspondence: Dr S. Sen, Rachel Upjohn Building, 4250 Plymouth Rd, Ann Arbor, MI 48109-5734, USA.

(Email: srijan@umich.edu)

(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 co-morbidities in a sample of 1015 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.

Assessment

All surveys were conducted through a secure online website designed to maintain confidentiality. Depressive symptoms were measured using the nine-item depression scale of the Patient Health Questionnaire (PHQ-9; Spitzer *et al*. 1999). The PHQ-9 is a self-report component of the Primary Care Evaluation of Mental Disorders (PRIME-MD) inventory designed to screen 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 over the past week and the occurrence of a series of stressful non-internship life events (marriage, childbirth, serious illness, death or serious illness in close family or friend, financial problems, end of a serious relationship, or becoming a victim of crime or domestic violence) during the past 3 months. Because the number of life events reported in each category was too low to use in subsequent analyses, we computed a sum score of all events for each measurement point for each subject and subsequently used it as the predictor.

Statistical analysis

We compared symptom severity at baseline with average symptom severity during the four measurements across the residency. This approach has been used in previous publications based on this dataset (Sen *et al*. 2010, 2013), and has the advantage of increased reliability of symptom assessment within internship 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.

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 χ^2 test, and model fit was assessed using the root mean square error of approximation (RMSEA; $\leqslant 0.06$ indicating a good fit) and the comparative fix index (CFI; $\geqslant 0.95$ 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
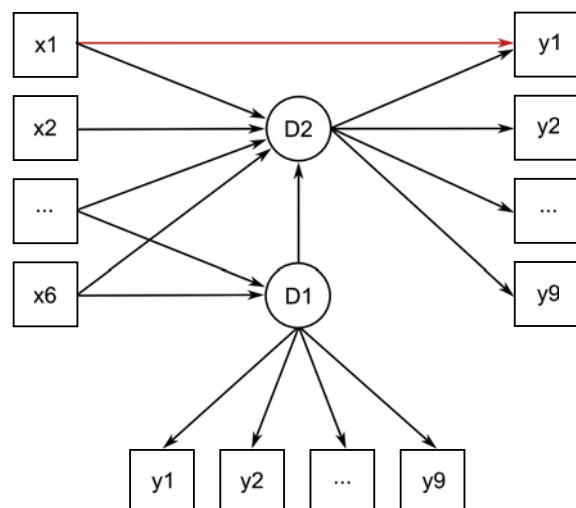
4 E. I. Fried et al.

Fig. 1. Visualization of the longitudinal multiple indicators multiple causes (MIMIC) models used in this study. D1, Latent depression factor at baseline; D2, latent depression factor during residency; y1–y9, depressive symptoms; x1 and x2, within-internship risk factors; x3–x6, baseline risk factors; the red arrow is an example of a significant direct effect of a risk factor on a symptom (in this case, x1 on y1) that is not mediated by D2.

Table 1. *Demographic characteristics of study participants*

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

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 a latent variable. Model I was constructed by an iterative process. First, all risk factors were allowed to have direct effects on all symptoms, except for the last PHQ-9 symptom suicidal ideation (for the purpose of model identification). Second, non-significant direct paths were removed and significant ones added for suicidal ideation until only significant paths from risk factors to symptoms remained. Model comparison and fit of the two MIMIC models were assessed analogous to the second analysis. A complication of the MIMIC model is the longitudinal nature of our data. Similar to our second analysis, we thus predicted the second measurement point by risk factors, controlling for the first. Although internship work hours and number of stressful life events naturally only affected the within-internship latent, baseline risk factors (sex, history of depression, neuroticism and childhood stress) were allowed to additionally affect the baseline latent variable. In both models, the residual of each symptom at baseline was allowed to be correlated with the residual of the same symptom during residency. Both MIMIC models were estimated using the MLE.

The first analysis was performed using R 2.13.0 (R Development Core Team, 2008), the second and third analyses were conducted in Mplus 7.0 (Muthén & Muthén, 1998–2012).

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).

Table 2. *Goodness-of-fit statistics and χ^2 difference tests for the two model comparisons (heterogeneity v. homogeneity, MIMIC models I v. II)*

	χ^2	df	RMSEA	CFI	χ^2_{diff}	df_{diff}	p
Test for heterogeneity							
Model I[a]	197.3	72	0.04	0.98			
Model II[b]	617.4	120	0.06	0.93	420.1	48	< 0.001
MIMIC models							
Model I[c]	868.5	199	0.05	0.94			
Model II[d]	1041.5	218	0.05	0.93	173.0	19	< 0.001

MIMIC, Multiple indicators multiple causes; df, degrees of freedom; RMSEA, root mean square error of approximation; CFI, comparative fit index; χ^2_{diff}, χ^2 statistic of the χ^2 difference test; df_{diff}, degrees of freedom of the χ^2 difference test; p, p value of the χ^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 3. *PHQ-9 symptoms predicted by risk factors (heterogeneity model)*

	Female sex β	Neuroticism β	Childhood stress β	History of depression β	Stressful life events β	Work hours β	R^2
Interest	−0.01	0.29***	0.05	0.13***	0.05	0.15***	0.20
Depressed	0.01	0.32***	0.04	0.15***	0.03	0.09***	0.23
Sleep	0.09***	0.20***	0.02	0.09**	0.04	0.07**	0.13
Fatigue	0.10***	0.20***	0.05	0.11***	0.05*	0.24***	0.20
Appetite	0.09***	0.24***	0.04	0.08**	0.05*	0.15***	0.20
Self-blame	0.04	0.36***	0.03	0.11***	0.02	0.10***	0.26
Concentration	0.05*	0.19***	0.08**	0.12***	0.08**	0.14***	0.17
Psychomotor	−0.04	0.22***	0.08**	0.03	0.06*	0.14***	0.11
Suicide	−0.08**	0.22***	0.10***	0.04	0.08**	0.05*	0.14

PHQ-9, Nine-item depression scale of the Patient Health Questionnaire; β, 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.

* $p<0.05$, ** $p<0.01$, *** $p<0.001$.

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).

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 χ^2 difference test indicated a strong differential impact of risk factors on symptoms (Table 2).

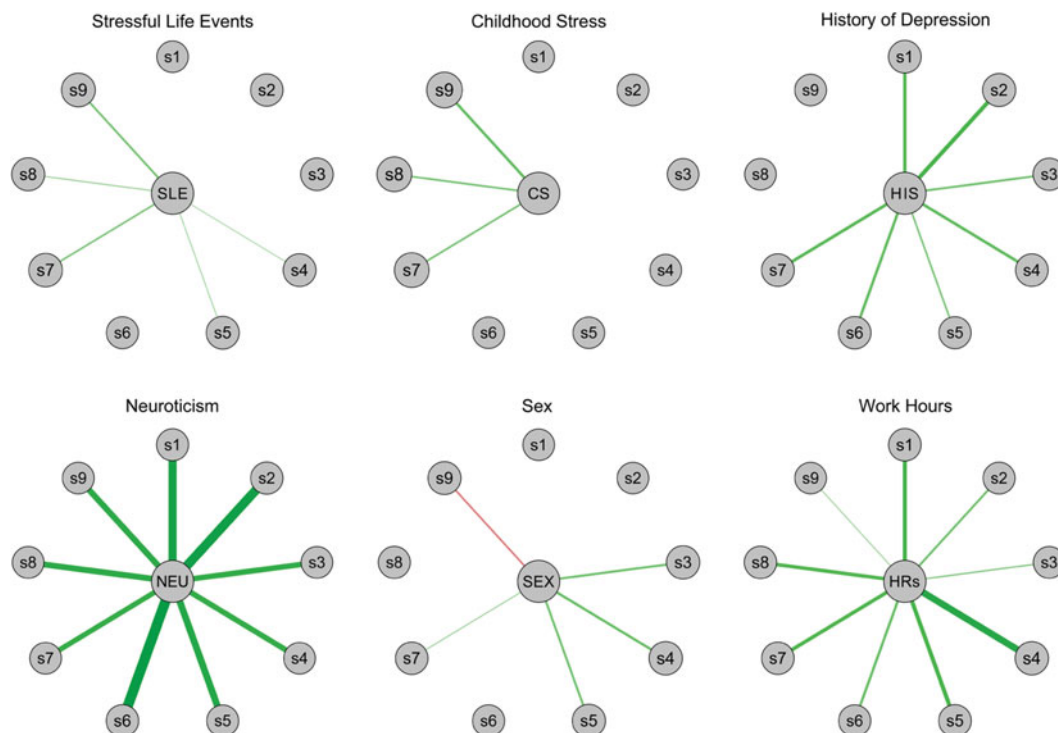The results for the heterogeneity model are summarized in Table 3 and visualized in Fig. 2. Of the various

Fig. 2. Standardized regression coefficients of six risk factors on changes of nine depression symptoms during residency, as estimated by the heterogeneity model. Thickness of lines indicates strength of regression weights. Green lines represent positive regression weights, red lines negative ones; sex was coded 0=male, 1=female. s1, interest; s2, depressed; s3, sleep; s4, fatigue; s5, appetite; s6, self-blame; s7, concentration; s8, psychomotor; s9, suicide.

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 χ^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

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*. 2006*a*) and were determined largely by clinical consensus instead of empirical evidence (Zimmerman *et al*. 2006*b*; 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 socioeconomic 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

Declaration of interest

None.

References

Angst J, Clayton P (1986). Premorbid personality of depressive, bipolar, and schizophrenic patients with special reference to suicidal issues. *Comprehensive Psychiatry* **27**, 511–532.

APA (2000). *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. American Psychiatric Association: Washington, DC.

APA (2013). *Diagnostic and Statistical Manual of Mental Disorders*, 5th edn. American Psychiatric Association: Washington, DC.

Baas KD, Cramer AOJ, Koeter MWJ, van de Lisdonk EH, van Weert HC, Schene AH (2011). Measurement invariance with respect to ethnicity of the Patient Health Questionnaire-9 (PHQ-9). *Journal of Affective Disorders* **129**, 229–235.

Baumeister H, Parker JD (2012). Meta-review of depressive subtyping models. *Journal of Affective Disorders* **139**, 126–140.

Beck AT, Steer RA, Garbin MG (1988). Psychometric properties of the Beck Depression Inventory: 25 years of evaluation. *Clinical Psychology Review* **8**, 77–100.

Beekman ATF, Deeg DJH, van Tilburg T, Smit JH, Hooijer C, van Tilburg W (1995). Major and minor depression in later life: a study of prevalence and risk factors. *Journal of Affective Disorders* **36**, 65–75.

Bollen KA (1989). *Structural Equations with Latent Variables*. Wiley: New York.

Borsboom D, Cramer AOJ, Schmittmann VD, Epskamp S, Waldorp LJ (2011). The small world of psychopathology. *PloS One* **6**, e27407.

Butterfield PS (1988). The stress of residency: a review of the literature. *Archives of Internal Medicine* **148**, 1428–1435.

Colman I, Naicker K, Zeng Y, Ataullahjan A, Senthilselvan A, Patten SB (2011). Predictors of long-term prognosis of depression. *Canadian Medical Association Journal* **183**, 1969–1976.

Coryell W, Winokur G, Shea T, Maser JD, Endicott J, Akiskal HS (1994). The long-term stability of depressive subtypes. *American Journal of Psychiatry* **151**, 199–204.

Costa PT Jr., McCrae RR (1997). Stability and change in personality assessment: the revised NEO Personality Inventory in the year 2000. *Journal of Personality Assessment* **68**, 86–94.

Cramer AOJ, Borsboom D, Aggen SH, Kendler KS (2011). The pathoplasticity of dysphoric episodes: differential impact of stressful life events on the pattern of depressive symptom inter-correlations. *Psychological Medicine* **42**, 957–965.

Cramer AOJ, Waldorp LJ, van der Maas HLJ, Borsboom D (2010). Comorbidity: a network perspective. *Behavioral and Brain Sciences* **33**, 137–150.

Crockett LJ, Randall BA, Shen Y-L, Russell ST, Driscoll AK (2005). Measurement equivalence of the center for epidemiological studies depression scale for Latino and Anglo adolescents: a national study. *Journal of Consulting and Clinical Psychology* **73**, 47–58.

Duffy TP (2005). Glory days: what price glory? *The Pharos of Alpha Omega Alpha-Honor Medical Society* **68**, 22–30.

Epskamp S, Cramer AOJ, Waldorp LJ, Schmittmann VD, Borsboom D (2012). qgraph: network visualizations of relationships in psychometric data. *Journal of Statistical Software* **48**, 1–18.

Furukawa TA, Streiner DL, Azuma H, Higuchi T, Kamijima K, Kanba S, Ozaki N, Aoba A, Murasaki M, Miura S (2005). Cross-cultural equivalence in depression assessment: Japan-Europe-North American study. *Acta Psychiatrica Scandinavica* **112**, 279–285.

Gilmer WS, McKinney WT (2003). Early experience and depressive disorders: human and non-human primate studies. *Journal of Affective Disorders* **75**, 97–113.

Gutman DA, Nemeroff CB (2003). Persistent central nervous system effects of an adverse early environment: clinical and preclinical studies. *Physiology and Behavior* **79**, 471–478.

Hamilton M (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry* **23**, 56–62.

Hasler G, Drevets WC, Manji HK, Charney DS (2004). Discovering endophenotypes for major depression. *Neuropsychopharmacology* **29**, 1765–1781.

Hasler G, Northoff G (2011). Discovering imaging endophenotypes for major depression. *Molecular Psychiatry* **16**, 604–619.

Hek K, Demirkan A, Lahti J, Terracciano A, Teumer A, Cornelis MC, Amin N, Bakshis E, Baumert J, Ding J, Liu Y, Marciante K, Meirelles O, Nalls MA, Sun YV, Vogelzangs N, Yu L, Bandinelli S, Benjamin EJ, Bennett DA, Boomsma D, Cannas A, Coker LH, de Geus E, De Jager PL, Diez-Roux AV, Purcell S, Hu FB, Rimm EB, Hunter DJ, Jensen MK, Curhan G, Rice K, Penman AD, Rotter JI, Sotoodehnia N, Emeny R, Eriksson JG, Evans DA, Ferrucci L, Fornage M, Gudnason V, Hofman A, Illig T, Kardia S, Kelly-Hayes M, Koenen K, Kraft P, Kuningas M, Massaro JM, Melzer D, Mulas A, Mulder CL, Murray A, Oostra BA, Palotie A, Penninx B, Petersmann A, Pilling LC, Psaty B, Rawal R, Reiman EM, Schulz A, Shulman JM, Singleton AB, Smith AV, Sutin AR, Uitterlinden AG, Völzke H, Widen E, Yaffe K, Zonderman AB, Cucca F, Harris T, Ladwig KH, Llewellyn DJ, Räikkönen K, Tanaka T, van Duijn CM, Grabe HJ, Launer LJ, Lunetta KL, Mosley TH Jr.,

Newman AB, Tiemeier H, Murabito J (2013). A genome-wide association study of depressive symptoms. *Biological Psychiatry* **73**, 667–678.

Hu L, Bentler PM (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling* **6**, 1–55.

Insel T (2013). Director's blog: transforming diagnosis. National Institute of Mental Health (www.nimh.nih.gov/about/director/2013/transforming-diagnosis.shtml). Accessed 6 June 2013.

Jang KL, Livesley WJ, Taylor S, Stein MB, Moon EC (2004). Heritability of individual depressive symptoms. *Journal of Affective Disorders* **80**, 125–133.

Jones RN (2006). Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination. Detecting differential item functioning using MIMIC modeling. *Medical Care* **44**, 124–133.

Jöreskog K, Goldberger A (1975). Estimation of a model of multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association* **10**, 631–639.

Katschnig H, Pakesch G, Egger-Zeidner E (1986). Life stress and depressive subtypes: a review of present diagnostic criteria and recent research results. In *Life Events and Psychiatric Disorders: Controversial Issues* (ed. H. Katschnig), pp. 201–245. Cambridge University Press: Cambridge.

Keller MC, Neale MC, Kendler KS (2007). Association of different adverse life events with distinct patterns of depressive symptoms. *American Journal of Psychiatry* **164**, 1521–1529.

Keller MC, Nesse RM (2005). Is low mood an adaptation? Evidence for subtypes with symptoms that match precipitants. *Journal of Affective Disorders* **86**, 27–35.

Keller MC, Nesse RM (2006). The evolutionary significance of depressive symptoms: different adverse situations lead to different depressive symptom patterns. *Journal of Personality and Social Psychology* **91**, 316–330.

Kendler KS, Aggen SH, Neale MC (2013). Evidence for multiple genetic factors underlying DSM-IV criteria for major depression. *Journal of the American Medical Association Psychiatry* **70**, 599–607.

Kendler KS, Kuhn J, Prescott CA (2004). The interrelationship of neuroticism, sex, and stressful life events in the prediction of episodes of major depression. *American Journal of Psychiatry* **161**, 631–636.

Kendler KS, Myers J, Prescott CA (2005). Sex differences in the relationship between social support and risk for major depression: a longitudinal study of opposite-sex twin pairs. *American Journal of Psychiatry* **162**, 250–256.

Kendler KS, Zachar P (2008). The incredible insecurity of psychiatric nosology. In *Philosophical Issues in Psychiatry* (ed. K. S. Kendler and J. Parnas), pp. 368–383. The Johns Hopkins University Press: Baltimore, MD.

Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR, Rush AJ, Walters EE, Wang PS (2003). The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication. *Journal of the American Medical Association* **289**, 3095–3105.

Kessler RC, Chiu WT, Demler O, Walters EE (2005). Prevalence, severity, and comorbidity of twelve-month DSM-IV disorders in the National Comorbidity Survey Replication (NCS-R). *Archives of General Psychiatry* **62**, 617–627.

Lichtenberg P, Belmaker RH (2010). Subtyping major depressive disorder. *Psychotherapy and Psychosomatics* **79**, 131–135.

Lopez AD, Mathers CD, Ezzati M, Jamiso DT, Murray CJL (2006). *Global Burden of Disease and Risk Factors*. Oxford University Press and The World Bank: New York.

Lux V, Kendler KS (2010). Deconstructing major depression: a validation study of the DSM-IV symptomatic criteria. *Psychological Medicine* **40**, 1679–1690.

Mazure CM (1998). Life stressors as risk factors in depression. *Clinical Psychology: Science and Practice* **5**, 291–313.

McGlinchey JB, Zimmerman M, Young D, Chelminski I (2006). Diagnosing major depressive disorder VIII: are some symptoms better than others? *Journal of Nervous and Mental Disease* **194**, 785–790.

Muthén LK, Muthén BO (1998–2012). *Mplus User's Guide. Seventh Edition*. Muthén & Muthén: Los Angeles, CA.

Myung W, Song J, Lim S-W, Won H-H, Kim S, Lee Y, Kang HS, Lee H, Kim J-W, Carroll BJ, Kim DK (2012). Genetic association study of individual symptoms in depression. *Psychiatry Research* **198**, 400–406.

Nierenberg AA, Trivedi MH, Fava M, Biggs MM, Shores-Wilson K, Wisniewski SR, Balasubramani GK, Rush AJ (2007). Family history of mood disorder and characteristics of major depressive disorder: a STAR*D (sequenced treatment alternatives to relieve depression) study. *Journal of Psychiatric Research* **41**, 214–221.

Oquendo MA, Barrera A, Ellis SP, Li S, Burke AK, Grunebaum M, Endicott J, Mann JJ (2004). Instability of symptoms in recurrent major depression: a prospective study. *American Journal of Psychiatry* **161**, 255–261.

Ostergaard SD, Jensen SOW, Bech P (2011). The heterogeneity of the depressive syndrome: when numbers get serious. *Acta Psychiatrica Scandinavica* **124**, 495–496.

Paykel ES (2003). Life events and affective disorders. *Acta Psychiatrica Scandinavica* **108**, 61–66.

Piccinelli M, Wilkinson G (2000) Gender differences in depression. Critical review. *British Journal of Psychiatry* **177**, 486–492.

Power RA, Keers R, Ng MY, Butler AW, Uher R, Cohen-Woods S, Ising M, Craddock N, Owen MJ, Korszun A, Jones L, Jones I, Gill M, Rice JP, Hauser J, Henigsberg N, Maier W, Zobel A, Mors O, Placentino AS, Rietschel M, Souery D, Kozel D, Preisig M, Lucae S, Binder EB, Aitchison KJ, Tozzi F, Muglia P, Breen G, Craig IW, Farmer AE, Müller-Myhsok B, McGuffin P, Lewis CM (2012). Dissecting the genetic heterogeneity of depression through age at onset. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics* **159B**, 859–868.

R Development Core Team (2008). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna.

Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, Kupfer DJ (2012). DSM-5 field trials in the United States and Canada, Part II: test-retest

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

Schmittmann VD, Cramer AOJ, Waldorp LJ, Epskamp S, Kievit RA, Borsboom D (2013). Deconstructing the construct: a network perspective on psychological phenomena. *New Ideas in Psychology* **31**, 43–53.

Sen S, Kranzler HR, Didwania AK, Schwartz AC, Amarnath S, Kolars JC, Dalack GW, Nichols B, Guille C (2013). Effects of the 2011 duty hour reforms on interns and their patients: a prospective longitudinal cohort study. *Journal of the American Medical Association Internal Medicine* **173**, 657–662.

Sen S, Kranzler HR, Krystal JH, Speller H, Chan G, Gelernter J, Guille C (2010). A prospective cohort study investigating factors associated with depression during medical internship. *Archives of General Psychiatry* **67**, 557–565.

Shafer AB (2006). Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *Journal of Clinical Psychology* **62**, 123–146.

Shanafelt T, Habermann T (2002). Medical residents' emotional well-being. *Journal of the American Medical Association* **288**, 1846–1847.

Solomon DA, Keller MB, Leon AC, Mueller TI, Lavori PW, Shea MT, Coryell W, Warshaw M, Turvey C, Maser JD, Endicott J (2000). Multiple recurrences of major depressive disorder. *American Journal of Psychiatry* **157**, 229–233.

Spitzer RL, Kroenke K, Williams JB (1999). Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. *Journal of the American Medical Association* **282**, 1737–1744.

Staner L (2010). Comorbidity of insomnia and depression. *Sleep Medicine Reviews* **14**, 35–46.

Tansey KE, Guipponi M, Perroud N, Bondolfi G, Domenici E, Evans D, Hall SK, Hauser J, Henigsberg N, Hu X, Jerman B, Maier W, Mors O, O'Donovan M, Peters TJ, Placentino A, Rietschel M, Souery D, Aitchison KJ, Craig I, Farmer A, Wendland JR, Malafosse A, Holmans P, Lewis G, Lewis CM, Stensbøl TB, Kapur S, McGuffin P, Uher R (2012). Genetic predictors of response to serotonergic and noradrenergic antidepressants in major depressive disorder: a genome-wide analysis of individual-level data and a meta-analysis. *PLOS Medicine* **9**, e1001326.

Taylor SE, Way BM, Welch WT, Hilmert CJ, Lehman BJ, Eisenberger NI (2006). Early family environment, current adversity, the serotonin transporter promoter polymorphism, and depressive symptomatology. *Biological Psychiatry* **60**, 671–676.

Widaman K (1993). Common factor analysis versus principal component analysis: differential bias in representing model parameters? *Multivariate Behavioral Research* **28**, 263–311.

Williams CD, Taylor TR, Makambi K, Harrell J, Palmer JR, Rosenberg L, Adams-Campbell LL (2007). CES-D four-factor structure is confirmed, but not invariant, in a large cohort of African American women. *Psychiatry Research* **150**, 173–180.

Wood AM, Taylor PJ, Joseph S (2010). Does the CES-D measure a continuum from depression to happiness? Comparing substantive and artifactual models. *Psychiatry Research* **177**, 120–123.

Zimmerman M, Chelminski I, McGlinchey JB, Young D (2006a). Diagnosing major depressive disorder X: can the utility of the DSM-IV symptom criteria be improved? *Journal of Nervous and Mental Disease* **194**, 893–897.

Zimmerman M, McGlinchey JB, Young D, Chelminski I (2006b). Diagnosing major depressive disorder I: A psychometric evaluation of the DSM-IV symptom criteria. *Journal of Nervous and Mental Disease* **194**, 158–163.