SYMPTOM-LEVEL ANALYSIS OF DEPRESSION AND FUNCTIONAL IMPAIRMENT: EVIDENCE FROM POPULATION-BASED STUDIES.

Regina García Velázquez

Doctoral Programme in Psychology, Learning and Communication,

Department of Psychology and Logopedics,

Faculty of Medicine

University of Helsinki, Finland

Doctoral dissertation, to be presented for public discussion with the permission of the Faculty of Medicine of the University of Helsinki, in Athena Building, room 302, on the 16th of October, 2019 at 12 o'clock.

Helsinki 2019

Supervisors

Professor Markus Jokela

Department of Psychology and Logopedics

Faculty of Medicine, University of Helsinki

Finland

Docent Tom Henrik Rosenström

Helsinki University Central Hospital

Finland

Reviewers

Professor Jouko Miettunen

Center for Life Course Health Research

Faculty of Medicine, University of Oulu.

Finland

Senior Researcher Klaas Wardenaar

Academic Centre of Psychiatry

Faculty of Medical Sciences, University of Groningen. The Netherlands

Opponent

Professor Peter de Jonge

Department of Developmental Psychology

Faculty of Behavioural and Social Sciences, University of Groningen. The Netherlands

ISBN 978-951-51-5538-2 (paperback)

ISBN 978-951-51-5539-9 (PDF)

Unigrafia, Helsinki, 2019

The Faculty of Medicine uses the Urkund system (plagiarism recognition) to examine all doctoral dissertations.

A mis queridos Sr K y Óscar, por enseñarme la importancia de medir lo intangible.

To those who suffer, no matter do they meet criteria and for what diagnosis.

In the end it is sorrow what we work for, not the labels.

Abstract

Depression is a common and disabling mental disorder. Empirical studies have shown that depressive symptoms vary substantially according to antecedents and outcome variables. Nevertheless, research is most often reduced to symptom counts and clinical settings. One of the limitations is the lack of knowledge about the presentation of individual symptoms and their association with clinically relevant outcomes, such as severe functional impairment. The heterogeneity of the symptoms may be an important source of information for a better understanding of depression. This thesis aims to produce empirical insights on how individual depressive symptoms relate to severe functional impairment in community-based samples.

In this doctoral dissertation, we used symptom-level data from two large epidemiologic studies that are representative of the population of the United States: the Collaborative Psychiatric Epidemiology Surveys (CPES), and the National Health and Nutrition Examination Survey (NHANES). Our results suggest that, in terms of statistical fit, sum-scores functioned almost as efficiently as symptom-level modeling in predicting self-rated functional impairment. However, examining symptoms individually offered a more detailed understanding of depression as a syndrome. For instance, symptom-level analyses revealed that age group moderated the associations of three symptoms with severe impairment. According to our results, middle-aged adults were more prone to feel severely impaired by these symptoms, as compared to adults aged 30 and below and in retirement age. The symptoms were depressed mood, self-criticism, and impaired concentration. Additionally, some symptoms did not have a significant association with high functional impairment in a fully adjusted model. This suggests that the association with impairment runs through other symptoms. In general, we found that cognitive-affective symptoms of depression related consistently to severe functional impairment. Among them, self-criticism emerged as particularly informative for its consistent association with high functional impairment across analyses, and for its effect on other symptoms as revealed by cross-sectional direction of dependence analyses. A practical implication of this is that supporting individuals' self-worth may protect against the development of depressive symptomatology and its corresponding impact on functioning.

Our findings motivate considering a wider range of symptoms, both in terms of severity and content, for further understanding the heterogeneity of depression. Similarly, functional impairment as a core component of severity calls for systematic exploration, and for a more refined measurement. A

better characterization of the severity of depression as a continuum is fundamental for theoretical developments in psychopathology, and potentially useful for planning more efficient interventions targeting the most disabling, dominant symptoms.

Tiivistelmä

Masennus on yleinen, toimintakykyä vakavasti rajoittava sairaus. Aikaisemmassa tutkimuksessa masennusta ja sen vaikeusastetta on tarkasteltu pääasiassa masennusoireiden yhteenlasketun lukumäärän mukaan. Rajoituksena tässä lähestymistavassa on se, että siinä sivuutetaan yksittäisten oireiden esiintyminen sekä yksittäisten oireiden ja masennuksen vaikeusasteen yhteys. Aikaisemman tutkimuksen perusteella kuitenkin tiedetään, että masennusoireiden riskitekijät ovat erilaisia ja masennusoireet ennustavat erilaisia lopputulemia. Onkin mahdollista, että tämän oiretasoisen variaation ymmärtäminen on avainasemassa masennuksen ymmärtämisessä.

Tässä väitöstutkimuksessa tarkasteltiin yksittäisten masennusoireiden ja vakavan toimintakyvyn menetyksen yhteyttä väestötasolla. Tutkimuksen aineisto koostui kahdesta laajasta yhdysvaltalaisesta väestötutkimuksesta. Tutkimuksen tulosten mukaan masennusoireiden lukumäärä ennusti vakavaa toimintakyvyn menetystä lähes yhtä hyvin kuin yksittäiset masennusoireet. Yksittäisten masennusoireiden tarkastelu paljasti, että varsinkin kognitiiviset ja affektiiviset masennusoireet olivat johdonmukaisesti yhteydessä heikkoon toimintakykyyn. Lisäksi, keski-ikäisillä masentunut mieliala, itsekriittisyys ja keskittymisvaikeudet olivat yhteydessä toimintakyvyn menetykseen voimakkaammin, verrattuna nuoriin tai iäkkäisiin. Näistä oireista itsekriittisyys nousi tutkimuksen tuloksissa merkittävimpään asemaan, sillä se oli selkeästi yhteydessä paitsi toimintakykyyn, myös muihin masennusoireisiin. Myötätuntoisen asennoitumisen kehittäminen itseä kohtaan saattaa siten suojata masennuksen kehittymiseltä ja masennukseen liittyvältä toimintakyvyn heikkenemiseltä.

Tutkimuksen tuloksia voidaan hyödyntää esimerkiksi masennuksen mallien ja hoidon kehittämisessä. Huomion kohdentamisesta keskeisimpiin masennusoireisiin voi olla hyötyä masennuksen hoitamisessa.

Acknowledgements

As any other outcome, this doctoral thesis is the result of many random variables that were involved throughout the years. All of them have contributed in unique ways to this work. As everything else, it would not be the same without their accumulating effects, additions, and interactions.

I would first like to express my sincere and deep gratitude to my supervisors Professor Markus Jokela and Docent Tom Rosenström, who gave me the opportunity to undertake the most exciting project. I am most grateful to Markus for his unwavering support. Always available and encouraging, he has been close enough for me to feel independent and, at the same time, very well taken care of. His attitude towards research inspires me: nothing is ever complicated within the walls of that office. Thanks, Markus, for being a wonderful antidote to the occasional bumps of this journey. I am greatly indebted to Tom for getting the most out of me, and for believing that I could manage to give it. I appreciate enormously his tireless patience and generosity to teach me and provide feedback, reflecting a genuine wish to participate in anything that could be of benefit for me. His approach to research is equally devoted and enthusiastic, in a way that becomes contagious. Thank you, Tom, for being the most patient and attentive tutor. Apart from their dedication to my dissertation, both my supervisors are to me a model to learn from in their own ways, for which I feel very fortunate.

I am very grateful to Professor Peter de Jonge for kindly agreeing to act as the opponent in the public defence of my thesis. I also thank Professor Jouko Miettunen and Dr. Klaas Wardenaar, the pre-examiners of my thesis. Your suggestions and ideas were of great value in improving the summary of my dissertation, and gave me a fresh angle to look from. I am obliged to the Eemil Aaltonen Foundation and the Academy of Finland for indirect funding. This gave me the possibility to focus exclusively on research.

I am particularly indebted to Docent Markku Verkasalo, who opened to me the doors and windows of Finland and the University of Helsinki. Not only did he welcome me, but keeps caring for me after these years in every possible way. Markku has been my guide, dear friend, and firm support since my first steps in Helsinki. I treasure every conversation between shoots, and warmly acknowledge that it is he who hides behind the wonders I have lived in the University of Helsinki. Markku, I probabily would not be where I am today without your trust –and these words may well be literal when coming from an immigrant :).

I wish to thank my current and former colleagues for giving me an academic home. I warmly thank Karolina Wesolowska, Kaisla Komulainen, Virpi Jouhki, Kia Glushkoff, Elli Oksman, Jaakko Airaksinen, Henrik Dobewall, and Kateryna Savelieva, among others, for sharing this journey. Your company and everlasting humor were a soothing balm for any workday whatever the circumstances. Karolina, our conversations have been a source of inspiration for me these years. I have learned new perspectives and found comforting ways to understand. Virpi and Kaisla, I thank you for giving warmth to this foreigner in Finland: I treasure the good vibes and support over

the years. Kia, thank you for being a fierce colleague, always willing to help and enthusiastic to tackle any challenge.

I am fortunate for my dear friends from the Department of Good Lunches and Peer Support. Saija Kankaanpää, Florencia Sortheix, Sanna Isosävi, Mette Ranta, and Elina Marttinen, who stayed close and tied despite of departing paths. I treasure the support that remains for years. Mi bonita Saija, eres irremplazable en este camino. Has sido una colega brillante, y aún mejor amiga. Gracias por hacerme sentir comprendida y cuidarme, has construido puentes para mí. Gracias, querida Flor, por inspirarme fuerza y ganas. Siempre firme y positiva, tu escucha y consejos son importantes para mí.

Tengo la suerte de contar con varios hogares en España y en Finlandia. Gracias a mis amigos hechos en Madrid, Alina y Manu, por ser mi familia en un período corto e intenso. Por continuar en la distancia compartiendo lo personal y lo profesional, esta tesis no sería lo mismo sin vosotros. Espero que volvamos a unirnos como colegas, como amigos nunca llegamos a separarnos. Mi más sincero agradecimiento a mis profesores de Metodología de la Universidad de Huelva, que me descubrieron una vocación. Esta tesis, sin duda, no existiría de no haber sido por vuestra influencia.

I warmly thank my dearest family and friends, your support is simply essential. Not for this dissertation, but for my own mental health :) You know that I treasure the laughter, discussions, and warmth each of you gives me: it helps me living with a foot in each country. You make it worth and hard, at the same time. Gracias, padres y titos, por creer en mí y sostenerme desde mis más tiernos recuerdos. I am specially indebted to my two beloved pillars, Gloria and Julius. Vosotros me enseñáis el significado de la palabra *incondicional*. No hay forma de expresar lo afortunada que soy por vosotros. Gloria, eres mi fuente de fuerza y aceptación, le sonrío a la vida porque mi hermana está en ella, y la comparte enteramente conmigo. Julius, me haces enormemente feliz. Gracias por cuidarme y derrochar puro amor, literalmente, a diario. Encontré en ti las mejores manos donde poner mi futuro. Sois el combustible de mis días.

Contents

Abbreviations

AIC	Akaike's Information Criterion
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CPES	Collaborative Psychiatric Epidemiology Surveys
DIF	Differential Item Functioning
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders, fourth version
DSM-5	Diagnostic and Statistical Manual of Mental Disorders, fifth version
HiTOP	Hierarchical Taxonomy Of Psychopathology
ICD-10	International Classification of Diseases
IRT	Item Response Theory
LiNGAM	Linear Non-Gaussian Acyclic Model
MD	Major Depression
MDE	Major Depressive Episode
MDD	Major Depressive Disorder
NHANES	National Health and Nutrition Examination Survey
OR	Odds ratio
PHQ-9	Nine-item version of the Patient Health Questionnaire
RMSEA	Root Mean Square Error of Approximation
SE	Standard Error
SRMSR	Standardized Root Mean Squared Residual
WHO	World Mental Health
WMH-CIDI	World Mental Health Survey Initiative's version of the Composite International Diagnostic Interview
WHODAS	World Mental Health's Disability Assessment Schedule

List of original publications

I. García-Velázquez, R., Jokela, M., & Rosenström, T.H. (2017). Symptom severity and disability in psychiatric disorders: The U.S. Collaborative Psychiatric Epidemiology Survey. *Journal of Affective Disorders, 222,* pp. 204-210. doi: 10.1016/j.jad.2017.07.015.

II. García-Velázquez, R., Jokela, M., & Rosenström, T.H. (2019). The varying burden of depressive symptoms across adulthood: Results from six NHANES cohorts. *Journal of Affective Disorders, 246,* pp. 290-299. doi: 10.1016/j.jad.2018.12.059.

III. García-Velázquez, R., Jokela, M., & Rosenström, T.H. (in press). Direction of dependence between specific symptoms of depression: a non-Gaussian approach. *Clinical Psychological Science.*

Introduction

Depression and its attributed burden

Depression is a disabling mental disorder. According to the World Health Organization (WHO), depressive disorders are responsible of an enormous amount of years of "healthy" life lost to disability For example, the last report of WHO attributed to depressive disorders the 5.8% of the overall of years lived with disability in the global population. The estimate was a sum of 44,175,000 years in 2016 across continents (WHO, 2018). Moreover, the WHO estimates that depressive disorders will rank first leading cause of loss of health by 2030 (WHO, 2004). Depression is not only disabling, but also common. More than 320 million of people live with depression, making it one of the most prevalent mental disorders globally (WHO, 2017). The estimated number of people suffering from depression increased by 18.4% between 2005 and 2015, presumably as a result of overall population growth and the proportionate increase in the age groups at which depression is more prevalent (Disease and Injury Incidence and Prevalence Collaborators, 2016).

It is very common for depression to overlap with a variety of other mental and physical illness, making it challenging to estimate its attributable overall burden (Alonso et al., 2011). Its association with mortality is indirect through comorbid diseases, but also independent (Cuijpers et al., 2014). At its worst, depression can lead to suicide. Persons suffering from depression are 20 times more likely to die to suicide than those not depressed (Ferrari et al., 2013). The WHO reports that, for every person who completes suicide, 20 or more may attempt to end their life (WHO, 2018). Similarly, an overwhelming majority of the patients diagnosed with an affective disorder report some kind of suicidal symptoms, from thoughts to repeated suicide attempts – some studies estimate that more than seven in ten (e.g. Aaltonen et al., 2016; Isometsä, 2014).

Thus, depression poses a substantial public health challenge with social, economic, and clinical implications. Psychotherapy appears to have only moderate effect, and seems to be equal beneficial across severity levels and types of intervention (Barth et al., 2013; Cuijpers, 2019). In what comes to pharmacotherapy, the effect size of antidepressant efficacy in trials is similarly modest, with little improvement over placebo (Cipriani et al., 2018; Khan & Brown, 2015; Khan, Faucett, Lichtenberg, Kirsch, & Brown, 2012). The effects of antidepressants are generally seen for the most severe cases of depression, and some experts argue that most of the patients with the disorder do not receive a clinically significant benefit out of medication (Fournier et al., 2010; Khan, Leventhal, Khan, & Brown, 2002;

Kirsch, 2014; Kirsch et al., 2008). Furthermore, the efficacy of clinical trials is not always generalizable to the clinical practice, since it has been shown that the majority of patients do not receive the same level of clinical attention than the participants in efficacy trials, and samples are not representative due exclusion criteria (Lorenzo Lorenzo-Luaces, Zimmerman, & Cuijpers, 2018; Sugarman, 2016). A recent meta-analysis on the efficacy of app-supported smartphone interventions found that these outperformed significantly control conditions in improving depressive symptoms (Linardon, Cuijpers, Carlbring, Messer, & Fuller-Tyszkiewicz, 2019), which suggests that some treatment gaps of depression could be reduced by remote interventions. The efficacy of smartphone interventions, however, does not seem to outperform that of active psychological interventions.

This problematic situation motivates a closer look at whether the diagnostic definition, Major Depression (MD[1]), accurately reflects the underlying depressive syndrome. Depression, as other mental disorders, is diagnosed based on a standard set of criteria that can be found in manuals or classification systems. Two classification systems dominate mental health diagnostics: the International Classification of Diseases (ICD), published by the World Health Organization (WHO, current version ICD-10), and the Diagnostic and Statistical Manual of Mental Disorders (DSM), which is produced by the American Psychiatric Association (APA, current version DSM-5). Depression, historically documented as *melancholia*, is an old malady that has been well described for centuries. The modern operational definition[2] of depression as a disorder has remained mostly unchanged since Feighner et al. proposed their diagnostic criteria in 1972, which included depression within the section of Primary Affective Disorders. Nowadays, it is found in the diagnostic manuals as Major Depression Disorder (MDD). In this dissertation we will generally refer to depression as a syndrome, varying from mild

[1] Major Depression is coded in the diagnostic systems as both Episode (MDE) and Disorder (MDD). For endorsing a MDD, the individual must have previously met criteria for MDE. The main difference between Episode and Disorder is duration: a two-week period is considered as an Episode, while a Disorder requires repeated episodes. The symptoms included and most of the other diagnostic criteria remain the same, and thus the description and further discussion of the symptom criteria in this dissertation involves MDE and MDD. This is the general case in psychiatric literature as well. In this dissertation, we will use the abbreviation MD to refer to clinical depression (both MDE and MDD).

[2] The term "operational definition" was coined by Percy W. Bridgman as a set of operations and assumptions stablished to define concepts which are not directly measurable. For example, length has no natural unit. Instead, it depends on the operational unit chosen, which may be inches or centimeters, and a set of assumptions. It occurs similarly with psychological concepts or constructs, such as intelligence or depression. An operational definition stablishes a common set of criteria to measure an intuitive concept. In their discussion about the consequences of the operationalism in Psychiatry, Parnas and Bovet (2014) argued that "Defining concepts by operations appeared to facilitate implementation of the verifiability criterion of meaning. It also seemed to solve the problem of potential theory-, habit-, or language-derived contaminations of scientific, observational statements about the world" (p. 192).

representations (e.g. subthreshold[3]) up to the clinical disorder. The term Major Depression (abbreviated as MD, and the corresponding specifiers of Episode and Disorder, MDE and MDD) will be used when referring specifically to clinical depression.

Both diagnostic systems in use, ICD and DSM, largely resemble each other in the operationalization of depression. In both, the symptoms considered diagnostic criteria[4] for an episode of MD are depressed mood, loss of interest or pleasure, decreased energy, feelings of guilt or worthlessness, disturbed sleep, changes in appetite or weight, poor concentration or indecision, psychomotor alterations, and death thoughts that may at their worst lead to planning and committing suicide. The episode should last at least two weeks and be discarded other causes to it (such as being under the effect of drugs or another mental disorder). The symptoms must cause substantial impairment in daily functioning, such as daily life activities, social participation, and self-care. The reader can find in Table 1 the operational definition of MDE according to DSM-IV. This definition covers the same symptoms included in the last edition, and is the one presented here for being used by the epidemiological datasets included in the original articles belonging to this dissertation. The changes introduced in DSM-5 relate to specifiers and exclusion criteria of MD, having a potential repercussion on prevalence rates of the disorder and its profiling. However, the aspects of depression examined here are not affected by the changes from the fourth to the fifth edition of DSM.

The diagnostic definition of MD, according to ICD-10, has the following differences with respect to that of DSM: (1) increased fatigability is a core symptom, and the requirement is endorsing at least two of these three core symptoms – to which anhedonia and depressed mood belong. Also two other symptomatic criteria are included: (2) reduced self-esteem and self-confidence, and (3) bleak or pessimistic views of the future. These two criteria are line with the cognitive perspectives on depression. (4) Only loss of appetite and weight are considered, not increased appetite or weight. While the symptom threshold of DSM-IV and DSM-5 was set at 5 five symptoms (including one core symptom), ICD-10

[3] Subthreshold depression refers to those symptom presentations that do not meet the diagnostic criteria but are, to some extent, clinically relevant. The person may endorse a number of associated symptoms but not the two-week duration criterion, or vice versa.

[4] The term *diagnostic criteria* refers actually to all criteria defining a given diagnosis, usually denominated with the Roman alphabet in the diagnostic systems (e.g. A, B, C… see Table 1). In psychiatric literature, the term diagnostic criteria most often refers to the associated symptoms contained, for example, in the criterion A of Major Depression (i.e. referred to with numbers and between parenthesis, see Table 1). We will use here the term "symptom criteria" to refer to them, and "diagnostic criteria" to refer to the whole set of criteria composing a diagnostic definition. We acknowledge focusing on symptom criteria is a simplification of the diagnostic definition.

requires four symptoms of which at least two must be core symptoms, and two or more associated symptoms. In ICD-10 it is defined a specific depressive syndrome called "somatic", which is described based on a certain presentation of the aforementioned symptoms plus marked loss of libido (WHO, 2004, p. 100). Psychotic features are described in both ICD-10 and DSM editions.

In the next sections of the Introduction we will present some challenges to the current diagnostic system of depression. The connecting thread for these challenges is the heterogeneity of symptom presentations contained within the diagnostic label of "depressed". We will first describe some implications, strengths, and pitfalls of treating depression as an aggregate[5]. Then, we will present some empirical evidence of heterogeneity across individual depressive symptoms. Last, we will consider several methodological foundations of this doctoral dissertation. These methodological aspects set the ground for better understanding what psychological measurement entails concerning our work.

Views on depression: on and off?

The statistics of depression are generally based on cases which are diagnosable with MD. A consequence of classifying persons between the conditions of *ill* and *healthy* (e.g. MD present or absent) is that all those labelled as healthy are considered as equally healthy, and likewise all those ill are analogously ill. Such assumption treats within-category cases as homogeneous (i.e. undifferentiated). Mental health professionals have questioned this simplification with the following argument: are all the cases diagnosed with clinical depression qualitatively the same? If so, are also those cases under the clinical threshold of MD equally healthy? To simplify, can we assume that MD turns *on and off* depending on whether a person meets the diagnostic criteria?

Researchers were skeptical of this perspective and argued that, instead of being a distinct and discrete category, depression expands along a continuum (or gradient) of severity, and that the threshold for "clinical caseness" (i.e. MDE or MDD) is an artificial split to this continuum. According to this perspective most of the population, who shows none or quite few depressive symptoms, would

[5] An aggregate can be defined as "a mass or body of units or parts somewhat loosely associated with one another", and also as "the whole sum or amount" (Merriam-Webster dictionary, 2019). These acceptations describe well what the diagnostic label of depression means in the context of this dissertation.

Table 1. Diagnostic criteria for Major Depressive Episode in DSM-IV

A. Five (or more) of the following symptoms have been present during the same 2-week period and represent a change from previous functioning; at least one of the symptoms is either (1) depressed mood or (2) loss of interest or pleasure.

(1) Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad or empty) or observation made by others (e.g., appears tearful). Note: In children and adolescents, can be irritable mood.

(2) Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation made by others).

(3) Significant weight loss when not dieting or weight gain (e.g., a change of more than 5% of body weight in a month), or decrease or increase in appetite nearly every day. Note: In children, consider failure to make expected weight gains.

(4) Insomnia or hypersomnia nearly every day.

(5) Psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down).

(6) Fatigue or loss of energy nearly every day.

(7) Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick).

(8) Diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others).

(9) Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide.

B. The symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.

C. The episode is not attributable to the physiological effects of a substance or to another medical condition.

D. The occurrence of the major depressive episode is not better explained by schizoaffective disorder, schizophrenia, schizophreniform disorder, delusional disorder, or other specified and unspecified schizophrenia spectrum and psychotic disorders.

E. There has never been a manic episode or a hypomanic episode.

belong to the low side of the continuum. The opposite extreme represents those suffering from very severe depression, and expectedly diagnosed with MD. This hypothesis is referred to as *dimensional,* as compared to the view of depression being a *categorical* trait (i.e. a "taxon"). The categorical view entails that depression can be easily differentiable between those who "have it" and those who do not. The taxon of depression, then, would have a boundary that naturally splits the population into groups. Despite some evidence in favor of MD as a taxon (Ruscio, Brown, & Ruscio, 2009; Ruscio, Zimmerman, McGlinchey, Chelminski, & Young, 2007), researchers came to agree that, like most of psychopathology, depression is best modeled as a dimensional continuum (Haslam, Holland, & Kuppens, 2012; Plomin, Haworth, & Davis, 2009; Ruscio, 2002; Ruscio et al., 2007).

Despite of a stable consensus on the dimensional nature of depression, classification based on categories is in extensive use. Both in clinical practice and research, professionals use categories for distinguishing depressed from non-depressed individuals. In clinical and applied contexts there are compelling reasons for it: the diagnostic classifications supply labels to ease communication among professionals, refer patients to treatment programs, assist training, facilitate forensic issues, and a long etcetera (Frances, 2016). Indeed, the epidemiological numbers about prevalence and burden of disease we provided in the previous section are possible due to classifying. In research, a common practice is dividing participants into depressed and non-depressed by using the cut-off score from depression inventories. This allows straightforward comparisons between groups and looking into, for instance, predictors or outcomes associated distinctively with the depressed group.

Issues concerning aggregated approaches in depression research

There are different ways to handle depression scores in research, apart from the aforementioned binary classification (i.e. MD yes/no). In this section, we will introduce in more detail the characteristics of aggregates in mental health, and elaborate on three of the main challenges posed by depression aggregates.

Grading depressive symptoms according to severity has become a standard. Some influential works in the late 90's found a monotonic increment (i.e. dose-response association) in validators of pathology as the number of depressive symptoms increased (Kendler & Gardner, 1998; Kessler, Zhao, Blazer, & Swartz, 1997). This means that, the more symptoms, the more severe were other correlating variables, such as risk of future depressive episodes, average length of longest episode, or impairment. These findings are compatible with the consensus on the dimensional nature of psychiatric traits, and

posited the intuitive rule that *the more symptoms* (most often the nine symptom criteria), *the more severe the disorder*. It is fairly intuitive that the more difficulties a person experiences, the more vulnerable they are and higher their risk of poor functioning. Thus, using sum-scores or symptom counts as aggregates of depression severity seemed a reasonable heuristic, which quickly took over the field.

A pragmatic compromise between binary classification and continuous sum-scores is ordinal grading (Kraemer, 2007). It consists on matching scores according to a severity arrangement, as for example "mild", "moderate", and "severe". This is the approach currently implemented in the DSM-5 for severity specification of MD, whom rating increases from mild to severe, depending on the number of symptom criteria and other indicators (APA, 2013). These categories can have practical consequences, such as recommending a specific treatment for a person according to the severity of their depressive scores (APA, 2010). The extended habit of using aggregated measures of depression as a proxy for disorder severity, was it ordinal grading or sum-scores, is backed by the literature showing that severity validators increase monotonically together with aggregated scores. Moreover, aggregates are extremely efficient variables for being easy to calculate, and for the wealth of information they contain in only one value. However, they also come with some caveats.

The first issue relates to the breadth of depressive **measurement** and its methodological implications. Depressive symptoms are an inevitable source of heterogeneity. Old characterizations of depression include affective, motivational, cognitive, and neurovegetative symptoms as diverse as derealization, constipation, headaches, crying spells or wretchedness (Kendler, 2016). The symptom criteria of today's DSM and ICD are much more concessive, but still include symptoms very different in nature (i.e. affective-motivational, cognitive, and somatic; see Table 1). The most popular inventories used to measure depression in research include sundry checklists of symptoms, each assessing depression based on their own description. Fried (2017) conducted a content analysis of seven commonly used depression scales. He identified 52 different symptoms, of which only 12% appeared in all scales.

Santor and colleagues (2006) examined the depression scales published in the previous 80 years. They concluded that scales vary importantly not only regarding content, but regarding the proportion of items allocated to assessing a given symptom domain (e.g. cognitive, somatic), and concerning item format. For example, some scales measure frequency of symptoms, while others query about intensity. In general, the choice of response format when designing rating scales has received little attention in research, and the reasons for choosing a given response format are rarely documented (Zimmerman,

Morgan, & Stanton, 2018). So far, there is not much evidence on which response format is more valid to measure depressive symptoms: frequency, recurrence, intensity? Regardless of the poor overlap in terms of content and format between scales, the standard is considering them all as measurements of depression, with no distinction. Aggregated depression measures are insensitive to all discrepancy: participants are simply given a value (e.g. "depressed", "moderate", or a score of 12), reducing the variegated symptoms to a single piece of information.

A second implication of the aggregated approach in depression research stems from the **variability** in symptom presentations. Depressive syndromes vary to such an extent, that it has been stated that "to declare that all those satisfying the DSM criteria for the diagnosis of major depression are suffering from the same disorder seems like magical thinking" (Goldberg, 2011). Research suggests that there is no such thing as "diagnostic prototype" in MD, with only a small amount of frequent combinations, that account for less than half of the patients diagnosed (Zimmerman, Ellison, Young, Chelminski, & Dalrymple, 2015). In a study of 3703 patients, the most common symptom profile was endorsed by less than a 2% (Fried & Nesse, 2015a). The variability existing within the label of "depressed" is so large, that some experts have attributed to it the low success of antidepressants (Khan, Mar, & Brown, 2018; Østergaard, Jensen, & Bech, 2011). According to them, it is possible that the etiology of the clinical syndrome does not correspond to the classification of MD.

Subtyping individuals with MD into more homogeneous categories (e.g. endogenous vs exogenous, or dominance of melancholic, psychotic, or manic features, etc.) has been tried and extensively considered (Parker, 2018). Some of these subtypes were present in DSM-IV. However, empirical studies in profiling of depression have not yielded conclusive evidence (Arnow et al., 2015; Harald & Gordon, 2012; van Loo et al., 2012). Recent research has revealed an additional source of heterogeneity: it seems that the structure of depression differs between general and depressed populations, and also between clinical subpopulations (Foster & Mohler-Kuo, 2018; Wanders et al., 2015). This means that the way depressive symptoms manifest in the case of pathologically depressed individuals may be different from in the general population, and even vary across depressed subpopulations according to variables such as the degree of chronicity, or comorbid anxiety. An example of this different presentation could be, for instance, that some symptoms are not just more prevalent in clinical compared to community samples, but have different co-occurrence patterns. Dividing participants into subsamples according to cut-offs on the sum-scores is similarly problematic (Muthén,

1989). To summarize, the inconsiderate use of aggregates seems open to doubt from a methodological standpoint.

A third aspect concerning the use of aggregates is the **boundary** drawn between "depressed" and "not depressed. This is an overarching challenge for mental disorders: there is no natural border separating illness from health (e.g. Kendell & Jablensky, 2003). In depression, there are no biological grounds for this diagnosis, and distinguishing between normal and pathologic levels of sadness is blurry and very complex. In these days, the threshold to classify cases and non-cases of MD in the diagnostic manuals is arbitrary. Arbitrary means there were no empirical grounds for setting them. A notable example is that the Feighner criteria relied strongly on a particular article by Cassidy and colleagues (1957). In this article, the authors used a threshold of six out of ten symptoms to diagnose clinical depression because "it sounded about right" (Kendler, Muñoz, & Murphy, 2010, p. 136). The DSM requirement has traditionally been to endorse five out of nine symptom criteria, with the condition of one of them being depressed mood or anhedonia. However, stablishing a threshold for MD has very important implications, especially because the number of symptoms is an indicator of severity of the disorder according to DSM-5. For example, a person not meeting the five symptom criteria in the last two weeks will have no access to public healthcare treatment nor be covered by their insurance in some countries. This decision is controversial: there is evidence that some cases of subthreshold depression show as high or higher impairment than diagnosable cases (Wakefield & Schmitz, 2017; Wanders et al., 2016; Zimmerman et al., 2008).

The threshold for MD is a delicate issue with broad implications. One of the reasons why experts have not reached a conclusion about when does depression become a disorder is ideological, namely the dilemma between the hazardousness of false positives and false negatives. Put differently, there is a moral choice in pathologizing intense but normal suffering, against leaving undiagnosed vulnerable individuals. This is naturally a source of deep disagreement (Frances, 2013; Horwitz & Wakefield, 2007; Kessler et al., 2003; Maj, 2014). Moreover, the empirical evidence is mixed. For example, a meta-analysis study showed that the increased risk of mortality in subclinical MD is as high as for clinical MD (Cuijpers & Smit, 2002). At the same time, it has been found that the milder cases of MDD do not respond to antidepressant medication better than to placebo (Khan, Brodhead, Kolts, & Brown, 2005). When

Mario Maj[6] posed the still timely question *when does depression become a mental disorder?* (2011), he concluded that "the threshold for a depressive state deserving clinical attention may be lower than that fixed by the DSM-IV, but the threshold for a depressive state requiring pharmacological treatment is likely to be higher. These thresholds may need to be based on the overall severity of depressive symptoms rather than, or in addition to, their number" (p. 86).

The cut-offs for severity labels are similarly problematic: what is the difference between "mild" and "moderate" depression? Research studying the degree of overlap between severity levels of different depression questionnaires have found that these differ substantially. This implies, for instance, that a person may receive the grading of severe depression according to a given scale, and moderate according to another. Moreover, there is no gold standard to judge whether it is the first scale which over-classifies, or the second that under-classifies (Zimmerman, Morgan, et al., 2018).

To sum up, demarcating what is or what is not clinical depression is an extremely challenging enterprise in research, especially for community-based, broad epidemiological studies in which limited resources and wide goals restrict the depth of the information collected. Aggregated measures of depression, such as labels or sum-scores, do not assist in easing the problem because they are rather opaque, and values are highly variable across instruments for the reasons elaborated above. Apart from those sources of variability, the symptoms themselves vary according to different factors. In the next section we will briefly review empirical evidence of the heterogeneity found across depressive symptoms.

Heterogeneity among depressive symptoms

In the previous section, we reflected on research approaches based on aggregated measures of depression. On one hand, they are informative and efficient ways to handle very complex information. On the other hand, there are a number of less desirable consequences. We reviewed some of them: the heterogeneity in measurement tools of depression, heterogeneity in symptom occurrence, and lack of golden standard to split depressive scores into categories.

[6] Mario Maj is an Italian psychiatrist and professor who was President of the World Psychiatric Association between 2008 and 2011, and of the European Psychiatric Association in 2003 and 2004. He is the founder and Editor of the journal World Psychiatry.

A common aspect of aggregated scores that makes them conceptually problematic is the implicit assumption that *all depressive symptoms are interchangeable*. Routinely, the symptoms added to the count weight the same (i.e. just "one symptom more"). Thereby it is common to symptom counts and labels in depression is that two persons with the same number of symptoms may have *totally different* symptom combinations – with all their potential consequences – which goes unnoticed by aggregating them (Fried & Nesse, 2015a; Østergaard et al., 2011). From the perspective of psychometrics this does not seem justified, since depression scales, apart from differing from each other in content and response format, tend to be multidimensional and/or not invariant across samples (e.g. Fried et al., 2016). Subscales from the same questionnaire may also show low intercorrelations (Jang, Livesley, Taylor, Stein, & Moon, 2004). Under these circumstances, compiling multidimensional information into a single aggregate as de facto routine seems questionable. In the next paragraphs we will present empirical evidence on how individual symptoms associate differently with antecedent and outcome variables.

Antecedents

Research investigating the association of individual symptoms with risk factors and antecedents has shown that particular adverse events may trigger different symptoms of depression. Among the adverse events studied are stress, failure, romantic loss, or social conflict. For example, appetite decrease is typically present in response to social loss (Cramer, Borsboom, Aggen, & Kendler, 2012; Keller, Neale, & Kendler, 2007). Moreover, there seems to be congruence to some extent between the adverse situation suffered and the functionality of the symptoms. Keller and Nesse (2006) studied that guilt, rumination, fatigue, and pessimism were prominent following failed efforts; while crying, sadness, and desire for social support were prominent following social losses. Depressive symptoms also differ in remote antecedents, like childhood stress or abuse, and parental coldness: anhedonia, weight loss, and trouble sleeping were linked to these environmental risk factors, but not to heritable factors.

Fried and colleagues (2014) found that other variables associated with individual symptoms are neuroticism (to self-blame and the core symptoms), sex (suicidal thoughts to being male), or work hours (to fatigue particularly). Oquendo and colleagues (2004) discovered that symptom occurrence patterns were instable in a clinical sample of recurrent MDD followed for 24 months, suggesting that individual symptom variability is more subject to contextual factors than to individual styles. This is an interesting piece of information, given that aggregate scores of depressive symptoms tend to be rather stable across

time (Struijs et al., 2020; Verduijn et al., 2017; Verhoeven, Wardenaar, Ruhé, Conradi, & de Jonge, 2018).

Few studies have identified specific neurobiological and metabolic correlates with individual depressive symptoms, revealing that symptoms also associate with some remote and proximal biomarkers apart from contextual factors (Jang et al., 2004; Jokela, Virtanen, Batty, & Kivimäki, 2016; Lamers, Milaneschi, de Jonge, Giltay, & Penninx, 2018; Lux & Kendler, 2010; Myung et al., 2012). In general, it seems that somatic symptoms are more heritable and consistently related to biomarkers than cognitive and affective symptoms.

Outcomes

Studies also show that depressive symptoms are heterogeneous in terms of outcomes. Outcomes are particularly interesting because they are indicators of the impact of a symptomatic presentation. Compared to identifying risk factors, it may be more cost-effective to identify which are the symptoms or combinations related to more severe impact on well-being. For instance, it seems that risk factors can predict the onset of different symptoms, but do not predict prognosis (Eaton et al., 2008; Hardeveld, Spijker, De Graaf, Nolen, & Beekman, 2010). Outcomes, in turn, may be more closely related to the prognosis of the disorder.

Functional impairment is one of the central outcomes of depression. Poor functioning associated with health can be assessed in different life domains, from personal relationships and work performance to autonomy to carry out self-care and daily activities. Impairment extends throughout the whole range of depression in varying intensity, ranging from no impairment to severe disability. It is also a standard in measuring clinical features like severity of the syndrome, quality of the remission of an episode, and global attributed burden. The branch of research devoted to study MD remission has traditionally examined how concrete symptoms relate to functional impairment. Remission studies research whether specific treatment options have a role in recovery from depression, such as the efficacy of concrete drugs or psychotherapy. However, there are several reasons for caution in generalizing their findings.

First and differently from observational studies examining symptom distribution in a given population, treatment efficacy studies examine cases *after* enrolling to a certain antidepressant or psychotherapy treatment, and thus exclude all those cases of depression who do not access any treatment – it has been estimated that a 56% of diagnosable MD do not seek for help (e.g. Kohn, Saxena, Levav,

& Saraceno, 2004). Second, results from efficacy trials may differ from those of the general population because of eligibility criteria. Efficacy trials have so stringent inclusion requirements that their samples differ substantially from the population seen in common clinical practice. A study estimated that only 17-25% of out-patients suffering from MDD would be eligible for antidepressant efficacy trials (van der Lem, van der Wee, van Veen, & Zitman, 2011). A recent study examining placebo-controlled studies across 20 years has estimated that the exclusion rate was over 76%, and varied across medications up to 99% (Zimmerman, Balling, Chelminski, & Dalrymple, 2020). Common reasons for exclusion are symptom severity, presence of other disorders, and suicidality (Zimmerman, Clark, et al., 2016b, 2016a; Zimmerman, Mattia, & Posternak, 2002; Zimmerman, Multach, et al., 2016).

Moreover, the results from clinical trials are not directly comparable among themselves for several reasons. First, there have been observed systematic differences in the sampling: psychotherapy studies include other comorbidities more often than antidepressant studies, and antidepressant trials include more severe and prolonged cases of MDD (Lorenzo-Luaces, Zimmerman, & Cuijpers, 2018). It also appears that depressive symptoms differ in their response to treatment (de Vries et al., 2018; Fournier et al., 2013; Olbert, Rasmussen, Gala, & Tupler, 2016; Taylor, Walters, Vittengl, Krebaum, & Jarrett, 2010). Some studies have also pointed to the side effects of antidepressants: symptoms like sleep disturbance, concentration problems and fatigue, anxiety, sexual dysfunction, and even suicidal ideation are among possible consequences of antidepressants, and thus particular symptoms may be reduced while others increased by a drug (Fried & Nesse, 2015b). Because of these reasons, there are no grounds to extrapolate the findings on impairment from clinical trials to the wider, general population.

Despite its far-reaching consequences, the impairment related to individual depressive symptoms has been described systematically by only few studies. Tweed (1993) interviewed 2687 individuals about concurrent and lingering social impairment as a consequence of individual symptoms. He found six symptoms which explained impairment: depressed mood, durable dysphoria, cognitive difficulties (concentration and slow thinking), suicidal ideation, fatigue, and lack of sexual interest. In another study, Faravelli et al. (1996) reported an association between number of symptoms and their severity. However, the qualitative categorizations the authors made (e.g., melancholic, somatic, etc.) usually attained better discrimination than the number of symptoms in explaining functional impairment: the clusters including cognitive-affective and psychosocial symptoms of depression were stronger indicators of impairment than the somatic symptoms (with exception of motor retardation).

In a more recent study, Fried and Nesse (2014) looked into the independent associations of depressive symptoms with several functioning domains in a clinical sample of MDD. The participants were queried about symptoms and functional impairment within the first week of antidepressant intake. They found that sad mood and concentration problems were the symptoms most strongly related to impairment across all domains. Anhedonia, retardation, fatigue, guilt, and suicidal ideation followed next. Tweed's study was based on a community sample, Faravelli and colleagues used a sample of patients referred to them for day-hospital treatment, and Fried and Nesse used data from the STAR*D study, a randomized clinical trial in which participants received citalopram. There are few studies so far in addressing directly the connection of impairment with individual symptoms of depression (at the symptom level and not through classes or clusters). Their results seem difficult to compare due to diverse sampling, choice of symptoms assessed, and conceptualization of impairment.

Suicidal behavior is another important outcome of depression. Perceived burdensomeness, hopelessness and loneliness have been traditionally linked to enhanced risk of attempting suicide (e.g. Van Orden, Lynam, Hollar, & Joiner, 2006), and among the somatic symptoms insomnia has been found to predict suicide attempt independently of the diagnosis of depression (Lin et al., 2018; Pigeon, Pinquart, & Conner, 2012). Most of the studies, however, study depression as an aggregate and thus do not adjust for other symptoms individually. In a population-based study adjusting for all symptoms simultaneously, Bolton et al. (2010) found that only the symptoms of anhedonia, worthlessness, and guilt, remained statistical significant predictors of suicidal attempt when controlling for all symptoms individually, symptom count, and comorbid psychiatric disorders together with sociodemographic predictors. In another epidemiological study they found that, among those with diagnosable MD, feelings of worthlessness was the symptom most strongly associated with history of suicide attempts (Bolton, Belik, Enns, Cox, & Sareen, 2008).

There is some emerging evidence of how individual depressive symptoms react to interventions, showing heterogeneity (Boschloo et al., 2019; van Eeden, van Hemert, Carlier, Penninx, & Giltay, 2019). This is undoubtedly in strong connection to the topic of depression outcomes. However, the association between individual symptoms with outcomes of depression, like persistence or recurrence, in an epidemiological context, is different for not being dependent on receiving a specific treatment (i.e. a given type of psychotherapy or a specific medication, to which understandably some symptoms could be more reactive than others). Thereby, both topics answer to different questions. Symptoms reacting to

intervention positively do likely alleviate depression severity and could contribute to better prognosis, but the question of which symptoms do predict relapse or chronicity themselves is different, and remains yet mostly unexplored.

A large proportion of individuals are at risk of developing depression at some point of their lives, but only half of those who reach clinical levels will fall into chronic and resistant forms (Lorenzo-Luaces, 2015). Chronicity, recurrence, suicidality, and hospitalization are, indeed, well-recognized outcomes of depression. These are most often seen when the depressive syndrome is severe enough, while functional impairment may occur much earlier. Studies have found that functional impairment is itself a risk factor for recurrence of depressive episodes (Gilmer et al., 2005; Solomon et al., 2004). Moreover, in their recent work examining functioning pre- and post-depressive episode, Bos et al. (2018) found evidence suggesting that it is pre-existing functioning which predicts future depressive episodes, instead of residual symptoms causing impairment by lingering. This conclusion was based on literature review and their own empirical findings, which examined mental and physical functioning. Thus, identifying which symptoms are most often related to severe impairment can be a first step for detecting vulnerable cases whose prognosis could be at risk of worsening. However, this is only a possibility given the incomplete scientific understanding of the associations among clinical outcomes of depression.

Methodological foundations

Main psychometric approaches to modeling depression

Psychiatry, a branch of medicine, has traditionally studied psychopathology as any other medical disease. The medical model of disease stablishes that symptoms root from physical dysfunctions (e.g. impaired language and memory, labile affect, etc., as result of neurological and degenerative processes). The usefulness of a diagnosis relies on treatment decisions: given a set of symptoms stemming from an identified dysfunction, a suitable treatment can be chosen. Such framework implies, in psychopathology, that a mental disorder causes the symptoms we observe. So far, most mental disorders lack a clear neurobiological mechanism and react variably to different treatments (Kendler, 2012; Singh & Rose, 2009), and MD is not an exception (Kennis et al., 2018). The difficulties with the medical model are not surprising in the case of psychiatric nosology[7]. This is because the current psychiatric classifications are the result of consensus among experts and were not built upon etiological evidence, nor have they undergone exhaustive empirical validation (Kendler & Parnas, 2012, 2014).

There is a wide research program aiming for better models of psychopathology. Among the main issues is the structure of psychopathology. Psychopathology models can target content-related questions (e.g. what are the factors underlying comorbidity across disorders? Is legal problems a valid diagnostic criterion for substance-related disorders?), and structural issues (e.g. do psychiatric disorders explain symptoms, or result from their interactions?). The classic methodology to address these questions comes originally from the field of Psychology, and it is based on latent trait modeling. Latent traits are unobserved, unmeasured variables which are most often modelled mathematically through confirmatory factor analysis, latent classes or profiles, or Item Response Theory models (Borsboom et al., 2016). Across these different methods, latent traits are used to model a common cause[8] which explains the correlation among observed variables (e.g. feelings of worthlessness ← depression → insomnia). A common cause model has the important implication that two items (e.g. symptoms) of the same trait (e.g. disorder) are independent from each other conditional on the trait (e.g. Cov[feelings of worthlessness,

[7] Nosology is a branch of medical science that deals with the classification of diseases. Diseases may be classified according to their pathogenesis (biological mechanisms causing it) or to the symptoms, in which case the term "syndrome" is appropriate. In practice the words disorder, disease, and syndrome are used interchangeably.

[8] Latent variables are not exclusive of common cause models. A common cause model uses most often a latent variable or trait in order to explain the covariation among the observed variables. A latent variable is any unobserved variable which is estimated within a model based on a group of indicators. For example, error terms in structural equation modeling are latent variables.

insomnia | depression] = 0; e.g. Thoemmes, Rosseel, & Textor, 2018). These models embody the *reflective* structure of psychopathology (i.e. the items reflect a latent trait). Such structure suits well the dimensional understanding of psychopathology: the higher the position on the latent trait continuum, the more likely is the occurrence of symptoms. The reflective framework became rapidly popular and has dominated psychopathology, via latent trait modeling, for decades (Duncan-Jones, Grayson, & Moran, 1986).

The interpretation of what a latent trait is in psychopathology remains an open question, and indeed a vastly discussed one. The latent trait framework estimates latent variables which explain the covariance in the data. Such latent framework is compatible with the medical perspective that mental disorders are biological-based phenotypes, by representing them as latent traits (Franić et al., 2013). The same framework is also suited to the viewpoint that latent traits reflect vulnerability or susceptibility to certain symptomatology (e.g. Caspi et al., 2014; Kotov, Krueger, & Watson, 2018). More recently, some authors have argued that a latent trait emerges when a set of random variables correlate positively with each other (van Bork, Epskamp, Rhemtulla, Borsboom, & van der Maas, 2017; Van Der Maas et al., 2006), as often happens with items in a questionnaire measuring related symptoms. This would imply that many latent variables found in empirical research were actually a mathematical artifact. This argument is at the core of a novel perspective on psychopathology, which has become increasingly popular in the last years under the name of *network approach* to psychopathology (Borsboom, 2017).

The network approach proposes a *formative* structure for mental disorders as opposed to the reflective structure characteristic of latent trait models. According to this theory, mental syndromes are constituted and maintained by networks of symptoms that affect each other. In MD, for instance, it could be that sleep problems cause fatigue and irritability the following day, which in turn dampens mood and affects performance, leading to frustration and self-reproach. The theory posits that individual differences in network structure and vulnerability affect symptom-to-symptom dynamics, which in turn lead to different symptom manifestations (e.g. some individuals suffer characteristically from somatic symptoms, while others have more cognitive symptoms). Therefore, the positive manifold results from mutual, direct connections between symptoms (van Bork et al., 2017). This theory has become popular for its emphasis in the concrete mechanisms operating in psychopathology, which smoothly connect to intervention planning (Cramer et al., 2016). The network approach has generated plenty of discussion and a number of insightful findings. However, being this approach relatively recent and mostly limited

to cross-sectional settings, the knowledge generated is still preliminary. For some stimulating thoughts on psychometric modeling of psychopathology we refer the reader to Fried (2017), Reise & Rodriguez (2016), or Kotov, Krueger & Watson (2018).

Validity in psychiatric disorders

A psychological measure is considered valid when it measures what it claims to measure. Likewise, conclusions made out of a study are valid when being coherent with the empirical evidence, and its resultant interpretations justified. Thus, validity involves not only measurement instruments but all the consequences that may follow measurement. In this section we will present some implications of validity for mental disorders.

Validity studies are abundant in psychometrics. One can examine the *content* validity of the operationalization of a construct[9], or how a rating scale measures such construct (e.g. what does the construct of intelligence involve? Is it covered in its different domains by our test?). Also the empirical information obtained with a test or another measure might show *convergent* and *discriminant* validity when it relates to other pieces of information in a coherent way (e.g. high depression scores correlating positively with number of days in sick leave, and negatively with life satisfaction). Then, evidence of *concurrent* and *predictive* validity implies that our information is actually consistent with a current or future criterion (e.g. an event, such as hospitalization related to severe scores of bipolar disorder). Depression, as many other psychological constructs, cannot be directly observed or measured in established units, such as grams or centimeters. This is why the notions of validity and reliability[10] are crucial: in order to better describe it, we need to gather systematic evidence. Importantly, validity is not a property inherent to a given test or concept, but we can *gather evidence* of a test score or a concept being valid in a given context.

Applying validity to psychiatric diagnoses is not straightforward. Because of the close relationship between mental health and the rest of medicine, some prominent researchers in psychiatry

[9] A construct is, according to the classic definition of Cronbach & Meehl (1955, pg. 283), "some postulated attribute of people, assumed to be reflected in test performance. In test validation the attribute about which we make statements in interpreting a test is a construct".

[10] Reliability is another central concept in psychometrics. It is related to the consistency of measurement under different circumstances, for example across time (test-retest reliability), instruments (internal consistency and parallel forms), and between judges (inter-rater reliability). Reliability tightly relates to the concept of measurement error, which causes unsystematic variation in measurement and prediction of scores. There are numerous coefficients to calculate reliability, and its quantification is somewhat more straightforward than that of some forms of validity.

seem to reduce validity to biological etiopathogenesis[11] (Jablensky, 2016; Kendell & Jablensky, 2003). Then depression (or any mental disorder) is valid only if it is a "disease entity" that is separated from the other diseases by "natural boundaries". This view of validity may be easily applied to physical disorders in other fields of medicine, but it seems a rather problematic conceptualization in psychiatry. The reason is that such perspective on validity emphasizes the value of neural circuitry, biomarkers, and genetic mechanisms. These are unarguably compelling validity criteria, as long as their role on the onset, maintenance, or prognosis of psychopathology is empirically demonstrated. Yet, demonstrating etiopathogenic mechanisms does not justify disregarding other factors involved in psychopathology.

Heritability and a variety biomarkers are only a part of the variance in psychiatric disorders, of more or less extent and still mostly unknown (Kennis et al., 2018). Indeed, there is a heated controversy around the genetic contribution to depression in current literature (Border et al., 2019; Smoller, 2019). Many factors linked to the development and course of psychiatric disorders are either not biological in nature, or very difficult to track directly from biomarkers. More or less complex psychological processes, such as attention, emotion, learning, or cognition are involved in psychopathology. In addition, cultural, socio-historical, and economic factors play an important role in the validity of mental illness. These factors may be as important as biological etiopathogenesis in explaining psychopathology, and thus validity studies shall reflect this etiological complexity.

Defining a diagnostic disorder or measuring it (e.g. with interviews or rating scales) is not particularly different from the rest of psychological measurement. Lending validity evidence of a mental disorder (i.e. MD) or attribute (i.e. depression) demands the same as from any other construct, such as empathy or consciousness. Why depression or narcissistic personality disorder, being psychological phenomena, should be validated under different standards? The famous five criteria for psychiatric validity[12] compiled by Robins and Guze (1970) were guidelines originally suggested for schizophrenia, and seem to be incomplete for most of the contemporary diagnostic disorders because their focus was mostly biological. Some authors have argued in favor of a wider perspective on validity of psychopathology (Krueger & Eaton, 2012; Lefere, De Rouck, & De Vreese, 2017).

[11] In medicine, etiopathogenesis refers to the determination or study of the cause and development of a pathology, in this case a mental syndrome (Merrian-Webster Medical Dictionary, 2018).
[12] Their so-called method for achieving diagnostic validity included five phases: clinical description, laboratory study, exclusion of other disorders, follow-up study, and family study.

A main argument for revising the validity of MD is that MD informs about depression, but it *is not* depression (and this duality applies generally to psychiatric classification). MD is an index of depression for classification purposes, and as such it should *measure* what it is intended to (Kendler, 2016). The MD diagnostic criteria do not constitute depression, constitute an index of it. As such, MD is an instrument that conveys plenty of information. This information should be useful for diagnostic and treating depression effectively. However, MD being an index for depression has influence on empirical reality. As far as DSM criteria serve for representing reality, they also convey what we understand as depression (van Loo & Romeijn, 2018). For instance, MD can be seen from many theoretical angles (e.g. cognitive, affective-motivational dysregulation, evolutionary…), which have an influence upon operationalization through the choice of diagnostic criteria. These theories shall be empirically tested and validated. Collecting evidence of how diagnostic criteria function serves the purpose of studying the validity of MD in its current definition, and may contribute to empirically-based theoretical developments to represent psychopathology and also non-pathological functioning (Costello, 1993).

The role of functional impairment in the severity of depression

So far there are neither biological nor structural markers to assess the severity of common mental disorder (Border et al., 2019; Zimmerman, Morgan, et al., 2018). Given the lack of "objective" indicators, the diagnostic manuals have included a requirement, known as the *clinical significance criterion*, to guarantee that a given combination of symptoms is harmful enough to justify diagnosis and consequent treatment. Diagnostic systems consider this criterion at the same level as the symptom, duration, and exclusion criteria. The wording of the criterion has to some extent variation across the disorders of DSM-5, but most often, it is described as "The symptoms cause clinically significant *distress* or *impairment* in social, occupational, or other important areas of functioning". The ultimate purpose of such criterion is to assist professionals in diagnosing more validly, by minimizing false positives (i.e. the mistake of diagnosing when no necessary) and false negatives (i.e. the mistake of not diagnosing when it would be the right choice). It is so that the clinical significance criterion makes it possible to diagnose subthreshold cases (i.e. fewer criteria than required) due to sufficient severity, and not to render diagnosis when criteria are met in a way it seems not pathological. Moreover, in the case of MD the clinical significance criterion plays a role in defining the severity of the syndrome, together with number of symptoms.

It was acknowledged explicitly in DSM-IV how challenging it is to evaluate this criterion: "Assessing whether this criterion is met, especially in terms of role function, is an inherently difficult

clinical judgment. Reliance on information from family members and other third parties (in addition to the individual) regarding the individual's performance is often necessary" (APA, 2000, p. 7). Some experts have criticized such definition for being too ambiguous and somewhat redundant (e.g Maj, 2014). This critique is supported by research evidence: full-blown cases are generally accompanied with disability or functional impairment in clinical samples (Zimmerman, Chelminski, & Young, 2004), and in epidemiological studies (Beals et al., 2004; Slade & Andrews, 2002; Wakefield, Schmitz, & Baer, 2010). Nevertheless, research has found that clinical significance-like specifiers reduced the community rates of mental disorders in DSM-IV and identified persons more likely to be using mental health services or having more severe symptoms (Narrow, Rae, Robins, & Regier, 2002). Some seminal studies on the impairment caused by depression showed a significant gradient of severity of impairment with number of symptoms, number of lifetime depressive episodes and comorbidities (Kendler & Gardner, 1998; Mojtabai, 2001).

On its behalf, distress as an indicator of clinical significance appears to discriminate poorly. The reason is redundancy to mental illness. Distress very often comes with the awareness of suffering from psychopathological symptoms, if is not simply intrinsic to the disorder itself (e.g. anxiety disorders are characterized by intense distress). In the case of MD, the symptoms are themselves generally distressing enough to cause "marked distress" not meeting the symptom or duration requirements (Spitzer & Wakefield, 1999; Zimmerman et al., 2004).

The challenges experienced with the clinical significance criterion have led to some suggestions. One of them, proposed by Mojtabai (2001), is to include the distress and impairment ratings as independent components of the diagnostic system, being measured with the Global Assessment of Functioning (GAF) scale. This scale was replaced in DSM-5 with the WHO Disability Assessment Schedule (WHODAS 2.0; Üstün, 2010). On the one hand, the benefit of standard scales is to systematically assess impairment, and to serve both clinical and research purposes. On the other hand, their limitation is being general-purposed and thus not symptom- or disorder-specific. This feature implies the risk, particularly in research, of being confounded by other co-occurring mental or physical conditions than the disorder being studied. This risk is worth to ponder when investigating disorders that tend to be comorbid with others, such as MD. Measures of global impairment require adjustment for co-occurring conditions if researching one disorder particularly.

Another suggestion, made by Wakefield and collaborators (Cooper, 2013; First & Wakefield, 2013; Wakefield, 2009), is to raise the symptom or duration criteria in order to increase their pathosuggestiveness. For example, given depressed mood is a relatively common symptom in the general population, diagnostic manuals could add an intensifying corrective to minimize false positives. Instead of a "blanket approach" of general impairment, which is unspecific and difficult to define for all disorders, the idea is to conduct disorder-specific modifications. These sort of suggestions are, nevertheless, difficult to carry out because of general resistance against changes to current systems in use.

Experts agree in severity not being reducible to symptom counts. Nevertheless, the amount of symptoms is predominantly used in research as single indicator of severity. Zimmerman et al. argued that "almost all research on severity is based on scores of depression symptom scales, though most scales have been developed without consideration as to how to best conceptualize and assess the severity of depression" (2018, pp. 84). It is logical that higher sum-scores are associated with higher severity. However, studies have shown that symptom scores are only a sphere of severity (Lux et al., 2010; Zimmerman, 2012; Zimmerman, Morgan, et al., 2018), and that the correlations of symptom counts and proxies of severity tend to be, at their best, modest (Kitamura, Nakagawa, & Machizawa, 1993; Lux et al., 2010). This simplification brings about two consequences. First, other indicators of severity are less studied than symptom counts; and second, the use of counts conceals finely tuned information of symptom characteristics. This information could be of clinical value, for instance in common severity correlates of specific symptoms. For example, it was mentioned earlier that feelings of worthlessness have a consistent and distinct association with suicide attempts (e.g. Bolton et al., 2008; Wakefield & Schmitz, 2015). This entails useful information, as compared to results claiming that a given sum-score in a depression scale, or a number of DSM symptoms, predict a higher probability of some outcome.

To summarize, it appears fundamental to gather information on how individual symptoms relate to functional impairment, which is a component of severity according to DSM-5 and undoubtedly an essential aspect of depression course and remission (Zimmerman et al., 2012, 2008). Evidence shows that the more symptoms the higher the probability of suffering from a clinically impairing syndrome. However, it is uncertain whether this impairment is due to the unspecific accumulation of symptoms, or explained to different extent by some symptoms more than by others. Furthermore, most of what we currently know about functional impairment of depression is based only on symptom presentation in

patients. Symptom distributions may differ substantially from their natural distributions in clinical samples, basically because clinical patients are a subset of the community. Patient samples show particular characteristics, not generalizable to the general population (Foster & Mohler-Kuo, 2018; Muthén, 1989). It would be critically important to better understand how individual symptoms relate to functional impairment in natural, general population samples. Knowledge on individual associations between symptoms and functional impairment would serve a variety of purposes, such as characterizing more impairing symptom profiles, delineating intervention programs targeting the most impairing symptoms, or informing which symptom (or duration) criteria show low ability to discriminate more pathological presentations.

Although we focus here in functional impairment as an indicator of severity, we acknowledge that conceptualizing the severity of depression, and in general of mental disorders, remains an open question which is much more complex than choosing one indicator or another (Zimmerman, Morgan, et al., 2018).

Aims of the study

The general purpose of this doctoral dissertation was illustrating novel ways to investigate the symptom criteria composing MD, and their association with functional impairment in the general population. Concretely, the study aimed to:

I. **Evaluate latent variable models of depressive symptom criteria**. Inspecting whether a psychometric model accommodates the data sets the baseline for further analyses and has important implications for the interpretation of results.

II. Assess the benefit of considering **individual diagnostic symptoms** to inform of the severity of MD, as **compared to sum-scores** (the 'standard' approach). Aggregates may be efficient, while individual symptoms may be more explanatory. These different levels of information are more or less justified depending on how convenient and useful they are when set side by side.

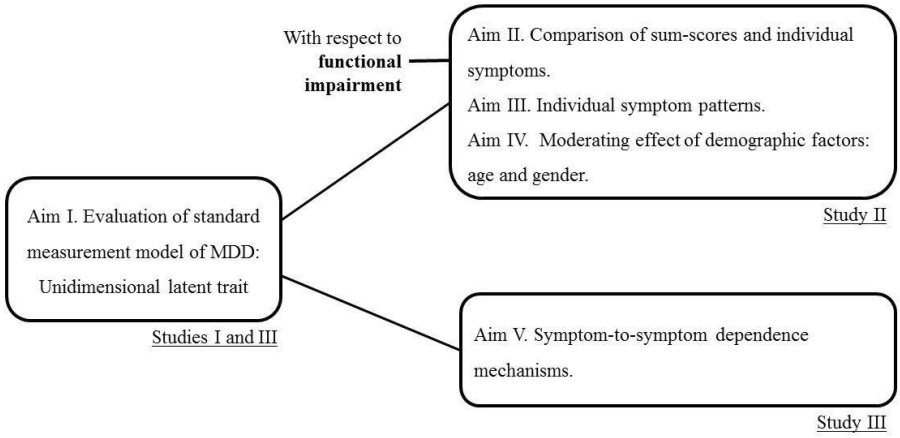
III. Study **symptom-specific patterns** of association with self-reported functional impairment in the general population. The evidence on naturalistic associations, throughout the whole depression spectrum, is mostly lacking.

IV. Investigate whether **age and gender moderate the association** between specific MD symptoms and self-reported impairment. If individual symptom-impairment patterns exist, are they generalizable to the community regardless of demographic variables? Demographic variables may help disentangling the heterogeneity observed in depression.

V. Bring **novel insights into the dependence mechanisms among MD symptoms**. It is useful to collect evidence on symptom-to-symptom dynamics for further characterizing the associations of depressive symptoms with functional impairment.

The aims of the study are outlined in Figure 1.

Figure 1. Aims of the study.



Methods

Collaborative Psychiatric Epidemiology Surveys (CPES)

Participants

The CPES (Alegria, Jackson, Kessler, & Takeuchi, 2015) is a large internationally available dataset of 20 130 residents of the United States, which comprises three large, nationally representative samples of the general population. The goal of the CPES is to collect data about the prevalence of mental disorders, impairments associated with these disorders, and their treatment patterns from samples of majority and minority adult populations in the United States. The three cross-sectional samples are composed of adults aged 18 and above, excluding institutionalized persons and those living on military bases. The National Comorbidity Survey Replication and the National Survey of American Life exclude non-English Speakers, but the National Latino and Asian-American Study of Mental Health focuses on participants who do not speak English. The response rate was 71% or higher for the three studies. For details on the analytic sample sizes and measures used in this doctoral dissertation, see Table 2.

The sampling methodology of the three CPES samples was highly similar, despite of some unique features and topical questionnaire modules. All the data collection was based on a four-stage area probability sampling. The primary stage sampling corresponded to counties as units or contiguous counties with small populations. The second stage of sampling contained geographically contiguous census blocks that were stratified by race/ethnicity composition of residents' households. The race/ethnicity stratification of area segments played a particularly important role in the NSAL and NLAAS sample designs, where it was used both to improve the sampling precision of the design and as a basis for more cost-effective oversampling in area segments with higher densities of households for targeted race and ethnicity subpopulations. The third and four stages of sampling involved housing units. The third-stage sampling rate was computed for each selected area segment in the CPES sample design. This rate was then used to select a systematic random sample of actual housing units from the area segment listing. Among the eligible household members, a random selection of the respondent was performed using a special adaptation of the objective selection method developed by Kish (1949). In general, the sampling strategy was designed to optimize the cost and error properties of the study-specific samples. The sampling weights for data analysis compensate for differences in the inclusion probabilities of population members who reside inside and outside high-density area domains, and were designed to

Table 2. Participants and measures included in the epidemiological studies used in this doctoral dissertation.

Study	**Subsample**	**N**	**Measures**		
			Depression	**Functional impairment**	**Other relevant variables**
NHANES (n=34 963)	2005-2006	5 334	PHQ-9 (referred to last 2 weeks)	How difficult the symptoms made it to carry out daily activities?	- Gender - Age - Ethnicity - Marital status - Medical conditions - Poverty
	2007-2008	5 995			
	2009-2010	6 360			
	2011-2012	5 615			
	2013-2014	5 924			
	2015-2016	5 735			
CPES (n = 20 130)	National Comorbidity Survey Replication	9 282	Depression, module, WHO CIDI (referred to most severe affective episode; n= 4 152)	- WHODAS - Days disabled in the last 30 days. - Degree of interference of symptoms - How often unable to carry out daily activities	- Gender - Age - Ethnicity - Number of symptoms - Post-Traumatic Stress Disorder-module - Generalized Anxiety Disorder-module - Mania module
	National Survey of American Life	6 199			
	National Latino and Asian-American Study of Mental Health	4 649			

be nationally representative. For more detailed information on the sampling characteristics of CPES, see Heeringa et al. (2004).

The CPES samples collect information on psychopathology based on the World Mental Health Survey Initiative's version of the Composite International Diagnostic Interview (WMH-CIDI, Kessler & Ustün, 2004), which is a modified version of the Composite International Diagnostic Interview (CIDI). The Initiative's version of the CIDI is a comprehensive, fully-structured interview designed to be used by trained interviewers for the assessment of mental disorders according to the definitions and criteria of ICD-10 and DSM-IV. It is suitable for clinical and research purposes, including measuring the prevalence and severity of mental disorders and assessing the burden of them. The CPES interviews were all computer-assisted, administered by trained personnel. Detailed information about the development and implementation of the CPES project can be found from Pennel et al. (2004).

Measures

Depressive symptoms

Depression has a specific section in the WHO-CIDI. Diagnostic criteria and symptoms from both DSM-IV and ICD-10 are included in the interview. The description of the symptoms, together with the terms used here to refer to them are described in Table 3. Due to its breadth, the interview is based on a skip-question system. For starting the depression module, the respondent had to answer positively to at least one of the screening items, reading "have you had episodes that lasted several days or longer when you felt…(symptom)" and one of the following: "sad, depressed or empty", "discouraged", "loss of interest", or "life had no meaning". If the respondent endorses one of these screening questions, the depression module is administered. It symptoms are asked concerning "think of the period when your sadness/discouragement/lack of interest and other problems were most severe and frequent. Did you…(symptom)". The answers are coded as "yes", "no", "does not know", and "refuses". While administering the depression module, several checkpoint filters determine whether to continue or not with the module. All the items corresponding to the DSM-IV symptom criteria were assessed rather precisely (e.g. compound items divided into several parts). However, the module was only continued when the participant had endorsed the stepwise screening process, and therefore information on sleep problem, for example, is missing if the participant did not endorse any of the screening symptoms. Likewise, the items concerning suicidal behaviour are only administered when the participant has endorsed items concerning death thoughts. This process leads to data missing not at random (MNAR),

and therefore imputation is not a feasible option. The size of the analytic sample for the depression module was n=4 152. The weighted frequencies of the DSM-IV symptom criteria are presented in Table 4. Note that the items concern the *most severe period* of depressive syndrome, and not lifetime symptom endorsement patterns.

Functional impairment

Functional impairment is assessed in the CPES with two general measures and other two which are specific of the affective episode. The World Health Organization Disability Assessment Schedule (WHODAS) is a 36-item questionnaire which measures health and disability as defined in the International Classification of Function, Disease and Health (WHO, 2018). Scores are a product of frequency and severity of problems (none, mild, moderate, severe) that respondents reported experiencing in the past 30 days, and are normalized to values ranging from 0 to 100, where higher numbers indicate worse functioning. The scale is divided into six domains of functioning: cognition, mobility, self-care, social interaction, role functioning, participation (Üstün, 2010). The scale is considered a generic assessment instrument for health and disability. This scale was referred to current functioning. In psychiatric literature both terms, impairment and disability, are used with no distinction (e.g. Sheehan, Harnett-Sheehan, & Raj, 1996). We may use the terms indistinctively as well.

A second indicator of general impairment is the item "Beginning yesterday and going back 30 days, how many days out of the past 30 were you totally unable to work or carry out your normal activities because of problems with either your physical health, your mental health, or your use of alcohol or drugs?". Participants quantified days from 0 to 30. Both of these generic indicators of functional impairment were collected of all participants of CPES.

The depression module of the interview contained two items querying the level of functional impairment and life interference caused by the depressive episode. The depressive episode reported is the most severe ever suffered by the participant, which is recalled in terms of symptom manifestation and impairing consequences. The first item concerns life interference of the symptoms, with the wording “You mentioned having [(two of/a number of)] the problems I just asked you about. How much did your sadness/discouragement/lack of interest and these other problems interfere with either your work, your social life, or your personal relationships during that episode?”, and response categories “not at all”, “a little”, “some”, “a lot”, and “extremely”.

Table 3. Depressive symptoms examined in the study.

Shortened term	WHO-CIDI descriptors	PHQ-9 descriptors
Depressed mood	Sad, empty, depressed, hopeless about future, nothing could cheer up	Down, depressed, hopeless
Anhedonia	Lost interest, nothing fun when good things happened	Little interest or pleasure in doing things
Appetite changes	Weight loss/gain not trying, much smaller/larger appetite	Poor appetite, overeating
Sleep problems	Trouble sleeping, slept much more than usual	Trouble falling asleep, staying asleep, sleeping too much
Psychomotor disturbances	Talk/move more slowly than usual, others noticed, so restless that could not stay still	So slowly that others could notice. Or the opposite - fidgety, restless
Fatigue	Low energy and tired for no reason	Tired, little energy
Self-criticism	Worthless feeling, felt guilty	Bad about yourself, failure, let down self or family
Difficulties to concentrate	Trouble concentrating, unusual indecisiveness	Trouble concentrating on things, such as reading newspaper or watching TV.
Thoughts of death or self-harm	Thought of death, better dead, suicide thoughts, plan, attempted	Better off dead, hurting yourself in some way

If the respondent stated that, at least, the symptoms had interfered "a little" with their life, a second items would follow: "How often during that episode were you unable to carry out your daily activities because of your sadness/or/discouragement/or/lack of interest?". The item has the following

response options: “never”, “rarely”, “sometimes”, and “often”. We dichotomized these items, so that they would be indicators of severe functional impairment by reflecting whether items had interfered “a lot” and “extremely” with life, and whether the respondent could not carry out her daily activities “often”. These items were only posed to participants who had endorsed the screening questions of the depression module, and were referred to the time of the worst affective episode. The 55.6% of the participants asked responded they had suffered a lot or extreme interference, while 28.9% reported being often unable to get by in daily life.

Other psychiatric disorders

Three other diagnostic categories were examined additional to MDE. These were General Anxiety Disorder (GAD, n=3 610), Manic Episode (ME, n=4 214), and Post-Traumatic Stress Disorder (PTSD, n=3 128). All of them followed the same skip-question system and therefore the modules were only administered when endorsing at least one of the screening symptoms. An index was computed for each of the disorders indicating how many DSM-IV criteria the participant had endorsed. These sum-scores ranged from 0 to 12 for MDE 0 to 7 for GAD, 0 to 16 for ME, and from 0 to 16 in the case of PTSD.

Other relevant variables

We adjusted our analyses for variables such as gender, years of age, ethnicity, marital status, and number of chronic medical conditions.

National Health and Nutrition Examination Survey (NHANES)

Participants

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. This is a major program belonging to the Centers for Disease Control and Prevention (CDC, 2017), which provides official statistics of the United States. The purpose of the data is to be used in epidemiological studies and health research, which help develop sound public health policy, and design health programs and services. In 1999, the survey became a continuous program that has a changing focus on a variety of health and nutrition measurements to meet emerging needs. The survey examines a nationally representative sample of about 5 000 persons each year. The NHANES project includes demographic, socioeconomic, dietary, and health-related matters including medical, dental, and physiological measurements.

A four-stage sample design was implemented in the NHANES samples included in this doctoral dissertation. The primary sampling units were counties, selected with probabilities proportional to size. The second sampling stage comprised area segments based on census blocks. The third stage were dwelling units. The fourth state of sampling involved households. The sampling procedure in the third and fourth stages sought at maximizing the sampling objectives of the studies (representativeness of race, origin, sex, income, age). The sampling weights were designed to meet the following objectives: (1) compensate for differential probabilities of selection due to race and Hispanic origin, income, sex, and age; (2) reduce biases arising from differences between respondents and non-respondents; (3) match an independent U.S. Census Bureau estimate of the population; (4) compensate for inadequacies in the sampling frame; and (5) reduce variances in the estimation procedure. Further details of the sampling and weighting methodology can be found from the website of NHANES (CDC, n.d.)

We selected the samples which include the depression screener questionnaire for this dissertation, the waves of 2005-2006, 2007-2008, 2009-2010, 2011-2012, 2013-2014, and 2015-2016. For details on the sample sizes and measures see Table 2. The analytic sample consisted of n=21 056 observations.

Measures

Depressive symptoms

The depression screener is the nine-item version of the Patient Health Questionnaire (PHQ-9; Kroenke & Spitzer, 2002). This questionnaire is administered to all participants aged 18 or older in the NHANES study, as part of computer-assisted personal interviews. The nine items of the questionnaire query how often the participant had been suffering from concrete depressive symptoms during the last two weeks, each self-rated on a four-point response scale (0=Not at all, 1=Several days, 2=More than half the days, 3=Nearly every day). The nine symptoms correspond to the MD symptom criteria in DSM-IV. The scoring system of PHQ-9 can be interpreted according to the cut-points of 5, 10, 15 and 20, indicative of mild, moderate, moderately severe and severe levels of depressive symptoms (Kroenke & Spitzer, 2002; Kroenke et al., 2010). The weighted frequencies of the PHQ-9 items (which correspond to the DSM-IV symptom criteria for MD) are presented in Table 2. It is worth to mention that PHQ-9 does not include any filtering conditional on symptom endorsement.

Table 4. Response frequencies of the DSM diagnostic symptoms of Major Depressive Disorder in the datasets included in the study.

DSM-IV symptom criteria	**CPES (WHO-CIDI)**	**NHANES (PHQ-9)**		
	Yes (%)	**Several days (%)**	**More than half the days (%)**	**Nearly every day (%)**
Depressed mood	90	16	3	3
Anhedonia	83	16	4	3
Appetite changes	70	15	4	4
Sleep problems	88	24	7	8
Psychomotor disturbances	10	7	2	1
Fatigue	81	35	8	8
Self-criticism	60	12	3	2
Difficulties to concentrate	85	11	3	3
Thoughts of death or self-harm	66	2	1	<.1

Functional impairment

The PHQ-9 contains an additional item addressing the question "if you checked off any problems, how difficult have those problems made it for you to do your work, take care of things at home, or get along with other people?". Functional impairment is therefore measured specifically as a result of the depressive symptoms the respondent has endorsed. Like the rest of PHQ-9 items, it is rated on a four-point Likert scale (0=Not difficult at all, 1=Somewhat difficult, 2=Very difficult, 3=Extremely difficult). Depending on the analyses, the four original ratings or two categories were examined (joining categories 0 and 1, and categories 2 and 3). This single item is considered a measure of functional impairment which correlates strongly with a number of quality of life, functional status and health care usage variables (Kroenke & Spitzer, 2002b, pp. 1–2). We dichotomized the item by merging the two lowest and two highest categories. The reason was achieving an indicator or functional impairment severe enough to

screen for clinical relevance (i.e. here having it "very" or "extremely" difficult to carry out normal activities).

Other relevant variables

Other variables considered were gender (reported as "Male" or "Female"), age (both in years and categorized into four groups: 18-30, 31-50, 51-65, and 65 and older), race/ethnicity (1=Mexican American, 2=other Hispanic, 3=non-Hispanic white, 4=non-Hispanic black, or 5=Other including multi-racial), marital status (1=married, 2=widowed, 3=divorced, 4=separated, 5=never married, 6=living with partner), ratio of family income to poverty (dichotomized with a threshold at 1, informing of whether the income of a household is below or above the poverty level), a count variable of self-reported medical conditions (including diabetes, heart disease, stroke, pulmonary disease, and cancer), and NHANES sampling cohort (2005, 2007, 2009, 2011, 2013, and 2015).

Statistical analyses

Regression models

Several regression models were used for predicting indicators of functional impairment in Studies I and II. In the regression models, predictors at the disorder level (i.e. sum-scores) and symptom level (e.g. most severe symptom, endorsement of symptom) were included. The results were adjusted for age, gender, and ethnic background. Some models were additionally adjusted for number of physical chronic diseases, living below the poverty threshold, and civil status. If not stated otherwise, statistically significant results are interpreted here with respect to a confidence level of $\alpha = .05$.

Functional impairment was a markedly skewed variable in the community datasets included in this dissertation. Consequently, the following models were used:

- *Censored regression model*. When outcome variables show high frequency of floor values, these models truncate the predictions at the lowest value of the observed distribution, while modelling the underlying variable according to a given distribution (e.g. normal, student, logistic). Censored regression models are a generalization of the tobit model (Tobin, 1958).

We defined the lower and upper boundaries of the models so that the predictions would range between the possible values of the outcome (e.g. days disabled in a month range usually from 0 to 30). The model assumes a latent variable y_i^*, and its predictions y_i depend on the following constraint:

$$y_i = \begin{cases} y_l & if\ y_i^* \leq y_l \\ y_i^* & if\ y_l < y_i^* < y_u \\ y_u & if\ y_i^* \geq y_u \end{cases}.$$

The latent variable y_i^* is estimated following the model:

$$y_i^* = \alpha + \beta x_i + u_i,$$

where α is the model intercept, x_i is the vector of linear predictors, and β the parameter vector of regression weights. The error term u_i is distributed normally with mean zero and dispersion σ^2. As with any other linear regression model, a link function $g(\mu)$ can be chosen to define the relationship between the linear predictor component and the mean of the distribution function, μ. The R package *crch* versions 1.0-0 and 1.0-1 were used (Messner, Mayr, & Zeileis, 2016). The multi-stage sampling weights from the CPES data were not used for not being implemented in the package.

- *Logistic regression model*. This model is often used when modeling a binary dependent variable. The model uses most often the logit function as the link function (i.e. $logit\ E(y) = \alpha + \beta x$). In this case, the regression weights contained in β are log-odds of the probability of an event (i.e. $y = 1$). The log-odds (i.e. natural logarithm of the odds) can be converted into odds by exponentiating them:

$$odds(Y = 1) = e^{\alpha+\beta x} = {}^{\pi_1}/_{1-\pi_1}.$$

Where π_1 is the probability of the event happening, and 1-π_1, also π_0, the probability of the complement of the event (i.e. the event not happening: $\pi_1 + \pi_0 = 1$). The interpretation in terms of odds ratios is straightforward by exponentiating the regression coefficient of a given predictor *x* (i.e. $OR = e^{\beta x}$). For example, for a particular depressive symptom with a response yes/no, we can estimate what are the odds of suffering from high functional impairment when endorsing such symptom (i.e. percentual difference expected to report high impairment having the symptom, as compared to not endorsing the symptom). For instance, an odds ratio of $OR = \exp\hat{\beta} = e^{1.89} = 6.62$ would mean that reporting the symptom *X* is associated with high functional impairment by a factor of 6.62, as compared to reporting no such symptom.

We provided some indices informing of the regression models. Aikaike's Information Criterion (AIC) is a fit index that penalizes for model complexity (Burnham & Anderson, 2002). AIC was used for comparing non-nested models. A lower AIC is indicative of a better fit or higher parsimony.

Nagelkerke's pseudo R^2 was also providad, which is asymptotically independent of the sample size and can be interpreted as the proportion of the outcome variation explained by the predictors. Pseudo R^2 values range between 0 and 1 (Nagelkerke, 1991).

There is a phenomenon known as *class imbalance*, which implies that statistical learning algorithms have estimation difficulties when a minority class has too few observations available. In the case of logistic regression, this problem seem to be worrisome when the absolute amount of observations is too little, while the magnitude of the imbalance itself does not play such a big role (e.g. the ratio). For instance, a binary outcome with probabilities 96 and 4% means about 1 398 observations severe reporting impairment in the overall NHANES data, which seems to be large enough to not to fall too short of information. In the case of very rare events, it has been recommended to collapse categories if theoretical considerations support the decision (Tabachnick & Fidell, 2013). We did therefore collapse the NHANES item assessing functional impairment into two categories, otherwise single categories would be problematic due to insufficient amount of observations compared to the amount of predictors in our models.

Severe class imbalance is important in predictive modeling. Its consequences in descriptive modeling, which is the case here, are not so far-reaching. The bias introduced in a model with severe class imbalance concerns mostly the intercept, and not particularly the slope estimates. This implies risk of deflated predictions, but no substantial risk of bias in terms of predictor-outcome associations (King & Zeng, 2001; Owen, 2007). In our data analysis approach, the potential effect of severe class imbalance is the underestimation of the likelihood of severe impairment (i.e. underestimation of the intercept), but the statistical differences seen between demographic groups, the key aspect for our aims, are not in risk of being underestimated. In sum, logistic models are rather robust to severe class imbalance (Alkhalaf & Zumbo, 2017), and therefore seem suitable for the characteristic variables in psychopathology. The package *survey* version 3.33-2 (Lumley, 2019) was used for regression models for its treatment of complex sampling weights of both CPES and NHANES surveys.

Item Response Models

Mental disorders are typically treated as single diagnostic syndromes, conveying unidimensionality. This assumption is characteristic for the reflective framework, meaning that a latent variable (e.g. depression liability) explains the covariation among the observed variables (i.e. items). The two-parameter logistic model was used:

$$p(x_j = 1|\theta, \alpha_j, \delta_j) = \frac{e^{\alpha_j(\theta-\delta_j)}}{1 + e^{\alpha_j(\theta-\delta_j)}},$$

where θ is the parameter corresponding to the level of trait for a person (e.g. depression), and α_j and δ_j are the item j's parameters. The δ parameter is said to measure the difficulty or severity of j (i.e. level of θ required for a .50 probability of endorsing an item). The α parameter informs of the discrimination of the item (i.e. the slope of the curve at its inflection point, δ_j). The δ parameter is especially interesting for psychopathology for estimating the level of the trait associated with endorsing a given symptom. Those symptoms with a higher δ imply a more severe expression of the underlying disorder or trait continuum (e.g. θ interpreted as depression). Item response models are well documented, for example, by de Ayala (2013) and Reckase (2009), and will not be further elaborated here. We may refer explicitly to *psychometric severity* of a symptom or item, meaning particularly this IRT-based estimate of severity.

The symptom criteria of MD present an exceptional case due to strong negative correlations among paired poles of symptoms (weight gain-loss, hypersomia-insomnia, motor retardation-agitation), and thus a two-tier bifactor model was fit, in which only the parameters corresponding to the general factor (i.e. those capturing the common variation) were used. See Cai (2010) for technical details on the Item Response Bifactor model. Item Response models were part of the Study I. The R package *mirt* version 1.21 (Chalmers, 2012) was used for the analyses.

Differential Item Functioning (DIF)

The purpose of DIF is testing psychometric bias. DIF methods assess whether some items (e.g. symptoms) behave differently across population sub-groups (e.g. age or gender) when adjusting for possible differences in the trait (e.g. depression). These analyses were used in Study II.

DIF testing consisted on fitting three logistic regression models, which were nested, for each item (i.e. symptom). For example, in the case of examining DIF with respect to gender in binary items:

$$\begin{gathered} Model\ 1: logit\ P(y_i = 1) = \alpha + \beta_i * trait \\ Model\ 2: logit\ P(y_i = 1) = \alpha + \beta_1 * trait + \beta_2 * gender \\ Model\ 3: logit\ P(y_i = 1) = \alpha + \beta_1 * trait + \beta_2 * gender + \beta_3 * trait * gender \end{gathered} .$$

DIF detection is carried out by comparing the three models (Models 1 vs. 3, 1 vs. 2, and 2 vs. 3). There are three two types of DIF. Uniform DIF corresponds to the case of a given group systematically more or less likely to endorse an item, regardless of their attribute level. Thus, the bias is constant across

all levels. Non-uniform DIF, differently, occurs when the probability of item endorsement varies according to the level of attribute of the groups, and thus the bias is not constant. Uniform DIF can be tested by comparing Models 1 and 2, and non-uniform DIF can be examined by comparing Models 2 and 3 (note it involves an interaction to model the non-constant bias, otherwise modeled with a main effect for uniform DIF). An overall test of DIF consists of comparing Models 1 and 3. The criterion for DIF-flagging was a change in Nagelkerke's $R^2 \geq 0.02$ (Gelin & Zumbo, 2003). A change in the index indicates that the model accommodates differently to the observed data, and thus implies DIF if the improvement introduced by the DIF-related regression estimates is substantial. The analyses were performed with the R package *lordif* (Choi, 2016).

Confirmatory factor analysis and tetrad constraints

Confirmatory factor analysis is a very common statistical method in psychological modeling. This technique estimates scores on a latent variable (i.e. directly unmeasurable), such as depression or verbal aptitude, based on observed scores on a test.

Only the essential information concerning the purposes of this dissertation will be presented here, for a deeper view refer, for example, to Brown (2015). The measurement equation for a unidimensional factor model is:

$$X_{ij} = \lambda_j \xi_i + \delta_{ij} ,$$

where the observed score X of participant *i* in item *j* (also known as indicator) is modeled as a result of the loading λ of the item *j* on the latent factor ξ (e.g. depression) and an error term δ. An assumption of the model is *local independence*, which entails that the covariations among the indicators of a latent factor (e.g. the items of a questionnaire measuring a single trait) are only due to the latent factor underlying them. Put differently, the indicators in a one-factor model are independent of each other conditional on the latent factor ($\rho_{X_j X_{j'}|\xi} = 0$). Spearman, who developed factor analysis to model intelligence in 1904, arrived to this solution in order to make the model identifiable.

The structural operationalization of local independence roots from the principle of *vanishing tetrads*. A tetrad is a difference between a product of one pair of covariances and a product of another pair, among four random variables (i.e. the determinant of a 2 x 2 sub-matrix of their full covariance matrix). Three tetrads (τ) can be derived from a set of four random variables (e.g. items in a questionnaire):

$$\tau_{1234} = \sigma_{12}\sigma_{34} - \sigma_{13}\sigma_{24},$$

$$\tau_{1342} = \sigma_{13}\sigma_{42} - \sigma_{14}\sigma_{32},$$

$$\tau_{1423} = \sigma_{14}\sigma_{23} - \sigma_{12}\sigma_{43},$$

where σ_{14} refers to the population covariance between the first and fourth variables (i.e. items). Of these three tetrads each one is linearly dependent of the other two. A vanishing tetrad is one that equals zero in its evaluated value[13]. Factor analysis requires a number of model equations that can be solved only when these tetrad constraints are included. A reflective structural model with one latent factor (i.e. unidimensional factor model) implies that all the tetrads are vanishing, whereas non-vanishing tetrads will be present if the variables fit to a mixed or purely formative measurement model (e.g. a network model). If the tetrads differ too much from zero, then the reflective model does not properly accommodate the data. For a thorough review, the reader can turn to Bollen and Bauldry (2011).

The principle of vanishing tetrads is structural, and independent of the distribution of the variables or the values of the parameters (i.e. does not depend on data being Gaussian or having any other specific distribution). In the general linear model, the covariance between the latent factor ξ and the residuals is $Cov(\xi, \delta_j) = 0$ for all variables *j*, and $Cov(\delta_j, \delta_k) = 0$ for $j \neq k$. λ_j is the factor loading of *j* onto ξ. Then, it follows that $Cov(X_j = \lambda_j\xi + \delta_j, X_k = \lambda_k\xi + \delta_k) = Cov(\lambda_j\xi, \lambda_k\xi) + Cov(\lambda_j\xi, \delta_k) + Cov(\delta_j, \lambda_k\xi) + Cov(\delta_j, \delta_k) = \lambda_j\lambda_k Cov(\xi, \xi) = \lambda_j\lambda_k Var(\xi)$.

This implies that all $Cov(X_g, X_h) \times Cov(X_i, X_j)$ and $Cov(X_g, X_i) \times Cov(X_h, X_j)$ are both equal to $\lambda_g\lambda_h\lambda_j\lambda_k Var(\xi)^2$, and the difference between them is zero for all tetrad equations under this model. If, for example, X_g and X_h are locally dependent (i.e. violate the model), this adds a unique component $Cov(\delta_g, \delta_h) \neq 0$ to the $Cov(X_g, X_h)$, and therefore it is very unlikely that $Cov(X_g, X_h) \times Cov(X_j, X_k) = Cov(X_g, X_j) \times Cov(X_h, X_k)$.

The tetrad tests were used in Study III under the following rationale. The symptom criteria in the operational definition of MD are very heterogeneous, covering somatic, affective-motivational and cognitive symptoms. It is rather unlikely that a single latent variable exhaustively underlies all these

[13] The determinant of a square matrix M of order 2 is denoted by $|M|$ and is calculated as: $|M| = \det\left(\begin{bmatrix} g & h \\ j & k \end{bmatrix}\right) = gk - hj$. The determinant of a matrix M will be zero if and only if there exists a λ such that $g=\lambda j$ and $h=\lambda k$.

symptoms, and therefore we chose a group of symptoms that more plausibly reflect a common trait, or directly relate to each other. These symptoms are specific from MD in the sense that they do not overlap with other disorders, and often feature core descriptions of the syndrome. The symptoms we focused on are depressed mood, anhedonia, self-criticism, and death thoughts. The used the model-implied tetrads for conducting a confirmatory bootstrapped test following the work of Bollen and Ting (1998; 1993). We wrote the script in R, being part of it based on functions of the R package *lavaan,* version 0.6-2 (Rosseel, 2012).

Confirmatory Factor Analysis (CFA) was also performed in Studies I and III with the package *lavaan*. We evaluated model fit according to the customary indices and thresholds. The following cut-off values indicate acceptable fit: RMSEA ≤ .06, CFI ≥ .95, and SRMSR ≤ .08 (Hu & Bentler, 1999; Kline, 2015). RMSEA (Root Mean Square Error of Approximation) and SRMSR (Standadized Root Mean Square Residual) are absolute measures of fit, which is better as the estimate decreases. The CFI (Comparative Fit Index) is an incremental measure and pays a penalty for every parameter estimated. An increase in the measure suggests better fit.

Linear Non-Gaussian Acyclic Model (LiNGAM)

The LiNGAMs were developed to infer dependence structures from cross-sectional, observational data. In the most basic setting, the LiNGAM models make the following assumptions: (i) the data generating process is linear, (ii) there are no unobserved confounders, and (iii) residuals have non-Gaussian distributions of non-zero variances, so that the dependence relations between variables can be modeled for a single observation (e.g., study participant) in matrix form as:

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{x} + \boldsymbol{\epsilon},$$

where the vector $\boldsymbol{\mu}$ contains the constants and the strength matrix **B** collects the regression coefficients. The vectors $\boldsymbol{x}$ and $\boldsymbol{\epsilon}$ include the observed variables and residual terms respectively. Assuming an acyclic structure (Bollen, 1989), it can be shown that it is always possible to perform simultaneous permutations on the connection strength matrix **B** to render it strictly lower triangular. "Acyclicity" entails the absence of feedback loops among variables, meaning their association pattern represents a directed flow of information.

In the example

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & 0 \\ 0 & -3 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix},$$

the variable x_2 is equal to its residual ϵ_2, and therefore "exogenous" since it is not affected by the other observed variables, neither x_1 nor x_3. In general, an exogenous variable is not predicted by other observed variables in the system, but in turn, it can be considered as an input to the system of variables. The next variable in the causal ordering is x_3, as it is directly influenced by the exogenous variable, and the last one in the causal chain is x_1. Once the matrices are permutated according to the causal ordering (**B** to lower triangular), we find:

$$\begin{bmatrix} x_2 \\ x_3 \\ x_1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ -3 & 0 & 0 \\ 0 & 2 & 0 \end{bmatrix} \begin{bmatrix} x_2 \\ x_3 \\ x_1 \end{bmatrix} + \begin{bmatrix} \epsilon_2 \\ \epsilon_3 \\ \epsilon_1 \end{bmatrix}.$$

This model is depicted in Figure 2. The goal of the algorithms applying LiNGAM is to estimate the **B** matrix from the data in order to infer direct causal associations. The **A** matrix, also computable, contains the total effects (direct and indirect) between the variables (Hoyer, Shimizu, Kerminen, & Palviainen, 2008).

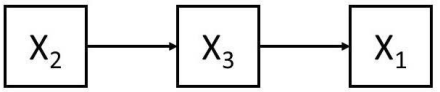
The DirectLiNGAM algorithm

Many LiNGAMs are related to Independent Component Analysis (ICA), which is a computational method used to separate a multivariate signal into additive components. These components are understood as latent factors that underlie the observed multivariate variable (Hyvarinen, Karhunen, & Oja, 2001). ICA-based and gradient-based LiNGAM algorithms are iterative, which poses some potential stability problems. In order to avoid these problems, Shimizu and colleagues introduced a more direct method, called DirectLiNGAM, which does not require potentially unstable iterative gradient ascent algorithms (Shimizu et al. 2006; Shimizu et al. 2011). The algorithm estimates the **A** and **B** matrices by first identifying one exogenous variable and removing its effect from the other variables by means of least squares regression, then again searches for the next exogenous variable and continues until a dependence ordering is established. When simplified to the pairwise setting in our illustration, we will generally refer to the independent variable as *X* and to the dependent as *Y*. The algorithm does not 'know' which is which and aims to distinguish between two systems of equations, either

$$\begin{cases} \mathrm{X} = \mu_X + \epsilon_X \\ \mathrm{Y} = \mu_Y + \beta\mathrm{X} + \epsilon_Y \end{cases} \quad \text{or} \quad \begin{cases} \mathrm{X} = \mu_Y + \beta\mathrm{Y} + \epsilon_X \\ \mathrm{Y} = \mu_Y + \epsilon_Y \end{cases}.$$

The residuals $\boldsymbol{\epsilon} = (\epsilon_X, \epsilon_Y)$ follow non-Gaussian distributions with zero mean and non-zero variance, and they are jointly independent. Based on the model assumptions and on the Darmois-Skitovich theorem, the independent or antecedent variable is the variable that is more independent of its residual (Shimizu et al. 2011). That is, we can define $\hat{MI}\ (X, \epsilon_y)$ as an estimate of mutual information

Figure 2.Example of Directed Acyclic Graph (DAG).



between X and the residual of Y when regressed on X, denoted ϵ_Y. The algorithm uses a nonparametric kernel-based estimate of mutual information. Bivariate mutual information quantifies how much information about one variable can be obtained through the other. This measure is based on the concept of information entropy, which quantifies the amount of information held in a random variable. Pairwise mutual information can be understood as a degree of departure from bivariate independence. It is important to notice that mutual information relates to *all* dependence, linear and nonlinear, whereas the Pearson product-moment correlation only entails linear dependence. For this reason, DirectLiNGAM can make use of higher statistical moments than covariances and surpasses traditional structural equation models in what comes to causal discovery. For a detailed review on the mathematical grounds of the algorithm, we refer the reader to the original papers by Shimizu and colleagues (Shimizu et al., 2011; Shimizu et al., 2006).

Simulation studies on the performance of DirectLiNGAM

DirectLiNGAM has shown promising performance in simulation studies and with real-world datasets (Rosenström et al., 2012; Rosenström & García-Velázquez, in press; Shimizu et al., 2011). We performed a brief simulation study to inspect how the DirectLiNGAM algorithm performs in circumstances realistic of research in psychopathology and psychiatric epidemiology.

We examined the detection success of DirectLiNGAM with ordinal variables. This is because the LiNGAM methods have been developed for continuous variables, which have more variability and more accurately defined units than the ordinal-valued (Likert-type) variables. Ordinal items are used very commonly in psychological questionnaires, such as the Patient Health Questionnaire included in NHANES studies. Rosenström et al. (2012) conducted a simulation study by categorizing continuous variables and preserving their scale, while varying several parameters like sample size and number of categories. The estimation performance of the algorithm was nearly flawless. More detailed simulation studies are needed so that users know the performance of the DirectLiNGAM algorithm with those research conditions typical of psychopathology. This is important, because using an ordinal level of measurement for a supposedly continuous underlying variable in survey questionnaires may alter the underlying 'true' scale or unit of that variable, which can lead to serious problems in statistical modeling of multiple variables (e.g. Kang & Waller, 2005). This possibility is especially alarming for distribution-based inference because transformations of scale also alter observed distributions.

Thus, in study III we addressed different aspects of ordinal variables, such as the number of categories (two to seven), and the type of categorization (even categorization, doubling the number of categories in either variable and adding noise to alter the grid for categorization). The study was conducted on several sample sizes (from 100 to 5000) and replicated 2000 times per setting to estimate the proportion of success.

Results

Psychometric modeling of the symptom criteria of depression

Two of the studies in this dissertation implemented latent variable models of depression. In Study I we presented a bifactor 2-parameter logistic item response (IRT) model. The bifactor model is characterized by all indicators loading on a general factor, and several specific factors accounting for the remaining shared variance among subsets of items. By means of a bifactor structure we modeled a general latent trait (i.e. depression), estimating separately the local dependencies between item pairs due to strong negative correlations between the opposed symptom criteria (e.g. excess and decrease in appetite or sleep, so-called compound criteria composed by the two extremes). The fit of this model was acceptable (RMSEA = .056, CFI = .937, SRMSR = .048) according to the customary thresholds for structural equation modeling.

The IRT models include an estimate of item severity that can be useful in studies of criterion validity. These severity estimates, in our analyses, corresponded to the general factor (i.e. were free from other sources of covariance, such as the aforementioned compound-criteria correlations). The latent trait (general factor in bifactor models, represented as θ) is generally estimated under the assumption of normality in IRT, implying that the 95% of the population scores are contained within the range $\theta = \{\pm 1.96\}$. A high severity estimate means that it is required as high θ to endorse the item with a 50% likelihood. The IRT models of the four disorders tested in Study I (GAD, ME, MDE, PTSD) reflected, in general, a rather clear dose-response association between psychometric severity and the impairment indicators (WHODAS scale and disability days). Moreover, psychometric severity was predictive of the impairment indicator (WHODAS scale) in regression analyses independently of the symptom counts. Such result can be understood as evidence of concurrent validity for the severity estimates. The same behaviour of psychometric severity estimates was seen for all diagnostic disorders with the exception of MD, which model showed some suboptimal characteristics worth to describe here.

First, the severity estimates of MDE symptom criteria in the model registered some very deviant values for three items (weight gain, hypersomnia, and motor agitation). The large magnitude of the estimates of these three items suggests there may be something wrong with the model (Reise & Waller, 2009). Second, the fact that the severity estimates do not predict the WHODAS scores in regression model of MDE does not support the use of the psychometric severity estimates. The unidimensional

bifactor IRT model of MDE did not perform as a useful model and showed some signs of computational inadequacy.

In Study III, we explicitly tested the plausibility of a reflective model in the NHANES data. Only the four items specific to the diagnosis of MD were included (i.e. anhedonia, depressed mood, feelings of self-criticism, and thoughts of death and self-harm). Across the separate analyses conducted on the six comparable NHANES samples, the bootstrapped confirmatory tetrad analyses indicated invariably that a reflective model did not accommodate the data well (highest bootstrapped p-value = .0025). This means that the items have associations among them, which are not captured adequately by the single latent factor.

Individual symptom criteria as compared to their sum-scores

After the difficulties observed in Study I for the latent estimates of MDE, we proceeded to examine observed variables only. Observed variables can be studied separately (individual depression criteria) or as aggregates (scale scores or symptom counts). We approached this topic in Study II by conducting several regression models, which included a wide variety of relevant covariates (among them the number of physical conditions and risk of poverty). We first computed a baseline model, where the PHQ-9 sum-score was specified as predictor of high depression-related functional impairment. In a posterior series of models, the PHQ-9 items were modeled separately and their independent association with the same dependent variable was estimated. According to the results, the model including all the individual symptoms fits the data better than the one using the sum-score in terms of AIC (4563.42and 4574.57, respectively) and Nagelkerke's pseudo R^2 (.331 and .316).

Associations of individual symptom criteria with self-reported functional impairment

This question was comprehensively targeted in Study II by means of several regression models, which were adjusted for relevant covariates to depression. Five symptoms (i.e. symptom criteria) showed a positive association with severe impairment after adjusting for all other symptoms simultaneously plus covariates. The symptoms were anhedonia, depressed mood, fatigue, concentration difficulties, and psychomotor problems. The higher the score on the symptoms, the most likely it was that a person suffered very or extreme difficulties to carry out normal life. The main effects in a generalized linear model can be added to each other to estimate the cumulative probability of endorsing the outcome. The unique contribution of each of these symptoms to the model's Nagelkerke's R^2 ranged from 3.0 to 4.3%,

suggesting that the association between individual symptoms and high functional impairment was mostly unspecific (i.e. symptoms mostly overlapped with each other in the variance shared with the outcome).

There were three symptoms that did not predict high functional impairment under a significance level of α=.05. The symptoms were sleep problems, appetite changes, and thoughts of self-harm and death. These symptoms showed a significant effect on the single-symptom regression models, which disappeared after adjusting for the effects of the other symptoms. This suggests that they lack association with functional impairment independently.

We will examine here the symptom criteria belonging to the CPES dataset in the same manner individual PHQ-9 items were analyzed in Study II. These analyses are not published elsewhere, and were conducted for complementing the information in this doctoral dissertation summary. The WHO-CIDI included more detailed information on symptoms, that we merged together to correspond to the compound criteria in DSM-IV and therefore to be comparable to other measurements (e.g. feelings of worthlessness and guilt were two different items, which we merged into one). The binary outcomes were high interference with life and being often unable to carry out daily activities. The logistic regression models for impairment are shown in Table 5. The models were adjusted for age, sex, marital status, ethnicity, and number of chronic conditions. The 56% of respondents reported that, while suffering their most severe depressive episode, their symptoms interfered "a lot" or "extremely" with their lives. Of those who reported some degree of interference, the 29% stated that they had been "often unable" to carry out daily activities due to that depressive episode. Such high rate of self-reported impairment stems from the fact that the interview refers to the worst affective episode the respondents remember.

The symptom-level associations in Table 5 are rather consistent between the two outcomes. Depressed mood, difficulties to concentrate or indecisiveness, and thoughts of death or self-harm showed significant associations with high interference with life, which were otherwise not statistically significant with regard to being often unable to carry out daily activities. The association with depressed mood was particularly large (OR=18.1). The symptoms bearing similar associations with both outcomes were anhedonia, appetite and weight changes, and self-criticism. Nagelkerke's R^2 was .106 for the interference model, and .105 for the inability model. The multi-stage probability weights were used for computing these models.

Table 5. Binary logistic regression estimates for symptom criteria of Major Depression, predicting two indicators of impairment in CPES (n= 4 152).

Symptom criteria of MDE (yes/no)	**Outcome: *a lot* or *extreme* interference with life[a]**				**Outcome: *often* unable to carry out daily activities due to symptoms[a]**			
	OR	**SE**	**Z**	**p-value**	**OR**	**SE**	**Z**	**p-value**
Depressed mood	18.102	.752	3.850	<.001**	1.000	1.025	.000	1.000
Anhedonia	3.184	.313	3.702	<.001**	5.877	.605	2.928	.003**
Appetite or weight changes	2.192	.309	2.543	.011*	2.455	.405	2.216	.027*
Sleep problems	1.946	.372	1.793	.073	2.487	.598	1.523	.128
Psychomotor disturbances	1.505	.309	1.325	.185	1.578	.259	1.759	.079
Fatigue	.905	.377	-.265	.791	.973	.373	-.071	.943
Self-criticism	2.489	.402	2.270	.023*	2.675	.491	2.005	.045*
Difficulties to concentrate	2.330	.373	2.266	.023*	2.625	.554	1.740	.082
Thoughts of death or self-harm	2.627	.261	3.707	.001**	1.201	.247	.738	.461

Footnote. Models adjusted for age, gender, ethnicity, marital status, and number of chronic conditions. Estimates are sample-weighted. [a]The symptoms reported concern the "worst" affective episode the respondent remembers. *: statistically significant at α=.05. **: statistically significant at α=.01. OR: Odds ratio, SE: standard error of the estimate.

The role of gender and age group in the association between symptom criteria of MD and self-reported functional impairment.

Aggregated sum-scores

Gender did not play a role in the association between severe self-reported impairment and the depressive symptoms causing it, according to the regression models in Study II (i.e. p-value >.05 in all models). Age in years predicted severe functional impairment, both as a main effect (additive; b = .02, Z = 3.04, p-value = .002) and moderating the association between PHQ-9 sum-score and very or extreme impairment (non-additive; b = -.001, Z = -2.60, p-value = .009). We further inspected this moderating effect by categorizing age into four age groups. The main and interaction terms remained statistically significant. The predicted values differed when comparing the oldest group (those aged 66 or above) to the two youngest groups (18 to 30 and 31 to 50). The associations can be seen in Figure 3, where the model predictions are displayed against age group. The model was adjusted for gender, number of co-occurring medical diseases, marital status, living below the poverty threshold, and NHANES cohort year.

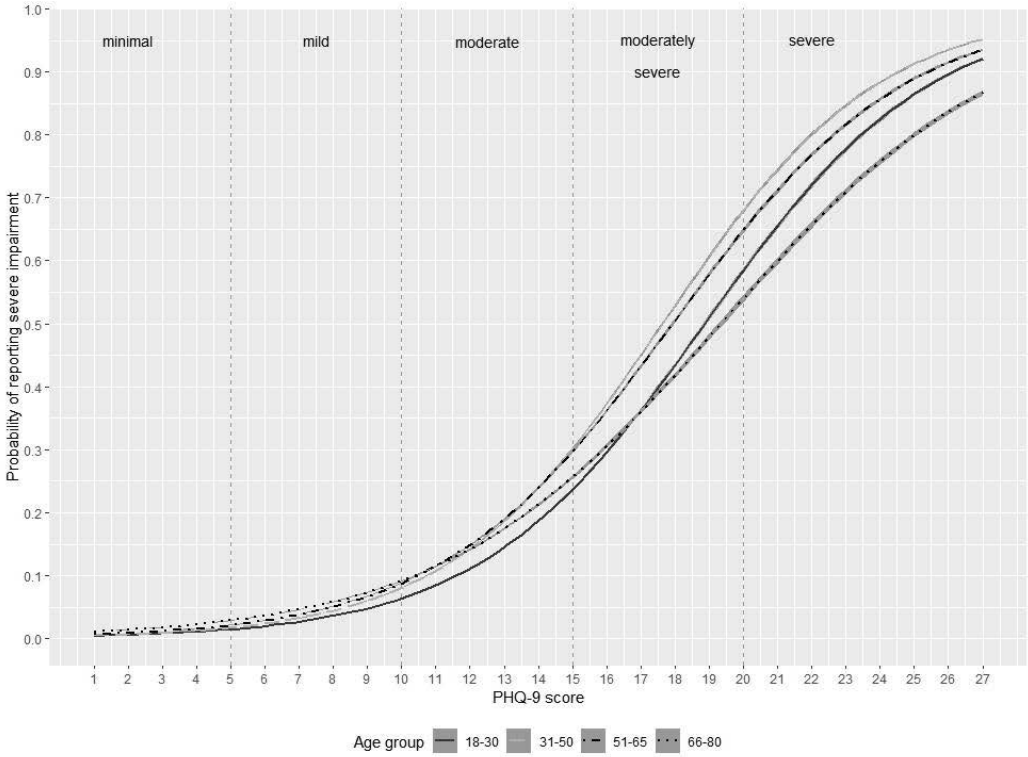
According to the model, being depression sum-scores mild (ranging from 1 to 9), older adults were slightly more likely to report severe impairment than were younger adults. However, as depression scores increased from moderate onward, participants aged 66 and older were increasingly less likely than middle-aged adults to report severe impairment. For example, at a score of 20 the probability of feeling severely impaired was 53% in the age group over 65, and 67% in the age group of 31 to 50 year-olds.

Individual symptoms

Gender did not play a role in the association between self-reported impairment and individual symptoms causing it. Age group did have a moderating effect on the association between individual depressive criteria and severe functional impairment. Three of the nine symptoms showed a significant interaction term with age group. These symptoms were depressed mood, concentration difficulties, and self-criticism. The interaction effects are displayed in Figure 4 indicating that individuals aged 31 to 65 were more likely to report that carrying out normal activities was very or extremely difficult when suffering from these symptoms.

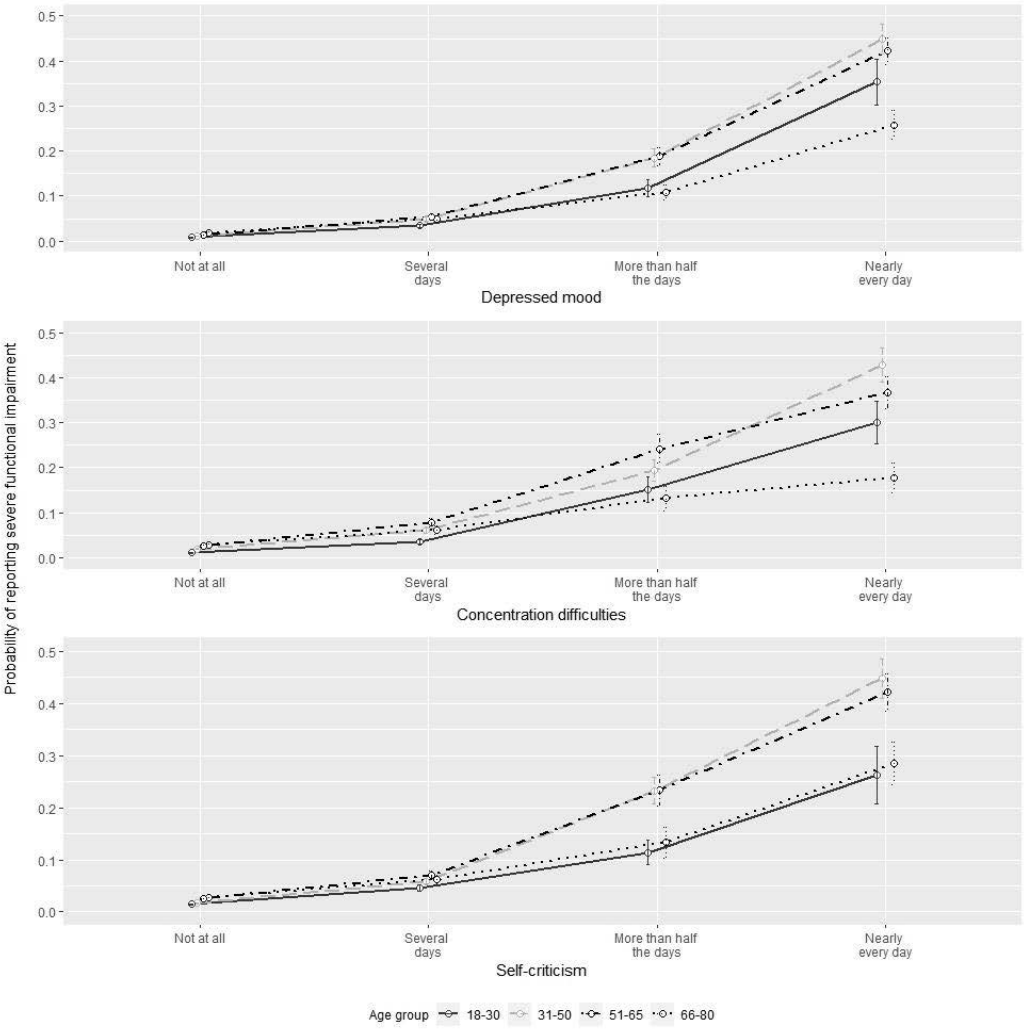
Post-hoc analyses revealed that these three symptoms showing interaction with age group explained the interaction effect seen at the aggregate level: once removed the three items from the sum-score, the moderating effect of age on sum-score was not statistically significant.

Figure 3. Logistic model fitted probabilities for experiencing as very or extremely difficult to carry out daily activities as predicted by PHQ-9 scores in NHANES (2005-2015, n= 34 963).



Footnote. The 95% confidence intervals are presented as a grey area around the lines. The annotations and dashed vertical lines correspond to previously established cut-off points for degree of depression in PHQ-9 scores.

Figure 4. Fitted probabilities for experiencing as very or extremely difficult to carry out daily activities, as predicted by the PHQ-9 symptoms which showed moderating effect by age group in NHANES (2005-2015, n= 34 963).



Footnote. The whiskers correspond to the 95% confidence intervals.

Direction of dependence among specific symptoms of depression.

Confirmatory tetrad analyses invariably rejected the unidimensional model, for each of the six NHANES samples. Consequently, we conducted direction of dependence analyses. Due to the novelty of this technique in psychopathology, we applied the algorithm separately to the six samples (NHANES year cohorts 2005-2006 to 2015-2016). This way, the replicability of the results of this exploratory method was shown. The algorithm DirectLiNGAM estimated a virtually identic dependence ordering for each of the six samples (Figure 5).

We estimated sampling-weighted, conditional probabilities for pairs of symptoms in the overall NHANES sample (Table 6). This descriptive information does not imply direction of dependence, but may assist in better characterizing which symptoms tend to be present when others occur. The tables are interpreted as the percentage of the population reporting a symptoms given that another symptom is present (e.g. for those endorsing low mood, what proportion reported anhedonia). Thus, the rates only imply bivariate associations, which means they are not adjusted for the presence of other relevant symptoms or variables that may explain the bivariate relationship. The conditional probabilities are lower than those reported by Zimmerman et al. (2006a) on otherwise similar analyses conducted on a clinical sample of outpatients, which seems a reasonable difference of general population as compared to clinical samples. The information contained in Table 6 has not been previously published in any original publication of this dissertation.

Figure 5. Directed Acyclic Graph displaying the direction of dependence estimated by DirectLiNGAM separately in six cross-sectional NHANES cohorts (2005-2015).

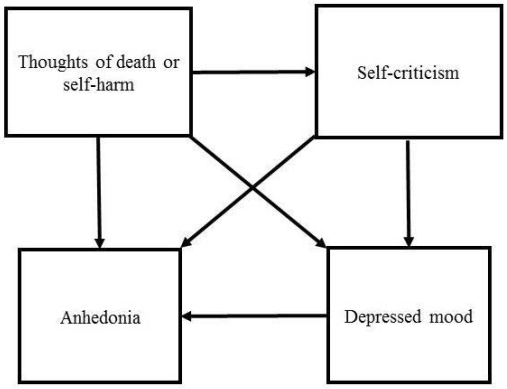


Table 6. Conditional probabilities between pairs of depressive symptoms as assessed by PHQ-9. NHANES samples 2005-2015.

	Anhedonia	Mood	Sleep	Fatigue	Appetite	Self-criticism	Concentration	Psycho-motor	Death	Severe impairment
Anhedonia	-	55[a]	25	27	34	47	41	42	55	53
Depressed mood	48[a]	-	24	26	31	58	43	43	76	55
Sleep problems	54	59	-	52	53	60	59	59	66	63
Fatigue	60	65	55	-	60	64	63	61	70	72
Appetite or weight changes	40	42	29	32	-	47	44	46	52	48
Self-criticism	33	46	20	20	28	-	38	38	71	46
Difficulties to concentrate or indecisiveness	32	37	21	22	29	42	-	49	54	46
Psychomotor disturbances	19	22	13	12	18	25	29	-	36	31
Thoughts of death or self-harm	7	12	4	4	6	14	9	11	-	14
Severe impairment	22	27	13	14	17	28	25	29	44	-

Note. [a] The table should be read as follows: 48% of patients with anhedonia reported depressed mood, whereas 55% of patients with depressed mood were anhedonic.

Discussion

This thesis examined the association of depressive symptoms and self-rated functional impairment in population-based samples of adults from the United States. Knowing more precisely how individual symptoms relate to relevant outcomes is valuable for disentangling the heterogeneity of depression. We will next briefly list the main findings of this dissertation study-wise. In the following sections, we will discuss the implications of these findings while bearing in mind the five aims of the dissertation, and interpret them in the context of other literature.

In Study I, we inspected whether the psychometric severity of diagnostic symptoms was useful as a parallel diagnostic feature. If so, this information could help clinicians working within the DSM framework. In three of the four disorders we examined, psychometric severity offered additional information over and above symptom counts. Results of the CPES data showed a dose-response relationship of psychometric severity with psychosocial disability for GAD, Mania, and PTSD, with the exception of MD. The latent variable model of MDE showed some suboptimal characteristics. Possible reasons are misspecification, and that the items were not suitable indicators of a common latent trait (i.e. depression). Thus, it did not appear as an adequate model to continue with in future analyses, and we consequently departed from modeling observed variables in later studies.

In Study II, we compared sum-scores and individual symptoms as predictors of severe functional impairment using NHANES data. The goodness of fit of these models suggested that modeling depression with individual symptoms marginally improved the predictive power of the model, as compared to modeling with the sum-score. In Study II, we also studied whether gender and age groups have implications for the impairment associated with MD symptom criteria. We found no evidence of the role of gender, and found that age group showed a moderating role for three symptoms: depressed mood, concentration difficulties, and self-criticism. Adults aged 31 to 65 were particularly prone to reporting severe functional impairment when suffering from these symptoms, as compared to younger and older adults. Other symptoms had a constant association with severe functional impairment across age groups, with the exception of sleep problems, appetite changes, and thoughts of self-harm and death, which did not hold any significant association with self-reported severe impairment in the mutually adjusted model.

In Study III, we assessed more exhaustively the suitability of a latent trait model to the four symptom criteria specific of MD (i.e. depressed mood, anhedonia, self-criticism, and thoughts of death or self-harm), by means of bootstrapped confirmatory tetrad test on the NHANES data. These four symptoms were selected because they seem to constitute a tighter symptom cluster, from which somatic symptoms differ consistently. Our analyses, based on six replications upon the large population-based cohorts, indicated unanimously that the unidimensional latent model does not fit the observed data. The six cohorts were then analyzed with the DirectLiNGAM algorithm, that aims to infer direction of dependence in non-Gaussian and cross-sectional data, based on their statistical distribution. The six replications indicated widely the same directed network: thoughts of death and self-harm → self-criticism → depressed mood → anhedonia. We discussed that the direction of dependence methods (among which LiNGAM is) are an untapped resource for psychiatric epidemiology because of exploding information commonly found in the field: non-Gaussian, cross-sectional data, from correlational studies. However, the performance of DirectLiNGAM with ordinal and noisy data remains underexplored. We therefore performed brief simulation studies assessing its estimation success under multiple conditions. The study revealed that DirectLiNGAM is a rather robust method for estimating direction of dependence in those circumstances, particularly with large samples and variables with the same number of categories.

Sum-scores as indicators of depressive symptoms

As mentioned above, we found that modeling depression with individual symptoms marginally improved the predictive power of the model. However, there is no standard rule for evaluating whether a change of 1.5% is indeed a significant improvement on model prediction, as compared to sum-scores. Nagelkerke's R^2 does not penalize for model complexity, and thus the value of such improvement in exchange to eight additional variables is, at least, questionable. Thus, from the statistical viewpoint we may conclude that **sum-scores functioned about as efficiently as individual symptoms** in predicting subjective severe functional impairment.

There are additional considerations from the substantive viewpoint. Examining individual symptoms offered in our study a **deeper understanding of depression as a syndrome.** For example, we found in post-hoc analyses that it was three symptoms that drove the interaction effect observed between the depression sum-scores and age on functional impairment. This information was concealed within a single value in the aggregate sum-score. If our finding held across studies, this would imply that adults from middle age to retirement are more prone to considering themselves as severely impaired by those

three depressive symptoms, in comparison with younger and older adults. However, it is also possible that the symptoms causing higher burden to younger adults and those at retirement age are not well captured within the standard MD symptom criteria. Notwithstanding, this finding was only revealed when analyzing and the level of individual symptoms. By analyzing aggregates, we take the risk of missing useful insights.

A second important caveat of using symptom aggregates, worth considering here, is lack of content overlap across scales. This was minimized in the studies of this dissertation by restricting the symptoms analyzed to the current DSM symptom criteria. However, plenty of studies based their results on depression scales, which scope and content may vary considerably. Reporting and analyzing mere sum-scores comes with the pitfall of ignoring **whether it is the accumulation of symptoms that explains an association** (for which sum-scores are suitable), **or is it particular symptoms** which drive the effect observed (as we saw in Study II). A combination of both is naturally possible. Information about these important etiological mechanisms are made accessible by comparing systematically individual symptoms and their aggregates.

Individual depressive symptoms and self-rated functional impairment

All symptoms were, to more or less extent positively related to impairment. This is presumable and has indeed little novelty as a research finding (e.g. Kendler & Gardner, 1998). The novelty was examining, in naturalistic samples, the association of individual symptoms on self-rated functional impairment while adjusting for all others. All symptoms are associated with impairment when examined one by one, but which of them are informative when taking all simultaneously into account is different. A regression estimate which is not statistically significant in a mutually-adjusted model tells that the variable (i.e. symptom) does not contribute to the predictive ability of the model over and above the rest of the variables included. In this sense, our approach is different to the more common study of depressive symptoms as aggregates. For a disorder being as heterogeneous as MD, a **symptom-based approach has the potential to reveal systematic associations of concrete symptoms that emerge while controlling for the presence of all others**. If these results were stable and replicated across different studies, specific symptoms could offer useful information in the presence of a variety of symptom presentations. **Those symptoms that relate to etiological or prognostic factors independently have obvious implications for treatment planning**, especially if they bear an influence on other symptoms too.

We found no clear trend between symptom prevalence and association with functional impairment. Put differently, both very prevalent and relatively infrequent symptoms showed associations with functional impairment. For example, depressed mood and anhedonia are core symptoms and tend to be endorsed easily, relative to other symptoms. Both emerged as significant predictors of functional impairment systematically in our analyses. However, sleep problems were practically as frequent in our data, yet the association of this symptom was never statistically significant after adjusting for the other criteria. Thus, a consistent finding was that **functional impairment was not simply approximated from prevalence distributions of symptoms** (e.g. very prevalent symptom expectedly more or less disabling than an infrequent symptom). This could explain, at least to some extent, the poorly informative estimates of psychometric severity in Study I. The severity parameter of IRT models is linked to prevalence rates, the more often the item is endorsed (i.e. here symptom), the lower the severity estimate (de Ayala, 2013). The statistically non-significant regression weight could be explained by the (prevalence-based) estimated severity not having a linear relationship with the outcome variable, WHODAS.

The regression models revealed two other insights. First, our models showed a rather low predictive power upon severe self-rated impairment, as measured by Nagelkerke's pseudo-R^2. This suggests that **functional impairment attributed to depressive symptoms was not a simple function of the symptoms or their counts**. Other previous studies of clinical MD have similarly argued that symptoms and functional impairment are rather different components of severity, not directly approximated by another (Lux et al., 2010; Zimmerman et al., 2008, 2012, 2018). Thus, measuring the functional impact of depression seems necessary for assessing the overall severity of the syndrome also in research.

A second finding related to the predictive power of the models is that the unique contribution of depressive symptoms was rather small and of similar size across symptoms, suggesting that **the association of individual depressive symptoms with impairment was ambiguous or overlapping**. These results are in line with those of a recent study by de Vries and colleagues (2018). They compared similarly sum-scores and individual symptoms in predicting response to antidepressants (monitored at second week, outcome registered on the sixth week), and concluded that individual symptoms did not add meaningful predictive information to the sum-score. They hypothesized the reason was that prediction of treatment response was not particularly symptom-specific (i.e. a wide range of symptoms involved, mixed in domain). However, clinical studies have also shown stronger symptom-specific

findings. For instance, Fried and Nesse (2014) calculated the relative importance of individual predictors over WHODAS scores and found that symptoms drastically varied on their effects on impairment. Their measure was based on average stepwise changes on a linear model, and estimated over unadjusted R^2. This disagreement could stem from differences in the measurement of symptom-specific contributions to the models, or from real differences in the association between symptoms and functional impairment across community-based and clinical samples.

The two core criteria of depression, depressed mood and anhedonia, are symptoms most commonly found to predict impairment, in our research and in others' (Faravelli et al., 1996; Fried & Nesse, 2014; Tweed, 1993). It is well known that depressed mood is among the most, if not the most common depressive symptom. Depressed mood is invariably present in clinical cases of varying severity, and also prevalent in subthreshold and dysphoric episodes which may transient on their own (Horwitz & Wakefield, 2007; Zimmerman, McGlinchey, Young, & Chelminski, 2006). Anhedonia and motivational symptomatology may be similarly described as rather prevalent and characteristic of clinical and less severe states (Rosenström & Jokela, 2017; Zimmerman, Chelminski, McGlinchey, & Young, 2006; Zimmerman, McGlinchey, Young, & Chelminski, 2006b). The **symptoms of depressed mood, followed by anhedonia, were dependent on death thoughts and self-criticism** according to the DirectLiNGAM-based results. Depressed mood and anhedonia were similarly likely to react to changes in other symptoms in a longitudinal study by Bringmann and colleagues (2015) on a clinical sample. They stated that sadness and loss of interest "do not play a large role in funneling the symptom spread themselves" (Bringmann et al., 2015, p. 7). In the same direction, a recent study by Boschloo and colleagues found that an internet-based intervention decreased depressed mood and anhedonia through the improvement of other symptoms, among which was self-criticism (Boschloo et al., 2019).

However, there is an additional interpretation of the LiNGAM-based network, which is estimated based on distributional properties. It has to do with sensitivity and specificity of the symptoms. Being depressed mood and anhedonia rather prevalent (and possibly reactive to other symptoms), they generally show high sensitivity and rather low specificity. Thereby, most of the severely impaired individuals suffer from depressed mood and anhedonia, but these symptoms themselves discriminate poorly between cases with more or less severe depression. This is because they tend to be present when others are, but the opposite may not hold (see Table 6). The Direct-LiNGAM algorithm uses a kernel estimate of mutual information, and so the direction of dependence estimated has to do with the informative of a variable

about another. This high sensitivity has to do as well with the fact that the two core symptoms registered consistently an association with severe functional impairment.

Across different analyses, there was systematic evidence that **feelings of worthlessness and guilt played a role in severe functional impairment, and are closely related to other symptoms too**. Self-criticism showed a significant association with all the indicators of severe functional impairment we tested in the regression models. Similarly, Faravelli at colleagues (1996) found that hopelessness and guilt were among the most correlated with the severity and global functioning scores. Moreover, in Study II self-criticism showed an age-related pattern, according to which adults aged 31 to 65 were more prone to report severe functional impairment than adults of other ages. This complements previous literature pointing at the informative value of this symptom. For example, a study analyzing symptom presentation and MD onset found that two symptoms predict onset before age of 60: feelings of worthlessness and guilt, and depressed mood (Heun, Kockler, & Papassotiropoulos, 2000). A recent follow-up study examining the stability of cognitive vulnerability and depressive symptoms found both are more stable than expected, indicating that affective symptoms tend to be maintained hand in hand with trait-like constructs, such as hopelessness, rumination, and worry (Struijs et al., 2020). Another study showed that subjects who had experienced earlier onset depression were more likely, after the age of 65, to report feelings of worthlessness and guilt than those with no previous depression (Gallagher et al., 2009). These findings suggest that self-criticism may reach clinical implications that are consistent with the association on severe impairment we found across adulthood. Furthermore, feelings of worthlessness and suicidal thoughts are considered criteria for complicated depression (Wakefield, Schmitz, First, & Horwitz, 2007), which is empirically related to a number of clinically relevant outcomes. Feelings of worthlessness also predicts concurrent and post-remission suicide attempts (Bolton et al., 2008; McGirr et al., 2007; Wakefield & Schmitz, 2015).

In the directed network, self-criticism was the second most dominating symptom after thoughts of death and self-harm. In the longitudinal study by Bringmann et al. (2015), feelings of worthlessness and guilt were among the most influential over other symptoms, while being very central to the network. In their study, suicidality was the symptom influencing others the most, while being the least influenced by the others. This is another similarity with our directed network, in which **thoughts of death and self-harm were a dependence indicator for the other cognitive-affective symptoms**. *Dependence* here implies that it more reliably and linearly implies the presence of other symptoms. This can stem from a

direct association, or from other etiological factors which underlie the directed chain observed –this would be a violation to the assumptions of the model, but a possibility.

A prospective study on suicidal acts during adolescence found that cognitive symptoms dominated the profile of suicide attempters, indicating the simultaneous activation of all these symptoms around suicidal acts (Nrugham, Larsson, & Sund, 2008). Other **studies of temporal antecedence**, understood as an indication of causal mechanisms, have pointed at these symptom dynamics via variability across time. In these studies, **suicidality is the most stable symptom, while depressed mood and anhedonia tend to vary the most.** Apart from the aforementioned study by Bringmann et al. (2015), other follow-up studies followed patients for a time-span of two to nine years (Karp et al., 2004; Oquendo et al., 2004; van Eeden, van Hemert, Carlier, Penninx, & Giltay, in press). Despite of the stability of suicidal thoughts in longer time intervals, there is some recent evidence of large variability across short timespans, like the course of one day (Kleiman et al., 2017). The authors found that other symptoms identified as risk factors for suicidality, like hopelessness, loneliness, and burdensomeness, were limited in predicting suicidal thoughts. These symptoms did also show high fluctuation. These intricate findings based on repeated measures illustrate how challenging it is to measure accurately the temporal dynamics of depressive symptoms. Note that these works are all based on clinical samples, to our knowledge no longitudinal study has yet addressed depressive symptom variability in the general population.

Overall, a question to further clarify **is whether suicidality is simply more stable over long periods of time, or has an additional role in reinforcing the other symptoms**, as DirectLiNGAM suggested. Kleiman et al. (2018) reached a similar point in an ecological momentary assessment study, in which participants who had recently attempted suicide experienced downward shifts in negative emotion after reporting suicidal thoughts. The authors hypothesize that suicidal thoughts may be reinforcing, and thereby tend to become more persistent. Experimental studies on cognitive reappraisal, a technique for cognitive reframing of emotional cues, have conversely found that applying reappraisal reduces self-reported negative emotions and the activity of the amygdala, a key element in emotional processing in the brain (McRae, 2016). The evidence on cognitive reappraisal is ample for healthy and depressed participants alike. In particular, some recent empirical evidence points at the protective effect of cognitive reappraisal against suicidal behavior (Forkmann et al., 2014; Kudinova et al., 2016; Ong & Thompson, 2019; Richmond, Hasking, & Meaney, 2017).

The **somatic symptoms we analyzed shed mixed results**. Very prevalent symptoms, like difficulties with sleep and fatigue, showed mostly no independent associations with severe self-reported impairment. Appetite and weight changes registered somewhat inconsistent pattern across datasets and analyses. Concentration difficulties emerged as the most strongly related to functional impairment through different analyses, and showed an additional interaction effect with age group in Study II. The impairing effects of concentration difficulties are well documented in studies of clinical depression for its link to cognitive impairment. A recent review examining the clinical relationship between cognitive impairment and psychosocial functioning in patients of MDD highlighted that "cognitive deficits are associated with short-term and longitudinal functional deficits", and that "older age and greater illness severity appear to increase susceptibility to and magnitude of psychosocial deficits in MDD" (Cambridge, Knight, Mills, & Baune, 2018, p. 170).

It is possible that the associations of somatic symptoms with functioning vary across particular clinical samples. Differences over the severity continuum of depression could help to disentangle the mixed results found for psychosomatic symptoms (Thombs et al., 2010), which undoubtedly belong to affective disorders (Simon, VonKorff, Piccinelli, Fullerton, & Ormel, 1999). The inconclusive pattern could also stem from the intrinsic challenge of research in psychosomatic symptoms. Somatic symptoms tend to be frequent in the general population, and stem from multiple causes including physical and mental illness. The use of population-based data may bring about very noisy measures of somatic symptoms, which in turn results in mixed findings. Such thing does also imply that somatic symptoms are challenging as diagnostic criteria as well, and motivate investing resources for studying their validity and how to measure them more accurately in the context of affective disorders.

Contributions of our work within current mental health

The traditional unidimensional model, based on a latent variable, failed to accommodate the data in the two datasets we utilized. At the same time, we found evidence of symptoms relating differently to subjective functional impairment. Our findings are in the same line with other literature that quite consistently finds arguments against the current "one size fits all" approach. This is a concern from the perspective of validity of the operational definition of MD. First, symptom presentation patterns; and second, their specific covariation with clinical validators, are basic for further advancing psychiatric disorders, and particularly for the challenges of MD as a diagnostic entity.

The fundamental argument for bottom-up approaches in research is the **clarity of handling individual symptoms**, as compared to handling depression scores. Depression aggregates may be insightful, but could hide important information more accessible with symptom-level approaches. This would ease reporting, comparing, and reviewing findings (Costello, 1993). Second, the use of clinically relevant criteria is important in mental health because most psychiatric disorders do not have natural boundaries nor stablished etiopathogenesis. A "good" symptom is not defined by biomarkers, but it is to prove being useful based on criteria defined by experts, such as informativeness and discrimination of psychopathological states. This knowledge on symptoms would bring useful output for psychopathology at no cost, with potential for theoretical and clinical developments. We consider three wide domains in which the approaches we have presented have a potential contribution.

The first one is the **evaluation and refinement of current diagnostic systems**. The diagnostic definition of MD, as most of the psychiatric diagnoses is not evidence-based, but the result of consensus among experts, tradition, and historical conjunctures (e.g. Kendler, 2016; Maj, 2012). The current diagnostic definition of depression has changed very little since the Feighner criteria were published in 1972 (Feighner et al., 1972). The Feighner criteria were the basis for developing the Research Diagnostic Criteria, which were in turn central to DSM-III. Thus, the Feighner criteria had a tremendous influence on the systematization of mental illness, an influence that remains alive today. A main goal that the authors of the Feighner criteria had in common was their concern about validity and the evidential support of psychiatric diagnosis (Kendler, Muñoz, & Murphy, 2010). Their motivation was promoting systematic and empirically-oriented psychiatric diagnosis. In what comes to depression, they reportedly looked into epidemiological surveys and were particularly influenced by an empirical article authored by a group of psychiatrists interested in comparative clinical descriptions (Cassidy et al., 1957). To the best of our knowledge, the specific process was never documented (Kendler et al., 2010).

Despite generating abundant research, the criteria themselves have undergone little change. We have accumulated plenty of empirical knowledge and the needs of the field have changed, yet we hardly ever question the validity of these criteria. They are accepted on the fly and used as a cornerstone for research. If revisions were to happen, this would call for a robust symptom-based body of literature to inform any decisions. The revisions undertaken in DSM-5 for the Substance Use Disorder are a good example of this. The current category combines the old diagnoses of substance use and substance dependence. The symptom descriptions were strengthened, the symptom *craving* was added, and the

symptom *legal problems* was excluded. The modifications were largely based on symptom-level considerations and research, and in the pursue of improving the validity of the diagnostic definitions (Hasin et al., 2013). This means, in practice, the aim of **making the diagnostic definitions more coherent with empirical data, external validators, and current symptomatic representations**. Changes of this magnitude have not yet taken place for MD. This is logical, because depression seems rather complex etiologically, as it is demonstrated by extensive literature with somewhat inconclusive findings and by our own research results (e.g. in Study I the concurrent validity of all other disorders was well approximated by severity estimates, and the IRT models functioned acceptably). Anyways, for the field to be prepared for any future changes, solid empirical evidence on the validity of individual diagnostic criteria would be required.

The second domain of potential application is **clinical practice**. It is not new for the field that clinical frameworks depart from a symptom-level basis. For instance, the cognitive theories of depression posit symptom-to-symptom mechanisms and stablish a workflow at the ground level: cognitive biases (through working memory, perception, reward processing, and thought) reinforce emotion dysregulation. Because resources are limited and depression is very burdensome, those intervention programs that minimize the impact of depression are particularly valuable to patients, to their families, and to society. This is why the role of external outcomes, like functional impairment, can be key in profiling the most influential symptom presentations. Clinical practices may become more evidence-based upon producing further empirical research on symptoms and their impact on functioning.

In contrast with the relevance of standardization for academics, it has been shown that professionals working in clinical psychology and psychiatry do not follow strictly standard guidelines, but mostly rely on their own experience when selecting a treatment (First et al., 2018). Kamenov at al. (2017) conducted a multi-informant study about treatment effectiveness in depression. They implemented a literature review, an expert survey, and interviewed patients. According to their results, current literature differs from clinicians in that, while research emphasizes the role of symptom aggregates (scores) in severity and recovery, clinicians highlight the importance of a variety of functioning domains and personal factors (e.g. self-concept, self-efficacy) in the treatment and recovery from depression. Moreover, patients and clinicians pointed at certain areas targeted by interventions that are not commonly addressed by the literature, like interpersonal relationships, problems in communication, or lack of social participation – areas among the most affected by depression according

to the authors (Kamenov et al., 2016). Kamenov and colleagues argued in favor of adapting research to clinical reality, because "a higher sum-score might mean a higher number of less affected functioning areas or a smaller number of domains with marked deterioration" (2017, p. 7).

The third domain is the **development of etiological models and theories.** Some examples of bottom-up developments have already taken place, driven by current research programs as the HiTOP and network approaches (Borsboom, 2017; Kotov et al., 2018). The HiTOP (Hierarchical Taxonomy Of Psychopathology) argues that most of the problems with current diagnostic taxonomies happen because the disorder classifications do not match empirical reality. There is a high degree of comorbidity among certain disorders, besides a lack of specificity of risk factors and correlates. Both may be explained by the fact that current diagnostic categories can be mapped within wider spectra, spectra that explains their covariation through a dependence hierarchy. They consider that signs and symptoms are the starting building block, from which to continue towards higher levels in the hierarchy (e.g. internalizing and externalizing factors, which group multiple disorders). The idea, as Markon put it, is to develop "model frameworks 'downward' as well as 'outward', by analyzing symptoms rather than diagnoses, and by integrating symptoms from Axis I and II disorders in a common framework." (2010, p. 273). On its behalf, the network approach to psychopathology has emphasized the role of direct symptom-to-symptom dynamics in the onset and maintenance of psychopathology. The role of symptoms is principal in the theory, which is characterized by statistical models based on only observed variables (symptom), although latent variables in network models could be implemented as well.

Another potential application is for etiological models that benefit from distribution-based casual inference. Distribution-based inferences point at the most likely direction of causation between pairs (or groups) of variables, under certain assumptions. These methods are a suitable addition to more general studies of **triangulation in etiologic epidemiology** (Lawlor, Tilling, & Smith, 2017). The principle behind triangulation is that an etiological model may be inferred for being robust across multiple methods that rely in non-overlapping assumptions. Thus, we may derive direction of causation upon an inference that holds across estimators, which make different key assumptions and have differing limitations to suit the observed data. Several methods used for similar purposes may be used in combination (Rosenström & García-Velázquez, in press). Direction of causation models, from behavioral genetics, use family data, assume normally-distributed variables, and estimate additive genetic, shared and non-shared environmental influences to determine of A on B (Heath et al., 1993). LiNGAM-based algorithms,

differently, require non-Gaussian distributions and use higher moments to determine direction of dependence between A and B (Shimizu, 2018). Methods estimating direction of dependence are an untapped resource for generating testable hypotheses and supporting insights from variety of methods in current psychopathology.

Methodological considerations and future directions

Population-based, cross-sectional data

We used representative, cross-sectional samples of the adult population of the United States. The breadth of the studies allowed to take into account a number of relevant covariates for, such as number of physical diseases, and to compute nationally representative estimates. Studying the epidemiology of psychopathology is essential for better characterizing pathological symptom presentations, since the general population offers a reference of the natural distribution of psychopathology. Psychopathology is a matter of degree, and therefore mapping the whole continuum of severity is important for advancing the understanding of common disorders. Moreover, general population studies cover a wider spectrum of depressive symptomatology, from spare to severe symptom presentations, as compared to clinical samples which comprise more developed forms of the syndrome. Thus, representativeness and size of the datasets are a strength of this dissertation work. At the same time, we identify a number of limitations in our data.

First, both studies are limited to the adult population of the United States, and thus the interpretation of our findings is not warranted for other populations. It is well known that depressive symptoms present differently across cultures (De Vaus, Hornsey, Kuppens, & Bastian, 2018; Ryder et al., 2008). Most research in mental health focuses on high-income countries, although the burden of depression and suicide are as high or higher in low- and middle-income areas (Guzmán, Cha, Ribeiro, & Franklin, 2019; Liu et al., 2019; WHO, 2017). Additionally, community samples may have different symptom presentation than clinical samples in terms of prevalence and structure as well (Foster & Mohler-Kuo, 2018). This entails that our results are derived from those at risk (in the case of CPES data) or for the general population (NHANES), and may not hold in strictly clinical populations, for which it would be necessary to carry out specific research. There were, however, some similarities between our findings and those based on clinical samples presented by Faravelli et al. (1996), Fried and Nesse (2014), and Tweed (1993). In terms of representativeness, it would be convenient to systematically compare

structural estimates and patterns of symptom associations with impairment across different samples as populations.

Another common feature in both studies is cross-sectional design. Cross-sectional data allows between-individual comparisons, but not to examine intra-individual symptom patterns across time, nor symptom changes. With respect to the implications of time, longitudinal studies and a wider time-span are needed to ascertain whether the patterns observed in Study II were actually age-dependent, and to disentangle possible cohort effects. Moreover, longitudinal data may support to some extent causal inference and would have conveniently complemented the analyses we conducted, for example, in the case of LiNGAM analyses in Study III. Our interpretation of the direction of dependence dynamics could be tested in a longitudinal context, in which temporal antecedence together with dependence among symptoms with each other and with functional impairment can be studied in more detail. The use of cross-sectional data is incomplete when aiming to study comprehensively the heterogeneity of mental disorders, which takes place across individuals and symptoms, but also over time (de Vos, Wardenaar, Bos, Wit, & de Jonge, 2015; Wardenaar & de Jonge, 2013).

Measurement of depressive symptoms

It is worth to consider the characteristics of the instruments used in the CPES and NHANES studies for measuring depressive symptoms. First, there are a number of characteristics in particular that these measurements do not have in common, and which are worth to take into account when interpreting our results.

The WHO-CIDI interview used in the CPES studies are based on a skip-question system. In the case of the depression module, the participant had to endorse having suffered an affective episode in order to administer the rest of the interview. Although most depressive episodes include mood symptoms or lack of interest (Zimmerman, McGlinchey, et al., 2006b), such filter-based design still forces a given dependence in the data and likely inflates the prevalence of the filter symptoms relative to the others. Second, the items concern the *most severe* depressive episode, and not lifetime symptom endorsement patterns. This conveys three problems. First, the most severe episode or a person is not necessarily representative of their depressive syndrome, and makes difficult to extract generalizable information about depression. On the other hand, ratings of the worst episode are most informative of the impairment depression may exert to depressed, non-clinical individuals. In this sense the association between the episode and impairment would be captured better than queries about current symptoms. Second, it is

likely that a worst episode joins together more symptoms, and therefore symptoms' joint (e.g. bivariate) occurrence patterns may be inflated too. Third, the worst episode could have happened decades ago or be still ongoing, which may have an influence in the reliability of recall. This aspect was not taken into account in our analyses. The responses to the interview were dichotomous, which in this case may decrease the noise of recall (i.e. it is easier to remember if something happened than its frequency or intensity). However, a drawback of dichotomous responses is the loss of information on the magnitude of symptom presentation, as compared with Likert-type response scales.

The PHQ-9 questionnaire (Kroenke & Spitzer, 2002) used in the NHANES study is a short form querying how often the participant had been suffering from depressive symptoms in the last two weeks. In this case no skip-question systems apply, nor loss of information due to dichotomization or difficulties in recall. However, the PHQ-9 questionnaire is not specific for affective episodes. As such, a person suffering, for instance, from sleep problems for any other reason (e.g. physical health, caring of small children, or irregular work shifts) may endorse the symptom. Also psychomotor disturbances may be a result of thyroid dysfunction instead of depression. The scale does not stablish clearly that the queries relate to an affective episode. This implies that the occurrence of the symptoms is not strictly attributable to a dysphoric episode, and that the prevalence rates may be biased to some extent with respect to those strictly caused by depressive manifestations. PHQ-9 is suitable for approximating how frequent the symptoms are in an epidemiological sense, but not so much about how frequent depression may be in the population – needless to state that diagnosing depression requires a thorough evaluation entailing much more information than meeting the thresholds of duration and symptom amount. Another important limitation of this questionnaire is using compound measures of symptoms. When querying about lack of sleep and oversleep at the same time, it is impossible to discriminate differential prevalence or patterns with associations with other symptoms or outcomes. These compound symptoms may be similar (e.g. worthlessness and guilt) or just opposite (e.g. increase and decrease in weight), but studying them separately and examining their validity as symptoms is not possible in these cases (Fried, 2015, 2017b).

The aforementioned dissimilarities across the depression instruments may contribute to some of the differences found in symptom-impairment associations. At the same time, the results are not directly comparable between them because of the different nuances of the instruments (i.e. *worst* episode against epidemiological screening of symptoms). There is one characteristic that both measures, the WHO-CIDI and PHQ-9, have in common. The skip mechanism in the WHO-CIDI limits the amount of symptoms

evaluated other than those of DSM-IV (i.e. in order to assess the ICD diagnostic symptoms, the respondent has to previously endorse a number and type of DSM-IV diagnostic symptoms). As to the PHQ-9, its content strictly corresponds to the DSM-IV criteria for MDE in a compound manner (i.e. two-week period). It is a strength of this thesis dissertation that the results conform to the DSM-IV symptom criteria, an internationally accepted taxonomy. It is unarguable that the replicability of our results, and their interpretation into the context of previous literature is facilitated because of studying the stablished diagnostic symptoms of MD.

Despite of its numerous benefits, the practice of translating mental disorders into operational terms sometimes involves an oversimplification of psychopathology (Maj, 1998; Parnas & Bovet, 2014). The extensive use of these operational criteria implies the risk that, at some point, research and clinical practice are lured away from considering other forms of symptomatic presentation than the diagnostic criteria in diagnostic manuals. Some of the factors underlying the shift away from the phenomenological tradition are, for instance, that grant-making bodies, scientific journals, or organizers of scientific meetings demand the use of the DSM criteria. In some countries in more applied settings, the diagnostic criteria are understood as the standard of experts' consensus, and therefore any other public institution, insurance company, or teaching syllabi tend to comply with them. Treatments and guidelines are developed using these criteria as cornerstone of what is a prototypic mental disorder (Hyman, 2010). It is more efficient, in sum, to use the official diagnostic criteria than anything else, which may be less cost-effective or subject to debate. This trend has lead through the decades to what experts call *reification* of mental disorders (Andreasen, 2007; Hyman, 2007, 2010; Kendler, 2016; Patten, 2015; van Loo & Romeijn, 2018). A drawback to the exclusive focus on symptom criteria is neglecting other symptoms which maybe be relevant for the disorder. Their implications for the onset, development, or prognosis of depression could remain poorly recognized, behind the emphasis on the standard diagnostic criteria.

The third edition of the DSM brought an enormous improvement to the reliability crisis of psychopathology, through the systematization of diagnostic descriptions with operational definitions. After some time, a far-reaching and unintended pitfall of this effective systematization is the tendency to conflate the DSM criteria with the disorders themselves (Andreasen, 2007). Not only the diagnostic criteria are assumed to be correct, but are assumed to *constitute* the disorder. According to Kendler, "we have confused an index of a thing with the thing itself" (2016, p. 777). On the contrary, depression has been traditionally described with a wealth of symptomatic manifestations, some of them still present in

scales (Fried, 2017a; Santor et al., 2006). In a thorough review of the phenomenology of depression across the last century, Kendler (2016) argued that a great amount of these symptoms did not become diagnostic criteria for reasons such as being too subtle, not succinct, time-consuming to evaluate, or lack specificity. The phenomenology of depression is much richer than the DSM criteria, involving areas as volition, metabolism, detachment experiences, and a variety of physical symptoms not covered partially or entirely in the official diagnostic definitions.

Thus, it is both a limitation and a strength of this thesis dissertation to be restricted to the DSM symptomatic criteria. A natural progression of symptom approaches to psychopathology is to broaden the symptomatic presentations studied, and to examine the performance of the current symptom criteria as compared to other symptoms which are currently ignored by the diagnostic manuals. Current research programs, such as HiTOP and the network approach, do benefit from a bottom-up paradigm for advancing knowledge on topics like comorbidity and the structure of mental disorders.

Comorbidity and other psychological factors

Comorbidity was taken into account only partially in this dissertation. It is common that depressive symptoms co-occur with other mental syndromes (Rush et al., 2005; van Loo, Schoevers, Kendler, de Jonge, & Romeijn, 2016). The symptoms reported in the studies may be due to, or aggravated by other mental syndromes. This was not controlled for in our analyses and brings potential risk of confounded findings. There is evidence of cross-diagnostic subtypes relating symptoms of anxiety and depression to disability (Wanders et al., 2016). The sub-types, found in a large Dutch cohort, include low severity classes with predominant cognitive ("worried") and somatic symptomatology, and subclinical and clinical classes showing a mixture of anxious and depressed presentations. The cross-diagnostic sub-types showed differential associations to clinical validators, including disability, psychiatric diagnoses, and physical health status.

There are other psychological phenomena that may play an important role in the relationship between depressive symptomatology and self-rated functioning. For instance traits such as self-efficacy, resilience, or neuroticism may moderate the actual or perceived impact of depressive symptoms on functioning. These variables may be particularly important to study in epidemiological surveys of depression because of the large proportion of non-clinical presentations. The psychological risk factors may emerge more clearly in analyses comprising the wide severity continuum of depression.

Physical medical conditions were adjusted for, although they were in both studies self-reported. Self-reported medical data may be less reliable than that extracted from registers. Further studies shall examine the role of comorbid symptomatology in the association between depressive symptoms and functioning.

Measurement of functional impairment

In this dissertation we used different measurements of functional impairment. This is a strength, since the use of several indicators reduces the impact of systematic measurement error, particularly in single item measurement. Yet the differences between the indicators may contribute to the differences found in symptom patterns. There are several methodological considerations worth to discuss. First, there were two generic measurements of inability to carry out normal life, the WHODAS scale and the number of days disabled in the last month. These indicators are not solely attributable to depression and thus are not reliable indicators of depressive symptoms' association with functioning. The two variables are self-reported, but a positive aspect is that the information they collect is rather straightforward to rate (i.e. they do not require much subjective judgement). Second, the studies also included indicators of functional depression which were depression-specific. These are more valid indicators in terms of interpretation of results, since they reduce by design confounding effects of comorbid disorders. This is particularly beneficial in community samples which are rather heterogeneous (i.e., general population samples may present important comorbidity with other disorders, which is reduced in clinical samples through exclusion criteria). However, these indicators share the limitation of being single items, and to some extent subjective to rate (e.g. participants may differ in what they consider high interference of symptoms).

The measurement of functional impairment implies a number of conceptual considerations. First, functional impairment is challenging to define or measure for being a complex construct, with no gold standard in mental health. McKnight and Kashdan (2009) defined three different domains in which depression could affect functioning: social, occupational, and physical. The indicators we used here were mixed in terms of what domains they measure, and sometimes left the participant the choice of what to consider as impairment or interference. For example, the PHQ-9 item and the interference item in CPES were clear in referring simultaneously to the three domains defined by McKnight and Kashdan, while it remains unclear what response process the participant followed. For example, whether all the functioning domains had the same weight on the participants' consideration, and whether the rating was based on

averaged perception of burden, or on the burden attributed to the most impaired domain. The item measuring the number of days "unable to work or carry out your normal activities" is very ambiguous regarding what normal activities are. Thus, the use of single items that collapse so complex information presents important methodological and conceptual challenges.

Future studies would benefit from more clear content and wording of the items measuring different domains of functioning, especially since there is some evidence that different symptoms impact different domains of life (Fried & Nesse, 2014). There is no standard scale at hand for measuring functional impairment of MD out of the field of effectiveness of interventions. Experts argue against the reductionist approach of effectiveness studies, which mostly focus on the alleviation of symptoms (through scores) while neglecting other important areas of impact that are themselves relevant for the patients (Kamenov et al., 2017; Lam, Parikh, Michalak, Dewa, & Kennedy, 2015; Zimmerman et al., 2012). The research gap seen in studies of intervention effectiveness is now well recognized, while there is a lack of studies addressing the impairment associated with MD symptoms in naturalistic settings.

Another important consideration is the directionality of effects. It is logical to assume that symptoms cause burden, while it is simultaneously possible that pre-morbid impairment is a risk factor for developing a given syndrome. The study by Bos et al. (2018) concluded that it is pre-existing functioning which predicts future depressive episodes, instead of residual symptoms that cause impairment by lingering. Naturally, subclinical symptoms may cause such impairment in the first place. The relationship between symptoms and impairment is rather complex, and thus we shall limit our interpretations to associations instead of causal directions.

A third aspect to bear in mind are the consequences of reification for the measurement of functioning. Current impairment associated with depressive symptoms may be due to unreported symptoms, apart from those modeled here. These missing symptoms contribute to model error and may cause misspecification (Tweed, 1993). Related to measurement is also the response bias resulting from the tendency towards negativity of currently depressed individuals, which has been argued to affect rating of symptom and functioning (Morgado, Smith, Lecrubier, & Widlöcher, 1991).

There is no consensus or standard on what is impairment as caused by psychiatric disorders in a general sense, neither how to operationalize it. Experts have argued in the last years about its practical consequences and stated the need for a clearer conceptualization. For example, Mario Maj (2012, p. 535) stated with respect to the current diagnostic systems, that "an attempt to operationalize in more objective

terms the impairment criterion appears timely, since that criterion, in its current formulation, depends too much on the clinician's and the patient's subjective judgement and is so generic to be actually redundant".

Conclusions and practical implications

There are some overarching findings of this doctoral dissertation, which are worth to highlight. Our results align with other empirical studies indicating that functional impairment is a core component of disorder severity, not accurately approximated merely by symptoms (Lux et al., 2010; Zimmerman, Balling, Chelminski, & Dalrymple, 2018; Zimmerman, Morgan, et al., 2018). Moreover, a symptom being more or less prevalent did not imply correlate of more or less severe functional impairment. It became apparent that the common practice of using sum-scores has the benefit of approximating relatively well individual symptoms, while helpful insights may be obscured about how concrete symptoms relate to self-rated impairment. This is a critical disadvantage to aggregate-based approaches, especially in what comes to advancement of etiological aspects of the disorder.

Knowledge at the symptom level seems conceptually more useful for understanding of depression than aggregates. For instance, symptom-level analyses revealed that age group moderated the associations of three symptoms with self-rated impairment. This information was concealed behind the sum-score. If our findings replicate across studies, this would imply that middle-aged adults are more prone to consider certain symptoms as severely impairing, as compared to adults aged 30 and below and in retirement age. These symptoms are depressed mood, self-criticism, and impaired concentration. Being resources limited, a practical implication of this is that prevention plans shall consider middle-aged adults are particularly vulnerable to certain depressive symptoms, as compared to other age groups.

The associations of cognitive-affective symptoms among themselves and with functional impairment emerged distinctively. However, most associations between depressive symptoms and severe functional impairment turned out mostly unspecific, and therefore the clinical relevance of individual symptoms upon functioning calls for further exploration. There is compelling empirical evidence that functioning is a relevant factor for both the onset and remission of depressive disorder (Bos et al., 2018; McKnight & Kashdan, 2009). If our results hold across studies, a concrete output of this doctoral dissertation is that cognitive symptoms reinforce affective-motivational ones, which is a vulnerability pathway for its consistent associations with severe functional impairment. A practical implication is that supporting individuals' self-worth may protect against the development of depressive symptomatology and its corresponding impact on functioning.

The studies in this doctoral dissertation shed light on the complex links between symptoms and functional impairment. Yet, the convoluted etiology of MD prevailed across analyses. Many aspects remained for further exploration and replication, such as the moderating effect of age, the dependence associations between cognitive-affective symptoms, or the inconsistent link of some symptoms with severe impairment. An important source of information, such as within-subject dynamics, was not explored. Our findings motivate considering a wider range of symptoms, both in terms of presentation (e.g., from mild to severe and longitudinally) and content, for further understanding the heterogeneity of depression. Similarly, functional impairment as a core component of severity calls for systematic exploration, and for a more refined measurement (Lam, Filteau, & Milev, 2011; McKnight & Kashdan, 2009; Zimmerman et al., 2008). The use of multi-item scales and instruments validated specifically for use in depression come as useful tools in the endeavor of better describing depression and its outcomes.

A more fine-grained analysis is undoubtedly beneficial in a clinical context (i.e. efficacy of treatment and remission), but also in epidemiological studies. It is necessary to reach systematic knowledge on the natural distributions of depressive symptoms and their links to clinically relevant correlates. A better characterization of the severity of depression as a continuum is fundamental for theoretical developments in psychopathology, and potentially useful for planning more efficient interventions targeting the most disabling, dominant symptoms.

References

Aaltonen, K., Näätänen, P., Heikkinen, M., Koivisto, M., Baryshnikov, I., Karpov, B., … Isometsä, E. (2016). Differences and similarities of risk factors for suicidal ideation and attempts among patients with depressive or bipolar disorders. *Journal of Affective Disorders*, *193*, 318–330. https://doi.org/10.1016/j.jad.2015.12.033

Alegria, M., Jackson, J. S., Kessler, R. C., & Takeuchi, D. (2015). *Collaborative Psychiatric Epidemiology Surveys (CPES), 2001-2003, United States*, pp. 12–19. Retrieved from http://doi.org/10.3886/ICPSR20240.v8

Alkhalaf, A., & Zumbo, B. D. (2017). The Impact of Predictor Variable(s) with Skewed Cell Probabilities on Wald Tests in Binary Logistic Regression. *Journal of Modern Applied Statistical Methods*, *16*(2), 40–80. https://doi.org/10.22237/jmasm/1509494640

Alonso, J., Vilagut, G., Chatterji, S., Heeringa, S., Schoenbaum, M., Üstün, T. B., … Kessler, R. C. (2011). Including information about comorbidity in estimates of disease burden: Results from the WHO World Mental Health Surveys. *Psychological Medicine*, *41*(4), 873–886. https://doi.org/10.1017/S0033291710001212

American Psychiatric Association. (2000). *Diagnostic and statistical manual-text revision (DSM-IV-TR)*. American Psychiatric Association.

American Psychiatric Association. (2010). *Practice guideline for the treatment of patients with major depressive disorder* (3rd ed.). Washington: American Psychiatric Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Andreasen, N. C. (2007). DSM and the Death of Phenomenology in America: An Example of Unintended Consequences. *Schizophrenia Bulletin*, *33*(1), 108. https://doi.org/10.1093/SCHBUL/SBL054

Arnow, B. A., Blasey, C., Williams, L. M., Palmer, D. M., Rekshan, W., Schatzberg, A. F., … Rush, A. J. (2015). Depression Subtypes in Predicting Antidepressant Response: A Report From the iSPOT-D Trial. *American Journal of Psychiatry*, *172*(8), 743–750. https://doi.org/10.1176/appi.ajp.2015.14020181

Barth, J., Munder, T., Gerger, H., Nüesch, E., Trelle, S., Znoj, H., … Cuijpers, P. (2013). Comparative Efficacy of Seven Psychotherapeutic Interventions for Patients with Depression: A Network Meta-Analysis. *PLOS Medicine*, *10*(5), e1001454. https://doi.org/10.1371/journal.pmed.1001454

Beals, J., Novins, D. K., Spicer, P., Orton, H. D., Mitchell, C. M., Barón, A. E., & Manson, S. M. (2004). Challenges in Operationalizing the DSM-IV ClinicalSignificance Criterion. *Archives of General Psychiatry*, *61*(12), 1197. https://doi.org/10.1001/archpsyc.61.12.1197

Bollen, K. A. (1989). Structural equations with latent variables. *Wiley Series in Probability and Mathematical Statistics*, *8*, 528. https://doi.org/10.1007/s13398-014-0173-7.2

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods*, *16*(3), 265–284. https://doi.org/10.1037/a0024448

Bollen, K. A., & Ting, K. (1998). Bootstrapping a Test Statistic for Vanishing Tetrads. *Sociological Methods & Research*, *27*(1), 77–102. https://doi.org/10.1177/0049124198027001002

Bollen, K. A., & Ting, K. (1993). Confirmatory tetrad analysis. *Sociological Methodology*, *23*, 147–175.

Bolton, J. M., Belik, S.-L., Enns, M. W., Cox, B. J., & Sareen, J. (2008). Exploring the correlates of suicide attempts among individuals with major depressive disorder: findings from the national epidemiologic survey on alcohol and related conditions. *The Journal of Clinical Psychiatry*, *69*(7), 1139–1149.

Bolton, J. M., Pagura, J., Enns, M. W., Grant, B., & Sareen, J. (2010). A population-based longitudinal study of risk factors for suicide attempts in major depressive disorder. *Journal of Psychiatric Research*, *44*(13), 817–826. https://doi.org/10.1016/j.jpsychires.2010.01.003

Border, R., Johnson, E. C., Evans, L. M., Smolen, A., Berley, N., Sullivan, P. F., & Keller, M. C. (2019). No Support for Historical Candidate Gene or Candidate Gene-by-Interaction Hypotheses for Major Depression Across Multiple Large Samples. *American Journal of Psychiatry*, *176*(5), 376–387. https://doi.org/10.1176/appi.ajp.2018.18070881

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13.

Borsboom, D, Rhemtulla, M., Cramer, A. O. J., van der Maas, H. L. J., Scheffer, M., & Dolan, C. V. (2016). Kinds versus continua: a review of psychometric approaches to uncover the structure of psychiatric constructs. *Psychological Medicine*, *46*(8), 1567.

Bos, E. H., ten Have, M., van Dorsselaer, S., Jeronimus, B. F., de Graaf, R., & de Jonge, P. (2018). Functioning before and after a major depressive episode: pre-existing vulnerability or scar? A prospective three-wave population-based study. *Psychological Medicine*, *48*(13), 2264–2272. https://doi.org/10.1017/S0033291717003798

Boschloo, L., Bekhuis, E., Weitz, E. S., Reijnders, M., DeRubeis, R. J., Dimidjian, S., … Cuijpers, P. (2019).

The symptom-specific efficacy of antidepressant medication vs. cognitive behavioral therapy in the treatment of depression: results from an individual patient data meta-analysis. *World Psychiatry*, *18*(2), 183–191. https://doi.org/10.1002/wps.20630

Bringmann, L. F., Lemmens, L. H. J. M., Huibers, M. J. H., Borsboom, D., & Tuerlinckx, F. (2015). Revealing the dynamic network structure of the Beck Depression Inventory-II. *Psychological Medicine*, *45*(04), 747–757. https://doi.org/10.1017/S0033291714001809

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (Second). Retrieved from https://books.google.fi/books?hl=en&lr=&id=JDb3BQAAQBAJ&oi=fnd&pg=PP1&dq=confirmatory+factor+analysis+brown&ots=-0FwhV4HZi&sig=P8t3DzPWg_zz52-Omc70efhRypQ&redir_esc=y#v=onepage&q=confirmatory factor analysis brown&f=false

Burnham, K. P., & Anderson, D. R. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd ed). In *Ecological Modelling* (Vol. 172). https://doi.org/10.1016/j.ecolmodel.2003.11.004

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*(4), 581–612.

Cambridge, O. R., Knight, M. J., Mills, N., & Baune, B. T. (2018). The clinical relationship between cognitive impairment and psychosocial functioning in major depressive disorder: A systematic review. *Psychiatry Research*, *269*, 157–171. https://doi.org/10.1016/J.PSYCHRES.2018.08.033

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., … Moffitt, T. E. (2014). The p Factor: One General Psychopathology Factor in the Structure of Psychiatric Disorders? *Clinical Psychological Science : A Journal of the Association for Psychological Science*, *2*(2), 119–137. https://doi.org/10.1177/2167702613497473

Cassidy, W., Flanagan, N., Spellman, M., & Cohen, M. (1957). Clinical observations in manic-depressive disease: a quantitative study of one hundred manic-depressive patients and fifty medically sick controls. *JAMA*, *164*, 1535–1546.

Centers for Disease Control and Prevention. (n.d.). NHANES Survey Methods and Analytic Guidelines. Retrieved January 23, 2019, from National Health and Nutrition Examination Survey: Sample Design website: https://wwwn.cdc.gov/nchs/nhanes/analyticguidelines.aspx#sample-design

Centers for Disease Control and Prevention. (2017). *National Health and Nutrition Examination Survey Data (NHANES)*. Retrieved from https://www.cdc.gov/nchs/nhanes/about_nhanes.htm

Chalmers, R. P. (2012). {mirt}: A Multidimensional Item Response Theory Package for the {R} Environment. *Journal of Statistical Software*, *48*(6), 1–29. Retrieved from http://www.jstatsoft.org/v48/i06/

Choi, S. W. (2016). *lordif: Logistic Ordinal Regression Differential Item Functioning Using IRT*. Retrieved from https://cran.r-project.org/package=lordif

Cipriani, A., Furukawa, T. A., Salanti, G., Chaimani, A., Atkinson, L. Z., Ogawa, Y., … Geddes, J. R. (2018). Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *The Lancet*, *391*(10128), 1357–1366. https://doi.org/10.1016/S0140-6736(17)32802-7

Cooper, R. V. (2013). Avoiding False Positives: Zones of Rarity, the Threshold Problem, and the DSM Clinical Significance Criterion. *Canadian Journal of Psychiatry*, *58*(11), 606–611. Retrieved from www.TheCJP.ca

Costello, C. G. (1993). The Advantages of the Symptom Approach to Depression. In C. G. Costello (Ed.), *Symptoms of Depression* (pp. 1–22). Wiley.

Cramer, A. O. J., Borsboom, D., Aggen, S. H., & Kendler, K. S. (2012). The pathoplasticity of dysphoric episodes: differential impact of stressful life events on the pattern of depressive symptom inter-correlations. *Psychological Medicine*, *42*(05), 957–965. https://doi.org/10.1017/S003329171100211X

Cramer, Angélique O. J., van Borkulo, C. D., Giltay, E. J., van der Maas, H. L. J., Kendler, K. S., Scheffer, M., & Borsboom, D. (2016). Major Depression as a Complex Dynamic System. *Plos One*, *11*(12), e0167490. https://doi.org/10.1371/journal.pone.0167490

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Cuijpers, P. (2019). Targets and outcomes of psychotherapies for mental disorders: an overview. *World Psychiatry*, *18*(3), 276–285. https://doi.org/10.1002/wps.20661

Cuijpers, P., & Smit, F. (2002). Excess mortality in depression: a meta-analysis of community studies. *Journal of Affective Disorders*, *72*(3), 227–236. https://doi.org/10.1016/S0165-0327(01)00413-X

Cuijpers, P., Vogelzangs, N., Twisk, J., Kleiboer, A., Li, J., & Penninx, B. W. (2014). Comprehensive Meta-Analysis of Excess Mortality in Depression in the General Community Versus Patients With Specific Illnesses. *American Journal of Psychiatry*, *171*(4), 453–462. https://doi.org/10.1176/appi.ajp.2013.13030325

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

De Vaus, J., Hornsey, M. J., Kuppens, P., & Bastian, B. (2018). Exploring the East-West Divide in Prevalence of Affective Disorder: A Case for Cultural Differences in Coping With Negative Emotion. *Personality and Social Psychology Review*, *22*(3), 285–304. https://doi.org/10.1177/1088868317736222

de Vos, S., Wardenaar, K. J., Bos, E. H., Wit, E. C., & de Jonge, P. (2015). Decomposing the heterogeneity of depression at the person-, symptom-, and time-level: latent variable models versus multimode principal component analysis. *BMC Medical Research Methodology*, *15*(1), 88. https://doi.org/10.1186/s12874-015-0080-4

de Vries, Y. A., Roest, A. M., Bos, E. H., Burgerhof, J. G. M., van Loo, H. M., & de Jonge, P. (2018). Predicting antidepressant response by monitoring early improvement of individual symptoms of depression: individual patient data meta-analysis. *The British Journal of Psychiatry*, 1–7. https://doi.org/10.1192/bjp.2018.122

Disease and Injury Incidence and Prevalence Collaborators. (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, *388*(10053), 1545–1602. https://doi.org/10.1016/S0140-6736(16)31678-6

Duncan-Jones, P., Grayson, D. A., & Moran, P. A. P. (1986). The utility of latent trait models in psychiatric epidemiology. *Psychological Medicine*, *16*(02), 391. https://doi.org/10.1017/S0033291700009223

Eaton, W. W., Shao, H., Nestadt, G., Lee, B. H., Bienvenu, O. J., & Zandi, P. (2008). Population-Based Study of First Onset and Chronicity in Major Depressive Disorder. *Archives of General Psychiatry*, *65*(5), 513. https://doi.org/10.1001/archpsyc.65.5.513

Etiopathogenesis Medical Definition. (2018). In *Merriam-Webster Medical Dictionary*. Retrieved from https://www.merriam-webster.com/medical/etiopathogenesis

Faravelli, C., Servi, P., Arends, J. A., & Strik, W. K. (1996). Number of symptoms, quantification, and qualification of depression. *Comprehensive Psychiatry*, *37*(5), 307–315. https://doi.org/10.1016/S0010-440X(96)90011-5

Feighner, J. P., Robins, E., Guze, S. B., Woodruff, R. A., Winokur, G., & Munoz, R. (1972). Diagnostic Criteria for Use in Psychiatric Research. *Archives of General Psychiatry*, *26*(1), 57. https://doi.org/10.1001/archpsyc.1972.01750190059011

Ferrari, A. J., Charlson, F. J., Norman, R. E., Patten, S. B., Freedman, G., Murray, C. J. L., … Whiteford, H. A. (2013). Burden of depressive disorders by country, sex, age, and year: findings from the global burden of

disease study 2010. *PLoS Med*, *10*(11), e1001547.

First, M. B., Rebello, T. J., Keeley, J. W., Bhargava, R., Dai, Y., Kulygina, M., … Reed, G. M. (2018). Do mental health professionals use diagnostic classifications the way we think they do? A global survey. *World Psychiatry*, *17*(2), 187–195. https://doi.org/10.1002/wps.20525

First, M. B., & Wakefield, J. C. (2013). Diagnostic criteria as dysfunction indicators: bridging the chasm between the definition of mental disorder and diagnostic criteria for specific disorders. *The Canadian Journal of Psychiatry*, *58*(12), 663–669.

Forkmann, T., Scherer, A., Böcker, M., Pawelzik, M., Gauggel, S., & Glaesmer, H. (2014). The Relation of Cognitive Reappraisal and Expressive Suppression to Suicidal Ideation and Suicidal Desire. *Suicide and Life-Threatening Behavior*, *44*(5), 524–536. https://doi.org/10.1111/sltb.12076

Foster, S., & Mohler-Kuo, M. (2018). New insights into the correlation structure of DSM-IV depression symptoms in the general population v. subsamples of depressed individuals. *Epidemiology and Psychiatric Sciences*, *27*(03), 288–300. https://doi.org/10.1017/S2045796016001086

Fournier, J. C., DeRubeis, R. J., Hollon, S. D., Dimidjian, S., Amsterdam, J. D., Shelton, R. C., & Fawcett, J. (2010). Antidepressant Drug Effects and Depression Severity. *JAMA*, *303*(1), 47. https://doi.org/10.1001/jama.2009.1943

Fournier, J. C., DeRubeis, R. J., Hollon, S. D., Gallop, R., Shelton, R. C., & Amsterdam, J. D. (2013). Differential change in specific depressive symptoms during antidepressant medication or cognitive therapy. *Behaviour Research and Therapy*, *51*(7), 392–398. https://doi.org/10.1016/J.BRAT.2013.03.010

Frances, A. (2013). The past, present and future of psychiatric diagnosis. *World Psychiatry*, *12*(2), 111–112. https://doi.org/10.1002/wps.20027

Frances, A. (2016). A report card on the utility of psychiatric diagnosis. *World Psychiatry*, *15*(1), 32–33. https://doi.org/10.1002/wps.20285

Franić, S., Dolan, C. V, Borsboom, D., Hudziak, J. J., van Beijsterveldt, C. E. M., & Boomsma, D. I. (2013). Can genetics help psychometrics? Improving dimensionality assessment through genetic factor modeling. *Psychological Methods*, *18*(3), 406.

Fried, E. (2015). Problematic assumptions have slowed down depression research: why symptoms, not syndromes are the way forward. *Frontiers in Psychology*, *6*, 309.

Fried, E. (2017a). The 52 symptoms of major depression: Lack of content overlap among seven common

depression scales. *Journal of Affective Disorders*, *208*, 191–197.

Fried, E.I. (2017b). What are psychological constructs? On the nature and statistical modelling of emotions, intelligence, personality traits and mental disorders. *Health Psychology Review*, *11*(2), 130–134. https://doi.org/10.1080/17437199.2017.1306718

Fried, E., Bockting, C., Arjadi, R., Borsboom, D., Amshoff, M., Cramer, A., … Stroebe, M. (2015). From loss to loneliness : the relationship between bereavement and depressive symptoms. *Journal of Abnormal Psychology*, *124*(2), 256–265. Retrieved from https://biblio.ugent.be/publication/8517055

Fried, E. I., Nesse, R. M., Zivin, K., Guille, C., & Sen, S. (2014). Depression is more than the sum score of its parts: individual DSM symptoms have different risk factors. *Psychological Medicine*, *44*(10), 2067–2076. https://doi.org/10.1017/S0033291713002900

Fried, E.I, & Nesse, R. M. (2014). The impact of individual depressive symptoms on impairment of psychosocial functioning. *PLoS One*, *9*(2), e90311.

Fried, E.I, & Nesse, R. M. (2015a). Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR* D study. *Journal of Affective Disorders*, *172*, 96–102.

Fried, E.I, & Nesse, R. M. (2015b). Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Medicine*, *13*(1), 72.

Fried, E.I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F., & Borsboom, D. (2016). Measuring depression over time... Or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment*, *28*(11), 1354–1367. https://doi.org/10.1037/pas0000275

Gallagher, D., Mhaolain, A. N., Greene, E., Walsh, C., Denihan, A., Bruce, I., … Lawlor, B. A. (2009). Late life depression: a comparison of risk factors and symptoms according to age of onset in community dwelling older adults. *International Journal of Geriatric Psychiatry*, *25*(10), 981–987. https://doi.org/10.1002/gps.2438

Gelin, M. N., & Zumbo, B. D. (2003). Differential Item Functioning Results May Change Depending On How An Item Is Scored: An Illustration With The Center For Epidemiologic Studies Depression Scale. *Educational and Psychological Measurement*, *63*(1), 65–74. https://doi.org/10.1177/0013164402239317

Gilmer, W. S., Trivedi, M. H., Rush, A. J., Wisniewski, S. R., Luther, J., Howland, R. H., … Alpert, J. (2005). Factors associated with chronic depressive episodes: a preliminary report from the STAR-D project. *Acta Psychiatrica Scandinavica*, *112*(6), 425–433. https://doi.org/10.1111/j.1600-0447.2005.00633.x

Goldberg, D. (2011). The heterogeneity of "major depression". *World Psychiatry : Official Journal of the World Psychiatric Association (WPA)*, *10*(3), 226–228. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21991283

Guzmán, E. M., Cha, C. B., Ribeiro, J. D., & Franklin, J. C. (2019). Suicide risk around the world: a meta-analysis of longitudinal studies. *Social Psychiatry and Psychiatric Epidemiology*, 1–12. https://doi.org/10.1007/s00127-019-01759-x

Harald, B., & Gordon, P. (2012). Meta-review of depressive subtyping models. *Journal of Affective Disorders*, *139*(2), 126–140. https://doi.org/10.1016/J.JAD.2011.07.015

Hardeveld, F., Spijker, J., De Graaf, R., Nolen, W. A., & Beekman, A. T. F. (2010). Prevalence and predictors of recurrence of major depressive disorder in the adult population. *Acta Psychiatrica Scandinavica*, *122*(3), 184–191. https://doi.org/10.1111/j.1600-0447.2009.01519.x

Hasin, D. S., O'Brien, C. P., Auriacombe, M., Borges, G., Bucholz, K., Budney, A., … Grant, B. F. (2013). DSM-5 Criteria for Substance Use Disorders: Recommendations and Rationale. *American Journal of Psychiatry*, *170*(8), 834–851. https://doi.org/10.1176/appi.ajp.2013.12060782

Haslam, N., Holland, E., & Kuppens, P. (2012). Categories versus dimensions in personality and psychopathology: a quantitative review of taxometric research. *Psychological Medicine*, *42*(05), 903–920.

Heath, A. C., Kessler, R. C., Neale, M. C., Hewitt, J. K., Eaves, L. J., & Kendler, K. S. (1993). Testing Hypotheses About Direction of Causation Using Cross-Sectional Family Data. *Behavior Genetics*, *23*(1), 29–50. Retrieved from https://pdfs.semanticscholar.org/5551/df22dedfa9668c2e8b0e59ee74244bc1755b.pdf

Heeringa, S. G., Wagner, J., Torres, M., Duan, N., Adams, T., & Berglund, P. (2004). Sample designs and sampling methods for the Collaborative Psychiatric Epidemiology Studies (CPES). *International Journal of Methods in Psychiatric Research*, *13*(4), 221–240.

Heun, R., Kockler, M., & Papassotiropoulos, A. (2000). Distinction of early- and late-onset depression in the elderly by their lifetime symptomatology. *International Journal of Geriatric Psychiatry*, *15*, 1138–1142. Retrieved from https://onlinelibrary.wiley.com/doi/pdf/10.1002/1099-1166%28200012%2915%3A12%3C1138%3A%3AAID-GPS266%3E3.0.CO%3B2-7

Horwitz, A. V, & Wakefield, J. C. (2007). *The loss of sadness: How psychiatry transformed normal sorrow into depressive disorder*. Oxford University Press.

Hoyer, P. O., Shimizu, S., Kerminen, A. J., & Palviainen, M. (2008). Estimation of causal effects using linear

non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, *49*, 362–378. https://doi.org/10.1016/j.ijar.2008.02.006

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hyman, S. E. (2007). Can neuroscience be integrated into the DSM-V? *Nature Reviews Neuroscience*, *8*(9), 725–732. https://doi.org/10.1038/nrn2218

Hyman, S. E. (2010). The Diagnosis of Mental Disorders: The Problem of Reification. *Annual Review of Clinical Psychology*, *6*(1), 155–179. https://doi.org/10.1146/annurev.clinpsy.3.022806.091532

Hyvarinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. Retrieved from https://www.wiley.com/en-us/Independent+Component+Analysis-p-9780471405405

Isometsä, E. (2014). Suicidal Behaviour in Mood Disorders—Who, When, and Why? *The Canadian Journal of Psychiatry*, *59*(3), 120–130. https://doi.org/10.1177/070674371405900303

Jablensky, A. (2016). Psychiatric classifications: Validity and utility. *World Psychiatry*, *15*(1), 26–31. https://doi.org/10.1002/wps.20284

Jang, K. L., Livesley, W. J., Taylor, S., Stein, M. B., & Moon, E. C. (2004). Heritability of individual depressive symptoms. *Journal of Affective Disorders*, *80*(2–3), 125–133. https://doi.org/10.1016/S0165-0327(03)00108-3

Jokela, M., Virtanen, M., Batty, G. D., & Kivimäki, M. (2016). Inflammation and Specific Symptoms of Depression. *JAMA Psychiatry*, *73*(1), 87. https://doi.org/10.1001/jamapsychiatry.2015.1977

Kamenov, K., Caballero, F. F., Miret, M., Leonardi, M., Sainio, P., Tobiasz-Adamczyk, B., … Cabello, M. (2016). Which Are the Most Burdensome Functioning Areas in Depression? A Cross-National Study. *Frontiers in Psychology*, *7*, 1342. https://doi.org/10.3389/fpsyg.2016.01342

Kamenov, K., Cabello, M., Nieto, M., Bernard, R., Kohls, E., Rummel-Kluge, C., & Ayuso-Mateos, J. L. (2017). Research Recommendations for Improving Measurement of Treatment Effectiveness in Depression. *Frontiers in Psychology*, *8*, 356. https://doi.org/10.3389/fpsyg.2017.00356

Kang, S.-M., & Waller, N. G. (2005). Moderated Multiple Regression, Spurious Interaction Effects, and IRT. *Applied Psychological Measurement*, *29*(2), 87–105. https://doi.org/10.1177/0146621604272737

Karp, J. F., Buysse, D. J., Houck, P. R., Cherry, C., Kupfer, D. J., & Frank, E. (2004). Relationship of

Variability in Residual Symptoms With Recurrence of Major Depressive Disorder During Maintenance Treatment. *American Journal of Psychiatry*, *161*(10), 1877–1884. https://doi.org/10.1176/ajp.161.10.1877

Keller, M. C., Neale, M. C., & Kendler, K.S. (2007). Association of Different Adverse Life Events With Distinct Patterns of Depressive Symptoms. *American Journal of Psychiatry*, *164*(10), 1521–1529. https://doi.org/10.1176/appi.ajp.2007.06091564

Keller, M. C., & Nesse, R. M. (2006). The evolutionary significance of depressive symptoms: different adverse situations lead to different depressive symptom patterns. *Journal of Personality and Social Psychology*, *91*(2), 316–330. https://doi.org/10.1037/0022-3514.91.2.316

Kendler, K S. (2012). The dappled nature of causes of psychiatric illness: replacing the organic–functional/hardware–software dichotomy with empirically based pluralism. *Molecular Psychiatry*, *17*(4), 377–388. https://doi.org/10.1038/mp.2011.182

Kendler, K.S. (2016). The Phenomenology of Major Depression and the Representativeness and Nature of DSM Criteria. *American Journal of Psychiatry*, *173*(8), 771–780. https://doi.org/10.1176/appi.ajp.2016.15121509

Kendler, K S, & Gardner, C. O. (1998). Boundaries of major depression: an evaluation of DSM-IV criteria. *The American Journal of Psychiatry*, *155*(2), 172–177. https://doi.org/10.1176/ajp.155.2.172

Kendell, R.S, & Jablensky, A. (2003). Distinguishing between the validity and utility of psychiatric diagnoses. *American Journal of Psychiatry*, *160*(1), 4–12. https://doi.org/10.1176/appi.ajp.160.1.4

Kendler, K.S., Muñoz, R. A., & Murphy, G. (2010). The Development of the Feighner Criteria: A Historical Perspective. *American Journal of Psychiatry*, *167*(2), 134–142. https://doi.org/10.1176/appi.ajp.2009.09081155

Kendler, K.S, Myers, J., & Zisook, S.M. (2008). Does Bereavement-Related Major Depression Differ From Major Depression Associated With Other Stressful Life Events? *American Journal of Psychiatry*, *165*(11), 1449–1455. Retrieved from https://ajp.psychiatryonline.org/doi/pdfplus/10.1176/appi.ajp.2008.07111757

Kendler, K.S., & Parnas, J. (2012). *Philosophical issues in Psychiatry II: Nosology* (K. Kendler & J. Parnas, Eds.). Oxford: Oxford University Press.

Kendler, K., & Parnas, J. (2014). *Philosophical issues in Psychiatry III: The nature and sources of historical change* (K. Kendler & J. Parnas, Eds.). Oxford: Oxford University Press.

Kennis, M., Gerritsen, L., van Dalen, M., Williams, A., Cuijpers, P., & Bockting, C. (2018). T149. Do We Have

Evidence for Predictive Biomarkers for Major Depressive Disorder? A Meta-Analysis and Systematic Review of Prospective Studies. *Biological Psychiatry*, *83*(9), S186. https://doi.org/10.1016/J.BIOPSYCH.2018.02.486

Kessler, R.C, Zhao, S., Blazer, D. G., & Swartz, M. (1997). Prevalence, correlates, and course of minor depression and major depression in the National Comorbidity Survey. *Journal of Affective Disorders*, *45*(1–2), 19–30. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9268772

Kessler, R.C., Merikangas, K. R., Berglund, P., Eaton, W. W., Koretz, D. S., & Walters, E. E. (2003). Mild Disorders Should Not Be Eliminated From the DSM-V. *Archives of General Psychiatry*, *60*(11), 1117. https://doi.org/10.1001/archpsyc.60.11.1117

Kessler, R.C, & Ustün, T. B. (2004). The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatric Research*, *13*(2), 93–121. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15297906

Khan, A., Brodhead, A. E., Kolts, R. L., & Brown, W. A. (2005). Severity of depressive symptoms and response to antidepressants and placebo in antidepressant trials. *Journal of Psychiatric Research*, *39*(2), 145–150. https://doi.org/10.1016/j.jpsychires.2004.06.005

Khan, A., & Brown, W. A. (2015). Antidepressants versus placebo in major depression: an overview. *World Psychiatry*, *14*(3), 294–300. https://doi.org/10.1002/wps.20241

Khan, A., Faucett, J., Lichtenberg, P., Kirsch, I., & Brown, W. A. (2012). A Systematic Review of Comparative Efficacy of Treatments and Controls for Depression. *PLoS ONE*, *7*(7), e41778. https://doi.org/10.1371/journal.pone.0041778

Khan, A., Leventhal, R. M., Khan, S. R., & Brown, W. A. (2002). Severity of depression and response to antidepressants and placebo: an analysis of the Food and Drug Administration database. *Journal of Clinical Psychopharmacology*, *22*(1), 40–45. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11799341

Khan, A., Mar, K. F., & Brown, W. A. (2018). The conundrum of depression clinical trials. *International Clinical Psychopharmacology*, 1. https://doi.org/10.1097/YIC.0000000000000229

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, *9*(02), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Kirsch, I. (2014). Antidepressants and the Placebo Effect. *Zeitschrift Fur Psychologie*, *222*(3), 128–134. https://doi.org/10.1027/2151-2604/a000176

Kirsch, I., Deacon, B. J., Huedo-Medina, T. B., Scoboria, A., Moore, T. J., & Johnson, B. T. (2008). Initial Severity and Antidepressant Benefits: A Meta-Analysis of Data Submitted to the Food and Drug Administration. *PLoS Medicine*, *5*(2), e45. https://doi.org/10.1371/journal.pmed.0050045

Kish, L. (1949). A Procedure for Objective Respondent Selection within the Household. *Journal of the American Statistical Association*, *44*(247), 380–387. https://doi.org/10.1080/01621459.1949.10483314

Kitamura, T., Nakagawa, Y., & Machizawa, S. (1993). Grading depression severity by symptom scores: is it a valid method for subclassifying depressive disorders? *Comprehensive Psychiatry*, *34*(4), 280–283. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8348808

Kleiman, E. M., Coppersmith, D. D. L., Millner, A. J., Franz, P. J., Fox, K. R., & Nock, M. K. (2018). Are suicidal thoughts reinforcing? A preliminary real-time monitoring study on the potential affect regulation function of suicidal thinking. *Journal of Affective Disorders*, *232*, 122–126. https://doi.org/10.1016/J.JAD.2018.02.033

Kleiman, E. M., Turner, B. J., Fedor, S., Beale, E. E., Huffman, J. C., & Nock, M. K. (2017). Examination of real-time fluctuations in suicidal ideation and its risk factors: Results from two ecological momentary assessment studies. *Journal of Abnormal Psychology*, *126*(6), 726–738. https://doi.org/10.1037/abn0000273

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Kohn, R., Saxena, S., Levav, I., & Saraceno, B. (2004). The treatment gap in mental health care. *Bulletin of the World Health Organization*, *82*(11), 858–866. https://doi.org//S0042-96862004001100011

Kotov, R., Krueger, R. F., & Watson, D. (2018). A paradigm shift in psychiatric classification: the Hierarchical Taxonomy Of Psychopathology (HiTOP). *World Psychiatry*, *17*(1), 24–25. https://doi.org/10.1002/wps.20478

Kraemer, H. C. (2007). DSM categories and dimensions in clinical and research contexts. *International Journal of Methods in Psychiatric Research*, *16*(S1), S8–S15. https://doi.org/10.1002/mpr.211

Kroenke, K., & Spitzer, R. L. (2002a). The PHQ-9: A New Depression Diagnostic and Severity Measure. *Psychiatric Annals*, *32*(9), 509–515. https://doi.org/10.3928/0048-5713-20020901-06

Kroenke, K., & Spitzer, R. L. (2002b). The PHQ-9: A New Depression Diagnostic and Severity Measure. *Psychiatric Annals*, *32*(9), 509–515. https://doi.org/10.3928/0048-5713-20020901-06

Kroenke, K., Spitzer, R. L., Williams, J. B. W., & Löwe, B. (2010). The Patient Health Questionnaire Somatic,

Anxiety, and Depressive Symptom Scales: a systematic review. *General Hospital Psychiatry*, *32*(4), 345–359. https://doi.org/10.1016/j.genhosppsych.2010.03.006

Krueger, R. F., & Eaton, N. R. (2012). Structural validity and the classification of mental disorders. In Kenneth S. Kendler & J. Parnas (Eds.), *Philosophical Issues in Psychiatry II: Nosology* (pp. 199–211). Oxford University Press.

Kudinova, A. Y., Owens, M., Burkhouse, K. L., Barretto, K. M., Bonanno, G. A., & Gibb, B. E. (2016). Differences in emotion modulation using cognitive reappraisal in individuals with and without suicidal ideation: An ERP study. *Cognition and Emotion*, *30*(5), 999–1007. https://doi.org/10.1080/02699931.2015.1036841

Lam, R. W., Filteau, M.-J., & Milev, R. (2011). Clinical effectiveness: The importance of psychosocial functioning outcomes. *Journal of Affective Disorders*, *132*, S9–S13. https://doi.org/10.1016/J.JAD.2011.03.046

Lam, R. W., Parikh, S. V, Michalak, E. E., Dewa, C. S., & Kennedy, S. H. (2015). Canadian Network for Mood and Anxiety Treatments (CANMAT) consensus recommendations for functional outcomes in major depressive disorder. *Annals of Clinical Psychiatry*, *27*(2), 142–149. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25954941

Lamers, F., Milaneschi, Y., de Jonge, P., Giltay, E. J., & Penninx, B. W. J. H. (2018). Metabolic and inflammatory markers: associations with individual depressive symptoms. *Psychological Medicine*, *48*(07), 1102–1110. https://doi.org/10.1017/S0033291717002483

Lawlor, D. A., Tilling, K., & Smith, G. D. (2017). Triangulation in aetiological epidemiology. *International Journal of Epidemiology*, *45*(6), dyw314. https://doi.org/10.1093/ije/dyw314

Lefere, S., De Rouck, R., & De Vreese, L. (2017). What to expect from reliability and validity claims? A pragmatic conception of psychiatric nosology. *Journal of Evaluation in Clinical Practice*, *23*(5), 981–987. https://doi.org/10.1111/jep.12686

Lin, H.-T., Lai, C.-H., Perng, H.-J., Chung, C.-H., Wang, C.-C., Chen, W.-L., & Chien, W.-C. (2018). Insomnia as an independent predictor of suicide attempts: a nationwide population-based retrospective cohort study. *BMC Psychiatry*, *18*(1), 117. https://doi.org/10.1186/s12888-018-1702-2

Linardon, J., Cuijpers, P., Carlbring, P., Messer, M., & Fuller-Tyszkiewicz, M. (2019). The efficacy of of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. *World Psychiatry*, *18*(3), 325–336. Retrieved from

https://onlinelibrary.wiley.com/doi/pdf/10.1002/wps.20673

Liu, Q., He, H., Yang, J., Feng, X., Zhao, F., & Lyu, J. (2019). Changes in the global burden of depression from 1990 to 2017: Findings from the Global Burden of Disease study. *Journal of Psychiatric Research*. https://doi.org/10.1016/J.JPSYCHIRES.2019.08.002

Lorenzo-Luaces, L. (2015). Heterogeneity in the prognosis of major depression: from the common cold to a highly debilitating and recurrent illness. *Epidemiology and Psychiatric Sciences*, *24*(466–472). https://doi.org/10.1017/S2045796015000542

Lorenzo-Luaces, Lorenzo, Zimmerman, M., & Cuijpers, P. (2018). Are studies of psychotherapies for depression more or less generalizable than studies of antidepressants? *Journal of Affective Disorders*, *234*, 8–13. https://doi.org/10.1016/j.jad.2018.02.066

Lumley, T. (2019). *survey: analysis of complex survey samples*.

Lux, V., Aggen, S. H., & Kendler, K. S. (2010). The DSM-IV definition of severity of major depression: inter-relationship and validity. *Psychological Medicine*, *40*(10), 1691–1701. https://doi.org/10.1017/S0033291709992066

Lux, V., & Kendler, K. S. (2010). Deconstructing major depression: a validation study of the DSM-IV symptomatic criteria. *Psychological Medicine*, *40*(10), 1679–1690. https://doi.org/10.1017/S0033291709992157

Maj, M. (1998). Critique of the DSM-IV operational diagnostic criteria for schizophrenia. *The British Journal of Psychiatry*, *172*, 458, 460. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9828982

Maj, M. (2011). When does depression become a mental disorder? *The British Journal of Psychiatry*, *199*(2), 85–86.

Maj, M. (2012). Development and validation of the current concept of major depression. *Psychopathology*, *45*(3), 135–146.

Maj, M. (2012). Validity and clinical utility of the current operational characterization of major depression. *International Review of Psychiatry*, *24*(6), 530–537. https://doi.org/10.3109/09540261.2012.712952

Maj, M. (2014). DSM-5, ICD-11 and 'pathologization of normal conditions.' *Australian & New Zealand Journal of Psychiatry*, *48*(2), 193–194. https://doi.org/10.1177/0004867413518825

Markon, K. E. (2010). Modeling psychopathology structure: a symptom-level analysis of Axis I and II disorders. *Psychological Medicine*, *40*(02), 273. https://doi.org/10.1017/S0033291709990183

McKnight, P. E., & Kashdan, T. B. (2009). The importance of functional impairment to mental health outcomes: A case for reassessing our goals in depression treatment research. *Clinical Psychology Review*, *29*(3), 243–259. https://doi.org/10.1016/J.CPR.2009.01.005

McRae, K. (2016). Cognitive emotion regulation: a review of theory and scientific findings. *Current Opinion in Behavioral Sciences*, *10*, 119–124. https://doi.org/10.1016/J.COBEHA.2016.06.004

Merriam-Webster Medical Dictionary (2018). "*Aggregate*". Retrieved from https://www.merriam-webster.com/dictionary/

Messner, J. W., Mayr, G. J., & Zeileis, A. (2016). Heteroscedastic Censored and Truncated Regression with crch. *The R Journal*, *8*(1), 173–181. Retrieved from http://web.b.ebscohost.com/ehost/detail/detail?vid=4&sid=6ece72e1-b8d6-46df-b9b1-3bfb4a07fe3d%40sessionmgr103&bdata=JnNpdGU9ZWhvc3QtbGl2ZSZzY29wZT1zaXRl#AN=118430358&db=a9h

Mojtabai, R. (2001). Impairment in major depression: Implications for diagnosis. *Comprehensive Psychiatry*, *42*(3), 206–212. https://doi.org/10.1053/COMP.2001.23142

Morgado, A., Smith, M., Lecrubier, Y., & Widlöcher, D. (1991). Depressed subjects unwittingly overreport poor social adjustment which they reappraise when recovered. *The Journal of Nervous and Mental Disease*, *179*(10), 614–619. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1919545

Muthén, B. O. (1989). Factor structure in groups selected on observed scores. *British Journal of Mathematical and Statistical Psychology*, *42*(1), 81–90.

Myung, W., Song, J., Lim, S.-W., Won, H.-H., Kim, S., Lee, Y., … Kim, D. K. (2012). Genetic association study of individual symptoms in depression. *Psychiatry Research*, *198*(3), 400–406. https://doi.org/10.1016/j.psychres.2011.12.037

Nagelkerke, N. J. D. (1991). A Note on a General Definition of the Coefficient of Determination. *Biometrika*, *78*(3), 691–692. Retrieved from http://links.jstor.org/sici?sici=0006-3444%28199109%2978%3A3%3C691%3AANOAGD%3E2.0.CO%3B2-V

Narrow, W. E., Rae, D. S., Robins, L. N., & Regier, D. A. (2002). Revised prevalence estimates of mental disorders in the United States: using a clinical significance criterion to reconcile 2 surveys' estimates. *Archives of General Psychiatry*, *59*(2), 115–123. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11825131

Nrugham, L., Larsson, B., & Sund, A. M. (2008). Specific depressive symptoms and disorders as associates and

predictors of suicidal acts across adolescence. *Journal of Affective Disorders*, *111*(1), 83–93. https://doi.org/10.1016/J.JAD.2008.02.010

Olbert, C. M., Rasmussen, A., Gala, G. J., & Tupler, L. A. (2016). Treatment outcome variation between depression symptom combinations in the STAR*D study. *Journal of Affective Disorders*, *201*, 1–7. https://doi.org/10.1016/J.JAD.2016.04.050

Ong, E., & Thompson, C. (2019). The Importance of Coping and Emotion Regulation in the Occurrence of Suicidal Behavior. *Psychological Reports*, *122*(4), 1192–1210. https://doi.org/10.1177/0033294118781855

Oquendo, M. A., Barrera, A., Ellis, S. P., Li, S., Burke, A. K., Grunebaum, M., … Mann, J. J. (2004). Instability of Symptoms in Recurrent Major Depression: A Prospective Study. *American Journal of Psychiatry*, *161*(2), 255–261. https://doi.org/10.1176/appi.ajp.161.2.255

Østergaard, S. D., Jensen, S. O. W., & Bech, P. (2011). The heterogeneity of the depressive syndrome: when numbers get serious. *Acta Psychiatrica Scandinavica*, *124*(6), 495–496. https://doi.org/10.1111/j.1600-0447.2011.01744.x

Owen, A. B. (2007). Infinitely Imbalanced Logistic Regression. In *Journal of Machine Learning Research* (Vol. 8). Retrieved from http://www.jmlr.org/papers/volume8/owen07a/owen07a.pdf

Parker, G. (2018). *The benefits of antidepressants: news or fake news?* https://doi.org/10.1192/bjp.2018.98

Parnas, J., & Bovet, P. (2014). Psychiatry made easy: operation(al)ism and some of its consequences. In Kenneth S Kendler & J. Parnas (Eds.), *Philosophical Issues in Psychiatry III: The Nature and Sources of Historical Change* (pp. 190–212). Oxford University Press.

Patten, S. B. (2015). Major depressive disorder: reification and (maybe) rheostasis. *Epidemiology and Psychiatric Sciences*, *24*(06), 473–475. https://doi.org/10.1017/S2045796015000682

Pennell, B.-E., Bowers, A., Carr, D., Chardoul, S., Cheung, G.-Q., Dinkelmann, K., … Torres, M. (2004). The development and implementation of the National Comorbidity Survey Replication, the National Survey of American Life, and the National Latino and Asian American Survey. *International Journal of Methods in Psychiatric Research*, *13*(4), 241–269. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15719531

Pigeon, W. R., Pinquart, M., & Conner, K. (2012). Meta-Analysis of Sleep Disturbance and Suicidal Thoughts and Behaviors. *The Journal of Clinical Psychiatry*, *73*(09), e1160–e1167. https://doi.org/10.4088/JCP.11r07586

Plomin, R., Haworth, C. M. A., & Davis, O. S. P. (2009). Common disorders are quantitative traits. *Nature*

Reviews Genetics, *10*(12), 872–878. https://doi.org/10.1038/nrg2670

Reckase, M. (2009). *Multidimensional item response theory* (Vol. 150). Springer.

Reise, S P, & Rodriguez, A. (2016). Item response theory and the measurement of psychiatric constructs: some empirical and conceptual issues and challenges. *Psychological Medicine*, *46*(10), 2025–2039.

Reise, S.P, & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48.

Richmond, S., Hasking, P., & Meaney, R. (2017). Psychological Distress and Non-Suicidal Self-Injury: The Mediating Roles of Rumination, Cognitive Reappraisal, and Expressive Suppression. *Archives of Suicide Research*, *21*(1), 62–72. https://doi.org/10.1080/13811118.2015.1008160

Robins, E., & Guze, S. B. (1970). Establishment of diagnostic validity in psychiatric illness: its application to schizophrenia. *American Journal of Psychiatry*, *126*(7), 983–987.

Rosenström, T., & García-Velázquez, R. (in press). Distribution-based causal inference: a review and practical guidance for epidemiologists. In W. Wiedermann, D. Kim, E. Sungur, & A. Von Eye (Eds.), *Direction Dependence in Statistical Models: Methods of Analysis*.

Rosenström, T., & Jokela, M. (2017). Reconsidering the definition of major depression based on collaborative psychiatric epidemiology surveys. *Journal of Affective Disorders*, *207*, 38–46.

Rosenström, T., Jokela, M., Puttonen, S., Hintsanen, M., Pulkki-Råback, L., Viikari, J. S., … Keltikangas-Järvinen, L. (2012). Pairwise Measures of Causal Direction in the Epidemiology of Sleep Problems and Depression. *PLoS ONE*, *7*(11), e50841. https://doi.org/10.1371/journal.pone.0050841

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Ruscio, A. M. (2002). Delimiting the boundaries of generalized anxiety disorder: differentiating high worriers with and without GAD. *Journal of Anxiety Disorders*, *16*(4), 377–400.

Ruscio, J., Brown, T. A., & Ruscio, A.M. (2009). A taxometric investigation of DSM-IV major depression in a large outpatient sample: interpretable structural results depend on the mode of assessment. *Assessment*, *16*(2), 127–144. https://doi.org/10.1177/1073191108330065

Ruscio, J., Zimmerman, M., McGlinchey, J. B., Chelminski, I., & Young, D. (2007). Diagnosing major depressive disorder XI: a taxometric investigation of the structure underlying DSM-IV symptoms. *The Journal of Nervous and Mental Disease*, *195*(1), 10–19.

https://doi.org/10.1097/01.nmd.0000252025.12014.c4

Rush, A. J., Zimmerman, M., Wisniewski, S. R., Fava, M., Hollon, S. D., Warden, D., … Trivedi, M. H. (2005). Comorbid psychiatric disorders in depressed outpatients: Demographic and clinical features. *Journal of Affective Disorders*, *87*(1), 43–55. https://doi.org/10.1016/J.JAD.2005.03.005

Ryder, A., Yang, J., Zhu, X., Yao, S., Yi, J., Heine, S., & Bagby, M. (2008). The Cultural Shaping of Depression: Somatic Symptoms in China, Psychological Symptoms in North America? *Journal of Abnormal Psychology.*, *117*(2), 300–313. Retrieved from https://ovidsp.uk.ovid.com/sp-3.31.1b/ovidweb.cgi?QS2=434f4e1a73d37e8c01e9bb09ab15b39284d7838b59b2a069ce51bd19da50c3c091fe5a27b7fa8ec813ac1cc44c0888eed0f459cf67e9c3f2d703f821c94c4cda5ca2e53da6e6f4e0f1f4e7c69f3e11f9e24a812f78eefb42fe9d9a52928710906d8929324

Santor, D. A., Gregus, M., & Welch, A. (2006). Eight Decades of Measurement in Depression. *Measurement* , *4*(3), 135–155. Retrieved from http://www.scalesandmeasures.net/files/files/Santor et al_ (2006) Eight Decades (1).pdf

Sheehan, D. V, Harnett-Sheehan, K., & Raj, B. A. (1996). The measurement of disability. *International Clinical Psychopharmacology*, *11 Suppl 3*, 89–95. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8923116

Shimizu, S, Inazumi, T., Sogawa, Y., Hyvarinen, A., Kawahara, Y., Washio, T., … Bollen, K. (2011). DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model. *Journal of Machine Learning Research*, *12*, 1225–1248.

Shimizu, S. (2018). Non-Gaussian Methods for Causal Structure Learning. *Prevention Science*. https://doi.org/10.1007/s11121-018-0901-x

Shimizu, S, Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2006). A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, *7*, 2003–2030. Retrieved from https://www.cs.helsinki.fi/u/ahyvarin/papers/JMLR06.pdf

Simon, G. E., VonKorff, M., Piccinelli, M., Fullerton, C., & Ormel, J. (1999). An International Study of the Relation between Somatic Symptoms and Depression. *New England Journal of Medicine*, *341*(18), 1329–1335. https://doi.org/10.1056/NEJM199910283411801

Singh, I., & Rose, N. (2009). Biomarkers in psychiatry. *Nature*, *460*(7252), 202–207. https://doi.org/10.1038/460202a

Slade, T., & Andrews, G. (2002). Empirical Impact of DSM-IV Diagnostic Criterion for Clinical Significance. *Journal of Nervous and Mental Disease.*, *190*(5), 334–337. https:// doi.org/ 10.1097/00005053-200205000-

00011

Smoller, J. W. (2019). Psychiatric Genetics Begins to Find Its Footing. *American Journal of Psychiatry*, *176*(8), 609–614. https://doi.org/10.1176/appi.ajp.2019.19060643

Solomon, D. A., Leon, A. C., Endicott, J., Mueller, T. I., Coryell, W., Shea, M. T., & Keller, M. B. (2004). Psychosocial impairment and recurrence of major depression. *Comprehensive Psychiatry*, *45*(6), 423–430. https://doi.org/10.1016/j.comppsych.2004.07.002

Spearman, C. (1904). "General Intelligence", Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201. https://doi.org/10.2307/1412107

Spitzer, R. L., & Wakefield, J. C. (1999). DSM-IV diagnostic criterion for clinical significance: Does it help solve the false positives problem? *American Journal of Psychiatry*, *156*(12), 1856–1864. https://doi.org/10.1176/ajp.156.12.1856

Struijs, S. Y., Lamers, F., Verdam, M. G. E., van Ballegooijen, W., Spinhoven, P., van der Does, W., & Penninx, B. W. J. H. (2020). Temporal stability of symptoms of affective disorders, cognitive vulnerability and personality over time. *Journal of Affective Disorders*, *260*, 77–83. https://doi.org/10.1016/J.JAD.2019.08.090

Sugarman, M. A. (2016). Are antidepressants and psychotherapy equally effective in treating depression? A critical commentary. *Journal of Mental Health*, *25*(6), 475–478. https://doi.org/10.3109/09638237.2016.1139071

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (Sixth). Retrieved from https://www.pearson.com/us/higher-education/program/Tabachnick-Using-Multivariate-Statistics-6th-Edition/PGM332849.html

Taylor, D. J., Walters, H. M., Vittengl, J. R., Krebaum, S., & Jarrett, R. B. (2010). Which depressive symptoms remain after response to cognitive therapy of depression and predict relapse and recurrence? *Journal of Affective Disorders*, *123*(1–3), 181–187. https://doi.org/10.1016/j.jad.2009.08.007

Thoemmes, F., Rosseel, Y., & Textor, J. (2018). Local fit evaluation of structural equation models using graphical criteria. *Psychological Methods*, *23*(1), 27–41. https://doi.org/10.1037/met0000147

Thombs, B. D., Ziegelstein, R. C., Pilote, L., Dozois, D. J. A., Beck, A. T., Dobson, K. S., … Abbey, S. E. (2010). Somatic symptom overlap in Beck Depression Inventory–II scores following myocardial infarction. *British Journal of Psychiatry*, *197*(01), 61–65. https://doi.org/10.1192/bjp.bp.109.076596

Tobin, J. (1958). Estimation of relationship for limited dependent variables. *Econometrica*, *26*, 24–36.

Tweed, D. L. (1993). Depression-related impairment: estimating concurrent and lingering effects. *Psychological Medicine*, *23*(02), 373–386.

Üstün, T. B. (2010). *Measuring health and disability: Manual for WHO disability assessment schedule WHODAS 2.0*. World Health Organization.

van Bork, R., Epskamp, S., Rhemtulla, M., Borsboom, D., & van der Maas, H. L. J. (2017). What is the p-factor of psychopathology? Some risks of general factor modeling. *Theory & Psychology*, *27*(6), 759–773. https://doi.org/10.1177/0959354317737185

van der Lem, R., van der Wee, N. J. A., van Veen, T., & Zitman, F. G. (2011). The generalizability of antidepressant efficacy trials to routine psychiatric out-patient practice. *Psychological Medicine*, *41*(07), 1353–1363. https://doi.org/10.1017/S0033291710002175

Van Der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842–861. https://doi.org/10.1037/0033-295X.113.4.842

van Eeden, W. A., van Hemert, A. M., Carlier, I. V. E., Penninx, B. W., & Giltay, E. J. (2019). Severity, course trajectory and within-person variability of individual symptoms in patients with major depressive disorder. *Acta Psychiatrica Scandinavica*, *139*(2), 194–205. https://doi.org/10.1111/acps.12987

van Loo, H. M., & Romeijn, J.-W. (2018). Letter to the Editor: Measuring and defining: the double role of the DSM criteria for psychiatric disorders. *Psychological Medicine*, *48*(05), 872–873. https://doi.org/10.1017/S0033291717001799

van Loo, H.M., Schoevers, R. A., Kendler, K. S., de Jonge, P., & Romeijn, J.-W. (2016). Psychiatric comorbidity does not only depend on diagnostic thresholds: an illustration with major depressive disorder and generalized anxiety disorder. *Depression and Anxiety*, *33*(2), 143–152. https://doi.org/10.1002/da.22453

van Loo, H.M, de Jonge, P., Romeijn, J.-W., Kessler, R. C., & Schoevers, R. A. (2012). Data-driven subtypes of major depressive disorder: a systematic review. *BMC Medicine*, *10*(1), 156. https://doi.org/10.1186/1741-7015-10-156

Van Orden, K. A., Lynam, M. E., Hollar, D., & Joiner, T. E. (2006). Perceived Burdensomeness as an Indicator of Suicidal Symptoms. *Cognitive Therapy and Research*, *30*(4), 457–467. https://doi.org/10.1007/s10608-006-9057-2

Verduijn, J., Verhoeven, J. E., Milaneschi, Y., Schoevers, R. A., van Hemert, A. M., Beekman, A. T. F., & Penninx, B. W. J. H. (2017). Reconsidering the prognosis of major depressive disorder across diagnostic boundaries: full recovery is the exception rather than the rule. *BMC Medicine*, *15*(1), 215. https://doi.org/10.1186/s12916-017-0972-8

Verhoeven, F. E. A., Wardenaar, K. J., Ruhé, H. G. E., Conradi, H. J., & de Jonge, P. (2018). Seeing the signs: Using the course of residual depressive symptomatology to predict patterns of relapse and recurrence of major depressive disorder. *Depression and Anxiety*, *35*(2), 148–159. https://doi.org/10.1002/da.22695

Wakefield, J. (2009). Disability and diagnosis: should role impairment be eliminated from DSM/ICD diagnostic criteria? *World Psychiatry*, *8*(2), 87–88. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19516926

Wakefield, J. C., & Schmitz, M. F. (2015). Feelings of worthlessness during a single complicated major depressive episode predict postremission suicide attempt. *Acta Psychiatrica Scandinavica*.

Wakefield, J. C., & Schmitz, M. F. (2017). Severity of complicated versus uncomplicated subthreshold depression: New evidence on the "Monotonicity Thesis" from the national comorbidity survey. *Journal of Affective Disorders*, *212*, 101–109.

Wakefield, J. C., Schmitz, M. F., & Baer, J. C. (2010). Does the DSM-IV Clinical Significance Criterion for Major Depression Reduce False Positives? Evidence From the National Comorbidity Survey Replication. *American Journal of Psychiatry*, *167*(3), 298–304. https://doi.org/10.1176/appi.ajp.2009.09040553

Wakefield, J. C., Schmitz, M. F., First, M. B., & Horwitz, A. V. (2007). Extending the Bereavement Exclusion for Major Depression to Other Losses. *Archives of General Psychiatry*, *64*(4), 433. https://doi.org/10.1001/archpsyc.64.4.433

Wanders, R. B. K., van Loo, H. M., Vermunt, J. K., Meijer, R. R., Hartman, C. A., Schoevers, R. A., … de Jonge, P. (2016). Casting wider nets for anxiety and depression: disability-driven cross-diagnostic subtypes in a large cohort. *Psychological Medicine*, *46*(16), 3371–3382. https://doi.org/10.1017/S0033291716002221

Wanders, Rob B.K., Wardenaar, K. J., Kessler, R. C., Penninx, B. W. J. H., Meijer, R. R., & de Jonge, P. (2015). Differential reporting of depressive symptoms across distinct clinical subpopulations: What DIFference does it make? *Journal of Psychosomatic Research*, *78*(2), 130–136. https://doi.org/10.1016/j.jpsychores.2014.08.014

Wardenaar, K. J., & de Jonge, P. (2013). Diagnostic heterogeneity in psychiatry: towards an empirical solution. *BMC Medicine*, *11*(1), 201. https://doi.org/10.1186/1741-7015-11-201

World Health Organization. (n.d.). Suicide data. Retrieved July 4, 2018, from Mental Health website: http://www.who.int/mental_health/prevention/suicide/suicideprevent/en/

World Health Organization. (2004). *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines* (2nd ed.; World Health Organization, Ed.). Retrieved from https://apps.who.int/iris/handle/10665/42980

World Health Organization. (2004). *The Global Burden of Disease: 2004 update.* Retrieved from http://www.who.int/healthinfo/global_burden_disease/2004_report_update/en/

World Health Organization. (2017). *Depression and other common mental disorders. Global health estimates.* Retrieved from http://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf

World Health Organization. (2018). *Global Health Estimates 2016: Disease burden by Cause, Age, Sex, by Country and by Region, 2000-2016*. Retrieved from http://www.who.int/healthinfo/global_burden_disease/estimates/en/index1.html

World Health Organization. (2018). *International Classification of Functioning, Disability and Health (ICF)*. Retrieved August 20, 2018, from http://www.who.int/classifications/icf/en/

Zimmerman, M. (2012). Symptom severity and guideline-based treatment recommendations for depressed patients: implications of DSM-5's potential recommendation of the PHQ-9 as the measure of choice for depression severity. *Psychotherapy and Psychosomatics*, *81*(6), 329–332. https://doi.org/10.1159/000342262

Zimmerman, M., Balling, C., Chelminski, I., & Dalrymple, K. (2018). Understanding the severity of depression: Which symptoms of depression are the best indicators of depression severity? *Comprehensive Psychiatry*, *87*, 84–88. https://doi.org/10.1016/j.comppsych.2018.09.006

Zimmerman, M., Balling, C., Chelminski, I., & Dalrymple, K. (2020). Applying the inclusion/exclusion criteria in placebo-controlled studies to a clinical sample: A comparison of medications. *Journal of Affective Disorders*, *260*, 483–488. https://doi.org/10.1016/J.JAD.2019.09.012

Zimmerman, M., Chelminski, I., McGlinchey, J. B., & Young, D. (2006). Diagnosing major depressive disorder X: can the utility of the DSM-IV symptom criteria be improved? *The Journal of Nervous and Mental Disease*, *194*(12), 893–897. https://doi.org/10.1097/01.nmd.0000248970.50265.34

Zimmerman, M., Chelminski, I., & Young, D. (2004). On the Threshold of Disorder: A Study of the Impact of the DSM-IV Clinical Significance Criterion on Diagnosing Depressive and Anxiety Disorders in Clinical Practice. In *Journal of Clinical Psychiatry* (Vol. 65). Retrieved from

http://www.psychiatrist.com/JCP/article/_layouts/ppp.psych.controls/BinaryViewer.ashx?Article=/jcp/article/Pages/2004/v65n10/v65n1016.aspx&Type=Article

Zimmerman, M., Clark, H. L., Multach, M. D., Walsh, E., Rosenstein, L. K., & Gazarian, D. (2016a). Symptom Severity and the Generalizability of Antidepressant Efficacy Trials. *Journal of Clinical Psychopharmacology*, *36*(2), 153–156. https://doi.org/10.1097/JCP.0000000000000466

Zimmerman, M., Clark, H. L., Multach, M. D., Walsh, E., Rosenstein, L. K., & Gazarian, D. (2016b). Variability in the substance use disorder exclusion criterion in antidepressant efficacy trials. *Journal of Affective Disorders*, *198*, 39–42. https://doi.org/10.1016/j.jad.2016.03.024

Zimmerman, M., Ellison, W., Young, D., Chelminski, I., & Dalrymple, K. (2015). How many different ways do patients meet the diagnostic criteria for major depressive disorder? *Comprehensive Psychiatry*, *56*, 29–34. https://doi.org/10.1016/j.comppsych.2014.09.007

Zimmerman, M., Martinez, J. A., Attiullah, N., Friedman, M., Toba, C., Boerescu, D. A., & Rahgeb, M. (2012). Why Do Some Depressed Outpatients Who Are in Remission According to the Hamilton Depression Rating Scale Not Consider Themselves to Be in Remission? *The Journal of Clinical Psychiatry*, *73*(06), 790–795. https://doi.org/10.4088/JCP.11m07203

Zimmerman, M., Mattia, J. I., & Posternak, M. A. (2002). Are Subjects in Pharmacological Treatment Trials of Depression Representative of Patients in Routine Clinical Practice? *American Journal of Psychiatry*, *159*(3), 469–473. https://doi.org/10.1176/appi.ajp.159.3.469

Zimmerman, M., McGlinchey, J. B., Posternak, M. A., Friedman, M., Boerescu, D., & Attiullah, N. (2008). Remission in depressed outpatients: More than just symptom resolution? *Journal of Psychiatric Research*, *42*(10), 797–801. https://doi.org/10.1016/j.jpsychires.2007.09.004

Zimmerman, M., McGlinchey, J. B., Young, D., & Chelminski, I. (2006a). Diagnosing major depressive disorder: II: is there justification for compound symptom criteria? *The Journal of Nervous and Mental Disease*, *194*(4), 235–240. https://doi.org/10.1097/01.nmd.0000207423.36765.89

Zimmerman, M., McGlinchey, J. B., Young, D., & Chelminski, I. (2006b). Diagnosing major depressive disorder IV: relationship between number of symptoms and the diagnosis of disorder. *The Journal of Nervous and Mental Disease*, *194*(6), 450–453. https://doi.org/10.1097/01.nmd.0000221425.04436.46

Zimmerman, M., McGlinchey, J. B., Young, D., & Chelminski, I. (2006c). Diagnosing major depressive disorder IX: are patients who deny low mood a distinct subgroup? *The Journal of Nervous and Mental Disease*, *194*(11), 864–869. https://doi.org/10.1097/01.nmd.0000244564.54694.87

Zimmerman, M., Morgan, T. A., & Stanton, K. (2018). The severity of psychiatric disorders. *World Psychiatry*, *17*(3), 258–275. https://doi.org/10.1002/wps.20569

Zimmerman, M., Multach, M., Walsh, E., Rosenstein, L. K., Gazarian, D., & Clark, H. L. (2016). Problems in the Descriptions of the Psychiatric Inclusion and Exclusion Criteria in Publications of Antidepressant Efficacy Trials: A Qualitative Review and Recommendations for Improved Clarity. *CNS Drugs*, *30*(3), 185–191. https://doi.org/10.1007/s40263-016-0314-y