



Trinity College Dublin

Coláiste na Tríonóide, Baile Átha Cliath

The University of Dublin

**Advancing Mental Health Research Using Data Science:
Investigating Vulnerability to Depression with Language
and Network Analysis**

Sean W. Kelley

School of Psychology and Trinity College Institute of Neuroscience

Trinity College Dublin, University of Dublin

A thesis submitted to

Trinity College Dublin, University of Dublin

For the degree of

Doctor of Philosophy

2023

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Signed: Sean Kelley

A handwritten signature in black ink that reads "Sean Kelley". The signature is written in a cursive, flowing style.

Date: 30/09/2022

Summary

Depression affects over 5% of the global population, yet treatments for depression are only effective in 30-50% of people. Changing the status quo requires an improved understanding of how depression manifests in real life. This thesis tackled methodological shortcomings in the literature to take the field beyond cross-sectional studies in order to develop new technology-based ways to gather rich and repeated data within individuals. We assessed language use patterns from social media and collected self-reported emotions through a series of ecological momentary assessment (EMA) studies. These methods were used to test predictions of depression as a complex and dynamic system allowing us to probe key aspects of network theory, language use, and co-morbidity. Language use on Twitter was shown to be only weakly predictive of depression and not suitable for individual predictions. By modelling mental health as a network, emotion network connectivity - based on EMA data – was found to be primarily related to fluctuations in depression, rather than simply severity. Finally, we used longitudinal time-series data from Twitter as a proxy for EMA to measure longer term changes in network connectivity. Networks constructed from depression-relevant language were found to be more connected during depressive episodes. The studies presented in this thesis, therefore, evaluated depression as a complex system and provide new ways of understanding how depression manifests and changes over time.

Chapter 2 sought to determine if people with depression use language differently on social media and if machine learning over these linguistic features could be used to accurately predict depression severity. Additionally, we tested how specific those predictions would be to depression, compared to other mental health conditions. The results showed that several aspects of language were correlated with depression severity, but machine learning models trained on language used on Twitter was only weakly predictive of depression. Furthermore, machine learning models were not specific to depression. When the depressed model was applied to 8 other mental health disorders, it was found to have similarly weak, but non-zero, predictive performance for most. We illustrate how circular methods for

defining depression (such as from tweet content itself) can lead to overly optimistic accuracy estimates. These results provide evidence that depressed people use language differently on social media, but the size of these effects is such that it cannot be used to make accurate predictions regarding an individual's mental health.

Chapter 3 used ecological momentary assessment (EMA) to estimate personalised emotion networks for 3 large independent community samples (Paid Students, Citizen Scientists, HowNutsAreTheDutch) and a smaller clinical one. Network theory posits that more tightly connected networks are related to increased depression severity. We tested this, in addition to a novel prediction that increased connectivity makes people vulnerable to changes in depression over time. Personalised networks were estimated using vector autoregressive models based on all possible combinations of 5-node networks. In line with predictions of network theory, mean contemporaneous network connectivity was positively associated with depression among Citizen Scientists, HowNutsAreTheDutch, but not Paid Students. In contrast, in all 3 datasets we found connectivity to be related to how changeable depression was over 8 weeks. When controlling for this week-to-week variation in depression, the association between connectivity and severity became non-significant in all samples. These findings demonstrate that elevated network connectivity leads to larger fluctuations in mental health symptoms, not necessarily more severity. This provides a mechanism by which to understand how people's depression varies naturally over time, transitioning into and out of depressive episodes.

Chapter 4 combined the methods used in chapters 2 and 3 to assess whether network connectivity increases during a depressive episode. Preliminary evidence suggests network connectivity increases prior to onset of an episode, i.e., a state transition. These results are based on small samples ($n = 1$) and are not conducted over a long enough time period to encompass the natural onset of depression. As shown in chapter 2, language features are associated with depression severity. Based on this, we used daily language features from Twitter posts, over the past year, as a proxy for EMA. Network connectivity was found to increase significantly within vs. outside a depressive episode. These results were not specific to any particular combination of nodes, instead generalising to a wide

variety of networks. This study demonstrated the usefulness of social media as a proxy for EMA to evaluate otherwise challenging predictions of network theory.

List of Publications and Presentations

This thesis incorporates material already published or currently under review in the following manuscripts:

Kelley, S. W., & Gillan, C. M. (2022). Using language in social media posts to study the network dynamics of depression longitudinally. *Nature communications*, 13(1), 1-11.

Kelley, S. W., Mhaonaigh, C. N., Burke, L., Whelan, R., & Gillan, C. M. (2022). Machine learning of language use on Twitter reveals weak and non-specific predictions. *NPJ digital medicine*, 5(1), 1-13.

Kelley, S.W., Fisher, A., Lee C.T., Gallagher, E., Hanlon, A.K., Robertson, I., & Gillan, C.M. (2022). Elevated Emotion Network Connectivity Leaves People Vulnerable to Fluctuations in Depression. <https://psyarxiv.com/agt5m/>

The following are presentations arising from this work:

Kelley, S. W., & Gillan, C. M. Using Twitter to Predict the Onset of Depression. Poster presentation at Society for NeuroEconomics 2019, Dublin, Ireland.

Kelley, S. W. Within-subject changes in network connectivity occur during an episode of depression: evidence from a longitudinal analysis of social media posts. Oral presentation at Conference on Complex Systems 2020, Virtual.

Kelley, S. W., Mhaonaigh, C. N., Burke, L., Whelan, R., & Gillan, C. M. Can the internet tell that you're depressed? Machine learning of language use on Twitter reveals weak and non-specific predictions. Poster presentation at Society for Biological Psychiatry 2021, Virtual.

Kelley, S. W. Is the Connectivity of Personalised Affect Networks Associated with Increased Propensity for Depression? Invited talk for the R15 Meeting Group of the Levinson Lab, University of Kentucky, Louisville, U.S.

Kelley, S.W., Gallagher, E., Hanlon, A.K., & Gillan, C.M. Personalised Affect Network Connectivity is Associated with Increased Propensity for Depression. Poster presentation at British Association for Psychopharmacology 2022, London, U.K.

Acknowledgments

This thesis would not have been possible without the ample support, mentorship, and guidance that I received throughout. Thank you to my supervisor, Prof. Claire Gillan, for all your advice, patience, and mentorship that you provided me. You have been a fantastic mentor and I'm deeply grateful for all the opportunities you gave me to grow during my PhD. Thank you for allowing me to explore my curiosity in a variety of research topics and methods.

I would like to thank Prof. Robert Whelan for all your help with machine learning and teaching me about novel aspects of data analysis. Thank you to Prof. Clare Kelly and Prof. Frédérique Vallières for serving as my examination appraisers. Thank you to Trinity College Dublin for financially supporting me through the Provost PhD Award. Thank you to Caoimhe Ní Mhaonaigh and Louise Burke for all your help with recruiting participants for the Twitter studies. Dr. Aaron Fisher, thank you for advice, feedback, and guidance regarding EMA in general and specifically for my personalised network study.

I would of course like to thank all past and present members of the Gillan Lab. A special thanks to the original members of the lab, thank you to Dr. Siobhan Harty, Dr. Andrew Pringle, and Dr. Tricia Seow for your advice, support, and generosity. Thank you to Kevin Lynch and Owen Lynch for your help with Neureka and making it a pleasure to come into work every day. To Eoghan Gallagher, thank you for all your help building Multi-Mood, troubleshooting it, and helping me with any and all questions that I had. Thank you to Sharon Chi Tak Lee for all of your questions that helped me to think more deeply about network theory. I would also like to thank Celine Fox, Kelly Donegan, Sara Rahmani, Anna Hanlon, Anna Rosicka, Steph Suddell, and Vanessa Teckentrup.

I would also like to thank my past mentors for their encouragement, time, and advice. Thank you to Dr. Jenny Mascaro for fostering my interest in neuroscience, teaching me about statistics, study designs, and providing extensive support while applying for graduate school. To Dr. Eric Feczko, thank you for teaching me

everything you know about network theory and inspiring a love of complex systems. Thank you to Dr. Lisa Parr, Dr. Bob Siegerink, Dr. Barbara Lundrigan, and Dr. Jack J McDowell for all your support and guidance in learning about research.

I would never have been able to complete this journey without the love and emotional support of my parents. Thank you for everything that you have done for me. Martin, thank you for being there for me through the highs and lows and always being there when I needed it most.

Table of Contents

DECLARATION	1
SUMMARY	2
LIST OF PUBLICATIONS AND PRESENTATIONS	4
ACKNOWLEDGMENTS	5
CHAPTER 1: GENERAL INTRODUCTION	11
1.1 RISK FACTORS FOR DEPRESSION	13
1.2 TREATMENTS FOR DEPRESSION	14
1.3 A NOVEL SOURCE OF DATA: SOCIAL MEDIA	16
1.4 LANGUAGE IS A TWO-WAY STREET	19
1.5 USING MACHINE LEARNING TO IDENTIFY MENTAL HEALTH PROBLEMS	21
1.6 EMOTION ARE NOT STATIC	24
1.7 NETWORK THEORY OF MENTAL HEALTH DISORDERS	25
1.8 KEY FINDINGS IN NETWORK THEORY	30
1.9 MOVING FROM CROSS-SECTIONAL TO PERSONALISED NETWORKS	31
1.10 CLINICAL FEASIBILITY AND UTILITY OF EMA AND PERSONALISED NETWORK MODELS	33
1.11 PERSONALISED NETWORK CONNECTIVITY AND MENTAL HEALTH	35
1.12 NETWORK CONNECTIVITY, RESILIENCE, AND VARIABILITY	37
EXPERIMENTS AND HYPOTHESES	42
CHAPTER 2: MACHINE LEARNING OF LANGUAGE USE ON TWITTER REVEALS WEAK AND NON-SPECIFIC PREDICTIONS	44
2.1 INTRODUCTION	44
2.2 METHODS	48
2.2.1 PARTICIPANTS	48
2.2.2 PROCEDURE	49
2.2.3 UNIVARIATE ASSOCIATIONS WITH MENTAL HEALTH SYMPTOMS	51
2.2.4 MACHINE LEARNING	51
2.2.5 COMPARISON OF VALIDATED SELF-REPORT VS TWITTER-DERIVED GROUND TRUTH	53
2.2.6 TRANSDIAGNOSTIC PSYCHIATRIC DIMENSIONS	53
2.2.7 COMPARISON OF TWEETS, RETWEETS, LIKES	54
2.2.8 SIMILARITY AND SPECIFICITY OF TEXT FEATURES ACROSS MENTAL HEALTH PHENOTYPES	54

2.2.9 THE EFFECT OF NUMBER OF WORDS PER USER ON PREDICTIVE PERFORMANCE	55
2.2.10 STATISTICAL POWER	55
2.3 RESULTS	57
2.3.1 UNIVARIATE ASSOCIATIONS WITH MENTAL HEALTH SYMPTOMS	57
2.3.2 MACHINE LEARNING	59
2.3.3 COMPARISON OF MODELS TRAINED ON TWEETS, RETWEETS, LIKES	63
2.3.4 SIMILARITY AND SPECIFICITY OF TEXT FEATURES ACROSS MENTAL HEALTH PHENOTYPES	63
2.4 DISCUSSION	68

CHAPTER 3: ELEVATED EMOTION NETWORK CONNECTIVITY LEAVES PEOPLE VULNERABLE TO FLUCTUATIONS IN DEPRESSION

3.1 INTRODUCTION	78
3.2 METHODS	80
3.2.1 EXPERIMENT 1: PAID STUDENTS AND CITIZEN SCIENTISTS IN NEUREKA	80
3.2.2 EXPERIMENT 2: REPLICATION IN A LARGE AND INDEPENDENT CITIZEN SCIENCE SAMPLE (HNATD)	83
3.2.3 EXPERIMENT 3: COMPARISON OF CLINICAL AND NON-CLINICAL NETWORK STRUCTURE AND STABILITY	84
3.3 RESULTS	85
3.3.1 NETWORK CONNECTIVITY AND DEPRESSION	85
3.3.2 DEPRESSION VARIABILITY	88
3.3.3 EXPERIMENT 2: REPLICATION IN A LARGE AND INDEPENDENT SAMPLE (HNATD)	90
3.3.4 EXPERIMENT 3: CLINICAL SAMPLE	92
3.3.5 THE IMPACT OF NODE VALENCE	93
3.4 DISCUSSION	94

CHAPTER 4: USING LANGUAGE IN SOCIAL MEDIA POSTS TO STUDY THE NETWORK DYNAMICS OF DEPRESSION LONGITUDINALLY

4.1 INTRODUCTION	98
4.2 METHODS	103
4.2.1 PARTICIPANTS	103
4.2.2 PROCEDURE	106
4.2.3 PRE-PROCESSING AND TEXT ANALYSIS	107
4.2.4 FEATURE SPECIFICATION	107
4.2.5 NETWORK ANALYSIS	108
4.2.6 NETWORK CONNECTIVITY OF A PRIORI NETWORK AND CURRENT DEPRESSION	109
4.2.7 CHANGE TO A PRIORI NETWORK WITHIN VS OUTSIDE DEPRESSIVE EPISODES	110
4.2.8 STABILITY CHECKS	110
4.2.9 GENERALISABILITY OF FINDINGS TO OTHER DEPRESSION NETWORKS	110
4.3 RESULTS	113
4.3.1 ASSOCIATION OF TWITTER TEXT FEATURES WITH CURRENT DEPRESSION SYMPTOMATOLOGY	113
4.3.2 OVERALL DEPRESSION NETWORK COMPOSITION	117

4.3.3 WITHIN-SUBJECT CHANGES IN NETWORK CONNECTIVITY DURING DEPRESSIVE EPISODES	118
4.3.4 GENERALISABILITY OF FINDINGS TO OTHER DEPRESSION NETWORKS	122
4.4 DISCUSSION	124
 CHAPTER 5: GENERAL DISCUSSION	 131
 5.1 SUMMARY	 131
5.2 IMPLICATIONS	135
5.2.1 EARLY WARNING SIGNALS FOR MENTAL HEALTH	135
5.2.2 CLINICAL USEFULNESS AND ETHICAL CONSIDERATIONS OF SOCIAL MEDIA	136
5.2.3 TREATMENT APPLICATIONS OF NETWORK THEORY	139
5.2.4 LIMITATIONS AND FUTURE DIRECTIONS	140
5.2.5 CONCLUSIONS	144
 6 REFERENCES	 144
 7 APPENDICES	 165
 7.1 APPENDIX I: SUPPLEMENTAL INFORMATION FOR CHAPTER 2	 165
7.2 APPENDIX II: SUPPLEMENTAL INFORMATION FOR CHAPTER 3	171
7.2.1 METHODS FOR EXPERIMENT 1: PAID STUDENTS AND CITIZEN SCIENTISTS IN NEUREKA	171
7.2.1.1 PROCEDURE	171
7.2.1.2 DATA PREPARATION	171
7.2.1.3 SENSITIVITY ANALYSES	172
7.2.1.4 PERMUTATION TESTS	172
7.2.1.5 NETWORK STABILITY TEST	172
7.2.1.6 OBSERVED POWER FROM EXPERIMENT 1	173
7.2.2 METHODS FOR EXPERIMENT 2: REPLICATION IN A LARGE AND INDEPENDENT CITIZEN SCIENCE SAMPLE (HNATD)	174
7.2.2.1 PROCEDURE	174
7.2.2.2 DATA PREPARATION	174
7.2.2.3 DATA ANALYSIS	175
7.2.3 METHODS FOR EXPERIMENT 3: COMPARISON OF CLINICAL AND NON-CLINICAL NETWORK STRUCTURE AND STABILITY	175
7.2.3.1 PROCEDURE	175
7.2.3.2 DATA PREPARATION	176
7.2.3.3 DATA ANALYSIS	176
7.2.4 SUPPLEMENTARY RESULTS	179
DIRECTED NETWORK CONNECTIVITY AND DEPRESSION IN EXPERIMENT 1.	179
7.2.4.1 DIRECTED NETWORK CONNECTIVITY AND DEPRESSION IN EXPERIMENT 2.	179
7.3 APPENDIX III: SUPPLEMENTAL INFORMATION FOR CHAPTER 4	188
7.3.1 CHARACTERISTICS OF PARTICIPANTS RECRUITED THROUGH PAID VS FREE CHANNELS	188
7.3.2 RELIABILITY CHECKS ON A 9-NODE NETWORK	189
7.3.3 SENSITIVITY OF OUR ANALYSIS TO THE REMOVAL OF 'SUPRA' CATEGORIES IN LIWC	191

7.3.4 THE EFFECT OF NUMBER OF DAYS ON GLOBAL NETWORK CONNECTIVITY AND ASSOCIATED ANALYSES	193
7.3.5 FREQUENCY OF OCCURRENCES OF THE 9 A PRIORI TEXT FEATURES IN THE 12-MONTH TIME SERIES	195
7.3.6 THE ASSOCIATION BETWEEN NETWORK CONNECTIVITY AND DEPRESSION SEVERITY IS NOT AFFECTED BY REMOVAL OF 3RD PERSON PRONOUNS	197
7.3.7 BOOTSTRAPPED CONTROL FOR UNEQUAL VARIANCES IN WITHIN-EPISODE NETWORK CONNECTIVITY	198
7.3.8 PERMUTATION TEST OF WITHIN-EPISODE IDENTIFIER	199
7.3.9 NO MEAN INCREASE IN THE USE OF 9 A PRIORI TEXT FEATURES OUTSIDE VS WITHIN A DEPRESSIVE EPISODE	201
7.3.10 ASSOCIATION BETWEEN ALL LIWC TEXT FEATURES AND CURRENT DEPRESSION SEVERITY	202

Chapter 1: General Introduction

Depression is one of the leading causes of disability globally (Vos et al., 2020), affecting approximately 5% of the world's population at any given time. Among mental health and substance abuse disorders, depression alone accounts for 40.5% of disability-adjusted life years, significantly more than the second, anxiety (14.6%), and third, illicit drug use (10.9%), largest sources (Whiteford et al., 2013). According to the DSM-V, major depressive disorder is characterised by low mood and/or loss of interest and pleasure in activities for at least 2 weeks. A diagnosis also requires that patients report at least 5 of 9 additional symptoms including: change in weight or appetite, insomnia or hypersomnia, psychomotor agitation or retardation, guilt or worthlessness, difficulty concentrating or indecisiveness, or suicidal ideation. The burden due to depression rises sharply during childhood, peaks in adolescence and early adulthood (Hankin et al., 1998; McLaughlin & King, 2015), and then gradually declines throughout midlife (Sutin et al., 2013). When considering the lifetime prevalence, the probability of experiencing an episode of major depression rises to 16.8% (Kessler et al., 2012). In 2018 in the U.S. alone, the economic burden of depression was estimated at \$326.2 billion, which is an increase of 37.9% over the cost in 2010, despite the proportion of people receiving treatment remaining relatively stable (Greenberg et al., 2021). Roughly 61% of this cost is incurred in the workplace, i.e., from missed days and reduced productivity, while 35% is attributable to direct costs, i.e., inpatient, outpatient, and pharmaceutical costs. In primary care settings, patients with depression used substantially more healthcare services than patients without depression (Pearson et al., 1999). Identifying means to reduce depression's substantial burden would deliver considerable social and individual benefits.

Despite this, the development of novel mental health treatments has stagnated over the past several decades. Research into antidepressants has not produced a new pharmacological agent since the 1980s. While the core principles of cognitive behavioural therapy, developed by Aaron Beck in the 1960s, remain largely unchanged. One explanation for the lack of progress and innovation in this space is the challenge in designing and implementing one-size-fits-all treatments. There exists tremendous individual variation in how people experience the world and how those experiences affect their mental health. Some people become depressed easily, after a minor stressor, while others can seemingly bounce back from highly traumatic events (Alim et al., 2008; Collishaw et al., 2007; Wingo et al., 2010). These sort of adverse childhood experiences (ACEs) are a well-established risk factor for mental health disorders, but only explain only a small proportion of overall variance (Kendler et al., 1998, 1999), much like polygenic risk, or brain imaging markers (Winter et al., 2022). More ACEs lead to a 1.14x greater risk of developing a mental health condition later in life, yet this translated into poor predictive accuracy (AUC of 0.58) (Baldwin et al., 2021). The emerging picture in mental health science is that individuals with depression are exceptionally complex, they not only arrive at a diagnosis in unique ways, but actually experience depression in unique ways too (Eiko I Fried & Randolph M Nesse, 2015a). Among patients seeking treatment for depression in the STAR*D study, there were 1,030 unique depression symptom profiles. Even the most common profile occurred in less than 2% of people, while fully 85% of total symptom profiles were endorsed by 5 or fewer patients. This complexity makes depression challenging to understand and predict from studies with small sample sizes that measure a relatively small set of variables. To overcome these challenges, a data science revolution in mental health research is underway, aiming to leverage large samples and multivariate datasets to bring new interventions to clinical practice. It is thought that big data coupled with longitudinal monitoring can be used in the near future to personalise treatment and predict future changes in healthy people. This thesis aimed to progress this field through a series of studies that use new methods for tracking rich and complex mental health data through time and at - scale. These include analyses of language use on social media and smartphone-based assessments. We applied these methods to answer important questions about how depression manifests in real life, including whether we predict an

individual's mental health status from language on social media, how language changes during depressive episodes, and understanding how symptom dynamics and interactions can cause some individuals to experience more fluctuations in depression severity.

1.1 Risk Factors for Depression

A substantial body of evidence has identified demographic factors associated with an elevated risk of depression, including: age, gender, socioeconomic status, and marriage. Depression is roughly twice as prevalent in women compared to men (Kessler et al., 1993). Lower socioeconomic status is associated with an elevated risk of developing major depression (Lorant et al., 2003; Sareen et al., 2011; Wang et al., 2010). Weaker social ties, lower perceived support, and greater loneliness are all related to depression (George et al., 1989; J. Wang et al., 2018). While higher rates of education have a protective effect, particularly in young adults, against depression and anxiety (Bjelland et al., 2008). Older adults generally have a lower prevalence of depression than younger ones, although there is an uptick in depression rates during late-life (Kessler et al., 2010; Luppá et al., 2012; Sutin et al., 2013). Consequently, even relatively basic risk factors like age and gender provide a useful starting point for understanding what influences the incidence and future course of depression onset and severity.

Childhood adversity is another well-established risk factor for the development of major depression (Felitti et al., 1998; Kessler et al., 1997; Kessler & Magee, 1993). Exposure to adverse childhood experiences is relatively common in the population, with prevalence rates of 52.1% (Felitti et al., 1998) to 63.6% (Dube et al., 2001) for experiencing any type of abuse. Adverse events are thought to cause depression by lowering the threshold for transitioning between health and mental illness, by making people less resilient. Individuals with more adverse experiences in childhood were more likely to become depressed when exposed to lower stress levels than those with fewer adverse experiences (Hammen et al., 2000). Evidence suggests that there is a dose-response relationship between the

number of childhood adverse events and probability of developing depression (Afifi et al., 2008; Chapman et al., 2004; Merrick et al., 2017). The relationship between adverse events and depression is then moderated by psychological resilience. Individuals with lower resilience experience more depression symptoms than more resilient people with the same number of adverse events (Poole et al., 2017). Adverse childhood experiences not only increase the likelihood of depression but make it more treatment resistant (Tunnard et al., 2014). These effects are, of course, not limited to childhood. Stressful life events later in life, e.g., financial problems, also contribute to the onset of major depression (Kendler et al., 1999).

Despite known risk factors and a defined set of symptoms, diagnosing depression is not straightforward. General practitioners are responsible for treating the majority of mental illness cases (Coyne et al., 2002; Olfson et al., 2002; Regier et al., 1978), but have a poor track record; only correctly diagnosing depression in 30% of patients (Ani et al., 2008). When physicians considered functional impairment together with depression symptoms, the rate of correct diagnosis rose marginally to 37%. Many have suggested that by using these risk factors, it may be possible to implement routine digital screenings for depression in primary care. To achieve this, we need to develop reliable, accurate, and valid models that can be easily administered in the clinic.

1.2 Treatments for Depression

Once a person is diagnosed with depression, they will typically receive psychotherapy, e.g., cognitive behavioural therapy, or a pharmaceutical drug, e.g., selective serotonin reuptake inhibitors (SSRIs), or some combination thereof. Antidepressants are most effective in patients with severe depression, and are no different from placebos for mild to moderate depression (Fournier et al., 2010). Despite an increase in antidepressant usage, first line antidepressants lead to remission in only 30-50% of patients (Rush, Trivedi, et al., 2006). Furthermore, a substantial proportion of patients, up to 30%, do not respond to any treatment (Al-Harbi, 2012; Fava, 2003). Side effects caused by antidepressants are a primary reason as to why patients prematurely discontinue treatment (Anderson & Tomenson, 1995; DEWAN & Anand, 1999; MacGillivray et al., 2003). Because of

this, treatment-seeking patients report a strong preference for psychotherapy over antidepressants, by a margin of 3 to 1, with women and young people expressing the greatest preference (McHugh et al., 2013). Even with this preference, the proportion of people receiving psychotherapy has declined over time with an increasing number receiving only medication (Marcus & Olfson, 2010; Olfson & Marcus, 2009; Olfson et al., 2002). While pharmacological and therapy-based solutions are the primary evidence-based treatments for depression, behavioural health interventions, in particular physical exercise, also significantly improve mental health for mild to moderate depression (Chekroud et al., 2018; Craft & Landers, 1998; Dunn et al., 2005).

If a person can access and complete treatment, that does not mean they are free of depression, since it is rarely 'cured'. One year after treatment, 40% of people will have a recurrence of depression. The subsequent risk of re-occurrence then increases with each additional depressive episode (Burcusa & Iacono, 2007; Solomon et al., 2000). Patients with residual symptoms after treatment are at a higher risk of relapse compared to patients without any residual symptoms (Paykel et al., 1995). Delaying treatment exacerbates these issues and results in lower remission rates (Bukh et al., 2013), more severe depression symptoms (Hung et al., 2017), more frequent relapses (Altamura et al., 2007), and a longer overall course of depression (Altamura et al., 2007). There's been relatively little research on the potential for targeting depression interventions prior to depression onset, due to difficulties in predicting the occurrence of depressive episodes. But enhanced care after treatment, beyond basic management, has been shown to improve mental health outcomes. Identifying which patients are at risk is crucial for treatments to be effective. Among healthy adolescent students, providing mindfulness training during the COVID-19 pandemic did not significantly change mental health outcomes at any point (Dunning et al., 2022). Patients who received a relapse intervention treatment, with active monitoring of depression symptoms, benefited by experiencing two additional depression free weeks over the subsequent year (Simon et al., 2002). While patients in primary care who received a relapse prevention programme experienced both significantly fewer depression symptoms (Katon et al., 2001) and increased remission rates (Rost et al., 2002).

Despite a high prevalence in the community and high cost to healthcare systems, depression is frequently either not treated at all or inadequately treated. Of people with depression, only 33% received some form of mental health care in high-income countries (Moitra et al., 2022). However, as previously mentioned, even evidence-based treatments are not effective for everyone. Because of this, and limited resources due to the cost and difficulty of providing treatment, it is important to personalise treatment allocation to the people most likely to benefit. Depression treatment personalisation should be only one part of a data-driven pipeline for improving mental health outcomes. By monitoring patients' relapse risk after treatment cessation, depressive episodes could be identified before they ever occur. At which point prophylactic interventions could be delivered just in time to mitigate the episode's occurrence or severity. A key challenge we face is that all of these efforts to personalise depression care and management require large amounts of within-individual data for both the development of these models and their implementation in routine care.

1.3 A Novel Source of Data: Social Media

Achieving the sample sizes and rich datasets required to make data science happen for mental health is difficult using traditional methods. Studies with large samples take years to recruit, often at enormous expense, and still can only monitor participants sporadically. We need to develop novel approaches to data gathering that are less burdensome for patients, richer, and more widely available. Many research groups are doing just that, investigating wearable devices (Jacobson & Chung, 2020; Pratap et al., 2019), electronic medical records (Coley et al., 2021; Nemesure et al., 2021), and other passive means of data gathering to dramatically scale up our access to data (Torous et al., 2016; Vaidyam et al., 2019). In this thesis, we examined another promising approach that focuses on the repurposing of enormous existing open datasets with millions of users – that is, social media data. Social media has exploded in use over the past 20 years; in 2021, 72% of Americans used at least one social site up from only 5% in 2005 (Auxier & Anderson, 2021). Although most popular among young adults, over time the percentage of all adults on social media has risen. Amidst this changing communication landscape, people now widely share and convey their personal

thoughts, images, and videos through a variety of platforms. Every post offers a potential reflection of a person's internal mental health and thought processes. Through sites like Facebook and Twitter, users post and archive records of their thoughts and feelings over time. By extracting posts from millions of users, social media can leverage large samples and longitudinal data to test hypotheses that are otherwise difficult or impossible using traditional lab-based methods.

Although social media provides several potentially useful variables, e.g., the time of day that people post, their social contacts, interaction frequencies, features of the images they post, and more (Chancellor & De Choudhury, 2020), perhaps the richest and most informative information comes from the language that people use. Accumulated evidence across a variety of data sources has found that both individual and population-level traits can be predicted using digital traces of language. Language used on Facebook and Twitter can predict age and gender to high degree of accuracy, Pearson correlation of 0.83 for age and 91% accuracy for gender (Sap et al., 2014). Personality traits that comprise the unique behaviours, actions, and thoughts that make people individuals are also linked to particular patterns of language. The Big Five personality traits, i.e., agreeableness, extraversion, neuroticism, conscientiousness, and openness, are associated with distinct patterns of language features (Kern et al., 2014; Schwartz et al., 2013). Besides personality, sensitive and stable individual traits such as political orientation, sexual orientation, and drug use are also predicted by Facebook language (Kosinski et al., 2013). Along with making individual-level predictions, social media can monitor macroscale changes in health at the population-level. One novel application of this is a hedonometer which uses a stream of real-time English-language Tweets to provide a continually updated picture of global happiness (Dodds et al., 2011). Furthermore, language not only relates to mental health but physical health as well. Within the U.S., language used on Twitter can predict the incidence of heart disease mortality at the county level (Eichstaedt et al., 2015). These studies demonstrate that language on social media can passively monitor a variety of health and psychological well-being metrics.

Recent research has established it may also prove useful for monitoring mental health, because of well-established systematic differences in language use across psychiatry, akin to a digital footprint. For example, several studies have shown that

people with depression use significantly more 1st person singular pronouns when writing and speaking. University students who were currently depressed used significantly more 1st person singular pronouns, e.g., “I”, “me”, “my”, than students without depression when asked to write about their thoughts and feelings regarding college (Rude et al., 2004). First person singular pronouns are also predictive of future depression severity, even after controlling for baseline depression (Zimmermann et al., 2017). A meta-analysis found that 1st person singular pronouns have a consistent, albeit relatively small ($r = 0.13$), positive association with depression (Edwards & Holtzman, 2017). Elevated use of 1st person singular pronouns is thought to reflect greater self-focus and rumination, which depression causes through changes in cognitive processes. First person pronoun usage is not only a downstream effect of altered cognitive processing, but may actively play a causal role in influencing how individuals process external stimuli. Kross et al. (2014), studied the effect of first-person pronoun use on negative affect and threat appraisal (Kross et al., 2014). In the first part of the study, prior to giving a speech, participants were asked to prepare themselves by reflecting on their thoughts and feelings using either as many first-person or non-first-person pronouns (person’s own name) as possible. People in the first-person group had a significant increase in negative affect after the speech compared to those in the non-first-person group, which buffered against any changes in negative affect. As an extension of the first study, participants were asked to assess their anticipatory anxiety and threat prior to giving a speech. Participants randomised to the non-first-person group experienced less challenge and anticipatory threat in response to their speech than the first-person group. Flexibility in writing style, especially frequency of particles and pronouns, is associated with fewer medical visits among people asked to write about their emotional thoughts and sentiments (Campbell & Pennebaker, 2003). These results demonstrate that language does not just passively reflect a person’s current mental health status.

Language differences in people with depression are not restricted to elevated use of first-person singular pronouns. Within English-language mental health forums, individuals in depression relevant forums express significantly more absolute words, e.g., certain, always, than those in general interest groups (Al-Mosaiwi &

Johnstone, 2018). While the topics, i.e., words that co-occur together, most associated with future depressive status were hostility, depressed mood, somatic complaints, and hopelessness (Eichstaedt et al., 2018; Schwartz et al., 2014). Perhaps unsurprisingly, individuals with depression also use more negative emotion words (Coppersmith et al., 2014; Coppersmith, Dredze, Harman, & Hollingshead, 2015) and express more distorted thinking (Bathina et al., 2021). Relative to healthy controls, people with depression exhibit more personalising (belief that others behave negatively due to one's actions), emotional reasoning (thinking based on feelings rather than evidence to the contrary), and overgeneralizing (making broad negative conclusions). The association between depression and cognitive distortions held even after removing 1st person singular pronouns.

1.4 Language is a two-way street

While social media opens a window into an individual's emotional well-being, communication on social media is never one-sided. Interactions with posts through comments, likes, and direct messages by friends, family, and strangers can have a profound positive, or negative, influence on our mental health. Self-disclosure of an illness is the first step towards receiving help and signifies an openness towards receiving support. Approximately 40% of Facebook users sought out health-related support on the platform, demonstrating that social media is not only a viable forum for identifying, but also receiving, support (Oh et al., 2013). People who asked for more health-related support also reported receiving greater social and tangible support from their Facebook friends. By removing inhibition, building trust and providing emotional release, self-disclosure produces a range of benefits to mental and physical health (Pennebaker, 1997). Reddit users provide high quality emotional, informational, prescriptive and instrumental feedback to users who self-disclose on mental health subreddits (De Choudhury & De, 2014). Posts with more content related to relationships, social aspects, and health and less inhibition were the most likely to receive support via comments. The benefits of self-disclosure extend to conditions beyond depression. People with schizophrenia had elevated linguistic complexity, topical coherence, and greater orientation

towards the future after, compared to before, they disclosed their diagnosis on Twitter (Ernala et al., 2017).

Another important way that language can affect, and not just measure, depression is in therapeutic settings. When therapists engage in cognitive behavioural therapy with their clients, their language has a significant effect on treatment outcomes. Therapeutic praise and planning for the future is associated with the largest improvement in treatment outcomes, while phrases related to risk assessment and therapeutic empathy are primarily related with worse outcomes (Ewbank et al., 2020). Monitoring therapist language and ensuring it adheres to evidence-based practices could potentially improve treatment outcomes, since approximately 8% of variance in treatment outcome can be attributed to the therapist (Kim et al., 2006; Lutz et al., 2007). Therapists tend to drift away from the core elements of psychotherapy as they apply their training in clinical practice, which often means that patients do not receive key interventions, e.g., therapists become hesitant to push for behavioural changes (Waller, 2009; Waller & Turner, 2016). As a consequence, experienced therapists can sometimes underperform newer ones (Shapiro & Shapiro, 1982).

To understand why some therapists, achieve better outcomes than others, it may be important to look at changes in language use during the therapeutic process. Therapist language does not remain static, but rather changes dynamically throughout a session, in response to their patients, and these patterns are stable across sessions within a therapist (Miner et al., 2022). Crisis-line counsellors have more successful conversations if they are more adaptable, react more to ambiguous messages, and respond more creatively (Althoff et al., 2016). Over the course of a conversation with a counsellor, people generally tend to talk less about the past and more about the present and future (Althoff et al., 2016). People who reported feeling better after the conversation, however, were less focused on the present and more on the future. They also used fewer 1st person singular pronouns than those who did not feel better at the end, reflecting a reduced self-focus. Given these observations, it is thought that monitoring language used dynamically during therapy, by patients and therapists, could improve psychotherapy delivery, adherence, and therefore overall efficacy. However, privacy concerns, patient confidentiality, and logistical challenges make it difficult

to acquire rich language datasets of this sort and test these novel ideas. Social media can circumvent some of technical challenges by providing researchers and therapists with access to rich timeseries of language data, at low burden to the participant, drawing on data that they are comfortable sharing with others online. There are many approaches to the analysis of such data, but perhaps none as popular as machine learning.

1.5 Using Machine Learning to Identify Mental Health Problems

Consistent evidence over the past 10 years shows that mental health status can be predicted from language on social media. In 2013, De Choudhury *et al*/ accurately classified people with and without depression using language on Twitter from one year prior to the reported onset (De Choudhury, Gamon, et al., 2013). Prediction models trained on Facebook data were comparable in performance to models trained on medical records (AUC of 0.69), and were able to predict a depression diagnosis up to 3 months prior (Eichstaedt et al., 2018). Similarly, in a study of Twitter users, the probability of being diagnosed with depression steadily increased in the 3-month period prior to receiving a diagnosis (Reece et al., 2017). Besides depression, social media language use can predict other mental health conditions including: anxiety, schizophrenia (Birnbaum et al., 2017; Ernala et al., 2017), suicide, eating disorders (Chancellor et al., 2016; De Choudhury, 2015), and post-traumatic stress disorder (Benton et al., 2017; Coppersmith, Dredze, Harman, Hollingshead, et al., 2015). Depression is the most commonly studied mental health condition on social media, accounting for almost one third of research studies (Chancellor & De Choudhury, 2020). Among these studies, Twitter is the most popular platform for studying mental health. For example, Twitter was used to assess the impact of antidepressant medication on positive and negative affect, cognition, and provide insights into which individuals, based on language use patterns, may benefit most from particular types of antidepressants, i.e., SSRIs vs Tricyclics (Saha et al., 2019). These results show that social media can model mental health status at scale and therefore hold great promise to address hypotheses that are challenging with traditional lab-based recruitment strategies.

Although language on social media can predict depression and other mental health conditions with a high degree of accuracy, few studies have investigated how specific these models are to depression versus other conditions. In mental health, comorbidity is the rule rather than the exception. Comorbidity estimates show that depression is highly correlated with panic disorder, agoraphobia, generalised anxiety disorder, post-traumatic stress disorder, obsessive-compulsive disorder, and seasonal affective disorder (Kessler et al., 2005). Having a mental health condition not only leads to poor health outcomes, but it exacerbates existing chronic illness. Having depression comorbid with a chronic illness results in worse overall health than only having either a chronic illness or depression (Moussavi et al., 2007). People with a mental health disorder and comorbid general health problem are 5.90 times more likely to die than the general population and live on average 11.35 years fewer than those without a diagnosis (Momen et al., 2022). This may be partially explained by lower medical treatment compliance in patients with depression (DiMatteo et al., 2000). A comparison of language across multiple mental health conditions found little specificity among individual text features (Coppersmith, Dredze, Harman, & Hollingshead, 2015). For example, 1st person singular pronouns were elevated in people with not only depression, but anxiety, eating disorders, and OCD. Some evidence for specificity comes from Gkotsis *et al* (2016), who found that mental health related subreddits have distinct language patterns with their own topic-specific vocabularies (Gkotsis et al., 2016). The problem here of course is that language used on a Reddit support forum for depression is necessarily different in *content* compared to another forum focused on schizophrenia, which creates circularities that can bias results. To put this another way, language used when talking about a specific topic may be quite different to language used daily life and this makes findings lack generalisability. Even in patients with co-morbidity, when engaging in a specific online support forum pertaining to one of their mental health problems, individuals are selectively choosing to speak about that aspect of their life experience. Language used by individuals within these forums is therefore not representative of their language in everyday life, but particular to one context.

Similar issues with establishing a 'ground truth' of depression were highlighted in a paper that compared Twitter and Facebook language model performance using

datasets trained on 4 types of mental health outcome measures for schizophrenia: (i) affiliation (i.e., followers of a schizophrenia related account on Twitter), (ii) self-disclosure of schizophrenia diagnosis (e.g., “I was diagnosed with Schizophrenia”, (iii) external validation of self-reported disclosure using clinical judgement, and (iv) patients with a confirmed diagnosis of schizophrenia – used to establish a ‘ground-truth’ state of mental health (Ernala et al., 2019). Machine learning classifiers trained on proxies of mental health (i.e., affiliation, self-report and appraised self-report) performed well on measures of internal validity but had poor external validity compared to the patient trained model. Notably the patient data trained model had similar performance when tested on unseen data as it did during cross-validation, with the largest discrepancy in model performance occurring between the affiliation and patient model. Even when compared to self-disclosure and clinically appraised self-disclosure models, this study found that measuring mental health through community affiliation is a particularly poor method of capturing a person’s ‘true’ mental health status. This is a problem for many large-scale studies that rely on readily (and publicly) available information to establish diagnosis and train models. An accumulation of evidence suggests that these proxy signals for mental health have poor construct validity and should be treated with caution.

But what is the alternative? Part of the reason research has relied on proxy signals of mental health is due to the difficulty in acquiring high quality alternatives.

Typically, studies involve a trade-off between the sample size and the quality of these measures. Although self-report questionnaires are the gold-standard, they are difficult to administer to large sample sizes. As a result, studies that require large numbers of participants typically default to using a self-disclosure, community membership, or keywords to establish cases of mental illness.

Aside from the quality of ground-truth measures, another consideration for research of this sort is more existential in psychiatry. There is broad agreement that our current diagnostic system does not correspond to biological mechanism nor is it predictive of treatment response (Insel et al., 2010). Moving towards a neuroscience-based approach will shift the focus away from symptom presentations because people do not fit neatly into diagnostic categories. They present with symptoms that co-occur across several conditions and the expression of symptom severity appears more continuous than categorical (Haslam, 2003;

Markon et al., 2011). This presents a problem for the binary classification scheme that is frequently used in predictive models. To the former issue, classification models trained on case-control data cannot assess the specificity of language features to any one condition. For that, machine learning models need to be trained on one disorder and then have their predictive performance evaluated on multiple other conditions. An alternative way to achieve that, and move beyond binary classification more generally, is to shift focus towards gathering self-report data for several conditions for each person in a study. Chapter 2 of this thesis therefore adopted a self-report approach to estimate how accurate and specific language models are at predicting the severity of mental health conditions.

1.6 Emotion Are Not Static

The first sections of this introduction focused primarily on using summary measures of language to predict mental health status. While this has considerable practical advantages when we think about implementation in a clinical setting as part of a decision-tool, it fails to capture the fact that emotional states are not static, they ebb and flow and respond adaptively to changes in an individual's environment. Downregulation and upregulation of emotions are crucial to reacting appropriately to changing environmental factors. Disruptions to emotional regulation can lead to maladaptive behaviours, poorer psychological well-being, and reduced flexibility (Houben et al., 2015; Kashdan & Rottenberg, 2010; van de Leemput et al., 2014; van Os et al., 2017). A range of emotion dynamics have been studied in the context of mental health, including emotional inertia, variability, and instability. Emotional inertia is the tendency for emotions to carry over from one time to another, which is operationalised as the autocorrelation of a particular emotion (Suls et al., 1998). Elevated emotional inertia has been linked to both current depression and as a prospective risk factor for its future development (Koval et al., 2012; Kuppens et al., 2012). A consequence of elevated emotional inertia is that when our mood worsens following a negative event, it can take longer to 'recover' back to baseline (Koval et al., 2015). Along with elevated emotional inertia, people with depression have more variance in their emotions. Koval et al., 2013 found that the variance (and autocorrelation) of negative affect

were significantly associated with depression in two independent studies, an EMA one and a separate film-clip task (Koval et al., 2013).

It may be surprising that both inertia and variance are associated with more protracted states of depression. This can be explained by the fact that instability (mean successive squared difference; MSSD), variability (standard deviation; SD), and emotional inertia (autocorrelation; AR) are not mathematically independent as shown by (Jahng et al., 2008). For example, associations between depression and emotional inertia do not survive controlling for greater variance in negative affect (which also correlates with depression severity). This result was subsequently replicated by Bos, de Jonge, and Cox 2019 (Bos et al., 2019). Both studies also found that the significant relationship between depression and negative affect variability disappeared after controlling for mean levels of negative affect. Whether dynamic emotional features add additional predictive value for predicting psychological well-being, borderline symptoms, life-satisfaction, and depression beyond mean levels of affect was investigated in a meta-analysis of 15 EMA studies (Dejonckheere et al., 2019). Controlling for mean affect dramatically reduced the effect size of the association between measures of emotional dynamics and depression. What these results show is that emotional dynamics do not appear to add significant explanatory power beyond mean levels.

Although this line of research on emotion and depression dynamics is an interesting framework for starting to think about the evolution and expression of symptoms in real life, one limitation is that it does not account for potential interactions among emotions. Studies have begun to highlight the importance of multivariate, dynamic models, with studies showing for example that an increasing correlation strength between positive and negative affect is related to future changes in depression severity in a between-participant analysis (van de Leemput et al., 2014). In the next section, I will cover this area in some depth as I outline the network theory of depression.

1.7 Network Theory of Mental Health Disorders

Mental health conditions have traditionally been modelled using a latent-cause framework. In a latent cause model, all symptoms of a condition are due to a

singular disease state (Figure 1.1A). For example, an infection of the influenza virus can lead to a fever, sore throat, and cough; treating the disease will eventually cause symptoms to disappear, while treating symptoms does not alter the underlying disease state. In many conditions related to physical health, this is a reasonable way to model the emergence and maintenance of symptoms. When applied to mental health conditions, however, several of the common cause model's primary assumptions begin to breakdown. Early-stage cancer is an example in which a person has a disease without the presence of any symptoms. Yet depression, along with other mental health conditions, is diagnosed exclusively by the presence of symptoms. While a latent cause model assumes that symptoms are completely independent of each other, and have no causal role in the activation of other symptoms, we know that this is not true in the case of depression. Depressive symptoms influence each other over time, e.g., having a low mood can lead to feeling more tired (Borsboom & Cramer, 2013). Individual depression symptoms also have different effects on psychosocial functioning, e.g., home management, which demonstrates the importance of considering individual symptoms above and beyond information provided by a summed scores (Fried & Nesse, 2014; Eiko I Fried & Randolph M Nesse, 2015b).

Recent work on a network theory of depression posits that depression emerges due to causal interactions among symptoms. Positive feedback loops between symptoms lead to either the development or maintenance of a depressive episode (Figure 1.1B). In a depression network, symptoms are 'nodes' while edges between symptoms are estimated. Node properties are quantified through a series of centrality estimates including: strength, closeness, and betweenness centrality. Strength centrality is the sum of the absolute edge strengths into and out of a node. Closeness centrality is the sum of the shortest paths from one node to all other nodes in a network. Unlike strength centrality, which only accounts for a node's influence on its most proximate neighbours, closeness centrality measures a node's influence on all other nodes in a network. Finally, betweenness centrality is the number of shortest paths that passes through a given node. Nodes with a high betweenness centrality act as 'bridges' in a network, in that they determine the possible pathways for information to flow from one node to another.

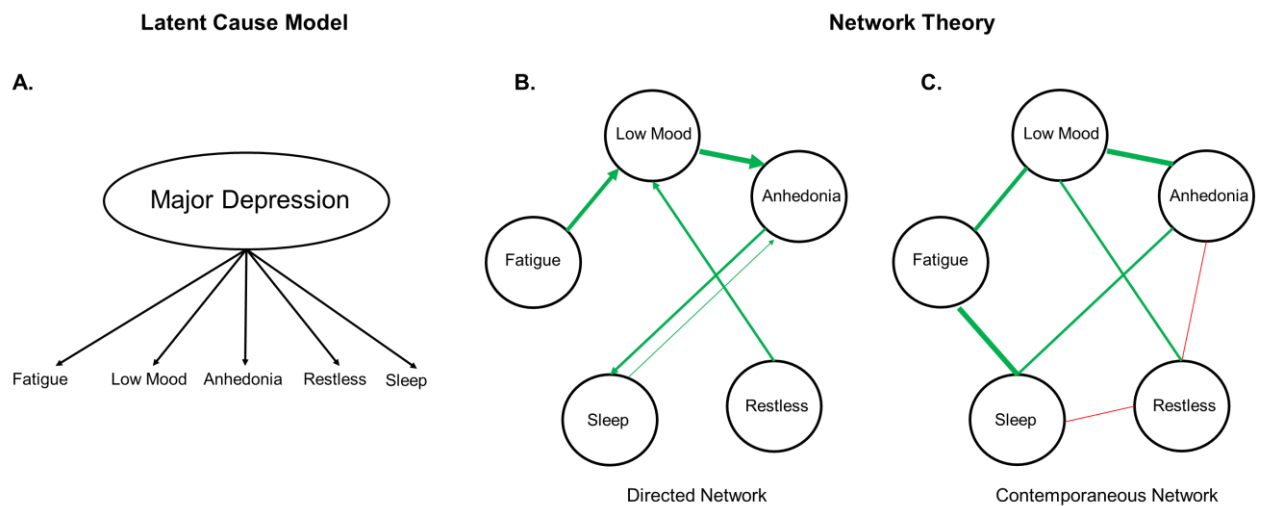


Figure 1.1: Schematic diagram to illustrate the difference between a latent cause (A) and network theory (B&C) model of depression.

Panel B is a directed network, which estimates all possible the lag-1 associations between symptoms. While panel C is a contemporaneous network, which represents the association between symptoms at the same point in time after controlling for any temporal effects.

Centrality estimates are used to rank a node's relative importance within a network. Nodes that are strongly connected to other symptoms, i.e., high centrality, are thought to exert stronger effects on other symptoms. Identifying influential symptoms could potentially improve clinical treatments by targeting the symptoms that are most likely to lead functional improvement. How useful central symptoms are as potential treatment targets, however, remains an open question. Early evidence suggests that symptom centrality is related to how strongly a change in an individual symptom predicts changes in other symptoms within a network (Robinaugh et al., 2016; Rodebaugh et al., 2018). For example, in cross-sectional network studies, sadness is typically the most central symptom and changes in it correlate most strongly with reductions in depression severity. While promising, concerns still surround the use of centrality indices as a proxy for a node's causal influence on other symptoms. Central nodes are frequently not the nodes with the greatest causal impact on the rest of the network (Dablander & Hinne, 2019). In a simulation study, Dablander and Hinne (2019) generated a true causal network structure, based on a directed acyclic graph (DAG), among a given set of nodes and then found the undirected partial correlation networks, essentially

a cross-sectional network, based on that ‘true’ casual structure. The authors found that centrality indices vary in their capacity to recover casual relationships between nodes depending on several factors including: network model (e.g., small-world network), density, and number of nodes. For example, in networks with less than 20 nodes, a node’s centrality indices correlated modestly with its true causal impact. As both the network density and number of nodes in a network increased, the correlation between a node’s centrality (i.e., degree, strength, closeness, betweenness, and eigenvector) and causal impact decline, and even become negative, for all centrality measures except eigenvector centrality. These results indicate that centrality measures may not accurately reflect a node’s true causal impact. An important limitation, however, is that by modelling causal interactions under DAG framework do not allow for feedback loops, i.e., autocorrelation, which is a known property of psychological symptoms.

While we now have the statistical tools to model symptom networks, there remains the problem of how to intervene on these networks to affect change. Psychological interventions, unlike medical or surgical ones, are not specific and affect many variables simultaneously – a condition termed “fat-handed” (Eronen, 2020). Symptoms have multiple causal factors that influence their current expression, e.g., whether insomnia severity is high, such that it is not possible to completely control its level of activity. For example, drinking a cup of coffee is not the only cause of feeling alert, much like the death of a loved one is not the only cause of feeling sad. As a result, psychological interventions violate two assumptions of causal inference: i) only affect a single target variable, X, on outcome Y and ii) all other variables not on the causal pathway between X and Y are held fixed.

Although these centrality indices provide a framework for understanding symptom importance and functionality within a network, there remain unresolved methodological limitations from their use in mental health research. Network theory assumes that all nodes are interchangeable and do not have unique properties, i.e., a node is identical to all other nodes in the network. In a psychological network, nodes are not interchangeable because they measure qualitatively different symptoms, e.g., weight gain versus low mood (Snijders, 2011). Furthermore, nodes derived from questionnaire items also tend to be less distinct than nodes in other types of networks, e.g., unique individuals in a social

network, due to multicollinearity between items (e.g., sadness tends to be correlated with guilt). As a consequence, the relative rank ordering of symptoms becomes less clear among conceptually similar nodes (Bulteel et al., 2016). An implicit assumption of centrality indices is that they are capturing some type of flow process through a network (Borgatti, 2005). Different types of flow processes that have been identified include: serial, transfer, and parallel. Information processes tend to flow in parallel, where the information is sent to many nodes at once. For example, sending a Tweet goes out to all of your followers simultaneously without intermediate carriers. Serial processes involve sequential transfer from one node to another, such as the spread of COVID-19 through a person's social network. Transfer flow is the actual distribution of a physical item between nodes, e.g., airplanes flying into and out of airports. Compared to serial and parallel flow processes, transfer flow processes are optimised to travel along the shortest path lengths. Closeness and betweenness centrality are measures that assume that the flow process takes the shortest path possible between nodes, akin to a physical transfer of an object across a network. Closeness centrality, on the other hand, makes no assumptions regarding the type of flow, i.e., serial or parallel, as long as it travels via the shortest path. Due to these assumptions, closeness and betweenness centrality have less relevance to mental health networks than simpler indices like strength centrality.

One of the primary differences between psychological networks and other types of networks is that in psychological networks edges are estimated rather than observed. Edges are typically estimated with either regularised partial correlations when using group-level data (undirected edges) or lag-1 vector autoregressive (VAR) coefficients in intra-individual data (directed edges). Psychological networks also contain a mixture of positive and negative edges. Though one thing we have observed in the literature is that often when centrality measures are applied to psychological networks, the absolute value of edge strength is used. This may be problematic as taking the absolute value ignores valuable information about the direction of associations between nodes. In a network of depression symptoms, a negative edge between symptoms would indicate a lower risk of symptom activation, e.g., sleep reducing tiredness, thereby providing a protective effect by preventing the activation of other symptoms in the network. Considering the

direction of edge strength provides a better overview of network connectivity and the interplay between symptoms and is therefore an approach adopted throughout this thesis.

1.8 Key Findings in Network Theory

The network theory of mental health makes several predictions regarding the relationship between network connectivity and mental health and growing body of literature has begun to test these. Firstly, elevated network connectivity is hypothesised to precede a transition from one state to another (Dakos et al., 2010) and is thought to be a possible risk factor for future deteriorations in mental health. A depression network simulation study showed that individuals with more strongly connected networks are expected to spend the majority of their time in a depressive state, i.e., experience more depression symptoms, even in the absence of external perturbations (Cramer et al., 2016). Interestingly, individuals with moderate levels of network connectivity are not restricted to either a healthy or depressed state but instead experience transitions between them. Depending on network connectivity strength, external perturbations, e.g., stress, lead to the emergence of different patterns of depression symptoms. Weakly connected networks can respond gradually to increasing stress resulting in a continuous increase in the number of depression symptoms. In more tightly connected networks, increasing stress levels instead cause a sudden and sharp increase in the number of symptoms without passing through a zone of intermediate symptom severity first.

Preliminary evidence suggested that people with greater depression severity have elevated depression network connectivity (Boschloo et al., 2015; Lee Pe et al., 2015; Claudia van Borkulo et al., 2015). Although a large study of participants with (N = 595) and without (N = 5,998) a diagnosis of major depression found no significant difference in network connectivity and only minor differences in network structure between groups (Hakulinen et al., 2020). More recent studies have consistently shown an increase in depression network connectivity after treatment, even though depression severity decreases (Beard et al., 2016; Berlim et al., 2020; F. M. Bos et al., 2018; Snippe et al., 2017). There are yet to be any definitive explanations for why treatments cause an increase in network

connectivity. One account is that over the course of treatment, the variance of individual symptoms changes, which then affects the covariance between symptoms (Fried et al., 2016; Terluin et al., 2016). Another possible explanation is that individuals included in depression studies have high baseline depression and then regress towards their mean score over time.

Most studies in the literature are based on cross-sectional data, which limits the ability to make causal inferences at the individual level (Robinaugh et al., 2020). Between-subjects effects do not generalise to individual statistical estimates unless the underlying process is ergodic, i.e., have a stable mean and variance over time (homogeneous) and no systematic change over time (stationary) (Molenaar, 2004). A comparison of intra- vs inter-individual positive and negative affect correlations found consistent mean, but not variance (Fisher et al., 2018). These findings suggest that inter-individual results cannot readily generalise to claims about individual effects. Additionally, many interesting processes in psychology are likely to violate stationarity assumptions. Throughout treatment for depression, a patient's symptoms are expected to decline over time leading to changes in the mean and variance of their symptoms. If treatment affects the covariance between depression and another facet of mental health, then stationarity would also be violated. A related issue is that clinicians are tasked with applying insights about treatment established at the group-level (i.e. nomothetically) to their individual patients, who are treated ideographically, a challenge termed the therapist's dilemma (Howard et al., 1996; Levine et al., 1992). Network analysis has the same issue - insights regarding which symptoms are the best treatment targets are unlikely to generalise to individuals if derived from group-level data. For example, cross-sectional networks estimate a different rank ordering of symptom importance and network structure than a dynamic personalised network (Bos et al., 2017). In the following sections we will outline the importance and practicalities of moving towards personalised networks in mental health science.

1.9 Moving from Cross-Sectional to Personalised Networks

Longitudinal studies that repeatedly measure symptoms within-individuals are needed in order to better estimate the causal interactions between symptoms and

can allow for the construction of what are called 'personalised networks'. Ecological Momentary Assessment (EMA) studies are one way to gather this data and involve participants repeatedly answering a series of questions multiple times per day over a period that can range from only a few weeks to a few months or even longer. Since 2000, the number of publications using EMA has taken off as researchers realised the importance of longitudinal data and widespread access to smartphones made running studies easier than ever (Hamaker & Wichers, 2017). Causal relationships are inferred using Granger Causality, which uses temporal precedence as a necessary condition for a causal relationship to be present (Granger, 1969). If elevated levels of depressed mood cause an increase in tiredness, for example, then a change in depressed mood must occur prior to any corresponding change in tiredness.

Personalised, or idiographic, networks are estimated from longitudinal within-person time-series data to model the dynamics of mental health symptoms. These networks offer the possibility to personalise psychotherapy by identifying treatment targets and relevant associations between symptoms. There are two types of personalised networks that can be estimated: directed (Figure 1.1B) and contemporaneous (Figure 1.1C). Directed networks use VAR models to estimate causal associations between symptoms. While contemporaneous networks estimate associations between symptoms using the correlated residuals from a VAR model, after controlling for all temporal effects (S. Epskamp, C. D. van Borkulo, et al., 2018). This modelling technique is referred to as a graphical vector autoregressive (GVAR) model (Sacha Epskamp et al., 2018). GVAR models have several assumptions including: stationarity, equal spacing of measurements in time, and multivariate normally distributed data. In particular, normally distributed data is unlikely to occur when measuring many mental health conditions. Contemporaneous networks can detect associations between symptoms that occur on a different time-scale than what is measured. Mental processes with a rapid onset and termination, e.g., a panic attack, will most likely be missed if they occur outside relatively infrequent measurement times. An advantage of contemporaneous networks is that they can still recover some of these associations, which would not show up in a directed network. Simulations of personalised networks found a low power to detect the true network structure

when less than 100 measurements are measured within a person (Mansueto et al., 2020). While the full network structure cannot be accurately measured with small sample sizes, the global network structure can be when the network analysis is restricted to only 6 nodes (Mansueto et al., 2020). Additionally, contemporaneous networks are less sensitive to the number of nodes than temporal networks.

1.10 Clinical feasibility and utility of EMA and personalised network models

The feasibility and usefulness of EMA has not yet been thoroughly evaluated in clinical practice, although early results show promise for improving mental healthcare. Patients who use EMA in a clinical setting report greater self-awareness of their mood and symptoms (Bos et al., 2020; Morris et al., 2010). Among the several studies that investigated clinical feasibility, there is generally a discrepancy between what a patient finds helpful and what a therapist does. In one small-scale pilot, the majority of therapy-seeking clients reported that personalised models made them more aware of their mood and symptoms (Frumkin et al., 2021). While therapists tended to agree that personalised networks could be useful for treatment planning and were easy to understand, they were uncertain whether these models provided additional information beyond what they already know about their clients. Therapists and patients generally express concern that EMA assessments will be too burdensome to complete, yet patients in EMA studies tend to have high compliance rates. In a large survey of U.S. based therapists, EMA was rated as one of the least helpful resources, compared to traditional resources, for a difficult therapy case (Ellison, 2021). Another study of within-subject personality dynamics also found therapists did not perceive EMA as useful in everyday practice due to similarities with their own expectations (Zimmermann et al., 2019).

Assessing the feasibility of EMA requires an understanding of what factors affect compliance. Studies have found that compliance is affected by several factors including: study duration, age, gender, and rates of physical activity. Despite all these competing influences, overall compliance with EMA remains high (McLean et al., 2017; Ono et al., 2019; Rintala et al., 2019). Longer study durations were consistently associated with lower compliance rates, with a caveat of most studies

lasting fewer than 7 days. Attrition may stabilise after a certain number of days as the sample becomes restricted to highly motivated participants. Another issue is that while participants are able to comply with a high frequency of assessments, i.e., 6 or more per day, different symptoms likely change at different timescales. While positive and negative affect are known to change rapidly over short-timescales (Trull et al., 2008), anhedonia – for example – likely only meaningfully changes over longer periods of time. Other psychological processes, such as eating behaviours, are likely to reflect more stable trait-like characteristics and exhibit less day-to-day variation. The ideal measurement frequency for different disorders remains an open question. In the case of depression another important factor to consider is low motivation, which may negatively affect long-term compliance with studies as intensive as EMA. People with an affective disorder report higher momentary burden from EMA compared to healthy controls, although this did not affect their adherence (van Genugten et al., 2020). For both participants with and without depression, the most commonly cited reason for missing a prompt was “being busy with another activity”.

Who then would benefit most from EMA? In contrast to the opinion of therapists cited above (Ellison, 2021), treatment resistant patients, or those with a chronic depression, may actually find EMA especially helpful. In a randomised control trial, incorporating EMA-feedback as an adjunct to pharmacological treatment resulted in a larger, and clinically relevant, decrease in depression symptoms, than antidepressants alone (Kramer et al., 2014). An ongoing study aims to determine whether EMA feedback in psychotherapy results in larger changes in depression than a no feedback condition (Riese et al., 2021). In such cases, EMA could provide a richer dataset to try and understand why treatment was not effective and help select new intervention targets. Practical web-based applications, such as PETRA (F. Bos et al., 2022) and PREMISE (Burger et al., 2021), are designed to facilitate customised EMA reporting and statistical interpretations of data.

Therapists, patients, and other relevant stakeholders need to be actively involved in the development of EMA monitoring to provide meaningful insights into symptom dynamics. EMA is still a novel technique that most clinicians are not familiar with or trained to interpret. Developing a larger evidence base for

personalised models and EMA is crucial for clinical adoption and integration into existing workflows. One way to establish efficacy will be to perform randomised control trials to evaluate the core purported benefits of personalised models and monitoring, e.g., whether intervening on the most central symptoms is superior to intervening on randomly selected ones. In the next section, we will review what observational studies have revealed to date about the relationship between personalised network measures and clinically meaningful outcomes.

1.11 Personalised Network Connectivity and Mental Health

Preliminary evidence suggests that network connectivity changes prior to the onset of a depressive episode. A longitudinal n of 1 study found that affect network connectivity increased, driven by stronger autoregressive loops, prior to a significant increase in depression symptoms (Cabrieto et al., 2018; Wichers et al., 2016). A replication study with one participant also found an increase in network connectivity prior to a sudden increase in symptom severity (Wichers et al., 2020). These findings are not, however, restricted to depression. A patient with severe psychosis rated her psychotic symptoms along with positive and negative affect on average 5 times per day over 201 days (Bak et al., 2016). Personalised directed networks were estimated for three clinical phases: stable, impending relapse, and full relapse. Node strength increased for most symptoms during the impending and full relapse phases compared to the stable state. Taken together, these studies provided preliminary support for the prediction that network connectivity should increase prior to a phase change within an individual. More recent evidence in patients with bipolar disorder found that when early warning signals were detected, i.e., increase in autocorrelation or standard deviation, there was an elevated probability of experiencing a transition to either depression or mania (F. M. Bos et al., 2022).

Personalised emotion network connectivity has been shown to be associated with depression severity. People with major depression have greater emotion network connectivity than healthy controls, primarily due to stronger connections among negative emotions (Lee Pe et al., 2015). Replicating these results in an adolescent sample, Lydon Staley et al., (2019) found emotion network connectivity to be positively associated with depression (Lydon-Staley et al., 2019). These two

studies measured emotion dynamics on very different time-scales, 8 times per day over 7 days in (Lee Pe et al., 2015) vs once per day over 21 days in (Lydon-Staley et al., 2019). This may be important as the time-scale of measurement can have a profound effect on the association between emotions. Watson (1988) showed that negative and positive emotions were more strongly correlated at shorter compared to longer time-intervals (Watson, 1988). A study by Shin et al., (2022) sought to resolve two issues within the emotion network literature: 1) the effect of measurement interval on association between network connectivity and depression and 2) the added predictive values of network connectivity beyond mean and variance. To resolve this, the authors conducted two EMA studies in individuals with and without depression. In an EMA study, participants rated their emotions 9 times per day for 8 days, while in a daily-dairy study participants rated their emotions once a day for 50 days. Emotion network density significantly predicted diagnostic status, i.e., depressed vs. non-depressed, over and above the mean and standard deviation of negative and positive affect but only in the EMA study (Shin et al., 2022). Differences in network connectivity may not be due exclusively to mental health risk, but could reflect skewed variance when subgroups are selected for severity. Wigman et al., (2013) found evidence of increasing autocorrelation and cross-lagged associations among affect items among people with greater depression severity (Wigman et al., 2013). A re-analysis of this data showed that these results could be explained by differences in variances between severity groups (Terluin et al., 2016). However, this study did not repeatedly measure symptoms over time and, as a result, cannot say whether these effects are due primarily to symptom severity or variability in symptoms. Due to substantial floor effects for negative affect, i.e., a tendency for healthy people to consistently rate negative items as absent, people with the greatest depression severity also tend to have the most variance in negative affect. Using an inverse Gaussian regression model, which can handle skewed data, there was no significant interaction between depression severity and the autoregression of negative affect.

Despite broadly promising results showing that network connectivity is related to psychological well-being, there remains an unresolved question regarding what exactly network connectivity measures. The vast majority of studies done to date

used a point estimate of depression severity when determining the association with network connectivity. However, depression symptoms, like emotions, change over time in response to daily stressors. Depending on an individual's underlying vulnerability, even perturbations can cause sudden and dramatic changes in symptoms. Because of this, depression symptom variation is highly heterogeneous at the individual-level. A meta-analysis of 25 studies found 4 general classes of symptom trajectories including 3 generally stable groups with either low (largest class), moderate or severe symptoms (smallest class)(Musliner et al., 2016). Several studies also identified an unstable group with either increasing or decreasing symptoms with large variation in the proportion of participants that fall into this category. During treatment for depression, up to 50% of people experience large sudden improvements in their pre-treatment levels of depression (Aderka et al., 2012; Hayes et al., 2007; Tang & DeRubeis, 1999). Sudden changes in symptoms do not just occur within the context of treatment but also naturally over time. A group of adolescents at-risk for developing psychopathology completed a daily diary related to positive and negative mental states (Schreuder et al., 2022). A large minority of participants, ~24%, had changes in symptoms that were large enough to change their diagnostic status, either no longer meeting criteria or meeting criteria for a new diagnosis. Collectively, these results demonstrate substantial within-individual variability in symptom trajectories over short and long time-scales. To date, little research has examined how connectivity relates to these changes in depression over time.

1.12 Network Connectivity, Resilience, and Variability

While much of the previous literature has linked connectivity with severity, elevated connectivity should theoretically drive increased symptom variability (i.e., change in both directions). Indeed, simulations of psychiatric networks predict that more connected networks will cause greater symptom variability, rather than increase symptom severity (Lunansky, Van Borkulo, et al., 2021). Despite this clear and testable prediction, no study to our knowledge has evaluated this, testing whether elevated network connectivity makes individuals more prone to depression variability than severity.

Kuranova et al., (2021) attempted to answer a different, although related, question: Does greater connectivity between affective states lead to increasing symptom severity (Kuranova et al., 2021)? In this study, 159 participants, from the longitudinal East-Flanders Prospective Twin Study, rated their current affect and thoughts 10 times per day for 6 days and completed the Symptom Check-List 90 (SCL-90), as a measure of general psychological well-being, at baseline and 1-year follow-up. Participants were split into two groups based on whether their symptoms increased or remained stable over time. No significant differences in network connectivity emerged between these groups. However, the group with increasing symptoms generally had stronger connections among negative affect items and negative affect states were more positively associated with and positive affect ones. In another study of adolescents at-risk of psychopathology, two groups with different depression symptom trajectories were identified: an increasing and stable group (Kuranova et al., 2020). The increasing group experienced a slower recovery in negative affect after unpleasant events than the stable group. Slower recovery time, after a perturbation, is a prediction of network theory because systems with lower resilience are unable to quickly downregulate negative emotions. Typically, as in this study, recovery time is measured by autoregressive lags. A longer recovery time means that negative emotions remain more similar to their previous values, indicating a lingering effect of the perturbation on the system. Because network connectivity is composed of cross-lagged and autoregressive effects, these results provide indirect evidence that elevated network connectivity is related to symptom stability. Perhaps most compelling, a recent EMA study of university students during COVID-19 showed that directed network connectivity was significantly positively associated with the absolute, but not signed, change in depression severity (Lunansky, Hoekstra, et al., 2021). These studies point to elevated network connectivity (i.e., reduced resilience) as a general marker of the magnitude of symptom fluctuations over time. Chapter 2 of this thesis used repeated assessments of emotional states in large independent community samples to test whether network connectivity is associated with depression severity, and if so, whether it can be explained better by fluctuations in depression.

An important point to clarify here is that emotional variability and instability do not necessarily lead to poorer psychological outcomes. Accumulating evidence among patients receiving treatment for a mood disorder demonstrates that unstable emotional dynamics are associated with greater reductions in depression (Hayes & Strauss, 1998; Lichtwarck-Aschoff et al., 2012; Olthof et al., 2020; Schiepek et al., 2014). Destabilisation of emotion dynamics suggests that patients are amenable to incorporating novel information leading to a greater likelihood of transitioning from a pathological to healthy psychological state. In people with high mean levels of negative affect, greater variability in negative emotions is associated with fewer depression symptoms (Maciejewski et al., 2022). Elevated emotion variability reflects a greater ability to respond to changing environmental contexts. Among patients with high mean levels of negative affect, more variability was related to a reduction in depression symptoms - previously described as “mood-brightening” (Ong et al., 2006). In contrast, among people with high levels of positive affect or low negative affect, greater variability is associated with greater depression severity. Yet unstable emotional dynamics, without appropriate guidance and supervision, can also lead to impaired functioning (Kuranova et al., 2020). It is thought that the study of these dynamics could help researchers understand how psychotherapy exerts its therapeutic effects, in presumably heterogenous ways across patients (Gelo & Salvatore, 2016). One possibility is that psychotherapy acts via a series of perturbations that eventually leads to a reconfiguration of existing emotional dynamics. For example, exposure therapy forces a confrontation with an aversive stimulus, which changes a person’s emotional processing with regards to the stimulus (Carey, 2011). By thinking of networks connectivity as a two-way street, we can potentially reconcile why a great many studies have shown that network connectivity increases rather than decreases after treatment (Berlim et al., 2020; Fionneke M Bos et al., 2018; Schweren et al., 2018; Snippe et al., 2017). Emotional variability should increase during the initial phase of treatment as part of the therapeutic process, leading to a concurrent increase in network connectivity (Hayes et al., 2015). Because changes in network connectivity have been typically assessed shortly after treatment (8-weeks) (Berlim et al., 2020; F. M. Bos et al., 2018), patients may not yet have fully transitioned into a new and stable state that might eventually have both reduced depression severity and reduced negative emotion variability.

With the explosion of research into EMA, ever more sophisticated methods for estimating personalised networks are being developed. Despite this progress, there remain significant methodological gaps in the field. EMA studies typically ask participants to rate numerous items related to positive and negative emotions yet construct networks from only a small subset of these items. Because of that, we often do not know to what extent results would generalise beyond a particular combination of items. Estimating many possible networks and estimating the association between network connectivity and depression can potentially circumvent this issue. If the relationship between connectivity and depression strongly depends on the nodes included in the network, then the effect is likely not robust. Another important, but often overlooked, aspect of network analysis is the stability of personalised networks. Network edges that are stable within an individual can be reliably estimated and reflect a true association between a given pair of nodes. Unstable networks, therefore, consist primarily of noise and any relationship with depression severity is likely spurious.

The burden on participants to complete multiple daily assessments over many months mitigates the long-term use of EMA. Because of this, high attrition rates among participants and relatively low sample sizes are common when compared to cross-sectional methods. Along with participant burden, the short duration (5-10 days) of most studies means that clinically relevant changes in mental health, such as transitions into and out of a depressive episode are not captured. Short-term studies can assess an individual's baseline network connectivity, but not how it varies over time in response to changing depression severity. For that, we need longitudinal data on short-term emotional dynamics along with slower-changing depression severity. One way to circumvent these methodological challenges is to use already available data that contains relevant mental health information. As previously discussed, on social media people write posts that inherently reflect their current emotional state archived in some cases over several years. This provides a rich data stream from which researchers can draw on to approximate the ecological validity of EMA while measuring slow-moving changes in mental health symptomology. Combining these approaches allows us to test network theory predictions regarding how network connectivity drives and maintains long-lasting depressive episodes. Modelling changes in depression symptom networks

over time will enable a better understanding of how causal relationships between symptoms lead to a depressive phenotype. By potentially targeting central symptoms, at the individual-level, networks may inform personalised treatments to improve remission rates from depression.

The first hypothesis addressed here is how language use changes with depression and whether those patterns are unique to depression or found trans-diagnostically. Prior evidence shows that language used on social media accurately predicts mental health, but that there may not be much specificity in these models due to substantial symptom overlap. Here, we investigated whether a language model trained on depression would generalise to other aspects of psychopathology and if by accounting for shared variance across these conditions we could identify unique patterns of language use. After establishing a relationship between language and mental health, we then attempted to interrogate how dynamic relationships among emotional states leads to depression. Network theory predicts that more tightly connected networks, i.e., stronger associations among emotions, predisposes individuals to worse depression than those with less connected networks. A flaw with this conceptualisation, as discussed above, is that elevated connectivity should lead to larger fluctuations in symptoms rather than necessarily more severe depression. We tested whether connectivity relates primarily to symptom severity or variability in several large community samples. A practical drawback of this approach is that it requires participants to self-report their emotions for long periods of time, which makes it difficult to recruit large samples. Consequently, other predictions of network theory become hard to evaluate. One such prediction is that network connectivity should increase within a depressive episode compared to otherwise healthy periods. For the final hypothesis addressed in this thesis, we will use language on social media as a proxy for EMA to estimate personalised networks. Based on networks constructed from language features, we can determine whether network connectivity increases during naturally occurring depressive episodes. Taken together, these different approaches aim to establish the ability to detect depression and identify causal factors that contribute to its development.

Experiments and Hypotheses

Experiment 1: Machine learning of language use on Twitter reveals weak and non-specific predictions

Experiment 1 examined the specificity of a depression trained machine learning (ML) model on 8 other aspects of mental health. We used Twitter language data from 1,006 participants along with questionnaire data from 9 self-report questionnaires that measured depression, schizotypy, social anxiety, eating disorders, generalised anxiety, obsessive-compulsive disorder, apathy, alcohol abuse, and impulsivity. We compared the depression classification performance from a model trained on self-report scores to one trained on depression keywords to illustrate how circular analyses can lead to overly optimistic estimates. Afterwards, we trained separate ML algorithms on each mental health outcome to i) determine model performance and ii) relative importance of each language feature. Finally, using a trans-diagnostic approach, we assessed the specificity of ML models for each trans-diagnostic dimension after controlling for the shared variance due to the other 2 dimensions.

Experiment 2: Connectivity of personalised emotion networks is associated with elevated depression variability

Experiment 2 examined the hypothesis that personalised emotion network connectivity is related to depression severity. We obtained EMA data from 4 independent samples: 155 Paid Students, 194 Citizen Scientists, 519 from HowNutsAreTheDutch, and 45 from a Clinical Sample. Within each sample, we estimated personalised contemporaneous and directed networks for each participant. First, we did this for a singular exemplar network using 5 emotions common across the 4 samples. We then estimated all possible combinations of 5-node networks based on the number of total items available: 4,368 networks for Paid Student and Citizen Scientists, 462 networks for HowNutsAreTheDutch, and 6,188 networks in a Clinical Sample. For each network combination, we correlated contemporaneous and directed network connectivity with baseline depression severity. Finally, in the Paid Student and Citizen Scientist arms, we determined whether the association between network connectivity and baseline depression

could be explained by variability in depression severity over the 8 weeks of the study.

Experiment 3: Using language in social media posts to study the network dynamics of depression longitudinally

Experiment 3 tested the hypothesis that network connectivity is elevated during a depressive episode. A large online sample 946 participants completed a depression questionnaire, self-reported any depressive episodes in the past year, and provided us with their Twitter data for the past 12 months. First, we determined whether a 9-node personalised depression-relevant language network's connectivity estimated from 12 months of tweets was associated with depression severity. We then constructed personalised depression-relevant language networks within and outside a depressive episode for each participant and tested if connectivity increased during periods of illness. We carried out a range of control analyses, including estimating 1,000 personalised networks based on different combinations of depression-relevant and depression-irrelevant language features to test whether our main findings effects were specific to depression-relevant language.

Chapter 2: Machine learning of language use on Twitter reveals weak and non-specific predictions

2.1 Introduction

Approximately 20% of adults will experience a mental illness in any given year (Abuse, 2019). But our ability to treat those affected is hampered by the fact that patients present to clinics relatively infrequently (Lépine et al., 1997), and when they do so, it is often belated, making their symptoms more difficult to treat (Ghio et al., 2014). For this reason, efforts to detect mental illness early and predict individual vulnerability is a key focus of research. Of course, this is challenging in real-world settings, because it is unclear what sources of data can and should be utilized to make these predictions. It has been proposed that one way to overcome this difficulty is to use other sources of data that the general public produce regularly, such as social media data, to detect, predict and better understand mental health in the population. Social media adoption is widespread with approximately 72% of U.S. adults using at least 1 social media platform, which offers a unique opportunity for gathering information about mental health (Perrin, 2015).

Recent studies have suggested that social media data can be used to recognise a broad range of mental health problems in the general public including depression (De Choudhury et al., 2014; De Choudhury, Gamon, et al., 2013; Eichstaedt et al., 2018; Reece & Danforth, 2017; Reece et al., 2017; Tsugawa et al., 2015), eating disorders (Chancellor et al., 2016; De Choudhury, 2015; T. Wang et al., 2018; Wolf et al., 2013), schizophrenia (Birnbaum et al., 2017; McManus et al., 2015; Mitchell et al., 2015), and suicide (Cheng et al., 2017; Coppersmith et al., 2018; Coppersmith et al., 2016; Nobles et al., 2018). A key premise of such work is that these data could be used to facilitate early intervention, for example by providing users with personalised risk scores for having a mental illness and/or developing one in the near future. Inherent in that is the assumption that such models are (i)

accurate enough to be clinically actionable and (ii) precise enough to detect one illness from another. In the present paper, we investigated the extent to which models based on social media data meet these criteria.

There is now a wealth of data supporting the notion that people with depression use language differently than those without depression. For example, depressed individuals use more first person singular pronouns (Rude et al., 2004), obscenities (De Choudhury et al., 2014) and express more negative emotions (De Choudhury, Gamon, et al., 2013) in their language. This language occurs in a variety of settings including semi-structured interviews (Zimmermann et al., 2017; Zimmermann et al., 2013), journal entries (Molendijk et al., 2010; Rude et al., 2004), and critically, social media posts (De Choudhury, Counts, et al., 2013; Reece et al., 2017; Tsugawa et al., 2015). However, it is not clear if these language patterns are specific to depression. Shared variance between disorders presents a challenge to identifying what aspects of language use are specific to a particular disorder. Indeed, because mental health disorders tend to co-occur in the same individuals and our existing diagnostic system lacks clear separation between disorders (Insel et al., 2010; Kessler & Magee, 1993), language-based models are unlikely to have high specificity when trained on summed scores or diagnostic categories. Studies typically compare data from depressed individuals to that of healthy controls, but do not test if language patterns discriminate among psychiatric disorders. There are few clear distinguishing features (Cohan et al., 2018; Coppersmith et al., 2014; Coppersmith, Dredze, Harman, & Hollingshead, 2015; Lyons et al., 2018) among multiple mental health conditions in the few papers that have studied multiple groups. For example, although elevated first person singular pronoun usage is considered to be a defining feature of language in depression, – reflecting an increase in self-focused attention – it is also elevated in people with obsessive-compulsive disorder (Cohan et al., 2018; Lyons et al., 2018), anxiety (Cohan et al., 2018; Lyons et al., 2018), eating disorders (Cohan et al., 2018; De Choudhury, 2015; T. Wang et al., 2018; Wolf et al., 2007; Wolf et al., 2013), and schizophrenia (Birnbaum et al., 2017; Cohan et al., 2018; Ernala et al., 2017; Mitchell et al., 2015; Zomick et al., 2019). Without accounting for comorbidity among disorders, it is not possible to discern whether first person singular pronouns are unique to depression, a transdiagnostic marker of mental

illness, or specific to another aspect of mental health entirely. Tackling the issue of specificity more directly, one study found greater evidence of third person plural pronouns (they, them) in those who participated in Schizophrenia discussion forums versus other sorts of mental health forums, a putative marker of persecutory delusions (Lyons et al., 2018). However, the topic of the discussion forum from which language-use was gathered is a major confounding factor. That is, the content in these forums may not reflect speech patterns of persons with schizophrenia in their everyday life, when not discussing their illness. Moreover, no clinically validated screening tools were used to define cases, rather, participation in these forums and explicit statements of self-diagnosis were used to identify patients.

As discussed in a recent review (Chancellor & De Choudhury, 2020), this is a common approach and has been applied to study language-use on more generic social media outlets like Twitter, using 'statements of diagnosis' e.g., "I have PTSD" (Coppersmith et al., 2014; Coppersmith, Dredze, Harman, & Hollingshead, 2015), to define cases of mental illness, rather than validated clinical instruments. In addition to the issue of diagnostic validity, this approach is limited by the fact that a person who openly reveals a diagnosis on Twitter is not someone trying to conceal it and is probably not part of the cohort of undiagnosed/untreated individuals that such methods may wish to identify. Moreover, they may be more likely to tweet about disorder-relevant topics, which could create circularity, inflating effect sizes, leading us to conclude that mental health status is more readily detected from social media data than it actually is. For example, Coppersmith et al., reported 85% precision at detecting generalized anxiety disorder on Twitter when allowing for a false positive rate of 10% (Coppersmith, Dredze, Harman, & Hollingshead, 2015) using this method. To remove these potential circularities, studies are moving toward less biased methods, where mental health status is not defined by the same or similar content that is ultimately used to study language-use. For example, separating the content used to define disorder status (e.g. membership of a mental health forum) from content used to characterize language use (posts by those users on other forums) (Ireland & Iserman, 2018). When this approach was taken, accuracy was substantially worse. Using a range of machine learning algorithms, the F1 score (average of

precision and recall) rarely exceeded 0.5 (Cohan et al., 2018). This poor performance could be because the signal is weak, or the diagnoses are not accurate. However, even in studies with a more rigorous definition of disorder status because the cases are still binary, evaluating how specific language use is to that disorder, and not another condition, remains an issue. This is due to the lack of multiple continuous measures of mental health in the sample participants. In the few studies that have administered self-report questionnaires to consenting participants, performance was again modest, albeit somewhat improved (Reece et al., 2017), but the specificity of the findings to the disorder of study (here, depression) was not examined.

Thus, while social media makes substantial amounts of language data available to researchers, a caveat to much of this research is the acquisition of high-quality mental health data. The notion that we can detect mental illness from social media posts presents opportunities for public health interventions, but with this comes with significant privacy concerns and potential for discriminatory practices to emerge (Brundage et al., 2018). But are these opportunities and concerns overstated? To date there is little evidence that mental health status can be detected accurately, and even less evidence for individuals who do not choose to openly disclose/discuss their mental health disorder status online. Moreover, if such predictions could be made with any fidelity, it is unclear if they can be in any way specific, which is crucial if these indicators are to be used to guide the choice of intervention. The present study sought to address these issues, determining (i) the specificity of language patterns to different aspects of mental health and (ii) providing an estimate of the performance of these models on unseen data. To do this, we acquired Twitter data over the past year from over 1000 individuals who completed 9 different self-report mental health related questionnaires and consented for us to link that to their Tweets. We tested the performance of a machine learning algorithm trained on a gold-standard ground truth measure of depression symptomatology when applied to unseen data. To test its specificity, we then applied this depression model to predicting scores of a range of mental health phenotypes. Finally, we trained a machine learning model to predict the residuals of 3 transdiagnostic dimensions of mental health (after controlling for

one-another), allowing us to identify text features that are specific to a particular dimension after removing the shared variance between disorders.

2.2 Methods

2.2.1 Participants

We recruited 1,450 participants for this study. The majority of participants were recruited on Clickworker (N = 1,395), an online worker platform, and were paid €2.5 for their participation. A smaller number participated voluntarily (i.e., without payment) and were recruited through general advertising on Twitter and in print media (N = 55). Participants were included for analysis if they were at least 18 years old and had a Twitter account with at least 5 days of tweets and if at least 50% of their tweets were in English. They were also required to pass an attention check, a combination of a captcha and an item with an obvious correct response (“Please select ‘A little’ if you are paying attention”). Of the 1,450 participants recruited, 99 were excluded due to failing the attention check and a further 345 participants were excluded for either not having at least 5 days of tweets or fewer than 50% of their tweets were in English. After excluding these participants, 1,006 participants were brought forward for analysis. Participants had a mean age of 30.5 years (SD: 10.1, range: 18-68), a majority were female (66.4%), currently employed (63.8%), and resided in either the U.K. (41%) or U.S. (46.9%). Participants tweeted an average of 21,126 words (SD: 30,204), median of 6,432 words, and a range of 43 – 163,700. In total, there were 21,252,845 words posted across the 1,006 participants.

Twitter Behaviour	Mean (SD)
No. of Tweets	231.3 (489.4)
No. of Retweets	173.5 (416)

	No. of Likes	768.4 (1041.3)
	Word count per day	96.5 (119.9)
Age	Mean (SD)	
	Years	30.5 (10.1)
Gender	N (%)	
	Male	312 (31%)
	Female	668 (66.4%)
	Transgender Male	6 (0.6%)
	Transgender Female	1 (0.1%)
	Non-Binary	16 (1.6%)
	Other	3 (0.3%)
Country	N (%)	
	Ireland	32 (3.2%)
	United Kingdom	412 (41%)
	United States	472 (46.9%)
	Canada	52 (5.2%)
	Australia	24 (2.4%)
	Other	14 (1.4%)
Education	N (%)	
	Less than high school	22 (2.2%)
	High School	220 (21.9%)
	Some University	301 (29.9%)
	Bachelor's degree	325 (32.3%)
	Master's degree	108 (10.7%)
	Professional degree	17 (1.7%)
	Doctorate	13 (1.3%)
Employment Status	N (%)	
	Currently Employed	642 (63.8%)

Table 2.1: Twitter Use and Demographics of Sample

2.2.2 Procedure

After providing informed consent, participants were asked to complete a self-report questionnaire and provide their Twitter handle which was used to collect the most recent (max. 3,200) tweets and (max. 3,200) likes from their account. Tweets were collected using a data collection app written in Python using the Twitter developer's Application Programming Interface. Participants were asked to

provide their age, gender, country of residence, current employment status, and highest educational attainment. Participants then completed 9 different psychiatric questionnaires including the Zung depression scale (SDS)(Zung, 1965), Short Scales for Measuring Schizotypy (SSMS)(Mason et al., 2005), Obsessive Compulsive Inventory Revised (OCI-R)(Foa et al., 2002), Eating Attitudes Test (EAT-26)(Garner et al., 1982), Barratt Impulsiveness Scale (BIS-11)(Patton et al., 1995), Alcohol Use Disorders Inventory Test (AUDIT)(Saunders et al., 1993), Apathy Evaluation Scale (AES)(Marin et al., 1991), Liebowitz Social Anxiety Scale (LSAS)(Liebowitz, 1987), State-Trait Anxiety Inventory (STAI)(Spielberger, 1983) (Figure S6 in 7.1 Supplementary Materials). This study was approved by the Trinity College Dublin Department of Psychology Research Ethics Committee (Approval ID: SPREC112018-32).

Pre-processing and Text Analysis

We restricted our analysis to tweets published in the 12 months prior to survey completion. Before text analysis, extraneous information was removed from tweets including reply symbol (@), hashtag symbol (#), emojis, punctuation, links (URLs), and all other non-alphanumeric characters. Periods, exclamation points, and question marks were the only punctuation retained because they are necessary to calculate the number of words per sentence. Tweets were aggregated into daily bins and text analysis was then performed on all tweets published per day per user. Daily observations were chosen to increase the amount of text for reliable estimation of text features. Text analysis of daily Tweets was carried out using the Linguistic Inquiry and Word Count (LIWC 2015) dictionary (Pennebaker et al., 2015). The LIWC is a dictionary comprised of approximately 6,400 words and word-stems with 90 different output variables including: linguistic characteristics (e.g., articles and pronouns), psychological constructs (e.g., sadness and positive emotions), and general text information (e.g., punctuation and word count).

In addition to text features, we carried out some additional analyses using Twitter metadata variables including number of followees and followers, replies per day, number of tweets per day, and the insomnia index. The insomnia index is the relative difference in percentage of tweets tweeted during the day (6:01AM to 8:59PM) versus the night (9PM to 6AM). Previous research has found that people

with depression tend to tweet more at night (Chen et al., 2018; De Choudhury, Counts, et al., 2013).

2.2.3 Univariate Associations with Mental Health Symptoms

We focus on depression at the outset because it is the most commonly studied disorder in this field of research and therefore several benchmark studies exist. To examine the specificity of text features to depression, in the first instance, we report univariate associations between the total score for each psychiatric disorder and the top 10 text features associated with depression severity including: word count, negative emotions, focus on present, verbs, auxiliary verbs, adverbs, tone, analytic, six letter words, and leisure words. Each linear model contained just one text feature and controlled for the effects of both age and gender (e.g., depression ~ adverbs + age + gender), both of which showed associations with mental health symptoms consistent with prior work (Figure S1 in 7.1 Supplementary Materials).

2.2.4 Machine Learning

Next, we trained a model to predict depression scores from LIWC text features using Elastic Net regularization (Zou & Hastie, 2005). Elastic Net is a combination of L-1 and L-2 norm regularization, performing both feature selection and regularization which results in a sparse solution when features are correlated with each other. We chose Elastic Net because the input text features are highly correlated and it has been shown to make accurate predictions with small effect sizes in samples larger than 400 (Jollans et al., 2019). Another advantage of Elastic Net is that the output (regression coefficients) is easily interpretable. It is thus possible to directly compare the relative importance of input features and see how they contribute to predictive performance.

We tested the model's performance in predicting out-of-sample depression scores, and to assess specificity, we also tested it on out-of-sample scores on 8 other psychiatric scales (which we not using in training). The data was split into training (70%) and test (30%) sets, stratified by gender to ensure equal proportions of gender categories between the two sets. Nested cross-validation was performed within the training data, using 10 outer loops stratified by gender and 5 inner loops, with optimization of elastic net hyperparameters (alpha and the l1 ratio)

within the inner loops. We repeated this process 100 times to select the best elastic net model to take forward to test on the 30% of data we held-out. We first tested it on depression scores to determine predictive power and then on the 8 other psychiatric scales. Random label permutation was used to determine the predictive value of these models. To ensure that the apparent predictive power of LIWC was not simply the result of confounded associations between age, gender, mental health, and language use. We performed control analyses that included: LIWC text features plus age and gender as features and compared it to a model with just age and gender (and randomly permuted LIWC features). This allowed us to determine if there was a marginal benefit of text features above and beyond the predictive ability of basic demographics (Dinga et al., 2020). Finally, we sought to identify reasons why our model made poor predictions for many individuals. To do this, we examined depression residuals and tested if these related to how participants use and interact with Twitter. We associated depression residuals from the held-out test set of the LIWC text feature only model with the z-score of i) mean word count, ii) total number of Tweets, iii) tweet volume, iv) total number of replies, v) number of followers, and vi) number of followees.

Selecting a machine learning method based in its performance is a form of overfitting. For this reason, we chose to work with the Elastic Net *a priori*, a method well suited to continuous prediction problems. However, for the purposes of comparison to other studies, and to ensure our results are not specific to the Elastic Net, we repeated our analysis pipeline using two alternative classification models. AUC, is the primary measure of predictive performance in most studies in the literature, so to enable comparisons between our results and those of previous studies, we binarised depression scores and classified participants as either depressed or non-depressed. Participants with depression scores above 50 were classed as depressed, while those with scores under 50 were non-depressed (Zung, 1965). We restricted our analysis to depression because that is the primary disorder of interest in the current study and several questionnaires used do not have established clinical cut-offs. Consequently, it would not be possible to directly apply the depression trained model to the other conditions as we did for the Elastic Net model. We next ran four separate models based on varying combinations of sample size and model type to assess performance. We used two classification

models: Random Forest (RF) and Support Vector Machine (SVM) with a radial kernel. Both models were validated with 10-fold cross validation and 100 experimental runs using either i) the full sample of participants ($n = 1,006$) or ii) the top 476 participants by word count. We chose the second sample size to be the same as that from de Choudhury et al., (2013)(De Choudhury, Gamon, et al., 2013), so we could make a direct comparison between our classification performances. Our sample had fewer mean posts over the previous year compared to that of de Choudhury et al., (2013), a mean of 4,533.4 posts per user compared to 1,173.2 posts per user. By selecting the top 476 users by word count the number of posts per user increased to 2,277.1, thereby making the samples more comparable.

2.2.5 Comparison of Validated Self-Report vs Twitter-Derived Ground Truth

We compared the performance of a model trained self-reported depression versus depression keywords extracted from Twitter. Using a regular expression, i.e., `depress*`, we identified any posts that contained depression relevant keywords and phrases. We found that approximately 2.0% of all days with Tweets had at least 1 keyword matching the depression regular expression. We then classified participants as either depressed or not depressed based on whether they had at least 1 post with a depression keyword present; with this approach, 44% of participants were classified as depressed. Days with Tweets that contained the depression keyword were omitted to ensure independence between the testing and training data. Subsequently, we trained a random forest classification model on the depression keyword outcome and compared it to the model trained on binarized self-reported depression. Finally, we evaluated all our models' performance using AUC, F1 score, accuracy, sensitivity, and specificity.

2.2.6 Transdiagnostic Psychiatric Dimensions

To test if performance might be more specific when using transdiagnostic psychiatric dimensions, rather than these questionnaire total scores, the individual answers to the 209 questions in our survey were transformed into 3 transdiagnostic dimensions. These were dimensions previously identified using factor analysis of these 9 questionnaires (Rouault et al., 2018), corresponding to 'anxious depression', 'compulsivity and intrusive thoughts', and 'social withdrawal'.

We then used the weights derived from that independent study to construct the 3 transdiagnostic dimensions in this study. The transdiagnostic dimensions are designed to reduce collinearity across these questionnaires and have been shown to relate to cognitive test performance and brain signatures in a stronger and more specific manner than the original questionnaire total scores (Gillan et al., 2016; Rouault et al., 2018; Seow & Gillan, 2020).

2.2.7 Comparison of Tweets, Retweets, Likes

We also ran additional analyses to understand the influence of Tweet type i.e., Tweet, Retweet, Like, and the amount of Twitter data on model performance. We trained depression models using text features that included only Tweets, Retweets, or Likes and then tested the models on all 9 questionnaire total scores. By training models separately on each type of Twitter data, we could determine whether each Tweet type is independently predictive of depression. We subsequently split the text feature data containing, with Tweets, Retweets, and Likes merged together, into quartiles based on the total number of Tweets. Then, we trained a depression model on data from each quartile and determined its predictive performance. Splitting the data into quartiles also allowed us to control for sample size, such that any differences in performance are solely caused by amount of text data. We expected that models trained on data from the upper quartiles (with the most twitter data) would have better performance than the lower quartiles.

2.2.8 Similarity and specificity of text features across mental health phenotypes

To probe specificity in more detail and with less of a central focus on depression, we generated predictive models for each psychiatric questionnaire separately, using the procedure outlined above. The Elastic Net model can assign a weight of 0 to variables that are not predictive. To better understand our results, we used feature selection derived from the Elastic Net regularization by summing each of the text feature's non-zero weights within each main fold and then averaged this value over the 100 iterations. The 'selection frequency' provides a useful heuristic for the importance of each text feature as a predictor of that clinical phenotype. Selection frequency is a good measure of a variable's importance since text

features that are frequently included are likely to be truly associated with the target outcome. After generating text feature selection frequencies for each questionnaire, we applied a hierarchical clustering algorithm, using Ward's method, to determine how similar language use is between mental health conditions (Murtagh & Legendre, 2014). To examine potential for specificity in a highly correlated space, we trained an Elastic Net model on i) each of the 3 transdiagnostic dimensions and ii) the residuals of each transdiagnostic dimension after controlling for the other 2 dimensions. That is, the anxious-depression residual is derived from a linear model as follows: anxious-depression ~ compulsivity + social withdrawal. We then used the selection frequencies in the same manner as above; to identify text features that were predictive of the residual and if so, whether or not that text feature was specific to that dimension, after removing the shared variance with the other dimensions.

2.2.9 The effect of number of words per user on predictive performance

As a control analysis, we tested the effect of the minimum number of words per user on predictive performance. For our main analyses we chose a relatively low word count threshold per user to maximize our sample size, including participants with at least 5 days of Tweets (Reece et al., 2017). However, there is evidence recommending that a minimum of 200 words per user be used in order to achieve stable predictive performance (Jaidka et al., 2018). While several other studies have used a minimum of 500 words per user (Eichstaedt et al., 2018; Schwartz et al., 2014). We thus tried three additional minimum word count per user threshold for inclusion: 200, 400, and 500 words per user. Sample size was not substantially affected by the additional inclusion criterion such that excluding participants with 200 words reduced the sample size to 945 participants, 400 words reduced the sample to 866 participants, and 500 words to 836 participants. However, to ensure that sample size differences across these minimum-word thresholds did not affect our results, we down-sampled our data to smallest sample size of 836 participants from the 500-word threshold and carried out all analyses on these subjects.

2.2.10 Statistical Power

Finally, we tested if we were sufficiently powered to find an effect size greater than the reported value. To interrogate this possibility, we simulated three types of

datasets with 99 input features and 1 continuous target outcome with a sample size of either 1,000 or 3,000. We set the correlation between either 1, 10, or 20 features with the target variable at $r = 0.32$, while the other input variables had no association with the target. We chose to set $r = 0.32$, because that is approximately twice the observed effect size obtained from our depression model. Furthermore, for the datasets with greater than 1 feature associated with the target, we simulated multicollinearity among the relevant features by setting the correlation between those features at $r = 0.50$. We then ran each dataset through our Elastic Net analysis pipeline, 10-fold nested cross validation with 100 experimental runs and reported both R^2 and the mean absolute error as measure of model fit. To test for the likelihood that the true effect size is larger than we report (i.e., that it is in fact $r = .32$) and we are missing it, we plot the proportion of cases in the 1,000-person sample that performed worse than our reported predictive performance (Figure S7 in 7.1 Supplementary Materials).

2.3 Results

Age was significantly negatively associated with all psychiatric questionnaires, except alcohol abuse (all $\beta < 0.07$, $p < 0.05$). Female participants had significantly elevated eating disorder symptoms ($\beta = 0.34$, $SE = 0.07$, $p < 0.001$), social anxiety ($\beta = 0.38$, $SE = 0.07$, $p < 0.001$), generalised anxiety ($\beta = 0.28$, $SE = 0.07$, $p < 0.001$), and depression ($\beta = 0.35$, $SE = 0.07$, $p < 0.001$) than men. Male participants had a significantly higher rate of alcohol abuse symptoms ($\beta = 0.31$, $SE = 0.01$, $p < 0.001$) (Figure S1 in 7.1 Supplementary Materials). As expected, all psychiatric questionnaires were positively correlated with each other (Figure S2 in 7.1 Supplementary Materials).

2.3.1 Univariate Associations with Mental Health Symptoms

The top ten text features associated with depression were word count, negative emotions, focus on present, verbs, adverbs, auxiliary verbs (all positively associated with depression severity, $\beta > 0.08$, $p < 0.05$) and tone, the analytic summary variable, number of six letter words, and leisure words (all negatively associated with depression severity, $\beta < -0.07$, $p < 0.05$). These effects were non-specific. Negative emotions (all $\beta > 0.08$, $p < 0.001$) were significantly positively associated with all aspects of mental health studied, except alcohol abuse ($\beta = 0.05$, $SE = 0.03$, $p = 0.05$) and obsessive-compulsive disorder ($\beta = 0.04$, $SE = 0.03$, $p = 0.11$). Schizotypy, social anxiety, and generalised anxiety were significantly associated with all 10 text features (all $\beta > |0.06|$, $p < 0.05$), except for the associations between social anxiety with tone ($\beta = -0.03$, $SE = 0.03$, $p = 0.33$)

and leisure ($\beta = -0.04$, $SE = 0.03$, $p = 0.12$), which were non-significant. None of the alternative questionnaires were significantly associated with a text feature in the opposite direction of depression. Individual text features were thus not specific to depression but broadly associated with other psychiatric dimensions (Figure 2.1).

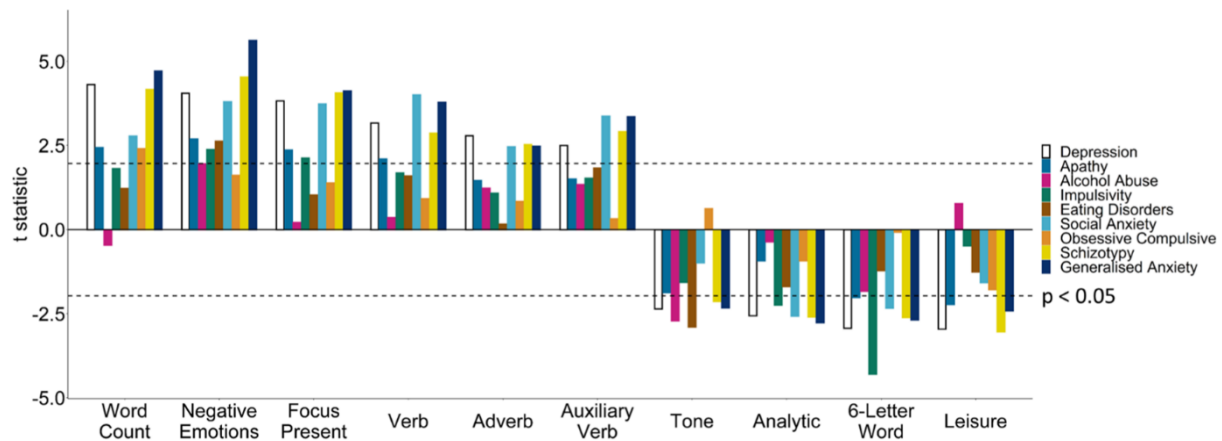


Figure 2.1: Associations between 9 self-reported psychiatric questionnaires and mean values over the past year of the top 10 LIWC text features associated with depression severity, controlling for age and gender (n = 1,006). Dashed lines indicate p-values below 0.05.

In terms of Twitter metadata, participants with elevated obsessive-compulsive symptoms followed more accounts ($\beta = 0.03$, $SE = 0.01$, $p = 0.01$), while participants who scored higher on eating disorder severity had a larger number of followers ($\beta = 0.02$, $SE = 0.01$, $p = 0.02$). Participants scoring high on depression, apathy, impulsivity, obsessive-compulsive disorder, and schizotypy tended to tweet more at night i.e. higher insomnia index (all $\beta < -0.06$, $p < 0.05$). Replies to Tweets (all $\beta > 0.08$, $p < 0.05$) and volume of Tweets (all $\beta > 0.03$, $p < 0.05$) were positively associated with all aspects of mental health recorded, except alcohol abuse and eating disorders (Figure 2.2).

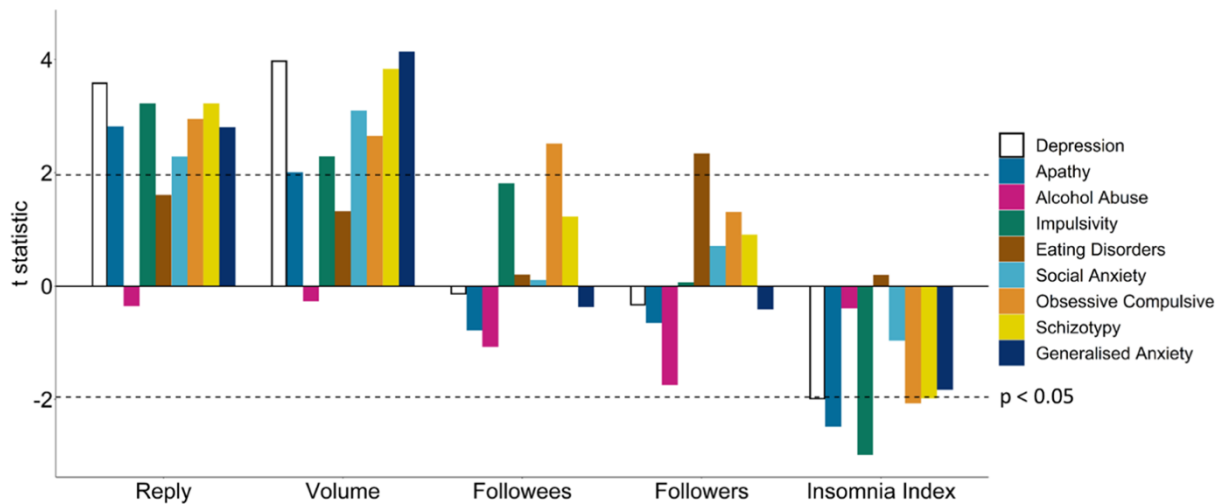


Figure 2.2: Associations between 9 self-reported psychiatric questionnaires and Twitter meta-data features.

Replies (all $\beta > 0.08$, $p < 0.05$) and volume (all $\beta > 0.03$, $p < 0.05$) of tweets were significantly elevated across all aspects of mental health studied, except for alcohol abuse and eating disorders ($n = 1,006$). Participants with elevated obsessive-compulsive symptomology tended to follow more accounts ($\beta = 0.03$, $SE = 0.01$, $p = 0.01$). While participants with more eating disorder symptoms had significantly more followers ($\beta = 0.02$, $SE = 0.01$, $p = 0.02$). Participants with greater depression, apathy, impulsivity, obsessive-compulsive, and schizotypy symptoms tweeted more at night than during the day (all $\beta < -0.06$, $p < 0.05$). Dashed lines indicate p-values below 0.05.

2.3.2 Machine Learning

We trained an Elastic Net model on depression symptoms and tested it on unseen data. The model of depression symptomatology had an R^2 of 0.025 ($r = 0.16$) vs. $R^2 -0.040$ ($r = -0.16$) for the null model. A model trained on text features plus age and gender ($R^2 = 0.045$, $r = 0.22$) performed better than a model with randomised text features plus age and gender ($R^2 = 0.039$, $r = 0.20$). Our simulation results also demonstrated that we were sufficiently powered to detect a larger signal, if it was truly present (see Supplementary Material).

After establishing there was modest, but non-zero signal, we applied the depression trained model using LIWC text features only to the other 8 psychiatric scales to test for specificity (Figure 2.3). The depression model had above zero

predictive power for all other aspects of mental health except impulsivity and alcohol abuse. Nominally, the depression model performed somewhat worse when tested on apathy ($R^2 = 0.008$, $r = 0.11$), alcohol abuse ($R^2 = -0.012$, $r = 0.04$), eating disorder symptoms ($R^2 = 0.011$, $r = 0.12$), and obsessive-compulsive disorder symptoms ($R^2 = 0.011$, $r = 0.12$), predictive ability was identical to depression for social anxiety ($R^2 = 0.025$, $r = 0.16$) and the model performed nominally better in predicting schizotypy ($R^2 = 0.035$, $r = 0.19$) and generalized anxiety ($R^2 = 0.041$, $r = 0.21$) scores. Alcohol abuse and impulsivity were the only aspects of mental health that had negative R^2 values for the non-random models. Increasing the number of words per user, with a constant sample size, did increase the depression trained model's predictive performance. Increasing the threshold from 5 days of Tweets (minimum of 43 words) to 500 words per user caused the R^2 to increase from 0.010 to 0.034 (Supplementary Table 1). However, there was substantial variation at the lower word count thresholds with $R^2 = -0.001$ at 200 words per user up to a maximum of $R^2 = 0.044$ at 400 words per user.

Prior research has suggested that partially dissociable transdiagnostic dimensions of mental health may be a better fit to the underlying neurobiology of mental illness. This was not true of its fit to language patterns in Twitter data assessed here. A model trained to predict 'anxious-depression' scores performed nominally worse ($R^2 = 0.016$) than the models trained and tested on the depression ($R^2 = 0.025$) or generalised anxiety ($R^2 = 0.045$) questionnaires, as reported above (Figure S3 in 7.1 Supplementary Materials). The anxious depression model was also non-specific, having modest but non-zero predictive power for both 'compulsivity and intrusive thought' ($R^2 = 0.025$) and social withdrawal ($R^2 = 0.014$). There were no significant associations between any aspect of Twitter use, e.g., number of Tweets, and depression residuals from the held-out test set (all $|\beta| > 0.02$, $p > 0.05$) (Figure S4a in 7.1 Supplementary Materials). Additionally, depressed residuals were normally distributed and centered on zero (Mean = -0.05, $t = -0.84$ ($df = 301$), $p = 0.40$) (Figure S4b in 7.1 Supplementary Materials). Deviations from true depression scores, i.e., depression residuals, within the LIWC text features model are not due to any systematic differences in participant engagement on Twitter.

Finally, we evaluated the predictive performance of a depression trained classification model (Table 2.2). The best performing model was a SVM trained on data from the top 476 users by word count which had an AUC and accuracy of 0.59. There was no difference in performance between the SVM and RF models, and both models performed worse than the 68% accuracy found by (De Choudhury, Gamon, et al., 2013). Decreasing the sample size to include only the top 476 participants also had no effect on predictive performance, although there was a substantial increase in sensitivity with a slight decline in specificity. The depression-keyword model had an 83.6% accuracy, 0.83 AUC, 76.9% sensitivity, and 88.9% specificity, which substantially outperformed the classification model trained and tested on self-reported depression (57% accuracy, 0.57 AUC, 52% sensitivity and 63% specificity) (Figure 2.4). While participants with depression relevant keywords do have greater depression severity ($\beta = 0.26$, $SE = 0.016$, $z = 4.00$ ($df = 1,004$), $p < 0.001$), the use of keywords within posts to define cases of mental illness demonstrates a substantial overestimation of model performance compared to validated self-report questionnaires.

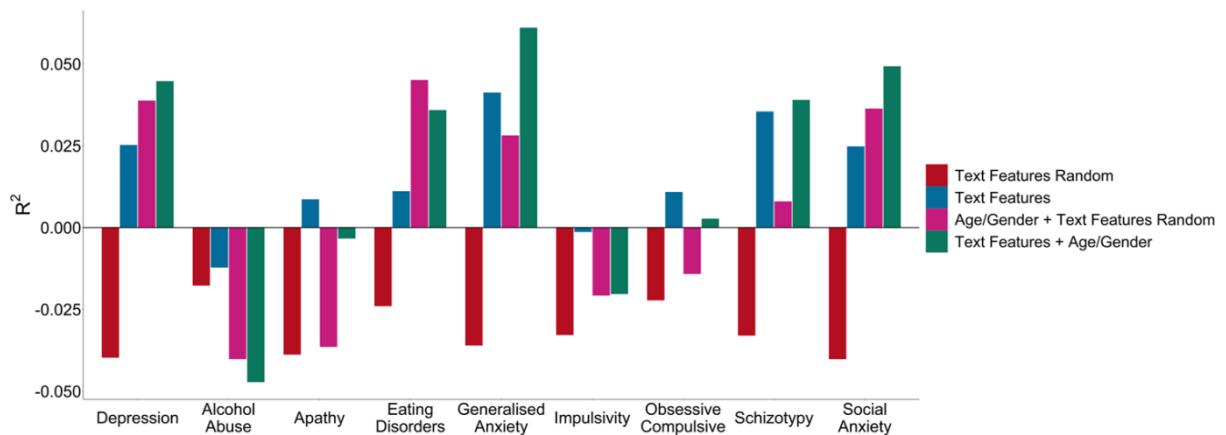


Figure 2.3: Elastic Net predictive performance of a depression model on itself, and 9 other aspects of mental health

Predictive performance (R^2) from an Elastic Net model trained on depression and tested on each of the other aspects of mental health recorded for randomised text features (red), text features only (blue), age and gender only (pink), and text features plus age and gender (green) ($n = 1,006$).

Model	AUC	Accuracy	F1	Sensitivity	Specificity
SVM (n = 1,006)	0.56	0.59	0.42	0.33	0.79
SVM (n = 476)	0.59	0.59	0.54	0.52	0.66
RF (n = 1,006)	0.56	0.58	0.44	0.38	0.75
RF (n = 476)	0.57	0.57	0.53	0.52	0.63
De Choudhury NA SVM (n = 476)		0.68	NA	0.58	NA

Table 2.2: Depression Classification Performance of a Support Vector Machine (SVM) and Random Forest (RF) model with varying sample size

Depression classification performance was similar between SVM and RF regardless of sample size. In the reduced sample size (n = 476) of participants with high word counts there was an increase in the F1 score and sensitivity for both the SVM and RF models. However, in both models the increase in sensitivity was accompanied by a reduction in specificity. Neither model improved over the classification performance of de Choudhury et al. (2013).

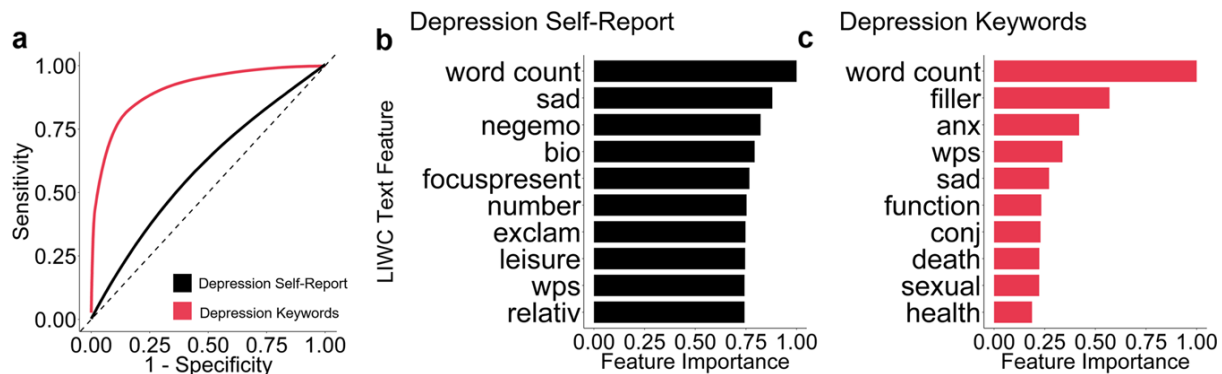


Figure 2.4: Comparison of text feature only random forest models trained on either i) a depression self-report questionnaire or ii) depression relevant keywords.

a) Receiver operator curves for depression self-report (AUC = 0.56) and depression keyword (AUC = 0.83) trained models, dashed line indicates chance level performance (n = 1,006). b) Top 10 text features for the depression self-report model. C) Top 10 text features for the depression keyword model.

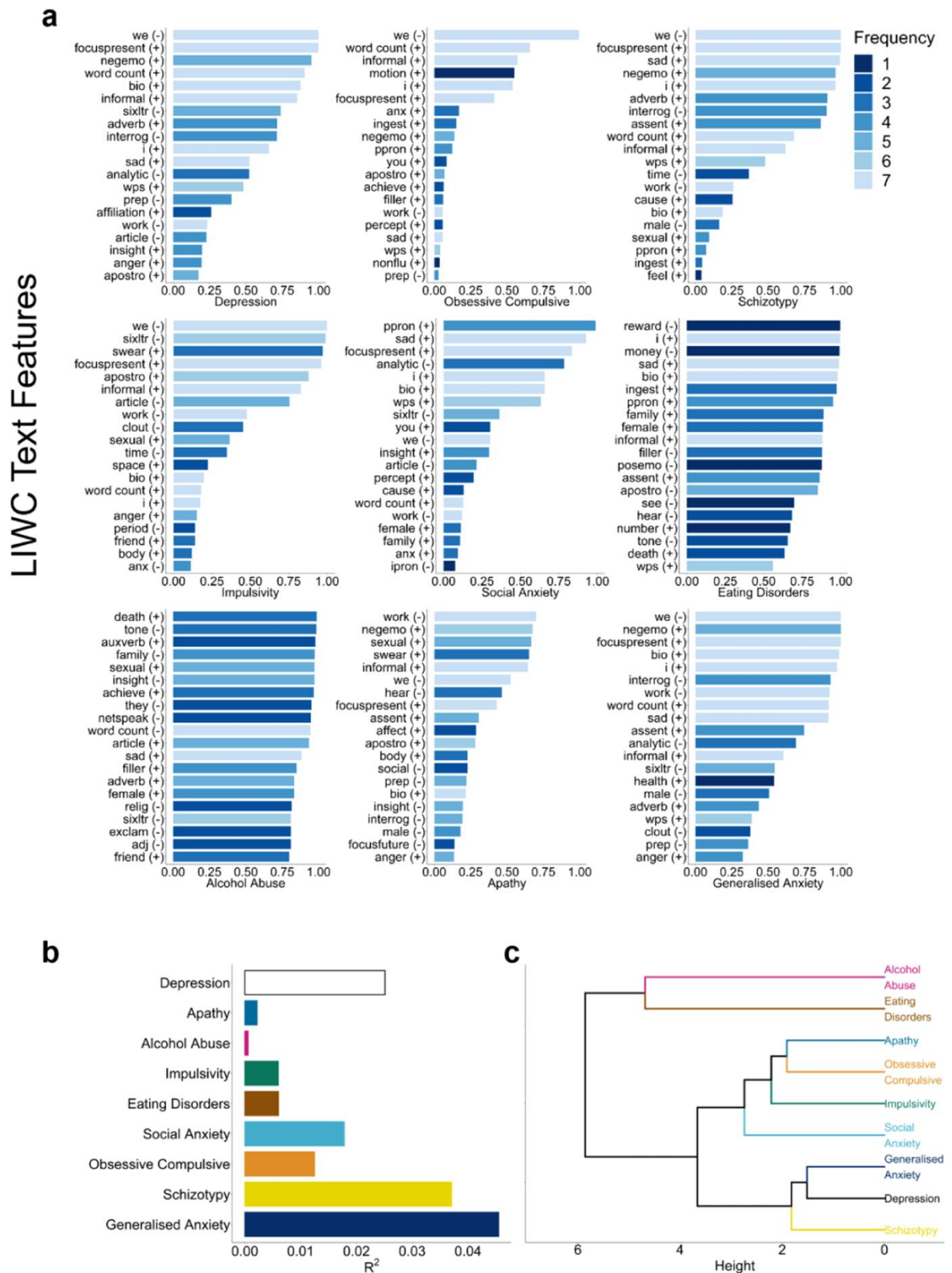
2.3.3 Comparison of models trained on Tweets, Retweets, Likes

By training the depression model on Tweets, Retweets, and Likes separately, we found that the Likes model had the greatest predictive value ($R^2 = 0.026$) compared with an R^2 of 0.010 for the Tweets only model (Figure S5a in 7.1 Supplementary Materials). The improved predictive power may be because there was simply more data for Likes ($M=798.4$, $SD=1041.3$) compared to Tweets ($M=231.3$, $SD=489.4$) and Retweets ($M=173.5$, $SD=416$) (Table 2.1). Indeed, when we split data based on quantity, predictive power was greatest in the 4th quartile of Tweet data ($R^2 = 0.043$) with negative R^2 values from models trained on the 1st and 2nd quartiles of Tweets (Figure S5b in 7.1 Supplementary Materials). Therefore, improved performance can be achieved when the corpus is particularly large, though it clearly remains modest, explaining just 4% of variance in depression.

2.3.4 Similarity and specificity of text features across mental health phenotypes

To compare the content of models developed for each of our 9 aspects of mental health, we generated predictive models for each separately and examined how similar the most predictive text features were. In the depression trained model, we found that 'focus on present' and 1st person plural pronouns were selected in 100% of models. While negative emotions and 1st person singular pronouns, text features previously found to be associated with depression, had selection frequencies of 0.951 and 0.659 respectively (Figure 2.5a). First person singular pronouns were selected more often in models related to schizotypy (0.963), social anxiety (0.658), eating disorders (0.999) and generalised anxiety (1.0) than depression. Negative emotions were slightly more specific to depression than 1st person pronouns with higher selection frequencies in only schizotypy (0.963) and generalised anxiety (1.0). Among the top 20 text features by selection frequency, only affiliation words were unique to depression. Eating disorders and alcohol abuse had the largest number of unique text features with five and six respectively. The generalised anxiety trained model had the highest predictive performance of all models ($R^2 = 0.045$), followed by schizotypy ($R^2 = 0.037$). For impulsivity, eating disorders, alcohol abuse, and apathy the percent of variance explained was below 1% (Figure 2.5b).

Hierarchical clustering revealed, unsurprisingly given the high correlation across these questionnaires, depression language use was most similar to generalized anxiety followed closely by schizotypy (Figure 2.5c). Obsessive compulsive disorder was most closely related to apathy in terms of language use patterns and are slightly more dissimilar to each other compared to the depression, generalised anxiety, and schizotypy cluster. Perhaps most interestingly, alcohol abuse and eating disorders formed their own cluster largely separate from other disorders, indicating substantial differences in language use both between those disorders and relative to the other disorders considered.



a) Model selection frequencies for the top 20 text features in models of 9 psychiatric questionnaires over 100 iterations. Darker colours indicate text features that appear among the top 20 text features across fewer questionnaires, i.e., are more specific. The mean direction of association between each text feature and the target outcome is denoted by a + (positive) or – (negative) next to the label. b) Predictive performance of models trained and tested on each self-report questionnaire c) Hierarchical clustering dendrogram from a model trained on mean text feature selection frequencies.

To examine specificity, we focused on a smaller set of three transdiagnostic dimensions of mental health that can be derived from the larger set of questionnaires: ‘anxious-depression’, ‘compulsivity and intrusive thought’ and ‘social withdrawal’. Much like the analysis of the original questionnaire total scores, the top text features by selection frequency in each transdiagnostic dimension were not specific to any one dimension (Figure 2.6a). Similar to models trained on each questionnaire individually, 1st person plural pronouns and focus on present words were among the top text features associated with not just anxious depression, but also compulsivity and intrusive thought dimensions. In each dimension, less than 45% of top 20 text features were specific to the top 20 for that particular dimension. However, after removing the shared variance between dimensions, at least 80% of text features were specific to each dimension (Figure 2.6b). No text feature was present in the top 20 of all 3 dimensions. Anger words, 1st person singular pronouns, and family words were the top text feature associated with anxious depression, compulsivity and intrusive thought, and social withdrawal respectively. Importantly each of these text features was specific to that particular dimension, i.e., not found in the top 20 for any other dimension.

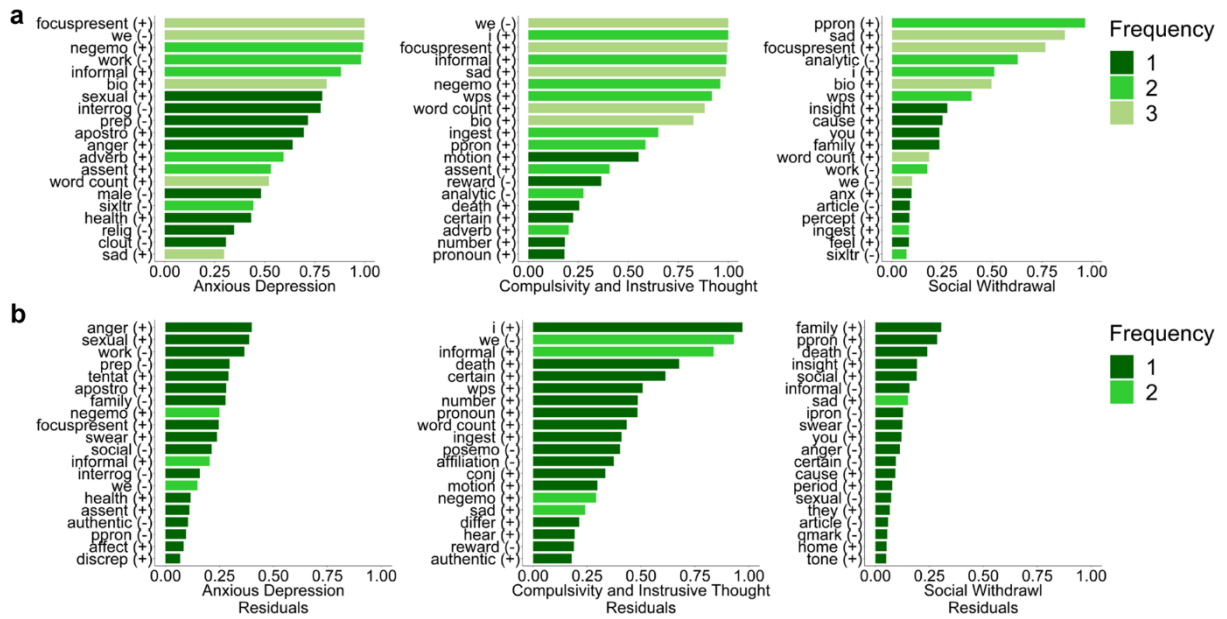


Figure 2.6: Transdiagnostic dimension text feature selection frequencies

a) Model selection frequencies for the top 20 text features in models of 3 transdiagnostic dimensions over 100 iterations. b) Model selection frequencies for the top 20 text features in models trained on the residuals of each transdiagnostic dimension after controlling for the shared variance due to the other dimensions.

2.4 Discussion

There is growing interest in the power of artificial intelligence for improving healthcare provision, early intervention, and diagnosis. But large amounts of data are needed to develop and train models, which can be arduous to gather. Social media data has been suggested to be a convenient and readily available source of such data. This is because social media platforms are in widespread use, users produce sometimes exceedingly high volumes of data, regularly and spanning many years, and these data often contains rich personal and emotional information of putative relevance to their mental state. Although several studies have examined this in recent years, there are substantial limitations to the methods in widespread use (Chancellor & De Choudhury, 2020; Guntuku et al., 2017), including but not limited to the validity of diagnostic classifications employed and the rigor of the machine learning methods employed. Here, we collected Twitter data and 9 validated self-report questionnaires from over 1,000 participants assessing their mental health. We used gold-standard machine learning methods with out-of-sample testing to establish the predictive power of models trained to predict depression and other aspects of mental health, using linguistic features derived from Tweets.

A model developed to predict individual differences in self-report depression explained 2.5% of variance when tested out of sample. The age and gender only model, however, slightly outperformed the text feature only model, illustrating that a similar level of depression prediction can be achieved using just these two data points. It is worth noting that age and gender are not routinely available on Twitter but were gathered as part of our survey. Furthermore, most studies in this area due not include age and gender into machine learning models for mental health prediction. Only 11 out of 78 studies even measured personal demographic features (Chancellor et al., 2020), typically through user settings or even inferred from posts through previous computational models. Consequently, models that demonstrated good performance based on only language use may be instead better explained by unmeasured demographic characteristics. When age, gender, and text features were included in the same model, it still only explained approximately 4% of variance in depression severity. As a result, including text

features into the age and gender model only resulted in an additional 0.5% increase in the variance explained of self-reported depression. Collinearity among text features may partially explain why linguistic features only add incremental predictive power to depression model performance. Furthermore, increasing the threshold for the number of words per user did not have a substantial effect on model performance. This indicates that the lack of unique variance contributed by text features alone is not due to a lack of data on the participant level. Overall, these results demonstrate that not only does a text feature model have limited power to prediction depression, but that it does not perform significantly better to a model composed of basic demographic characteristics.

We examined the specificity of this depression model, on 8 other questionnaire total scores gathered from the same participants. We found that although the model had some small predictive value for 6 other aspects of mental health studied here, generalized anxiety, schizotypy, obsessive-compulsive disorder, eating disorder, apathy and social anxiety, it was not able to explain variance in alcohol abuse and impulsivity. Furthermore, we found that there were no associations between any aspect of Twitter use, e.g., word count, and the residuals of the depression model's predictions. Failures in model performance, therefore, seem to be random and not explained by lower engagement nor number of social connections.

We tested if previously identified transdiagnostic symptom dimensions, which tend to perform better than these questionnaires in fitting cognitive test performance (Gillan et al., 2016), might improve signal and/or specificity. This was not the case. However, after controlling for shared variance among the transdiagnostic dimensions, we found that most text features were specific to the residuals of each dimension. Perhaps most strikingly, 1st person singular pronouns have been consistently found to be a key characteristic of depression relevant language (Edwards & Holtzman, 2017; Rude et al., 2004), but when controlling for shared variance, we found that an increased use of 1st person pronouns was actually most associated with the compulsivity and intrusive thought dimension. Overall, generalised anxiety and schizotypy were the best performing models while the alcohol abuse model had close to zero out of sample performance. Hierarchical clustering revealed that language use associated with alcohol abuse and eating

disorders were most dissimilar to the other disorders. Most prior social media research has focused on associations between language use and alcohol usage at the sub-national, rather than individual, level (Curtis et al., 2018). A possible explanation for the low predictive value of the alcohol model is that few people in our study scored high enough to qualify as alcohol dependent.

A depression classification model trained on the presence of depression relevant keywords had substantially better predictive performance compared to a model trained on dichotomized self-reported depression. Prior studies have shown that regular expression, i.e., keywords, can be used to identify depression with a high degree of accuracy (Nakamura et al., 2014; Prieto et al., 2014). To our knowledge, however, no studies have compared relative predictive performance of a depression-keyword trained model to one trained on depression self-report scores within the same sample. Although self-report measures are more difficult to pragmatically acquire from a large sample, they represent an important and clinically validated ground truth. Our results indicate the potential pitfalls of defining cases of mental illness through keyword-based methods, that is, a sort of content-based circularity can arise when social media posts are used to define caseness, train and evaluate machine learning models. Our data suggest that persons more likely to discuss depression in Tweets, have a distinct pattern of associated language use, but they do not necessarily suffer from clinical depression, with only 50% of these participants meeting the clinical cut-off for depression. These findings underscore the need to use valid ground truth estimates of mental health in developing models of clinical relevance.

Exploratory analyses found that elevated rates of replying and Tweeting were broadly associated with mental health, correlating all questionnaire total scores, except alcohol abuse and eating disorders. Inconsistent evidence exists around whether people with greater depression severity are more (Eichstaedt et al., 2018) or less (De Choudhury, Gamon, et al., 2013) active on social media. We found that participants with greater obsessive-compulsive severity had more followees while people with more severe eating disorder symptoms had more account followers. Depressed individuals have consistently been shown to Tweet more at night during the day (Chen et al., 2018; De Choudhury, Gamon, et al., 2013; Leis et al., 2019). Later Tweet times were associated with depression severity, but also

apathy, impulsivity, obsessive compulsive, and schizotypy. Impulsivity had the strongest association with the insomnia index, in line with prior research showing a positive association between impulsivity and sleep disturbances (Grant & Chamberlain, 2018; Van Veen et al., 2017). Besides findings related to the number of followees and followers, Twitter metadata, like language use, was generally not specific to any one mental health condition.

People with depression have been found to use language differently from healthy controls. Most studies, however, compare people with one mental health disorder with healthy controls (Bucci & Freedman, 1981; Fineberg et al., 2016; Hsuen et al., 2019; Molendijk et al., 2010; Rude et al., 2004); few have examined the specificity of different aspects of language use across disorders. The non-specific patterns of language use observed here, both in prior work (Cohan et al., 2018; Coppersmith, Dredze, Harman, & Hollingshead, 2015; Lyons et al., 2018) and the current study, is likely related to the high comorbidity rates among disorders. We found that only by removing the shared variance among disorders could we identify which aspects of language use were specific to each mental health dimension. Major depressive disorder is positively associated with a variety of other mental health conditions including: panic disorder, agoraphobia, generalised anxiety disorder, post-traumatic stress disorder, obsessive compulsive disorder, and separation anxiety disorder (Kessler et al., 2005). For example, a patient diagnosed with major depression is 8.2 times more likely to have a concurrent diagnosis of generalised anxiety than someone without depression (Kessler et al., 1996). In our study, we found that depression and anxiety had the most similar language use of any pair of disorders. Depression symptoms overlap strongly with other disorders and are associated numerous symptoms in other diagnostic categories (Boschloo et al., 2015). In a network of Diagnostic and Statistical Manual of Mental Disorders-IV symptoms, depression symptoms (insomnia, psychomotor agitation/retardation, and depressed mood) were the most connected symptoms with connections to over 28% of other symptoms in the network (Borsboom et al., 2011). The spread of symptoms across disorders makes it unlikely that individual text features or even combinations of text features could ever be specific to categorical disorders, a finding in line with growing

consensus that these diagnostic categories are overlapping and warrant revision (Insel et al., 2010).

Social media is not a one-way street. While the content of social media posts reflects the underlying mental health of the user, interactions, both passive and active, on the platforms can act to either improve or worsen mental health. When users experience a stressful event, they are more likely to disclose this information on social media. Self-disclosure was shown to subsequently moderate the adverse effects of a stressful event and led to enhanced life satisfaction and lower depression via enhanced social support (Zhang, 2017). However, in a separate study, Reddit users who transitioned to talking about suicide had elevated levels of self-disclosure but received less social support and engagement than users who did not (De Choudhury et al., 2016). Furthermore, specific types of social support are more likely to lead to improvements in mental health, e.g., use of the phrase 'be tough' (De Choudhury & Kiciman, 2017). Increasing awareness about these types of comments would help friends, family, and content moderators to know what to say to and what not to say to someone experiencing mental health difficulties. While there are benefits to self-disclosure these can only be realized if the user is able to communicate free of stigma and receive adequate support. The effects of self-disclosure on social media highlight the need to follow users longitudinally and consider factors beyond just language use, i.e., social network structure, when predicting mental health. Considering the availability of online social support could help triage users with the same predicted risk of mental illness; users with less social support should be prioritized for receiving help.

Most Twitter data is generated by a small subset of users, 80% of Tweets are written by only 10% of users (Wojcik & Hughes, 2019). We found some evidence that machine learning language models perform slightly better when trained on subsets of users with more Tweets. This might suggest that in an even more select sample, those in the top 10% of users overall, one could produce more reliable predictions. However, two things are important to remember here. First, even in our top quartile, the variance explained only rose to a high of 4.3%, additional gains are unlikely to take this to the realm of practice. Similarly increasing the minimum word count per user only slightly increased the percent variance explained. At a minimum threshold of 400 words, 6.4% of variance was

explained, while a threshold of 500 words was slightly worse at 3.4%. Second, those users are not representative of social media users in general, so even if such performance could be achieved, these models are unlikely to be generalizable. An interesting possibility is that the signal may be more meaningfully improved if private sources of text could be harnessed such as text messages. This would have the additional benefit of increasing the amount of data available for each user while simultaneously being more relevant to a user's true mental health status.

Although we demonstrated that social media data has low predictive power at an individual level, this should be contextualized as part of the broader landscape of effect sizes in mental health science. For example, well-established correlates of mental health problems such as adverse childhood experiences only yield an area under the curve of 0.58 in predicting mental health problems at age 18 (Baldwin et al., 2021). A recent preprint showed that resting-state and structural brain-wide associations to psychopathology are exceedingly small, with no reliable correlation exceeding 0.16 (Marek et al., 2020). Because these observations do not have value as individual predictors, does not make the observation devoid of meaning. Mental health is exceedingly complex and likely combinations of a range of sources of multimodal data will be required to take these small effects and transform them into meaningful N-of-1 predictions. Twitter data, by itself, has already proven an interesting testbed for nascent theories of mental health such as network theory, which for example, has struggled to acquire large enough longitudinal datasets to test some of its core predictions (Mansueto et al., 2020). We recently found for example that using social media posts as a proxy for experience sampling allowed us to study a large cohort of individuals through a transition to a depressed state, detecting subtle network signatures of depression vulnerability (Kelley & Gillan, 2020).

Mental health detection from social media offers the potential for generating continuous insights into mental health at the population and individual level, but also poses a unique set of ethical challenges. Large scale analyses of social media data are typically exempt from requiring participant consent due to the public nature of data and lack of experimental intervention. Because of this exemption, social media users are often unaware whether or not their data is

included in research and when asked tend to be uncomfortable with the idea that their Twitter data could be used for research purposes without their knowledge (Fiesler & Proferes, 2018). While it is impractical to ask for consent in all circumstances, requiring consent whenever possible ensures that participants have safeguards for how their data is used. Predicting an individual's mental health outside of a clinical context inherently poses the question of how to act on that information and whether there is in fact an obligation to act (Chancellor et al., 2019). Unlike clinicians, software developers are not obligated to intervene if their algorithm detects that a person is struggling with their mental health. If the developers are not obligated to intervene, would the burden fall on family members, friends or the individuals themselves? Even if a patient consented to having their social media feed monitored by their physician, a high rate of false positives would overwhelm a clinician and impede their ability to effectively allocate care. Furthermore, there is a potential for misuse of mental health predictions by bad actors who do not consider the best interests of the user. Passive and automatic detection of mental illness could lead to targeted advertisements of prescription medication (Ford et al., 2019) or result in an increase in health insurance premiums. A final concern relates to algorithmic bias based on the data used to train these models. Social media users tend to be younger, more affluent, and hold more left leaning political views than the general population (Mellon & Prosser, 2017a; Wojcik & Hughes, 2019). Furthermore, social media research is strongly focused on predominantly English-speaking countries yet there is evidence that people from different cultures behave differently online, for example users from China and India post questions online more frequently than users from the US and UK (Yang et al., 2011). Extrapolating models to very different users than the models were trained on could lead to systematic biases that impact the predictive performance for groups not included in the training data.

Prior studies have had larger Twitter datasets in terms of the number of posts per user. For example, de Choudhury et al (De Choudhury, Gamon, et al., 2013) had a mean of 4,500 posts per user in a 1-year period, while in our study had a mean of about 1,100 posts, including likes, per user. As mentioned above, models perform better when provided more training data per user. Indeed, this study

achieved greater predictive power than reported here. However, there were other differences across our studies too; our sample was twice as big, and we used an independent training set to build our model and then evaluate it on an independent test set. Compared to simple K-fold cross validation using the entire dataset (De Choudhury, Gamon, et al., 2013), this procedure is less likely to overfit the data and overestimate predictive performance. Another potential limitation to our study is that our text feature analysis was restricted to only using categories from the LIWC library. Some evidence exists that more data-driven approaches, e.g., topic analysis, could slightly improve predictive ability over closed libraries (Nobles et al., 2018; Resnik et al., 2015). More sophisticated machine learning models, such as convolutional neural networks have the potential to make superior predictions than more commonly used algorithms, although with the limitation of needing substantially more data (Orabi et al., 2018). While these methods might indeed yield improvements in performance, the use of LIWC has key advantages. LIWC is a closed library that has been well-validated and studied across a range of communication media from diary entries (Rude et al., 2004) to spoken word (Sun et al., 2020). This means that the numerical values and classifications assigned to individual words in LIWC does not change from dataset to dataset, as is often the case with topics and neural networks (Agrawal et al., 2016; Greene et al., 2014). This makes the insights derived here more reproducible and generalizable to new datasets that may be of keen interest in future, such as text messages and email communications.

Regarding the choice of social media platform, it is nonetheless a limitation that our study was confined to Twitter. Recent evidence has also shown that Facebook may be more predictive of mental health conditions than Twitter (Jaidka et al., 2018). We selected Twitter because it is the most used social media platform for studying mental health, comprising approximately 40% of studies on the subject, while Facebook makes up only about 8% (Chancellor & De Choudhury, 2020). It remains a limitation that these results could reflect a relative lack of predictive performance that is particular to Twitter. Because we did not have binary diagnostic information, we did not attempt to classify participants with either depression vs. anxiety, obsessive-compulsive disorder etc., i.e., multi-class classification of mental health diagnoses. Instead, we tried to continuously predict

a participant's score on a range of self-report questionnaires probing different aspects of mental health. Therefore, rather than differentiating users with one diagnosis or another, we instead attempted to quantify the similarity of language use between self-report symptoms of highly comorbid conditions. We think this dimensional approach has many advantages, but this creates a limitation in how directly these data can be applied to diagnoses assigned by a clinician. Finally, subjects in this study reported mental health symptoms at the point of study entry, and we analyzed data corresponding to the 12-month directly prior to this. This necessitates taking a 'trait' perspective on the mental health symptoms we assessed and it is likely that our model is diluted by variations in state/episodic features of depression. However, in a recent study we found that individuals' use of depression-relevant text features in fact didn't change significantly across within-subject periods of mental health and wellness, suggesting this may not be a major issue (Kelley, 2021).

We found that language use patterns on Twitter that relate to depression symptom severity cannot be used to develop predictive models with high accuracy on an individual subject-basis. A model trained to predict depression is also non-specific, being additionally predictive of several mental health symptom profiles. Although performance was poor at the individual subject level, the effect sizes observed are not out of proportion with other routinely studied cross-sectional observations in psychiatry. The addition of age and gender improved performance of our depression model, suggesting that the combination of various sources of multimodal data (with individually small effect sizes) is a viable path forward to improve predictive power of these class of models. Although it is important to emphasise that the inclusion of age and gender, rather than linguistic text features, contributed the most explanatory power to depression model performance. Overall, our findings cast doubt that language use on social media can add much value to the prediction of individual mental health, above and beyond more simple and widely available demographic characteristics. While better model performance from social media data may eventually be possible with the inclusion of other data types, e.g., socioeconomic status or posting frequency, text features are unlikely to play a substantial role in driving these results.

Furthermore, controlling for other mental health conditions and training models on the resultant residuals is a promising method for finding language use specific to that condition. To our knowledge, we are the first study to train machine learning algorithms on the residuals of mental health dimensions in order to identify unique patterns of language. This approach highlights the benefits of using self-report questionnaires to measure mental health since it is not possible for studies with a binary classification of cases, i.e., healthy control vs. case, to account for shared variance between disorders. Although classification studies are able to identify cases of mental illness, they are unlikely to be able to determine specifically what aspects of language are different and unique to a particular condition. Determining specific changes in language patterns and use is crucial for the utility of using text data for diagnostic purposes, regardless of data source.

Nevertheless, we do not believe that social media should be used in a diagnostic setting both for privacy concerns on behalf of the user and the relatively low quality of predictions that would limit clinical utility. Despite the low signal, by virtue of the availability of large amounts of data, the analysis of social media data remains a useful tool to test theories of mental health that are difficult to test using conventional means. Should people be concerned that their mental health status can be unintentionally revealed by the content of their Tweets? We think the data do not support this as a meaningful risk at present.

Chapter 3: Elevated Emotion Network Connectivity Leaves People Vulnerable to Fluctuations in Depression

3.1 Introduction

Throughout the day, our emotions are constantly changing in response to our surroundings (Cacioppo & Gardner, 1999; Carver, 2015). Within certain bounds, these changes are regulated to be adaptive (Kashdan & Rottenberg, 2010); in response to transient frustrations or disappointments, we feel sad temporarily, but that feeling fades, and doesn't necessarily impact our sense of self-worth, our sleep or motivation. Network theories posit that psychological resilience can be captured by using a dynamical systems framework, through the study of how emotions interact with one-another (Borsboom & Cramer, 2013; Cramer et al., 2016). These theories predict that connected networks of emotion allow negative states to propagate through the system more easily, with negative emotions activating other negative emotions and creating positive feedback loops. A hypothesis that emerges from this conceptualisation is that people with more connected emotion networks may be more prone to getting stuck in depressive states, leading to worse prognoses and poorer response to treatment (Cramer et al., 2016). Preliminary evidence for these predictions has come from some small-scale studies of within-subject personalised networks (Lee Pe et al., 2015; Lydon-Staley et al., 2019; Shin et al., 2022; van de Leemput et al., 2014; Wigman et al., 2015), with mixed support from between-subject cross-sectional networks (Lee Pe et al., 2015; Lydon-Staley et al., 2019; Shin et al., 2022; van de Leemput et al., 2014; Wigman et al., 2015). However, there are conceptual problems. Network

connectivity should not necessarily be related to worse symptom severity, but rather the changeability of depression over time (i.e., low resilience). According to complex systems theory, low resilience is characterised by elevated autocorrelation, variance, and cross-lagged relationships (Carpenter & Brock, 2006; Dakos et al., 2010; Scheffer et al., 2009). Systems with low resilience are not able to recover from small perturbations, such as common everyday stressors, which results in an accumulation in variance in the systems' underlying components (Chen et al., 2012). Network simulations predict this exact result; as network connectivity increases, symptoms become more variable (Lunansky, Van Borkulo, et al., 2021). Recent empirical evidence for this claim found that emotion network connectivity was positively related to the absolute, but not signed, change in depression severity (Lunansky, Hoekstra, et al., 2021).

Why then does the current literature tend to show worsening depression linked to network connectivity? A possible explanation for these conflicting results is the reliance on cross-sectional data. Bos et al (Bos et al., 2017) showed striking differences in network structure between cross-sectional and within-individual networks, indicating that cross-sectional networks do not accurately reflect causal interactions among symptoms, within-subject. Where personalised networks have been studied, depression is measured infrequently and so variability cannot be quantified. But perhaps a bigger issue stems from well-established confounds between depression variance and severity that arise due to positive skew of these symptoms with substantial floor effects (Dejonckheere et al., 2019; Ringwald & Wright, 2022).

To address this, we examined the dynamics of how emotions change throughout the day and tested the relationship with depression severity and fluctuations in depression over time. We gathered twice daily EMA data (positive and negative affect) and weekly depression questionnaires from two samples, paid students (N=155) and citizen scientists (N=154), for 8 weeks. We constructed personalised emotions networks for each participant and examined the association between the connectivity of those networks, depression severity and 8-week variability. We further tested if these findings generalised to a large independent community sample (N=519) and explored if we could also see evidence for this in a small patient sample (N=45).

3.2 Methods

3.2.1 Experiment 1: Paid Students and Citizen Scientists in Neureka

3.2.1.1 Participants

We recruited EMA data via a smartphone app, *Neureka*, from two independent samples. The first were ‘Paid Students’ gathered in multiple waves between September 2019 – April 2021, currently enrolled at universities in the Republic of Ireland. The second sample comprised members of the public who downloaded the *Neureka* app from the app store, and participated on a voluntary basis between June 2020 – June 2022. They are here on referred to as ‘Citizen Scientists’. After applying the exclusion criteria (detailed below), data from N=155 paid students and N=194 citizen scientists were analysed. Students in the final sample had a mean age of 23 years (SD: 4.7 years, range: 18-41), 74.8% of the sample was female, and 54.2% of participants were depressed. Citizen scientists were older, less female and less depressed: mean age of 49.5 years (SD: 13.6 years, range: 18-82), 64.9% female, and 32.5% depressed (Table S1 in 7.2.3.3 Supplementary Materials).

3.2.1.2 Procedure

Participants registered for ‘Multi-Mood’, an 8-week ecological momentary assessment study within Neureka. On sign-up, users completed the Center for

Epidemiologic Studies – Depression Scale (CES-D 20) and selected two 3-hour time intervals per day, between 6:00-11:30am and 6:00-11:30pm, when they would like to receive twice daily notifications to rate their mood. The notification appeared randomly within that 3-hour interval and participants could not rate their mood until the notification appeared. After receiving the notification, participants had until the end of the 3-hour window to complete their assessment. Participants rated their mood on 9 negative and 7 positive affect items, e.g., “I feel down” (Table S2 in 7.2.3.3 Supplementary Materials). Items were rated on a 7-point Likert scale from -3 (not at all) to +3 (very much so). Once a week, participants repeated the CES-D 20 (Andresen et al., 1994).

Of the 277 participants recruited for the student arm, 249 completed at least 1 assessment, and 164 completed at least 75% of assessments, which was the threshold for inclusion in data analysis applied to all datasets described in this paper. A further 9 participants were excluded for not completing at least 7 out of 9 weekly depression questionnaires, leaving data from a final N of 155 in this arm, with their characteristics displayed in Table S1 (7.2.3.3 Supplementary Materials).

In the citizen scientist arm, 3,854 participants signed up to Multi-Mood with 1,739 completing at least 1 assessment and 222 completing at least 75% of assessments. Of the participants who completed at least 75% of assessments, 28 were excluded for not completing at least 7 of 9 weekly depression questionnaires. After these exclusions were applied, data from a final sample of 194 Citizen Scientists were analysed (Table S1 in 7.2.3.3 Supplementary Materials). Further details regarding data preparation steps are available in 7.2.1.2 Supplementary Materials.

3.2.1.3 Network Estimation

Time series of EMA items were used to construct personalised directed and contemporaneous networks for each participant. Directed networks allow us to understand temporal dependence between emotions via time-lagged associations. Contemporaneous networks summarise the undirected associations between emotions, at the same moment in time. In directed and contemporaneous networks, all EMA items are considered “nodes”. The “edges” between these

nodes refer to their directed or contemporaneous (partial) correlation. To construct these networks, we used a vector-autoregressive (VAR) model using the *vars* package in R. In a VAR model, the dependent variable is the time-series of the EMA item and the independent variables are the lagged (t-1) version of this item's time-series, plus the lagged time series of all other EMA items. Analyses described in this paper are based on 5-node networks, that are based on 5 time-lagged regression models. An example of one of these regressions is as follows:

$$Worried_t = \beta_1 Worried_{t-1} + \beta_2 Not\ Enthusiastic_{t-1} + \beta_3 Not\ Energetic_{t-1} + \beta_4 Irritable_{t-1} + \beta_5 Not\ Content_{t-1}.$$

From this analysis, we can determine the autocorrelation of 'worried' onto itself (β_1), as well as the 'influence' of the other 4 nodes onto 'worried' ($\beta_2, \beta_3, \beta_4, \beta_5$). Repeating this regression with the other 4 nodes as the DV results in 25 regression coefficients which represent the edges of the directed network. The contemporaneous network is constructed from the correlation of residuals from these regressions. In both kinds of networks, overall network connectivity was operationalised as the sum of all (signed) edges, i.e., expected influence, rather than the sum of the absolute edges. Within a symptom network, a positive edge means that the activation of a node leads to the activation of another node. A negative edge then implies that activating a node would lead to the reduced activation of another node. Robinaugh et al, (2016) found that in networks with exclusively positive edges both strength centrality and expected influence were positively correlated with node influence, i.e., the effect of a node's deactivation on the rest of the network. But in networks with a large proportion of negative edges present, only expected influence continues to be correlated with node influence.

3.2.1.4 Network Stability and Generalisability

To test for the stability and generalisability of network structure, we focused on a single network (directed and contemporaneous) of 5-nodes that were common to each dataset (including in Experiments 2 and 3). These were 'worried', 'irritable', 'not enthusiastic', 'not content' and 'not energetic'. Within each study, we visualised the mean network structure of this exemplar network by averaging each edge across participants to produce an average for each dataset. We tested if

network structures were stable using a network stability test (Appendix II, SI methods).

3.2.1.5 Permuted Network Analysis

Although we gathered data on 16 EMA items, prior research showed that networks constructed from large sets of items are less stable than smaller networks (Mansueto et al., 2020). We considered engaging in both theory and data-driven selection to reduce the total number of EMA items. But without any strong *a priori* conviction, we considered all possible networks and focused on meta-results that would be more robust and generalisable. We therefore estimated all 4368 possible 5-node network compositions for every participant in the two samples. For each individual network, we calculated the Pearson's R correlation between network connectivity and baseline depression. This gave us an item-agnostic view of how all possible emotion networks behave, ensuring results were not driven by a single 5-node selection. We removed any outliers in network connectivity and depression greater than or equal to 3 standard deviations from the mean on a per analysis basis. To summarise the overall pattern of association, we plotted the distribution of these correlation coefficients.

3.2.1.6 Per-Participant Average Connectivity

We took the average of connectivity of these 4,368 networks for each person in the study, producing their 'per-person average connectivity' score. This value was then used as an individual difference measure that we correlated with depression. It was also used to determine the sample size needed to replicate the effects in external data (Appendix II, SI Results).

3.2.1.7 Week 8 Depression and Depression Variance

We tested if our baseline results generalised to depression scores reported at the end of the study. We tested if variance of depression (SD over 8 weeks) was a stronger correlate of network connectivity than these point estimates of depressions severity.

3.2.2 Experiment 2: Replication in a large and independent Citizen Science sample (HNATD)

3.2.2.1 Participants

To replicate these findings externally, we applied our analyses to the ‘How Nuts Are The Dutch’ (HNATD) dataset (Krieke et al., 2016). HNATD was a crowdsourced EMA study conducted in the Netherlands that recruited over 12,000 general participants between May 2014 – December 2018. We extracted complete datasets from N=519 participants who were unpaid, and had an average age in the 40s (M=40.3 years (SD:13.6, range:17-73); Table S1 in 7.2.3.3 Supplementary Materials). Participants were more female than the Citizen Scientists (83%) and with a higher proportion classified as depressed based on their self-report scores (51.3%; though note different depression instruments have different sensitivities and specificities in their thresholds). HNATD participants completed 3 assessments per day for 30 days with assessments spaced out at 6-hour intervals. Further details are available in the SI Methods.

3.2.3 Experiment 3: Comparison of Clinical and Non-Clinical Network Structure and Stability

3.2.3.1 Participants

The final sample studies was a ‘Clinical Sample’ diagnosed with anxiety or depression (Fisher et al., 2017). We classed this analysis as exploratory, as the power analysis from Experiment 1 suggests that at N=45, it is underpowered to estimate the effects observed in our non-clinical samples. Nonetheless, we felt it might be informative for future studies to apply our analysis to a Clinical Sample (95% depressed; Table S1 in 7.2.3.3 Supplementary Materials).

3.2.3.2 Cross-Experiment Comparisons

To assess consistency across our different samples, we constructed a single ‘exemplar’ network (directed and contemporaneous) of 5 EMA items that were present in each dataset (‘worried’, ‘irritable’, ‘not enthusiastic’, ‘not content’, ‘not energetic’). Within each study, we summarised and visualised the mean network structure of this exemplar network by averaging each edge across participants.

We tested if network structures were reliable using a network stability test and cross-sample correlations of edge strengths (SI Methods).

3.3 Results

3.3.1 Network Connectivity and Depression

For Paid Students and Citizen Scientists, we constructed $N=4,368$ (i.e., all possible) 5-node networks from 16 EMA items they completed twice a day for 8 weeks on the smartphone app, *Neureka*. A single exemplar contemporaneous network is presented in Figure 3.1A, with edges denoting the partial correlation of emotions experienced at the same time. This network was also internally stable, with correlation stability .85 and .80 for Paid Students and Citizen Scientists, respectively (Figure 3.1B). The structure of this network was highly consistent across the samples; edges were correlated at $r=0.97$ (Figure 3.1C). We tested for association between contemporaneous network connectivity of each of these networks and baseline depression scores. The resulting distribution of Pearson R correlation coefficients was nominally positive for both Paid Students (Median $r=0.07$) and Citizen Scientists (Median $r=0.21$) (Figure 3.1D). However, very few 5-node networks were significant in Paid Students (proportion p-values <0.05 : 0.64%). In contrast, almost all Citizen Scientists' networks showed a significant correlation, with greater connectivity linked to more severe depression symptoms (proportion p-values <0.05 : 86.42%). Next, we summarised information across these 4,368 networks to compute a per-participant average connectivity score. In

line with the results from the permuted 'network level' analyses, individuals' mean contemporaneous network connectivity was significantly positively associated with baseline depression in Citizen Scientists ($r(192)=0.23$, $p=0.001$) but not Paid Students ($r=.08$, $p=.34$) (Figure 3.1E). We tested the robustness of the association between connectivity and depression scores in the Citizen Scientists by testing if connectivity was also associated with their Week 8 depression rating. This was the case in Citizen Scientists at the network level (Median $r=0.18$, IQR: 0.13, 0.22; 68.66% p -values <0.05) (Figure 3.1F) and for the per-participant average connectivity ($r=0.19$, $p=0.007$) (Figure S1A in 7.2.4 Supplementary Materials). The Paid Students' distribution was centred on zero (Median $r=-0.01$, IQR: -0.05, 0.02; 0.48% p -values <0.05) (Figure 1F) and there was no association between the per-participant average connectivity and week 8 depression ($r=-0.01$, $p=0.89$) (Figure S1A in 7.2.4 Supplementary Materials). We repeated these analyses with directed networks (where edges reflect time-lagged partial correlations, rather than contemporaneous ones) and the results were replicated (see 7.2.4 Supplementary Materials, Figure S1B, S2). However, these directed networks, by their nature, are less stable (Figure S2B in 7.2.4 Supplementary Materials), and so effects were smaller, and a larger N is therefore required to replicate them (see 7.2.1.6 Supplementary Methods).

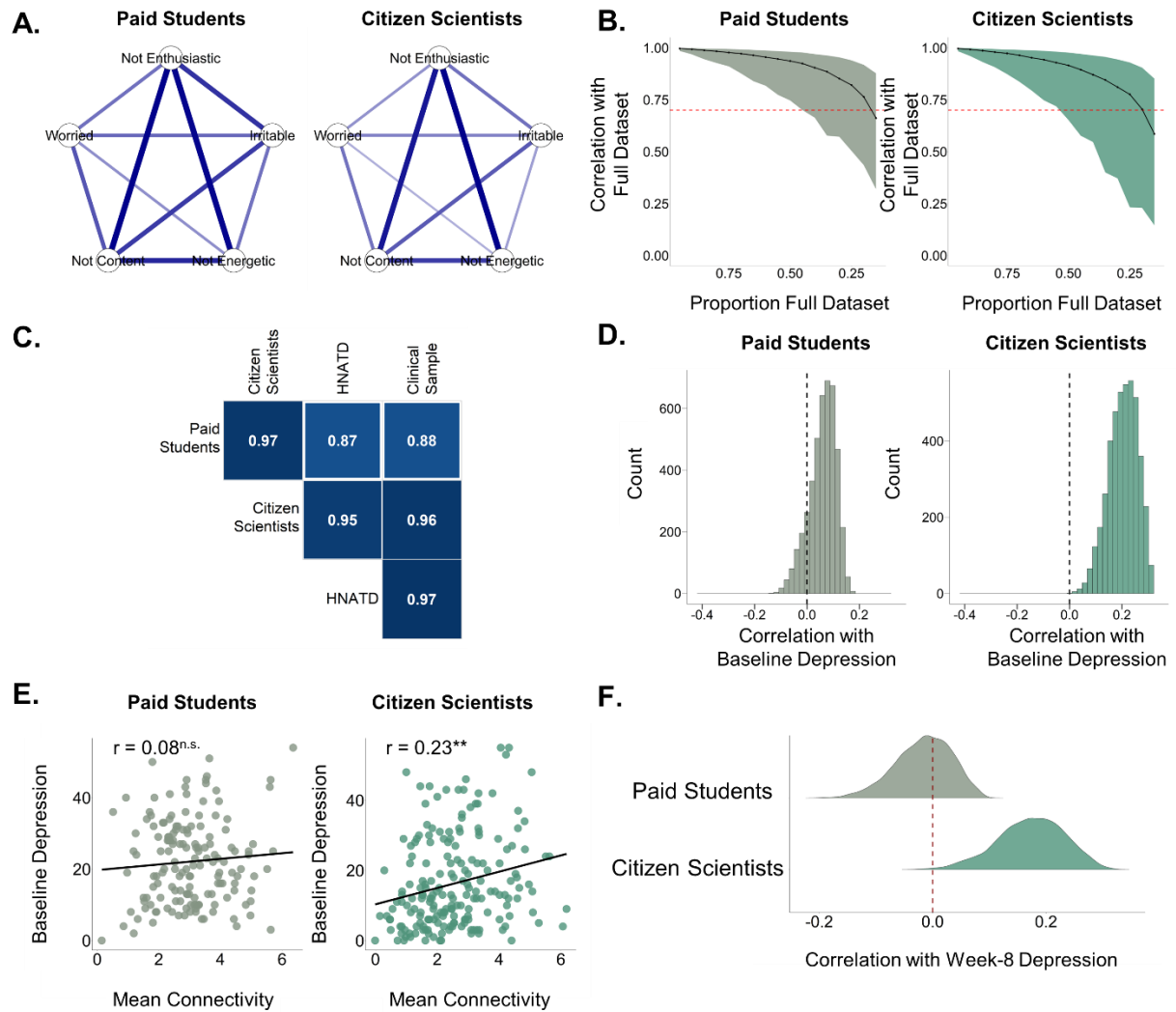


Figure 3.1. Contemporaneous Network Connectivity in Paid Students and Citizen Scientists.

A) Structure of the ‘exemplar’ network common to all datasets. Structure was highly consistent across samples (edges correlated at $r=.97$). B) Networks were stable in both samples, with correlation stability values of 0.85 and 0.8 for Paid Students and Citizen Scientists respectively (where 0.5 is recommended). C) Correlation among edge strengths in the exemplar contemporaneous network across all 4 independent samples. D) Histograms of association between contemporaneous network connectivity and baseline depression for all possible combinations of 5-node networks in Paid students and Citizen Scientists. Citizen Scientists, but not Paid Students, showed a positive association between network connectivity and depression. E) Association of per-participant mean network connectivity scores and baseline depression. Citizen scientists ($r=.23$), but not

Paid Students ($r=.08$) showed a significant association with baseline depression. F) Histograms of association between contemporaneous network connectivity and Week 8 depression.

3.3.2 Depression Variability

Moving beyond point estimates of depression severity, we assessed if network connectivity was related to the volatility of depression over the 8-week study period, operationalised as the standard deviation of CES-D scores. For illustration, we plot the top and bottom quartiles from each sample (Figure 3.2A,3.2B). In both samples, greater variability of depression was associated with higher mean depression (Paid Students: $r=0.18$, $p = 0.03$; Citizen Scientists: $r=0.52$, $p < 0.001$; Figure 3.2C,3.2D), but as is clear from Figure 3.2B, this was much more pronounced in Citizen Scientists. Indicating that mean and variance of depression were more confounded in Citizen Scientists compared to Paid Students.

Repeating our main analysis with SD depression, we found that contemporaneous network connectivity was positively associated with SD depression in both Paid Students (Median $r=0.26$, IQR: 0.22, 0.28; 98.53% p -values <0.05) and Citizen Scientists (Median $r=0.35$, IQR: 0.31, 0.39; 100% p -values <0.05) (Figure 3.2E). This result was also seen in the individual-level analysis, in both Paid Students ($r=0.29$, $p<0.001$) and Citizen Scientists ($r=0.40$, $p<0.001$) (Figure 3.2F,3.2G). For directed networks, the pattern was much the same (see SI appendix). Directed networks were also positively associated with SD depression in Paid Students (Median $r=0.20$, IQR: 0.18, 0.23; 36.29% p -values <0.05) and Citizen Scientists (Median $r=0.19$, IQR: 0.14, 0.24; 48.88% p -values <0.05) (Figure 3.2E). Per-participant average directed network connectivity was associated with SD Depression in both Paid Students and Citizen Scientists ($r>0.21$, $p>0.008$) (Figure 3.2H,3.2I), making this overall, by far, the most robust clinical correlate of network connectivity.

Given this, we tested if differences in SD of depression might explain away the depression severity findings reported for the Citizen Scientists. To test this, we ran a linear regression with per-participant average connectivity as the dependent variable and baseline depression severity and SD of depression as independent

variables. After controlling for SD depression, the association with baseline depression became non-significant in both contemporaneous ($r=0.23$, $p = 0.001$ to $r=-0.0004$, $p=0.99$) and directed networks ($r=0.19$, $p=0.01$ to $r=0.03$, $p=0.69$), while the SD effects remained significant in both network types ($r>0.10$, $p<0.001$). As a final step, we checked whether the significant association between depression variability and network connectivity is due to greater variability at the level of individual EMA items. After controlling for mean EMA item variance in Paid Students, the correlation between network connectivity and depression variability reduced to $p=.05$ ($r = 0.29$, $p < 0.001$ to $r = 0.17$, $p = 0.05$) (Table S4 in 7.2.4 Supplementary Materials). In Citizen Scientists it remained significant ($r = 0.40$, $p < 0.001$ to $r = 0.21$, $p = 0.007$).

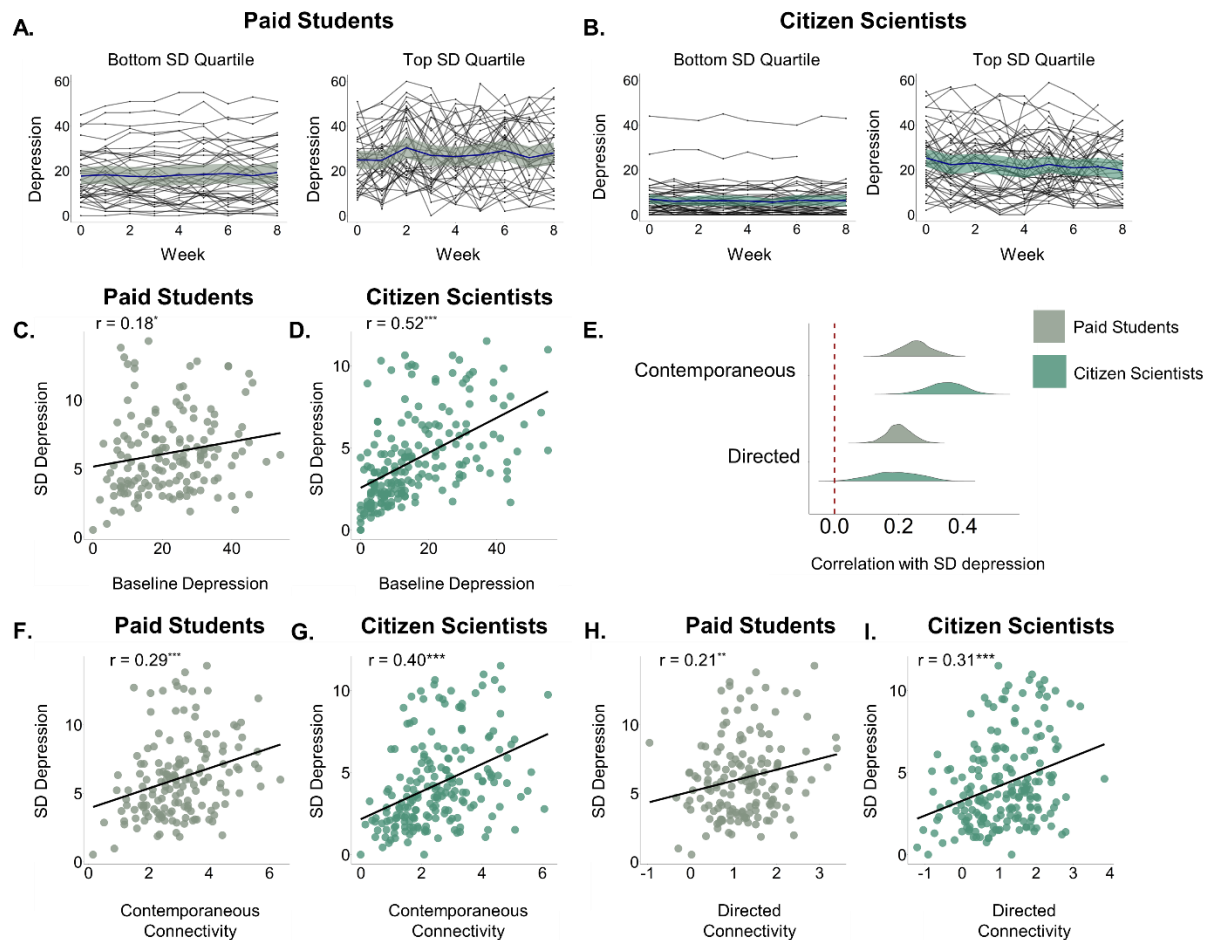


Figure 3.2. Depression variability is positively related to network connectivity.

A) Time series of depression scores over 8 weeks of the study, split by bottom (left) and top (right) quantiles of SD depression for Paid Students. B) Time series of depression scores over 8 weeks of the study, split by bottom (left) and top (right) quantiles of SD depression for Citizen Scientists. C) Association between baseline depression and depression variance for Paid Students. D) Association between baseline depression and depression variance for Citizen Scientists. E) Histograms of association between contemporaneous and directed network connectivity and 8-week depression variability (SD) for all possible combinations of 5-node networks. F) Association of per-participant mean contemporaneous network connectivity scores and depression variability for Paid Students. G) Association of per-participant mean contemporaneous network connectivity scores and depression variability for Citizen Scientists. H) Association of per-participant mean directed network connectivity scores and depression variability for Paid Students. I) Association of per-participant mean directed network connectivity scores and depression variability for Citizen Scientists

3.3.3 Experiment 2: Replication in a large and independent sample (HNATD)

3.3.3.1 Network Connectivity

We tested if these findings would replicate in an independent sample that used different EMA items (but included the items from the exemplar network), a different response modality (VAS 0-100 vs. Likert -3 to +3) and had a higher frequency of assessments (3x per day), over a shorter time frame (30 days versus 56 days). In HNATD, the exemplar contemporaneous network had a correlation stability of .70 (Figure S3A in 7.2.4 Supplementary Materials) and despite these differences across studies, the correlation of edges between HNATD and Paid Students was $r=0.87$ and was $r=0.95$ for Citizen Scientists. Among all combinations of 5-node networks in HNATD ($N=462$ networks), the median correlation between contemporaneous connectivity and baseline depression severity was positive (Median $r=0.12$) and the majority of 5-node networks were significantly associated (proportion p -values <0.05 : 73.81%) (Figure 3.3A). This translated to the per-participant average connectivity analysis, where network connectivity was significantly positively associated with baseline depression ($r(514)=0.14$, $p=0.002$)

(Figure 3.3B). Instead of weekly depression ratings, which were not available for this sample, we used a proxy for depression variability, which we operationalised as the SD of the 'I feel gloomy' EMA item – held out from all network analyses to avoid circularity (Top and bottom quartiles visualised in Figure 3.3C). Consistent with the Paid Student and Citizen Scientist samples, the relationship between depression severity and network connectivity was explained by variance in 'gloomy' over time. The SD of gloomy was positively associated with baseline depression ($r=0.34$, $p<0.001$) (Figure 3.3D) and per-participant average network connectivity ($r=0.27$, $p<0.001$) (Figure 3.3E). When entered in the model, it rendered the association between baseline depression and network connectivity non-significant ($r=0.14$, $p=0.002$ to $r=0.06$, $p=0.20$). Furthermore, the connectivity of virtually all networks (99.78% p -values <0.05) were significantly associated with the SD of gloomy. Like in Citizen Scientists, the association with SD of gloomy was still significant, although reduced, after controlling for mean EMA item variance (Table S4 in 7.2.4 Supplementary Materials). Despite being underpowered for some of the directed network analyses, considering their reduced stability (Figure S3B in 7.2.4 Supplementary Materials), we repeated all these analyses for directed networks. We did not find any significant effects (Figure S4 in 7.2.4 Supplementary Materials); most notably, the per-person average network connectivity was not significantly correlated with SD depression, though it went in a direction consistent with the prior results ($r=0.07$, $p=.13$).

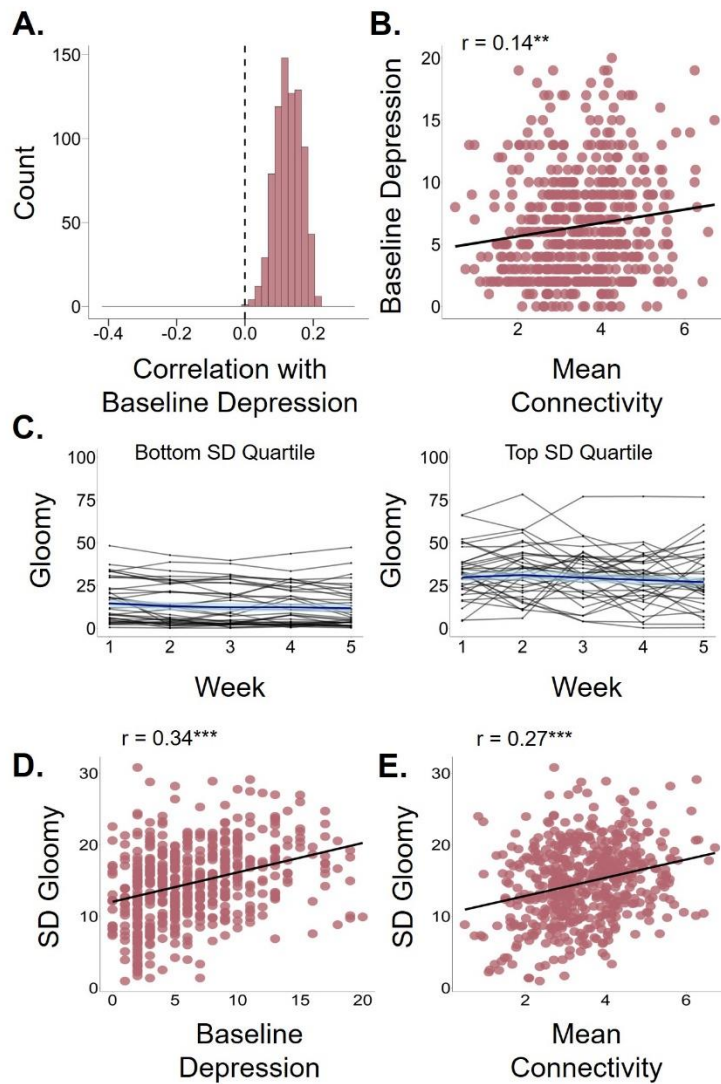


Figure 3.3. The association between network connectivity with baseline depression and SD of gloomy in a large and independent sample (HNATD).

A) Histograms of association between contemporaneous network connectivity and baseline depression for all possible combinations of 5-node networks in HNATD participants. B) Association of per-participant mean contemporaneous network connectivity scores and baseline depression ($r=.14$, $p<.001$). C) Time series of 'I feel gloomy' over 30 days of the study, averaged into weekly bins, split by bottom (left) and top (right) quantiles of SD gloomy. D) Association of baseline depression and SD gloomy ($r=.34$, $p<.001$). E) Association of per-participant mean contemporaneous network connectivity scores and SD gloomy ($r=.27$, $p<.001$).

3.3.4 Experiment 3: Clinical Sample

Finally, we examined data from a much smaller clinical sample to test, in an exploratory way, if a similar pattern of results would be observed. We emphasise that this analysis is underpowered (Appendix II, SI Methods), but nevertheless informative to explore as it is conceivable that effects may be much larger in more severe patients. The structure of the contemporaneous exemplar network from the Clinical Sample was highly consistent with the 3 community samples (Paid Students: $r=0.88$; Citizen Scientists: $r=.96$; HNATD: $r=0.97$). This network was also stable, with a CS coefficient of 0.80 (Figure S5A in 7.2.4 Supplementary Materials). This was also the case for directed networks; correlations were high (Paid Students: $r=0.81$; Citizen Scientists: $r=.84$; HNATD: $r=0.95$) and the network had a correlation stability of 0.55 (Figure S5B in 7.2.4 Supplementary Materials).

3.3.4.1 Association between network connectivity, baseline depression and variance

There was no association between contemporaneous [$r(43) = -0.22$, $p=0.15$] or directed connectivity [$r(43) = -0.04$, $p=0.79$] and baseline depression severity. The association between SD of 'down or depressed' and network connectivity was also not significant for either contemporaneous ($r=-0.005$, $p=0.97$) or directed ($r=0.13$, $p=0.41$) networks. Unlike the 3 community samples, which all showed significant relationships between variance and baseline depression severity, there was no association between the SD of the EMA item 'down or depressed' and baseline depression ($r=0.11$, $p=0.46$), supporting the view that mean and variance are less confounded in datasets where depression scores are elevated and less skewed.

3.3.5 The impact of node valence

We tested if the number of negative emotions in a network would increase the association between network connectivity and depression variance over the course of the study. For contemporaneous networks, we found positive associations in Paid Students, $r=.67$, $p < 0.001$, Citizen Scientists, $r=.16$, $p<.001$ and HNATD (for SD gloomy), $r=.25$, $p>0.01$ (Figure S6A in 7.2.4 Supplementary Materials). For directed networks, the results were similar, albeit somewhat smaller, networks with more negative nodes had stronger associations to depression variance (Paid Students, $r=.06$; Citizen Scientists, $r=.17$; HNATD, $r = .27$) (Figure S6B in 7.2.4 Supplementary Materials).

3.4 Discussion

Network theory posits that mental health disorders are the result of casual interactions among emotional experiences like sadness, guilt, or a lack of motivation. One of its key predictions is that people with more connected networks are less resilient, and therefore more vulnerable to depression. However, studies to date that have tried to test this assumption have had small samples or too few assessments per person to reliably estimate individual differences in personalised network connectivity. Moreover, the mechanism through which this process occurs has not been fleshed out. Here we tested the idea that associations between network connectivity and depression severity, when present, are explained by an increased variability of depression that arises from having a more connected network.

We estimated personalised network connectivity in 2 large samples who completed twice daily EMA assessments of positive and negative affect, alongside weekly standardised depression questionnaires for 8 weeks. Baseline depression severity was related to contemporaneous network connectivity in Citizen Scientists, but not Paid Students. In contrast, 8-week depression variability was related to network connectivity in both samples. After we controlled for depression variability, the association between contemporaneous connectivity and depression became non-significant. These findings were replicated in a third large independent community sample. Results were highly consistent for directed networks, though effects were smaller due to reduced stability. Across all datasets and network types, results were stronger when networks contained more negative affect items. We conclude that network connectivity is related primarily to fluctuations in depression over time, and in some settings (but not all), this manifests in higher severity.

One way to think of this conclusion is through a resilience framework. Elevated network connectivity means that changing one symptom is likely to cause downstream, coherent, changes in others. But sparsely connected networks are less likely to change and thus can be viewed as more resilient; however, depending on the set-point of depression, this 'resilience' can be for better or worse. In line with this, unstable emotional dynamics have been linked to better

treatment outcomes in mood disorders (Hayes & Strauss, 1998; Olthof et al., 2020), obsessive-compulsive disorder (Heinzel et al., 2014; Schiepek et al., 2014), childhood aggression (Lichtwarck-Aschoff et al., 2012), and personality disorders (Hayes & Yasinski, 2015). In people with high levels of negative affect, greater negative affect variability is associated with fewer depressive symptoms (Maciejewski et al., 2022). This implies that when negative emotions are already high, greater variability indicates the system can change in a positive direction. Another consistent finding in the cross-sectional network literature is that depression network connectivity increases, rather than decreases, after successful treatment (Beard et al., 2016; Berlim et al., 2020; F. M. Bos et al., 2018; Höller et al., 2022). Likewise, among patients diagnosed with a psychotic disorder, positive and negative symptom networks increase in connectivity when patients respond to treatment (Esfahlani et al., 2017).

Although there can be benefits, studies (including the present study) have shown that systems with low resilience are in some settings susceptible to worse outcomes and deterioration. Individuals with low trait resilience have stronger associations between positive and negative emotions, especially when stressed (Ong et al., 2006). Adolescents who exhibit slower negative affect recovery – a sign of low resilience – after an unpleasant event, tended to show a pattern of worsening depression severity over time (Kuranova et al., 2020). Longitudinal n-of-1 studies in person with depression found increasing variance of mental health states prior to a significant increase in depression severity (Wichers et al., 2016; Wichers et al., 2020). Collectively, these results suggest that network connectivity predisposes people to have more variable symptoms, without a particular directional change. Within psychotherapy, elevated emotional instability could enable the integration of new information, reduce the entrenchment of the current pathological state, and facilitate the transition into a new healthy state. Network connectivity may be a promising marker of treatment response and a way to understand the mechanism of action.

It is important to acknowledge limitations of our study. Personalised networks were estimated during vector autoregression, which assumes that emotion dynamics are stationary, i.e., constant mean and variance over time. This assumption may be reasonable in relatively healthy participants but would not hold in people

undergoing treatment for depression. It is possible that network connectivity could substantially change with the addition of other emotions. However, estimating all possible combinations of networks and using external data with a slightly different EMA composition helps to ameliorate this concern. In our replication dataset, we did not have access to a repeated and longitudinal assessment of depression using a validated scale; instead, we constructed a proxy for depression variance from the EMA item 'I feel gloomy'. This therefore serves as an indicative or conceptual replication of the results of the two samples in Experiment 1. In this study, we operationalised network connectivity as the sum of all signed edges. We chose this approach in order to account for the varying influence of positive versus negative edges within a network. However, a more common approach within the network analysis literature is to use the absolute value of edges to estimate network connectivity. When networks primarily consist of positive edges, there will be few differences in overall network connectivity between these two approaches. Yet the inclusion of different types of symptoms, i.e., positive versus negative affect items, could affect the relationship between estimates of network connectivity and depression. Specifically, this would result when nodes have qualitatively different meanings, i.e., an edge's influence on another node depends on the node valence. In a network of all negative affect items, activation of one node is likely to lead to the activation of another node in the same direction – most edges are positive. In a network with a mixture of positive and negative items, activation of a positive edge likely leads to the deactivation of a negative edge, leading to a network consisting of both positive and negative edges. To account for this potential issue, we recoded all positive affect items to be negative such that every item within a given network has the same valence.

For contemporaneous networks, the vast majority of edges across all possible networks are positive (87.5% in Citizen Scientists, 93.3% in Paid Students, and 92.4% in HGIN). Consequently, there is not likely to be a significant difference between the use of signed vs absolute edges when calculating overall network connectivity and its relationship with either baseline depression or depression variability. Unlike contemporaneous networks, directed networks have a lower proportion of positive edges (58.5% in Citizen Scientists, 60.7% in Paid Students, and 59.4% in HGIN). Given the higher proportion of negative edges in directed

networks, overall network connectivity will be larger when the absolute value of edges is taken into account. Thus, the relationship between directed network connectivity and depression is more sensitive to differences in how connectivity is calculated. Yet, directed networks are already more weakly associated with depression variability than contemporaneous networks, have fewer networks significantly associated with variability, and are substantially less stable. Additionally, increasing the number of negative nodes led to a stronger association between the variability and connectivity, although the difference in correlation strength was still small. Using absolute network connectivity could increase the sensitivity to detect associations with depression, by more accurately modelling edge influence among networks that contain both negative and - reverse coded - positive affect items. Overall, differences in how signed and absolute network connectivity affects the relationship between depression and connectivity is most evident in directed, rather than contemporaneous networks.

We found evidence that network connectivity is associated with symptom severity in some settings, but that this effect can be explained by depression variability. Given that recent studies produced mixed results regarding the role of network connectivity in mental health, shifting focus from severity to variance may resolve these inconsistencies. Network theory offers a powerful causal framework for explaining the development, maintenance, and evolution of mental health disorders.

Chapter 4: Using language in social media posts to study the network dynamics of depression longitudinally

4.1 Introduction

Network theories of mental illness propose that disorders like depression emerge from cascades of casual interactions that occur between symptoms (Borsboom & Cramer, 2013). In contrast to traditional frameworks that suggest symptoms are indicators of a single underlying disease state, network theories posit that these symptoms and their interactions are actually what drive these conditions. For example, diminished feelings of worth, compounded by insomnia, may lead to a loss of energy, resulting in weight gain and a decreased ability to think and concentrate. Positive feedback among these symptoms is thought to contribute to the maintenance of a depressive episode (Cramer et al., 2016; Smith et al., 2018).

Preliminary support for network theory has come from studies comparing the network structure of self-reported depressive symptoms between groups of individuals with and without a diagnosis, or before and after some intervention. Individuals with depression, compared to individuals without depression, are thought to have greater connectivity between depression symptoms, reflecting an elevated vulnerability to 'knock-on effects', that may result in fairly sudden and persistent changes in depression (Cramer et al., 2016). This has been partially born out in the data; studies have shown that patients with depression have increased connectivity among depression symptoms compared to healthy controls (Lee Pe et al., 2015; Santos et al., 2017; Wigman et al., 2015) and the same is true for several other mental health conditions (Heeren & McNally, 2017; Jimeno et al., 2020; Segal et al., 2020; van Rooijen et al., 2018). Moreover, participants who go on to have persistent depression have more strongly connected networks at baseline than those who later enter remission (McElroy et al., 2019; C. van

Borkulo et al., 2015) and the same appears to be true for patients with eating disorders who undergo treatment (Smith et al., 2019). As a change in a system's state approaches, e.g., onset of a depressive episode, network connectivity is expected to increase, reflecting elevated vulnerability, as the system becomes less and less able to recover from external stress (Chen et al., 2012; Dakos et al., 2009; Scheffer et al., 2009). One study found some evidence that network connectivity increases approaching a depressive episode, although these effects did not directly capture the onset of an episode and were not shown within-subject (van de Leemput et al., 2014).

Despite these results, recent studies have yielded some inconsistent findings. Another study failed to extend the findings of Van Borkulo and colleagues (2015) to an adolescent sample – finding instead no difference in baseline network connectivity in patients who went on to have worse outcomes (Schweren et al., 2018). Furthermore, several studies have failed to find evidence for one of the key predictions of network theory, that individuals who recover (e.g., following treatment) show reductions in their network connectivity. In fact, several studies have actually shown an increase in network connectivity following treatment for depression (Berlim et al., 2020; F. M. Bos et al., 2018; McElroy et al., 2019; Snippe et al., 2017).

One explanation for the lack of consistent findings is that the majority of prior work has been based on networks constructed from group-level symptom correlations (i.e., between-subject, cross-sectional). An alternative approach is to construct networks based on repeated assessments gathered from the same individual over time (i.e. within subject, longitudinal), which allows one to characterize each individual's network structure, sometimes referred to as a personalised network (S. Epskamp, C. D. van Borkulo, et al., 2018). This may be an important distinction, because it is unclear to what extent cross-sectional networks capture how symptoms causally relate to one-another over time, within an individual. This issue was addressed empirically when researchers analysed the same dataset in multiple ways, allowing them to directly compare the structure of cross-sectional networks versus personalized ones (Bos et al., 2017). They found that the two analysis approaches can sometimes yield different results, including different

associations between symptoms and finding that different symptoms were the most central.

We therefore considered if the inconsistent findings in the field of network analysis and mental health to date might stem from the over-reliance on cross-sectional methods for characterizing network structure. Longitudinal studies are needed to test some of the key predictions of network theory, such as whether network connectivity increases as someone transitions from a healthy to acutely ill state. We know of only two network studies that attempted to measure symptoms within the same individual over a long enough period to capture a naturally occurring change in mental health state. These two examples are both single-subject observational studies. Wichers et al., (2016) reported data from a single patient over 239 days, a period of time that spanned the transition into a depressed state, concluding that the patient's network connectivity increased prior to the onset of the depressive episode, though data were insufficient for a formal analysis (Wichers et al., 2016). In a longitudinal study of one patient with psychosis, researchers similarly observed a qualitative increase in network connectivity during both an impending and full relapse (Bak et al., 2016). While these studies are suggestive, larger samples and formal statistical tests are needed to address one of the most fundamental predictions of network theory of mental illness - does network connectivity increase during a depressive episode? The present study aimed to fill this gap by comparing the connectivity of networks within versus outside a depressive episode. Rather than using self-report symptoms, however, we used linguistic features associated with depression posted by users on the social media platform Twitter (Coppersmith, Dredze, Harman, & Hollingshead, 2015; De Choudhury, Gamon, et al., 2013; Edwards & Holtzman, 2017; Zimmermann et al., 2013).

One reason that studies addressing this question are lacking is because the data required is challenging to gather; multiple assessments are required per day, per participant, over several weeks or months. Challenges are compounded by the need for a naturally occurring depressive episode to have its onset during this period. To circumvent these challenges, we utilized an alternative to ecological momentary assessment (EMA). Instead of asking participants to report their mood, motivation, sleep etc. daily over a prolonged period of time, we analysed

depression-associated textual data already archived on Twitter, a social media platform. We analysed data over a 12-month period from nearly 1,000 participants that in some cases spanned the onset of a depressive episode. Central to our approach are prior observations that individuals with depression, both people with a clinical diagnosis and those with self-reported depression symptoms, have significant linguistic differences in both their writing and speech patterns compared to those without depression. For example, individuals with depression use significantly more 1st person singular pronouns than individuals without depression in personal essays (Molendijk et al., 2010; Rude et al., 2004) and semi-structured interviews (Zimmermann et al., 2017), which is thought to reflect enhanced self-focused attention that occurs in a depressed state (Edwards & Holtzman, 2017). Along with changes in pronoun use, depression is associated with negatively biased cognitive distortions, e.g., “everyone thinks that I am a loser” (Bathina et al., 2021). These findings are also observable in social media posts (Coppersmith et al., 2014; Coppersmith, Dredze, Harman, & Hollingshead, 2015; De Choudhury, Gamon, et al., 2013; Eichstaedt et al., 2018), and include increased use of swear and negation words, anger, references to death, and changes in the use of articles and other pronouns (Al-Mosaiwi & Johnstone, 2018; De Choudhury et al., 2014). People with depression are also less active on Twitter in the early morning (3am-6am) than healthy controls exhibiting an altered circadian rhythm, but also used significantly more personal pronouns, negative affect words, and rumination words during this time. Language usage on social media is thus not static, instead changing over time reflecting underlying changes in a participant’s mental health (Ten Thij et al., 2020). By examining longitudinal data, we can examine fluctuations in time within-subjects, allowing us to ask if these depression-associated linguistic features become more connected when someone is in the midst of a depressive episode. For example, when a person is more self-focused, using words like “I”, “me” and “my”, is that person also more anxious, angry or sad? Is that association stronger when someone is currently depressed than when they feel well? This sort of data has the benefit of being objective and relatively plentiful, but the mapping of these linguistic features onto to specific self-reported symptoms diagnostic of depression (such as sadness, sleep disturbances and motivation) has never been formalised and needs considerable further study. With these limitations in mind, we tested if, similar to what has been predicted for self-

report symptoms, depression-associated linguistic features are more inter-connected in individuals who have greater self-report depression severity and become more connected, within-subject, during a depressive episode. We first constrained our analysis to networks constructed from 9 text features (Coppersmith et al., 2014; De Choudhury, Gamon, et al., 2013; Reece et al., 2017) that previously studies have linked to depression in the literature to ensure the independence of our analyses, but we subsequently test if our results generalize to a range of networks constructed from depression-associated linguistic more broadly.

In this work, we show that participants with greater depression severity have higher overall network connectivity among a network of 9 a priori selected depression-relevant linguistic features. Among participants with self-reported depressive episodes, we found that network connectivity is higher within vs. outside an episode. These results were not dependent on our chosen network; networks constructed from random samples of depression-related linguistic features are significantly more connected during a depressive compared to networks of depression-irrelevant linguistic features. Our study illustrates that Twitter data, albeit noisy, can be used as an alternative to ecological momentary assessment to study depression longitudinally and in this case, test key predictions of network theory.

4.2 Methods

4.2.1 Participants

We recruited 1,713 participants for this study. The majority were recruited on Clickworker (N = 1,395), an online worker platform, and were paid €2.5 for their participation. A smaller number participated voluntarily (i.e., without payment) and were recruited through general advertising on Twitter and in print media (N = 318). Participants were included for analysis if they were at least 18 years old and had a Twitter account with at least 30 days of tweets and if at least 50% of their tweets were in English. They were also required to pass an attention check, a combination of a captcha and an item with an obvious correct response (“Please select ‘A little’ if you are paying attention”). Of the 1,713 participants recruited, 99 were excluded due to failing the attention check and a further 668 participants were excluded for either not having at least 30 days of tweets or fewer than 50% of their tweets were in English. After excluding these participants, 946 participants were brought forward for analysis. Participants had a mean age of 29.6 years (SD: 10.6, range: 18-66), a majority were female (65.2%), currently unemployed (51.6%), and resided in either the U.K. (35.9%) or U.S. (50.7%).

Participants reported more than half (59.0%) of the sample reported at least one depressive episode in the past year (mean: 1.56 episodes, SD: 0.81) with an average duration of 104.06 days (SD: 97.06) and 45.7% reported being diagnosed

by a physician with depression at some point in their life. Participants were asked to self-report the dates of any depressive episodes in the past year; a depressive episode was defined as a period of at least two weeks with low mood and loss of interest or pleasure in activities every day or nearly every day. Participants that reported at least one depressive episode tweeted ($\beta = 84.7$, $SE = 38.4$, $p = 0.03$) and liked the Tweets of others ($\beta = 193.8$, $SE = 70.0$, $p = 0.006$) more frequently than participants without a depressive episode, but did not retweet significantly more ($\beta = 49.8$, $SE = 32.6$, $p = 0.13$). Individuals who reported a depressive episode in the past 12 months were younger ($\beta = -2.2$, $SE = 0.70$, $p = 0.002$), more female ($\chi^2(5, N=946) = 26.0$, $p < 0.001$), less likely to be employed ($\chi^2(1, N=946) = 4.3$, $p = 0.04$), and more likely to have been diagnosed with depression by a physician ($\chi^2(1, N=946) = 178.5$, $p < 0.001$). Individuals that reported a depressive episode were also more likely to have a lower educational attainment than those without a depressive episode ($\chi^2(6, N=946) = 28.0$, $p < 0.001$). There was no significant difference in country of residence for individuals with versus without a depressive episode ($\chi^2(5, N=946) = 8.1$, $p = 0.15$) (Table 4.1).

Participants recruited through Clickworker tweeted ($\beta = -217.8$, $SE = 41.5$, $p < 0.001$), retweeted ($\beta = -162.0$, $SE = 35.4$, $p < 0.001$), and liked other posts ($\beta = -227.0$, $SE = 76.5$, $p = 0.003$) significantly less than people who were not paid for their participation (Table S1 in 7.3.1 Supplementary Materials). Furthermore, although there was no difference between groups in the percentage of depressive episodes in the past year ($\chi^2(1, N=946) = 0.42$, $p = 0.52$), paid participants were significantly less likely to have been ever been diagnosed by a physician with depression ($\chi^2(1, N=946) = 6.9$, $p = 0.01$).

	Full Sample N = 946	No Depressive Episode N = 388	Depressive Episode N = 558	p-value
Twitter Behaviour				
Tweets	356.1 (581.4)	306.1 (517.2)	390.8 (620.2)	0.03*
Retweets	276.0 (494.1)	246.7 (468.0)	296.5 (510.8)	0.13
Likes	1097.7 (1062.7)	983.4 (996.6)	1177.2 (1100.2)	0.006**
Recruitment Type (% Paid)	680 (71.9%)	274 (70.6%)	406 (72.8%)	0.52 ^a
Word count per day	131.6 (127.7)	115.2 (114.8)	143.1 (134.9)	0.001**
Age (years)	29.6 (10.6)	30.8 (11.3)	28.7 (10.1)	0.002
Gender				
Male	304 (32.1%)	147 (37.9%)	157 (28.1%)	<0.001*** ^a
Female	617 (65.2%)	240 (61.9%)	377 (67.6%)	-
Transgender Male	6 (0.6%)	0 (0%)	6 (1.1%)	-
Transgender Female	1 (0.1%)	1 (0.3%)	0 (0%)	-
Non-Binary	15 (1.6%)	0 (0%)	15 (2.7%)	-
Other	3 (0.3%)	0 (0%)	3 (0.5%)	-
Country				
Ireland	44 (4.7%)	24 (6.2%)	20 (3.4%)	0.15 ^a
United Kingdom	339 (35.8%)	147 (37.9%)	192 (34.4%)	-
United States	480 (50.7%)	188 (48.5%)	292 (52.3%)	-
Canada	42 (4.4%)	13 (3.4%)	29 (5.2%)	-
Australia	15 (1.6%)	4 (1.0%)	11 (2.0%)	-
Other	25 (2.6%)	12 (3.1%)	14 (2.5%)	-
Education				
Less than high school	16 (1.7%)	4 (1.0%)	12 (2.2%)	<0.001*** ^a
High school	190 (20.1%)	66 (17.0%)	124 (22.2%)	-
Some university	317 (33.5%)	108 (27.8%)	209 (37.5%)	-
Bachelor's degree	285 (30.1%)	142 (36.6%)	143 (25.6%)	-
Master's degree	104 (11.0%)	53 (13.7%)	51 (9.1%)	-
Professional degree	15 (1.6%)	4 (1.0%)	11 (2.0%)	-
Doctorate	19 (2.0%)	11 (2.8%)	8 (1.4%)	-
Currently Employed (% Yes)	458 (48.4%)	204 (52.6%)	254 (45.5%)	0.04* ^a

Physician diagnosed depression (% Yes)	432 (45.7%)	76 (19.6%)	356 (63.8%)	<0.001***a
--	-------------	------------	-------------	------------

Table 4.1: Demographic and Twitter use characteristics of sample

Twitter and demographic characteristics of all participants along with differences between participants with and without a depressive episode.

*p < 0.05, **p < 0.01, ***p < 0.001 ^a Chi-square test

4.2.2 Procedure

After providing informed consent, participants were asked to complete a self-report questionnaire and provide their Twitter handle which was used to collect the most recent (max 3,200) tweets and (max 3,200) likes from their account. Tweets were collected using a data collection app written in Python using the Twitter developer's Application Programming Interface. Participants were asked to provide their age, gender, country of residence, current employment status, and highest educational attainment. They were also asked if they have ever been diagnosed by a physician with depression and if yes to provide the approximate date of diagnosis. Next, they completed a self-report depression questionnaire to establish their current symptom severity levels. In the first wave of recruitment, 263 participants completed the Centers for Epidemiologic Studies Depression scale (Turvey et al., 1999) (CES-D 8). In subsequent recruitment waves, the remaining 1,450 participants completed the Zung Self-Rating Depression Scale (SDS) (Zung, 1965) instead. We combined scores from the two depression scales by standardizing each scale by its mean and standard deviation. Finally, participants were asked to report up to five depressive episodes in the past year. A depressive episode was defined as a period of at least two weeks in which the participant felt both low mood and loss of interest or pleasure in hobbies and activities nearly every day for most of the day. We chose this definition to increase the sensitivity for detecting depressive episodes and to reduce participant burden by only requiring the two essential components of a depression diagnosis. Episodes were recoded to be "not depressed" if they were shorter than 2 weeks in duration and were merged together if separated by fewer than 2 weeks (effectively recoding intervening days as also being depressed).

4.2.3 Pre-processing and Text Analysis

We restricted our analysis to tweets published in the 12 months prior to survey completion. Before text analysis, extraneous information was removed from tweets including: reply symbol (@), hashtag symbol (#), emojis, punctuation, links (URLs), and all other non-alphanumeric characters. Periods, exclamation points, and question marks were the only punctuation retained because they are necessary to calculate the number of words per sentence. Tweets were aggregated into daily bins and text analysis was then performed on all tweets published per day per user. Daily observations were chosen to increase the amount of text for reliable estimation of text features. Text analysis of daily Tweets was carried out using the Linguistic Inquiry and Word Count (LIWC 2015) dictionary (Pennebaker et al., 2015). The LIWC is a dictionary comprised of approximately 6,400 words and word-stems with 90 different output variables including: linguistic characteristics (e.g., articles and pronouns), psychological constructs (e.g., sadness and positive emotions), and general text information (e.g., punctuation and word count). The LIWC has been used in prior studies that reported a relationship between Twitter sentiments, text features, and depression (De Choudhury, Gamon, et al., 2013; Reece et al., 2017). As an initial step to verify that these features were picking up depression symptomology, we averaged each feature over the past 12 months and then tested for correlation with current depression severity.

Among the 9 averaged LIWC text features, any value more than 3 standard deviations from the group mean for that text feature was subsequently removed. Approximately 1.1% of data in the full sample of participants was excluded using this criterion.

4.2.4 Feature Specification

We selected 9 LIWC text features a priori for network analysis based primarily (but not wholly) on the findings of de Choudhury et al. (2013), who found that the following text features had relevance to self-reported depression severity: 1st person singular ("1st Pers. Sing.", 24 words incl. "I", "me", "mine"), 1st person plural ("1st Pers. Pl.": 12 words incl. "we", "our"), 2nd person ("2nd Pers.", 30 words incl. "you", "your"), and 3rd person ("3rd Pers.", 28 words incl. "she", "they") pronouns,

negative (“Neg. Emo.”: 744 words incl. “hurt”, “ugly”) and positive emotions (“Pos. Emo”: 620 words incl. “love”, “nice”), swear (“Swear”: 131 words incl. “damn”), articles (“Articles”: 3 words: “a”, “an”, “the”), and negation (“Negate”, 62 words, incl. “not”, “never”) words. Specifically, these words were found to have either a significant change in mean, variance, momentum, or entropy in their sample at a stringent correction for multiple comparisons. Based on findings from prior work, however, we did not average the two 1st person pronouns plural and singular together into a single 1st person pronoun category. Prior work has shown they have bidirectional associations to depression and indeed, we confirmed this in our data as well (Coppersmith, Dredze, Harman, & Hollingshead, 2015; De Choudhury et al., 2014; Lyons et al., 2018). Using the LIWC, we calculated the proportion of text on each day with tweets in the past year that included words from each of the LIWC’s 87 categories. This resulted in a time-series for each of the 87 text feature categories for each participant. Days without tweets were not considered or assigned any value and consequently participants who tweet less often had fewer days in their time series than participants who tweeted every day.

4.2.5 Network Analysis

Networks were constructed by examining the correlation between these text feature time series (nodes), using regularized partial correlations to determine the contemporaneous association between text features (S. Epskamp, D. Borsboom, et al., 2018). The contemporaneous association is based on the residuals of the lag-1 correlation and removes any temporal effects due to other variables measured at the same time point (S. Epskamp, C. D. van Borkulo, et al., 2018; Wild et al., 2010). Individual node strength, the sum of the absolute values of partial correlations into a node, and global network strength, average of node strength across all nodes, are the primary indicators of network connectivity in psychological networks. Personalised networks were estimated for each participant using the *graphicalVar* (version 0.2.4) package with LASSO regularisation. Regularized partial correlations control for associations between all other nodes in a network with high specificity. Consequently, an edge that is present in a regularized network likely presents a true edge rather than a false positive. However, as we were not focused on particular edges between nodes, but rather the broader characterisation of ‘connectedness’, we set the

hyperparameter (gamma) to 0, which, although still regularised, causes the model to prefer *more* connections over fewer. This avoided a situation where many edges would be returned as 0 and is the same approach applied in a prior study using cross-sectional networks (C. van Borkulo et al., 2015). A range of 10 tuning parameters (lambdas) was considered for each person's model (nLambda=10).

4.2.6 Network Connectivity of a priori Network and Current Depression

Using this method, we first estimated personalised networks for all participants (N = 946) and created a mean of these personalised networks to describe the network's overall composition regarding strength, closeness, and betweenness centrality. Any individual node strength value, among the 9 LIWC text features, that was greater than 3 standard deviations from that node's group mean was excluded from analysis. Using this exclusion criterion, approximately 2.5% of all node strength values were omitted. In order to test for the reliability of edge strengths in the network, we split our sample into two equal halves, calculated personalised networks for all participants, and then correlated the mean edge strengths between the two halves. We found that among the 36 unique edges in the network, there was a high degree of reliability between the split halves ($r(36) = 0.99$, $p < 0.001$). The edge strength between Neg. Emo. and Swear was much stronger than between other edges, reliability was $r(35) = 0.97$, $p < 0.001$ when we exclude this edge. Because individual networks tended to be sparse, the average of most edges tended towards zero leading to a high correlation between split halves (Figure S1a in 7.3.2 Supplementary Materials). Global network strength was not normally distributed (Shapiro-Wilk test, $W = 0.95$, $p < 0.001$) and had a rightward skew (Figure S1b in 7.3.2 Supplementary Materials). The strength centralities of most nodes, except Swear and Neg. Emo. had a strong right skew due to the relative sparsity of those nodes (Figure S1c in 7.3.2 Supplementary Materials). Note we did not calculate closeness and betweenness centralities for individual networks due to edge sparsity and the strong correlation with strength centrality. To test if network connectivity based on the entire 12-month dataset was related to depression symptom severity, we correlated each individual's network characteristics (i.e., network node strengths) with their current depression severity.

4.2.7 Change to a priori Network Within vs Outside Depressive Episodes

Among participants who reported a depressive episode, we estimated two separate personalised networks for each person, representing periods when they were within and outside a depressive episode, hypothesising that networks would be more tightly connected within compared to outside of an acute episode. Individuals were required to have at least 15 days of tweets both within and outside an episode. We compared network connectivity using regression with “depressive episode” (1 = within episode, 0 = outside episode) as a within-subjects variable predicting node strength. We did the same analysis predicting global network strength.

4.2.8 Stability Checks

Network analysis is made more robust by having fewer nodes and this was why the networks presented here are limited to 9 nodes (Mansueto et al., 2020). The number of nodes in the current study is well within the typical range (5-11) included in other network papers on depression (Berlim et al., 2020; McElroy et al., 2019; Schweren et al., 2018; Snippe et al., 2017; Wichers et al., 2016). We quantified network stability for each personalised network using the *bootnet* package (version 1.5). Stability was assessed by repeatedly dropping up to 75% of cases in the sample and correlating the resultant strength centrality to the estimate based on the full sample. Stability was quantified with the correlation stability coefficient (CS coefficient), with 1,000 bootstrapped samples, i.e., the maximum proportion of the sample which can be dropped to retain a 0.7 correlation with the full sample in 95% of cases. A simulation study by Epskamp, Borsboom, and Fried (2018) proposed using a threshold for CS coefficients above 0.50 to ensure that the ordering of centralities is interpretable. Personalised networks of all participants were found to be highly stable (mean strength CS coefficient: 0.65, SD: 0.12) along with networks constructed from within (mean strength CS coefficient: 0.50, SD: 0.23) and outside (mean strength CS coefficient: 0.64, SD: 0.13) a depressive episode.

4.2.9 Generalisability of findings to other depression networks

We based our initial analysis on an a priori network of text features previously linked to depression. The idea behind this was to be as conservative as possible and ensure independence of the selection of features from this dataset. Following this proof of principle, we tested if this finding would extend to other depression-associated linguistic networks (i.e., networks constructed from other text features that were significantly associated with depression severity). Crucially, this analysis controls for the possibility that networks of any kind would be more strongly connected within vs outside an episode. We thus hypothesised that networks comprising text features significantly associated with depression would show greater within subject changes in connectivity compared to those not associated with depression. To test this, we constructed a list of all depression-relevant and 1000 depression-irrelevant text features from LIWC, based on an arbitrary threshold of $p < .05$ for their bivariate association with depression severity (see Figure S5 for full list of associations in 7.3.10 Supplementary Materials). We then selected 1000 random sets of 9 features from each of the depression-relevant and irrelevant lists and estimated personalised networks for all 2000 sets. We did this twice for each participant, once based on Tweets that were published within an episode and once based on Tweets outside an episode. This allowed us to test if global network connectivity was greater within versus outside an episode for each of the 2000 networks, using regression with “depressive episode” (1 = within episode, 0 = outside episode) as a within-subjects variable predicting node strength (exactly the same analysis as for the a priori network). Finally, we took the 2000 betas for the within-subjects variable ‘depressive episode’ in these analyses (Figure 4.4a: 1000 depression-relevant and 1000 depression-irrelevant betas) forward to a general linear regression to determine if the extent to which episodes became more connected during an episode was greater in depression-relevant compared to depression-irrelevant networks of language use.

Control Analyses

A range of control analyses are presented in the online supplement. These examine several potential confounding influences in network estimation. First, we noted that participants had more data (i.e., more days) outside an episode than within one. To control for the possibility that differences in the number of days might be driving our results, we (i) conducted a permutation test that randomised

the identifier “within episode” versus “outside episode” 1000 times within subject and (ii) ran an additional within-subjects regression analysis that included the number of days within and outside an episode as a covariate. Because 3rd person pronouns (she/he, they) were selected a priori, but not significantly associated with current depression severity in this sample, we repeated our analyses omitting this node to ensure our results were not affected by its inclusion. In the LIWC library, certain ‘supra-categories’ are inclusive of multiple sub-categories. For example, within the personal pronoun category are: 1st person singular/plural, 2nd person, 3rd person, and impersonal pronouns. To test if this had a material effect on our results, we repeated our analyses excluding these supra-categories and the results were unchanged (Figure S2 in 7.3.3 Supplementary Materials).

Mean personalised networks were visualised using the qqgraph package (version 1.6.9). Between and within-subjects regression were performed using the glm (version 3.6.1) and lmer packages (version 3.1-3). All statistical analyses were performed in R (3.6.1).

4.3 Results

We tested whether global personalised network connectivity, constructed from an a priori set of 9 depression-relevant linguistic features, is associated with baseline depression severity using Twitter data from 946 participants. In a subset of 286 participants, we sought to determine whether within-subject personalised network connectivity is greater within vs. outside a self-reported depressive episode. Finally, we checked whether changes in within-episode network connectivity generalised to 1,000 other combinations of 9-node networks.

4.3.1 Association of Twitter Text Features with Current Depression Symptomatology

As an initial step, we verified whether Linguistic Inquiry and Word Count (LIWC) text features averaged over the past year in our sample were significantly associated with current depression symptom severity (Table 4.2). Language pertaining to negative emotions (Neg. Emo), use of 1st person singular (1st Per. Sing), use of 2nd person pronouns (2nd Pers.), swear words (Swear), and negations (Negate) were significantly positively associated with current depression symptom severity. While use of 1st person plural pronouns (1st Pers. Pl.), articles (Articles), and words pertaining to positive emotions (Pos. Emo) were negatively associated with depression severity. There was no significant association with 3rd

person pronouns (3rd Pers.) (Figure 4.1). The proportion of days with Tweets within-subject that contained each of the 9 text features is presented in the online supplement (Table S2 in 7.3.5 Supplementary Materials). Swear words were the least frequently occurring text feature (Mean: 0.30, SD: 0.22), while articles were the most frequent (Mean: 0.80, SD: 0.14).

TEXT FEATURE	EXAMPLES	WORDS PER FEATURE	BETA	SE	P-VALUE
NEG. EMO.	Hurt, Ugly	744	0.17	0.03	<.001
1ST PERS. SING.	Me, I, Mine	24	0.14	0.03	<.001
2ND PERS.	You, your	30	0.08	0.03	0.02
SWEAR	Damn	131	0.11	0.03	<.001
NEGATIONS	Not, Never	62	0.12	0.03	<.001
1ST PERS. PL.	We, Our	12	-0.11	0.03	<.001
ARTICLES	A, The, An	3	-0.11	0.03	<.001
POS. EMO.	Love, Nice	620	-0.07	0.03	0.03
3RD PERS.	She, They	28	0.06	0.03	0.06

Table 4.2: Association of average use of 9 a priori text features over a 12-month period with current depression symptom severity.

In terms of consistency with the prior literature, negative emotions (Coppersmith et al., 2014; Coppersmith, Dredze, Harman, & Hollingshead, 2015; De Choudhury, Counts, et al., 2013; Eichstaedt et al., 2018; Lyons et al., 2018; Tsugawa et al., 2015), 1st person singular pronouns (Coppersmith, Dredze, Harman, &

Hollingshead, 2015; De Choudhury, Counts, et al., 2013; De Choudhury et al., 2014; Edwards & Holtzman, 2017; Eichstaedt et al., 2018; Lyons et al., 2018; Rude et al., 2004), swear words (Al-Mosaiwi & Johnstone, 2018; Coppersmith et al., 2014; De Choudhury, Counts, et al., 2013; De Choudhury et al., 2014; Reece et al., 2017; Rodriguez et al., 2010), and negations (Al-Mosaiwi & Johnstone, 2018; De Choudhury, Counts, et al., 2013; De Choudhury et al., 2014; Leis et al., 2019) were shown to be positively associated with depression severity. While, positive emotions (Capeceelatro et al., 2013; De Choudhury, Counts, et al., 2013; De Choudhury et al., 2014; Lumontod III, 2020; Molendijk et al., 2010; Rodriguez et al., 2010), 1st person plural (Coppersmith, Dredze, Harman, & Hollingshead, 2015; De Choudhury et al., 2014; Lyons et al., 2018), 2nd person (De Choudhury, Counts, et al., 2013; De Choudhury et al., 2014; Leis et al., 2019), and 3rd person pronouns (De Choudhury, Counts, et al., 2013; De Choudhury et al., 2014; Leis et al., 2019; Lyons et al., 2018; ODea et al., 2018) have been found to be negatively associated with depression. Article use has been found to be significantly associated with depression severity, although there is inconsistency regarding the direction of the effect (Al-Mosaiwi & Johnstone, 2018; Coppersmith, Dredze, Harman, & Hollingshead, 2015; De Choudhury, Counts, et al., 2013; De Choudhury et al., 2014; Reece et al., 2017). We therefore replicated previously established directional associations for 6 (negative emotions, 1st person singular, 1st person plural pronouns, swear words, negations, and positive emotions) of 9 LIWC text features, but were unable to replicate negative associations for 2nd person (we found a significant positive association) and 3rd person plural pronouns (trend-level in the direction of a positive association). Given conflicting evidence surrounding the direction of associations between article use and depression in the existing literature, our finding of a negative association can be neither confirmatory or dis-confirmatory at this point. Directionality notwithstanding, we were thus broadly assured that the text features we selected showed relevance to depression. Therefore, we used this set of linguistic features as nodes upon which to construct personalised depressive networks.

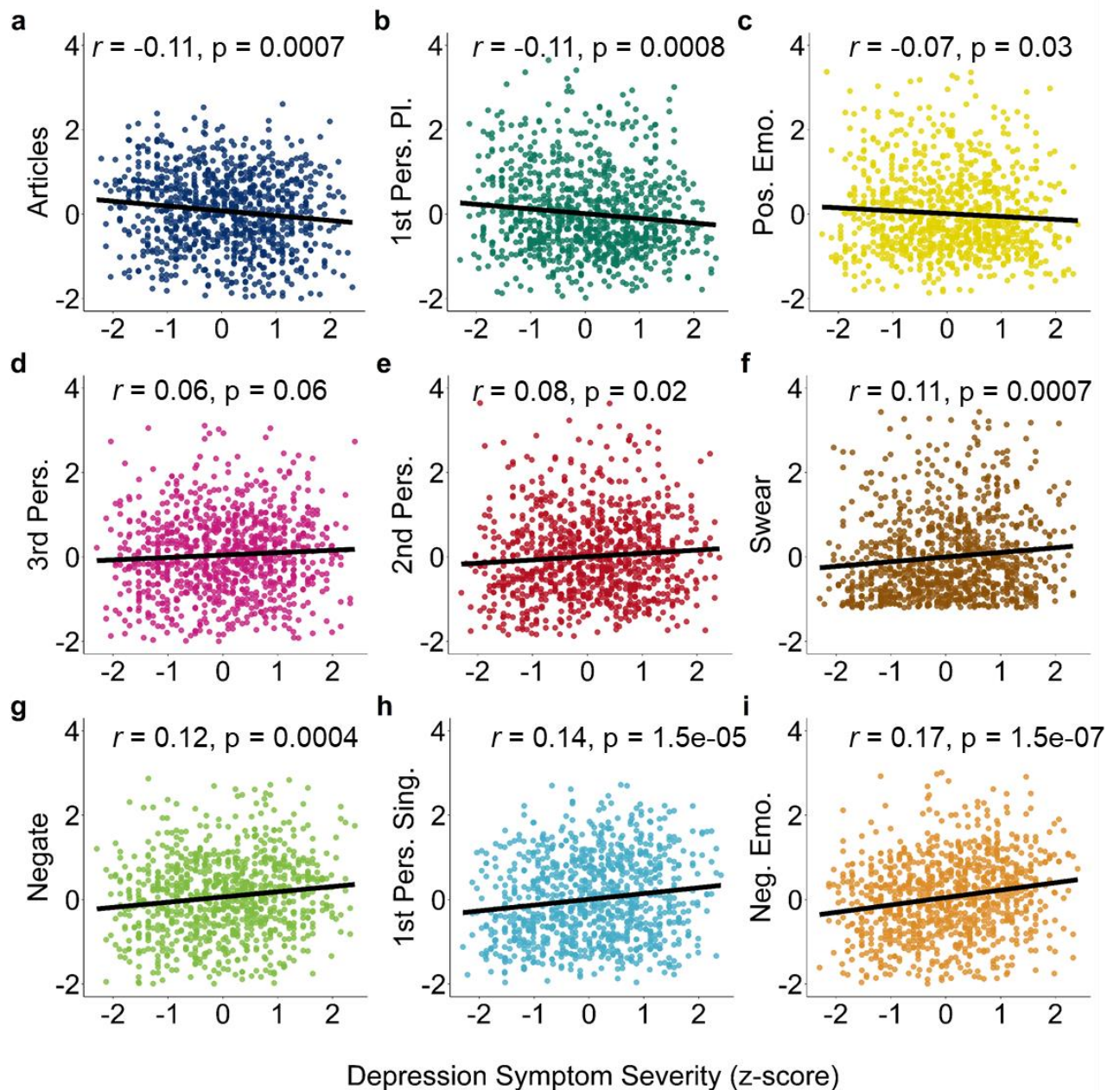


Figure 4.1: Self-reported depression severity is associated with several text features derived from Tweets

a-i) Association between self-report depression symptom severity and the mean of LIWC text features over the past year. There is a significant association between self-reported depression and every LIWC text feature, except 3rd person pronouns (Figure 4.1d). As we decided to construct networks based on these textual features a priori based on the work of de Choudhury et al. (2013), we nonetheless retained 3rd person pronouns in the subsequent analyses. Uncorrected two-sided Pearson correlations.

4.3.2 Overall Depression Network Composition

We constructed personalised networks for each participant, based on these 9 depression-associated text features derived from Tweets posted over the 12 months preceding study enrolment. From these individual networks, we tested how network structure differed as a function of depression severity. In support of a central hypothesis of network theory, we observed a significant positive association between depression severity and global network strength ($\beta = 0.008$, $SE = 0.003$, $p = 0.002$) (Figure 4.2a). That is, those individuals with the highest levels of depression had the most tightly connected networks in the sample. Participants with higher depression severity had significantly larger node strength of negative emotions, swear words and articles [Neg. Emo: $\beta = 0.02$, $SE = 0.007$, $p = 0.007$; Swear: $\beta = 0.02$, $SE = 0.007$, $p = 0.009$; Articles, $\beta = 0.01$, $SE = 0.003$, $p < 0.001$] (Figure 4.2b). The overall network of depression-related linguistic features was characterised primarily by several weak positive connections, and one strong positive connection between negative emotions and swear words (Figure 4.2c). There was a significant positive association between number of days and global network connectivity ($\beta = 0.00009$, $SE = 0.00002$, $p < 0.001$) (Figure S3a in 7.3.4 Supplementary Materials). However, there was no significant association between number of days in the time-series with current depression severity ($\beta = 0.0002$, $SE = 0.0003$, $p = 0.43$) (Figure S3b in 7.3.4 Supplementary Materials) and the significant association with current depression severity remained after controlling for the number of days in each personalised network (Table S3 in 7.3.5 Supplementary Materials). Furthermore, our findings were not affected when networks were constructed without 3rd person pronouns, which had no significant association with current depression severity (Figure S4 in 7.3.6 Supplementary Materials).

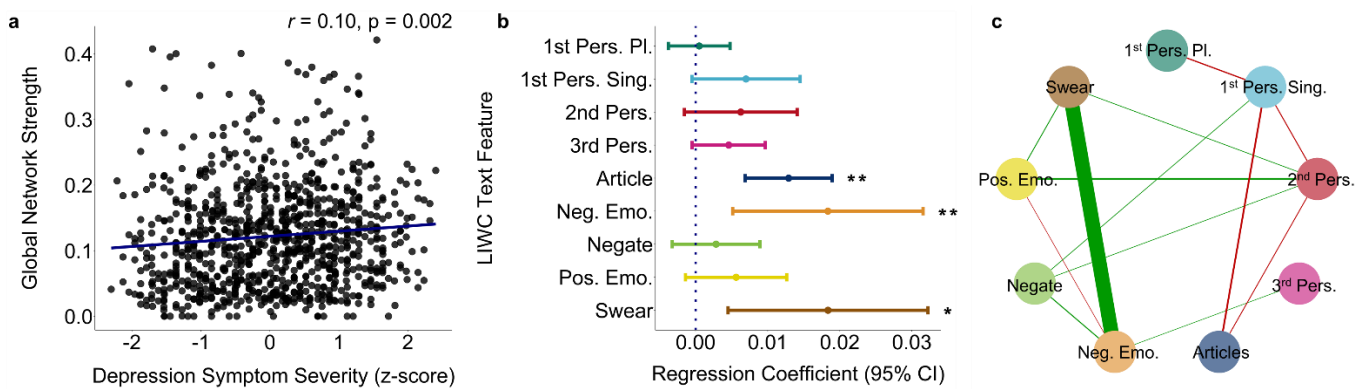


Figure 4.2: The connectivity of personalised networks of depression-relevant language is associated with individual differences in self-reported depression severity

a) There was a significant positive association between global network strength and depression severity ($\beta = 0.008$, $SE = 0.003$, $p = 0.002$) ($N = 946$). b) Mean regression coefficients with 95% CIs for individual network node strengths, positive coefficients indicate increased node strength with greater depressive severity. There was a significant association between depression severity and the node strength of Neg. Emo ($\beta = 0.02$, $SE = 0.007$, $p = 0.007$), Swear ($\beta = 0.02$, $SE = 0.007$, $p = 0.01$), and Articles ($\beta = 0.01$, $SE = 0.003$, $p < 0.001$). c) Mean personalised network of all participants. Green and red lines indicate positive and negative associations respectively. Line widths represent edge strength between nodes, with larger widths corresponding to greater edge strength. Associations were not affected after adjusting for the number of days, a proxy for network stability, (Table S3 in 7.3.5 Supplementary Materials) nor were they altered by omitting 3rd person pronouns from the network (Figure S4 in 7.3.6 Supplementary Materials). ^aUnadjusted two-sided Pearson correlation, ^bTwo-sided general linear regression unadjusted for multiple comparisons.

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

4.3.3 Within-Subject Changes in Network Connectivity during Depressive Episodes

To test the hypothesis that networks of depression-associated linguistic features become more tightly connected during an episode, we compared network

connectivity within-subject for periods when participants were depressed versus non-depressed over the preceding 12-month period. This required the construction of two personalised networks per person – one ‘within episode’ and another ‘outside episode’ among a subset of the sample who reported an episode in the past 12 months (N=286). In line with our hypothesis, the networks of our participants had a significantly higher global network strength if constructed using language data gathered during an episode (‘within’) versus a time when they were not currently depressed (‘outside’) ($\beta = 0.03$, SE = 0.009, $p = 0.005$, Figure 4.3A). We found our results were robust to unequal variances in the distribution of global network strength with the Wilcoxon-Signed Rank Test ($V = 16,840$, $p = 0.009$). We also performed the analysis using a bootstrapped sample of 80% of the data and re-did the within-subject analysis 1,000 times as a strong control for skewed strength centrality distributions. We found that the distribution of within-episode regression coefficients was significantly above zero ($\beta = 0.03$, SE = 0.0001, $p < 0.001$) (Figure S5 in 7.3.7 Supplementary Materials). In terms of the specific nodes of the network, during a depressive episode, 1st person singular (1st Pers. Sing., $\beta = 0.03$, SE = 0.01, $p = 0.03$), 1st person plural (1st Pers. Pl., $\beta = 0.04$, SE = 0.01, $p = 0.002$), 2nd person (2nd Pers., $\beta = 0.03$, SE = 0.01, $p = 0.04$), 3rd person (3rd Pers., $\beta = 0.04$, SE = 0.01, $p < 0.001$), use of articles (Articles, $\beta = 0.04$, SE = 0.01, $p = 0.01$), and negation words (Negate, $\beta = 0.05$, SE = 0.01, $p = 0.001$) all had significantly larger node strengths than the same networks constructed during times when the participants were not currently depressed (Figure 4.3b).

Changes in node strength were not due to mean increases in the text features themselves, because there were no significant differences among any of the text features within versus outside an episode (Table S4 in 7.3.9 Supplementary Materials). However, within-and outside episode networks of these participants had an average duration of 80.8 days (SD: 61.7) and 171.5 days (SD: 85.6) respectively, meaning that on-average participants spent considerably more time in a non-depressed state than in a depressed one ($\beta = -90.8$, SE = 6.2, $p < 0.001$, Figure S3c in 7.3.4 Supplementary Materials). This gave us cause for concern as we noted a significant negative association between global network connectivity and the number of days that within-episode networks were based on ($r = -0.16(286)$, $p = 0.01$) (Figure S3d in 7.3.4 Supplementary Materials). The

relationship is non-linear with the largest global connectivity values found during short (i.e., under 30 days) within-episode periods despite the removal of outliers. Additionally, there was a significant interaction such that the direction of association between the number of days of data and network strength depended on whether the data was from within vs. outside an episode ($\beta = 0.0005$, $SE = 0.0001$, $p < 0.001$). We reasoned, therefore, that the difference in number of days upon which within vs outside episode networks were based presented a potential confound to interpretation. Indeed, a permutation test that randomly shuffled the identifier (within-episode with outside-episode) within each participant showed a greater bias towards elevated connectivity for those (fake) within-episode periods than would be expected by chance ($\hat{\beta} = 0.007$, $SE = 0.0002$, $p < 0.001$; Figure S6 in 7.3.8 Supplementary Materials). Importantly, however, 99.3% of the betas observed were smaller than in the real unshuffled data, meaning that over and above any bias introduced by differences in the number of days within/outside episode, the true designation of being within an episode led to higher network connectivity. Consistent with this, after adjusting for the number of days in our regression, the significant increase in global network connectivity within versus outside an episode was reduced, but still statistically significant ($\beta = 0.02$, $SE = 0.01$, $p = 0.02$). Of the individual node strengths examined, only articles (Articles, $\beta = 0.03$, $SE = 0.02$, $p = 0.04$) still had a significantly larger node strength within an episode. Thus, the finding of increased network connectivity within versus outside an episode for our *a priori* depression network survived correction for the number of days within episode.

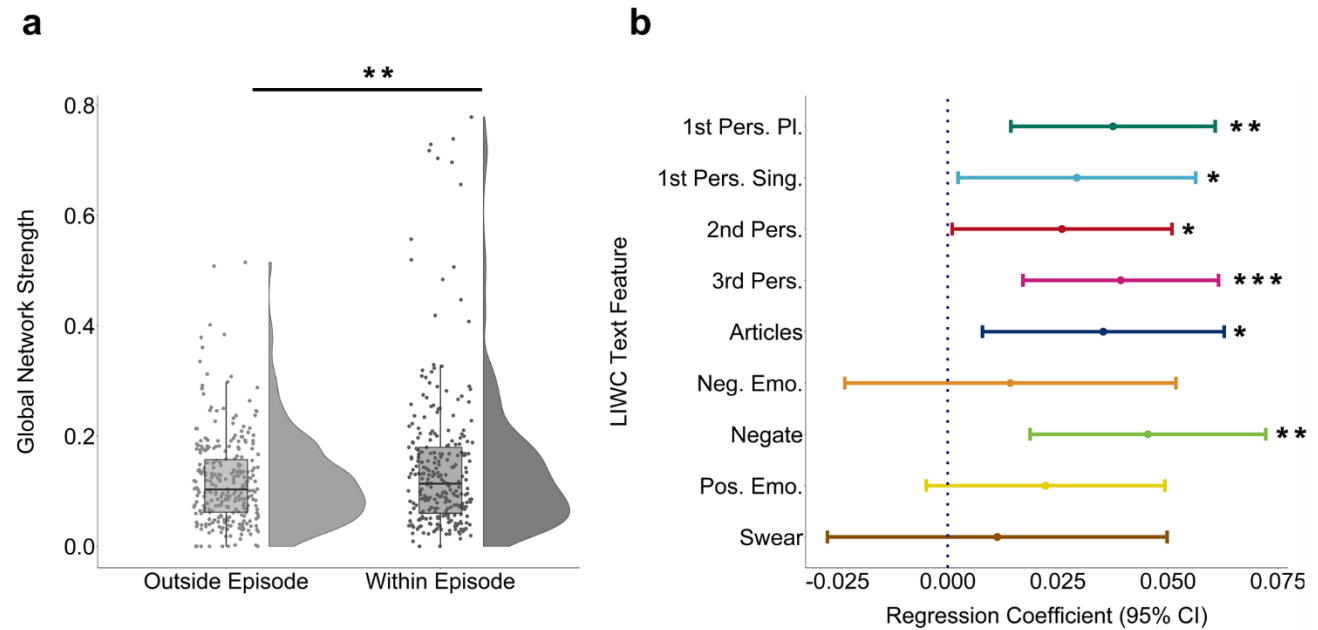


Figure 4.3: Personalised network connectivity increases during a depressive episode for specific symptoms

a) Regression coefficients for global network strength and b) mean individual network node strengths with 95% CIs for personalised networks of participants ($N = 286$) with a depressive episode in the past year for periods within and outside an episode, corrected for number of days. There was a significant increase in global network strength ($\beta = 0.03$, $SE = 0.009$, $p = 0.005$) and, among individual nodes, a significant increase in 1st person singular (1st Pers. Sing., $\beta = 0.03$, $SE = 0.01$, $p = 0.03$), 1st person plural (1st Pers. Pl., $\beta = 0.04$, $SE = 0.01$, $p = 0.002$), 2nd person (2nd Pers.), $\beta = 0.03$, $SE = 0.01$, $p = 0.04$), 3rd person (3rd Pers., $\beta = 0.04$, $SE = 0.01$, $p < 0.001$), use of articles (Articles, $\beta = 0.04$, $SE = 0.01$, $p = 0.01$), and negation words (Negate, $\beta = 0.05$, $SE = 0.01$, $p = 0.001$) node strength with an episode. Regression coefficients are derived from a within-subject linear model and no adjustments were made for multiple comparisons. The boxplot depicts the median (centre line), upper and lower quartiles, i.e., interquartile range, whiskers, i.e., $\pm 1.5 \times$ interquartile range. ^{a,b}Unadjusted two-sided within-subject regression model.

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

4.3.4 Generalisability of findings to other depression networks

The network of depression-associated features that we constructed was based on features previously described in the literature, and designed to ensure the independence of our analysis from mean-level effects or indeed noise in the present dataset. But it is important to note that this is not the only depression-related linguistic network that can be constructed from these data, nor is it necessarily the best. Of the 87 LIWC text features at our disposal, 59% were significantly associated with current depression severity at an uncorrected $p < 0.05$ level. Bivariate correlations between all LIWC text features and current depression severity can be found in the supplementary material (Table S5 in 7.3.10 Supplementary Materials). We thus tested if our results held when networks were constructed from different sets of 9 randomly selected text features associated with current depression severity. Networks of text features associated with depression have significantly larger within-episode connectivity than those of networks *not* associated with depression ($\beta = 0.01$, $SE = 0.0005$, $p < 0.001$, Figure 4.4a). The network with the largest increase in within-episode connectivity (Figure 4.4b) included the following depression relevant features: 1st person singular pronouns (“1st Pers. Sing”), clout (“Clout”: non-transparent summary variable indicating social status/leadership), personal pronouns (“Pers. Pron.”, e.g. you, they), function words (“Function” e.g. on, and), tentative (“Tentative”, e.g. maybe), negative emotions (“Neg. Emo”., e.g. ugly), power (“Power”, e.g. superior), negation (“Negate”, e.g. not), and achieve (“Achieve”, e.g. win). In the top 100 depression relevant networks, time and tentative words were found in 30% of networks (Figure 4.4c). Our a priori selected network is consequently not the only network with elevated within-episode connectivity nor is it the network with the largest increase in connectivity. But rather is part of a general trend that networks constructed from depression-relevant language features have greater connectivity when in the midst of a depressive episode.

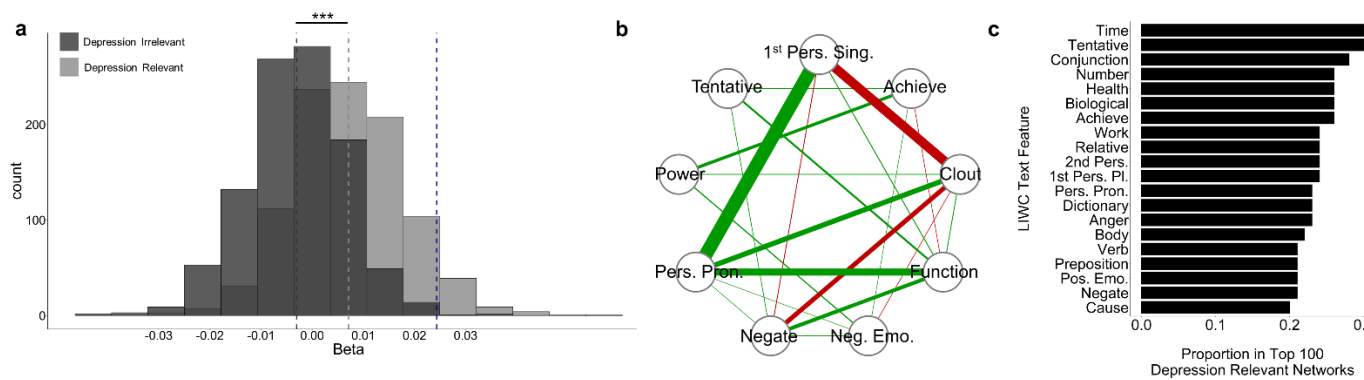


Figure 4.4: The generality of within-subject changes in network connectivity to other depression-relevant networks

a) Effect of episode on global network connectivity in 2000 random networks constructed from text features either significantly (light grey, N=1000) or not significantly (dark grey, N=1000) associated with current depression severity in bivariate correlations at the $\alpha = 0.05$ level. The dark grey and light grey dashed lines indicate the means of the depression irrelevant and relevant distributions respectively. Approximately 59% of LIWC text features were significantly associated with current depression. After adjusting for the number of days, networks of text features associated with depression have a significantly larger regression coefficient indicating elevated within-episode network connectivity ($\beta = 0.01$, $SE = 0.0005$, $p < 2e-16$). The effect of within-episode period in our network selected a priori is shown by a dashed blue line. Unadjusted two-sided general linear regression model

b) For illustration purposes, we show the personalised network with the largest increase in within-episode connectivity (N = 1 network). Nodes correspond to 1st person singular pronouns ("1st Pers. Sing."), clout ("Clout": non-transparent summary variable indicating social status/leadership), personal pronouns ("Pers. Pron.", e.g. you, they), function words ("Function" e.g. on, and), tentative ("Tentative", e.g. maybe), negative emotions ("Neg. Emo.", e.g. ugly), power ("Power", e.g. superior), negation ("Negate", e.g. not), and achieve ("Achieve", e.g. win).

c) Text features most likely to appear in networks sensitive to within-subject changes in depression status. Tentative and time related words were found in

30% of the 100 depression relevant networks found to be most sensitive to episode.

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

4.4 Discussion

The network theory of mental illness posits that causal interactions between symptoms result in positive feedback loops that lead to the development and maintenance of poor mental health episodes. This theory generates a range of predictions that have been difficult to examine in self-reported depression symptom data due to the difficulty in collecting large volumes of longitudinal self-report data. We adopted an approach to test these predictions by studying time series of linguistic features that are associated with depression, extracted from the social media platform Twitter. These linguistic features are outwardly observable indicators of a range of internal states that prior work has shown to be relevant to depression. While these linguistic features of depression cannot be directly mapped to individual clinically recognised symptoms, we nonetheless posited that they might interact and serve to reinforce one-another just as has been predicted by network theory for classic symptoms of depression. We predicted that networks constructed from these depression-relevant language features would be more strongly connected in those with higher levels of depression severity and moreover that they would become even more tightly connected when people were in a depressed state.

To test these initial predictions, we took a conservative approach in using 9 a priori text features with previously established relevance to depression from archival Twitter data. We found significant associations between 8 of 9 text features selected and current depression severity, of which 6 were consistent with now well-established directionality in the literature. These included positive associations between the use of 1st person singular pronouns and negative emotions and depression symptom severity. Next, we constructed personalized networks from these 9 features and found that higher levels of current depression severity were associated with greater connectivity of our a priori depression-associated linguistic network across participants. Crucially, we then leveraged the longitudinal nature of this dataset to study how connectivity changes within-subject as their mental health changes. Participants retrospectively reported periods of time when they had a depressive episode in the past year and we constructed networks for 'within' and 'outside' these dates. We demonstrated that the connectivity of depression-related linguistic networks increased within-subject as participants moved into periods of depression. This was true of our a priori

network, but crucially also for a range of 9-node networks constructed from randomly selected text features that had related to overall cross-sectional depression symptom severity. That is, networks constructed from depression-relevant text features were more likely to become tightly connected during an episode than networks constructed from depression-irrelevant text features.

Network theory offers a compelling explanation for the heterogeneity of disorders (E. I. Fried & R. M. Nesse, 2015) and are supported by patients' experiences of causal relationships between symptoms (Frewen et al., 2012; Frewen et al., 2013) and the efficacy of cognitive behavioural therapy, which aims in part to diminish associations between symptoms (Beck, 1979). However, there has been conflicting evidence in the literature regarding whether individuals with a mental illness have greater symptom network connectivity than healthy participants (Hakulinen et al., 2020). These results can partially be explained by the over reliance on cross-sectional data, which potentially averages out individual differences in network connectivity. Two prior studies found preliminary within-subject evidence of an increase in network connectivity during an acute phase of mental illness (Bak et al., 2016; Wichers et al., 2016). However, both involved only one participant. In this study, we established an increase in within-subject connectivity of a depression-relevant network during a depressive episode in a large sample. Crucially, we also leveraged Twitter text features as a tool for estimating personalised depression networks, rather than using self-report data. While that can be construed as a strength of our investigation, it is also a major weakness; there remains a critical gap in testing if self-reported symptoms would behave in a similar manner.

Using network analysis to understand individual vulnerability to depression is a promising avenue for potentially developing novel and personalized interventions. This is because symptoms with a high strength centrality are thought to have a disproportionate ability to activate or deactivate other symptoms. For example, evidence from a cross-sectional social anxiety disorder network suggests that changes in the most central symptoms in anxiety networks are predictive of more distributed changes in symptoms (Rodebaugh et al., 2018). In a prospective study of anorexia nervosa patients, higher levels of the most central symptoms at baseline were negatively associated with successful recovery (more so than less

central symptoms) (Elliott et al., 2019). However, a major caveat of research thus far is the use of cross-sectional networks to derive key insights – in some cases these align with personalised networks, but in others not (Bos et al., 2017). It is hoped that a push towards the development of more individualised approaches to network estimation will allow us to translate these basic findings into clinical practice. This might take the form of targeting symptoms that are, for the individual, most central, thereby preventing an individual from developing a disorder in the first place (Fried et al., 2017). To realise this potential, interpretable depression features will likely be essential. While we believe our data shows an interesting generalisation of network theory beyond self-report symptoms, more work will be needed to extract clinically actionable insights, if they exist, from the study of linguistic features of depression.

Our study was not without limitations and caveats. First, we are not suggesting that Twitter posts will ever (or should ever) be used to make clinical decisions. People on social media tend to selectively express their emotions, i.e. impression management, which obfuscates their true emotional state (Newman et al., 2011). Some use these platforms for work, to sell things, for self-promotion and in some cases, to vent their emotions. Therefore, the indirect assessment of depression-relevant language through text analysis will always lead to data that is substantially noisier than otherwise obtained via ESM and would never be of sufficient quality, in our view, to make individualised predictions. Indeed, the effect sizes reported here are low. This is in part because the linguistic features in tweets have low correlations with overall depression severity and also because our definition of a depressive episode was broader than is typical and based on a retrospective report. It therefore remains of key importance to establish if networks of goal-standard assessments of self-report depression symptoms display the same characteristics of the linguistic features studied here and to establish if effect sizes are clinically meaningful in such datasets. That said, we believe the present findings are of significant theoretical importance in two key ways.

First, in a large enough sample, we can use noisy data like this to test key aspects of network theory. The broad alignment of our findings with the prior literature (e.g., overall association of network connectivity and depression severity) and predictions of network theory (e.g., within subject changes in connectivity during

episodes) affirm there is clear signal in these data. There is significant potential, we believe, in using such data to answer questions that are otherwise practically impossible using EMA. Second, the proof of principle established here suggests that other sources of linguistic data that are potentially more indicative of current mood (e.g., text messages, speech) could be mined to help deliver personalized warning signs to individuals. In this context, it is noteworthy to also acknowledge that our ground-truth measure with respect to depressive episodes is a retrospective report and likely subject to errors in recall. This would only serve to diminish our effect size further, which in our view, makes our results more compelling. In a real-world context, it is likely that changes in connectivity associated with episodes of depression are stronger than reported here.

Other limitations to our data include the fact that social media users tend to be younger, better educated, wealthier, and politically more liberal than the general population, posing potential generalizability problems (Mellon & Prosser, 2017b; Wojcik & Hughes, 2019). Additionally, online workers on platforms similar to ClickWorker have their own unique sociodemographic profile and crucially, endorse higher rates of a range of mental health problems than the general public (Ophir et al., 2020; Shapiro et al., 2013). In the present study, a large proportion of participants had received a mental health diagnosis in the past and more than half reported a depressive episode in the past year. The high rates of the latter were likely partially inflated by our decision to require just 2 symptoms of depression (low mood and reduced interest) to have been present consistently for a two-week period (instead of the usual 5 of 7) (Association, 2013), but are likely to be partially due to the known profile of online workers. It remains to be tested if the findings from this study will generalize to individuals recruited via other means and indeed through clinical settings. The majority of text and sentiment analysis libraries used to examine the association between language features and mental health are only available in English (Medagoda et al., 2013). Significant differences have been shown in use of negative, positive emotions, personal pronouns, articles, and other lexical attributes between social media users in western (U.S. and U.K.) and non-western (India and South Africa) countries (De Choudhury et al., 2017). The vast majority of our sample came from predominantly English-speaking countries. Because of this, we do not know how these associations may generalise to different

languages and cultural settings. Similar to several other personalised network papers (Aalbers et al., 2019; Cuevas et al., 2020), we found that none of the nodes in our a priori network was normally distributed. Network analysis assumes that all nodes are multivariate normally distributed (Sacha Epskamp et al., 2018), however it is not yet known the extent to which edge and centrality estimates are affected by deviations from normality. Finally, the LIWC is only able to account for the proportion of words in a particular category, e.g., proportion of 1st person pronouns in a passage of text. Any context or more subtle usage of language, such as irony, that would change the underlying emotional meaning of a text are not captured by this method. More broadly, there are a range of more sophisticated analytical approaches when it comes to the content and sentiments in tweets that may prove stronger indicators of depression (but see (Chancellor & De Choudhury, 2020)) and thus better candidate nodes to construct depression-relevant networks. We chose instead to use an established library and to focus on language features previously shown to relate to depression to keep a degree of independence in the datasets used to derive depression-relevant features and the one (here) used to study how their network compositions changes through time. Future work might draw on alternative methods and have greater power to interrogate network dynamics in these datasets.

We found support for two of the principal predictions of network theory using a proxy for longitudinal (historical) EMA. Specifically, we found that the connectivity (partial correlation) between a set of pre-defined linguistic features of depression relates to an individual's current depression symptom severity. Moreover, we found that this network connectivity increases within-subject during a depressive episode. Future work can utilize this methodological approach to test and refine key aspects of network theory. Elevated network connectivity within an episode was not specific to the a priori LIWC text features chosen; they generalised to a broader set of linguistic features that are associated with depression severity and future research might elect to utilize the best performing network we identified here. Whether these findings generalise to other aspects of mental health is not yet known. Recent work suggests that there are a host of commonalities across various aspects of mental health in their use of language on Twitter. Regardless of whether there is some degree of specificity of the nodes that comprise such

networks, it will be interesting to determine if network connectivity increases, within-subject, occur during the acute phase of other mental health illnesses, such as bipolar disorder. Given the vast amount of data available and its longitudinal archival nature, social media network analysis is a promising method for testing some of the tougher predictions of network theory, albeit using very different 'markers' of depression.

Chapter 5: General Discussion

5.1 Summary

Although depression is a common mental health condition, we continue to have a relatively rudimentary understanding of the mechanisms underlying how it emerges and is maintained. A significant proportion of people treated for depression do not recover, and we do not know why (Souery et al., 2006). People can be exposed to similar stresses but have vastly different outcomes (Fried & Nesse, 2014). This heterogeneity might be explained by complexity in mental health dynamics that current cross-sectional models fail to capture. Big data, which can be gathered through sources like social media and (to a lesser extent) EMA, is needed to test the ideas and bring them into practice. The aims of this thesis were centred on these issues. First, I aimed to explore novel methods for gathering rich and repeated depression-relevant data, at a scale amenable to a range of data-intensive analyses approaches like machine learning and network analysis. Secondly, I aimed to use these methods to advance our understanding of depression, testing key predictions of the network theory of mental health and clarifying how social media data can be utilised for testing theories, but not for making specific and accurate mental health predictions.

The aim of the first experimental chapter, Chapter 2, was to determine the accuracy of machine learning models trained to detect depression self-report severity scores using language features on Twitter and to assess the specificity of these models to scores of 8 other mental health conditions. Prior work has found differences in univariate aspects of language use between individuals with and without depression. A well-replicated example is that people with depression use significantly more first-person singular pronouns, e.g., “I”, than healthy controls. Extending these results, machine learning models trained on language from social media have demonstrated a high degree of accuracy in classifying individuals with depression, but also other mental health conditions. A caveat to this work is that it does not assess the specificity of models in predicting depression compared to another condition, e.g., obsessive-compulsive disorder. Here, we collected self-

reported questionnaire data on 9 mental health conditions along with all available Twitter data, i.e., posts, retweets, likes, from the past year from 1,006 participants. An elastic net regularisation model was trained on language features derived from Twitter and used to predict out-of-sample depression severity. The model had weak predictive power, $R^2 = 0.025$, and when tested on the 8 other mental health conditions had above zero predictive power for all conditions, except impulsivity and alcohol abuse. To compare to previous studies, the model was also trained on binarized depression scores resulting in a maximum AUC of 0.59. The weak performance of these models stands in contrast to other work which found moderately high classification accuracy - AUC of 0.70 to 0.89 (De Choudhury et al., 2014; De Choudhury, Gamon, et al., 2013; Eichstaedt et al., 2018; Reece et al., 2017).

One potential reason for the inconsistency in model accuracy is that a common methodological approach in the literature is to use self-disclosure, e.g., "I feel depressed", as a measure of mental health status. Typically, this is done out of convenience to acquire a large sample size because consenting and gathering self-report gold-standard depression data from more than a few hundred participants becomes quickly infeasible. The major drawback to this approach is that it threatens the validity of diagnosis and can introduce circularities. Although some studies have tried to obviate this concern by asking clinicians to rate the validity of self-disclosure posts (Shen et al., 2013), without directly assessing a person's symptoms it remains challenging to determine which posts signify true cases of mental illness. Second, even if when diagnoses assigned using this method have good specificity, there are serious challenges to sensitivity and generalisability. Many are not willing to disclose mental health problems on social media and those who do disclose publicly may not be representative.

Problematically, those who disclose online might be more likely to discuss depression-relevant topics in their posts, which would emerge (circularly) in their language. Thus, compared to self-report questionnaires, proxies for mental health on social media are not reflective of a person's 'true' health and likely lead to overestimates of classification accuracy. We illustrated this in Chapter 2; when the model was instead trained on relevant depression keywords, i.e., "depressed" or "depression", performance was substantially higher, AUC of 0.83.

Other studies have also found that language use patterns are not specific to any particular mental health condition (Coppersmith, Dredze, Harman, & Hollingshead, 2015). However, because these studies use binary outcomes of mental health status, e.g., depressed vs. not-depressed, they cannot account for the relative contribution of other mental health symptoms. An advantage of the self-report scores used here is that we can assess the specificity of language use patterns, after controlling for shared variance due to other symptoms. When models were trained on the residuals of previously established trans-diagnostic dimensions (Gillan et al., 2016; Rouault et al., 2018) (anxious-depression, compulsivity and intrusive thoughts, and social withdrawal), unique patterns of language use emerged.

Chapter 2 focused on assessing mental health using aggregate measures, inferred via social media. But we know that emotional states change dynamically over time and interact with each other. Chapter 3 sought to examine these dynamics, testing key tenets of network theory. The network theory of mental health predicts that emotion networks that are more connected will lead to greater depression severity. However, mixed empirical evidence exists for this prediction due to the reliance on cross-sectional studies (Rodebaugh et al., 2018; Spiller et al., 2020). Personalised networks overcome this limitation by estimating causal associations between emotions at the level of the individual. To test this, we gathered two independent community samples (Paid Students and Citizen Scientists) and tested our findings on one large community sample and a smaller clinical one. We identified comparable networks across datasets and illustrated that the network structure was highly similar regardless of the sample, and each produced stable networks. In Citizen Scientists, we found that elevated contemporaneous network connectivity was positively correlated with depression severity. This finding generalised to a wide range of network types, composed of different emotions and varying numbers of negative emotions. When tested using directed networks, similar results were found. However, no significant relationship between connectivity and severity was found in the Paid Students. In trying to uncover why this was, we examined variance of depression severity over the 8 weeks and discovered that it was more confounded with baseline depression severity in Citizen Scientists than Paid Students. That is, the relationship between

baseline depression and its 8-week variance was much larger in Citizen Scientists ($r=.52$) compared to Paid Students ($r=.18$).

This is relevant, because some have suggested that elevated connectivity of emotion networks should lead to large fluctuations in symptom severity – both increases and decreases – rather than being specifically linked to greater severity. This is because covariance in the constituent components of a network confers lower resilience to external perturbations to the system (Cramer et al., 2016; Dakos et al., 2009), making it more changeable. We posited that the findings in Citizen Scientists might therefore truly reflect depression variance, not severity at all. Taking advantage of 8 weeks of depression self-report assessments, we tested this possibility. After accounting for fluctuations in weekly depression symptoms, the association between connectivity and baseline severity disappeared. We replicated this fully in a third datasets, where we also observed correlations with baseline severity, which were better explained by variance in depression. These findings demonstrate that emotion network connectivity is primarily related to symptom variability rather than severity, which also generalised to different types of samples and network compositions. We tested these ideas on a very small clinical sample of diagnosed patients, but the results did not extend, possibly due to the small sample size and high rates of depression severity.

The final experimental chapter, Chapter 4, combined aspects of chapters 2 and 3 to demonstrate that social media can function as a proxy for EMA and be used to evaluate a challenging prediction of network theory. While EMA allows for a direct interrogation of a person's momentary emotional states over time, it can be difficult for participants to remain adherent for long periods (Froehlich et al., 2007). For this reason, most EMA studies are typically no more than several weeks, in duration in order to minimise attrition rates. A systematic review found that most studies were either 7 or 14 days long with a mean of 22 days (Hall et al., 2021). Few studies, thus, have been able to monitor long-term changes in depression and because of that there is limited knowledge about changes in network connectivity during naturally occurring depressive episodes. Some evidence exists, from n-of-1 studies, that symptom network connectivity increases prior to the onset of a depressive episode as well as during the episode (Wichers et al., 2016; Wichers et al., 2020). To test these ideas quantitatively in a larger sample,

we used text features associated with depression as proxies for depression symptoms to construct personalised networks. We tested if the connectivity of these networks were associated with depression and if that connectivity changes when someone is acutely unwell, i.e., during periods when participants reported being depressed versus healthy. Network connectivity was shown to be associated with severity and elevated during self-reported depressive episodes. Importantly, these elevated network connectivity findings were depression-specific and generalised to networks constructed from any number of depression-relevant language features.

5.2 Implications

5.2.1 Early Warning Signals for Mental Health

A principal goal of precision psychiatry is to develop early warning signals to detect the onset of depression and other mental health conditions. Depression relapse is the norm rather the exception, at least 50% of people will relapse after one episode (Burcusa & Iacono, 2007). An early warning system would enable patients to get interventions at the time when they will have the largest impact (Abuse, 2019). Consistent with a non-linear system, the majority (>60%) of people with depression experience two stable states (healthy and depressed) with rapid transitions between them (Hosenfeld et al., 2015). Tightly connected networks produce sudden changes in symptom severity, such that the system is switching between two equilibria (van Borkulo et al., 2011). This thesis found that elevated connectivity is not only related to greater depression severity but also whether someone is currently experiencing a depressive episode. Because depression episodes have distinct network properties, i.e., elevated within-episode connectivity, it may be possible to identify signals that precede the transition into depression. Network theory posits that the transition from one state to another should be preceded by increasing variance, autocorrelation, and network connectivity among a system's components, which are collectively referred to as critical slowing down (Dakos et al., 2009; Scheffer et al., 2009). This is because prior to a state transition, a perturbation will result in slower recovery back to baseline as the system is unable to rapidly mitigate its effect. Elevated network

connectivity means that a perturbation's effects on one component of a system do not remain isolated, instead affecting a broad swath of the network.

While not all state transitions are preceded by critical slowing down signals (Boettiger & Hastings, 2012), preliminary evidence suggests that they may be able to forecast upcoming depressive episodes. In one study, a person rated their momentary emotional states 10 times per day for 239 days. The authors found that prior to a significant increase in depression severity, i.e., a state transition, there was elevated variance, autocorrelation, and network connectivity among the mental states (Wichers et al., 2016). More recent and larger studies ($n > 1$) can some doubt on whether these signals are reliable markers of psychological transitions. An EMA study of 31 participants diagnosed with depression who rated their positive and negative emotions for 8-weeks found rises in autocorrelation was associated with worsening depression, but not variance nor network connectivity (Curtiss et al., 2021). The presence of early warning signals in patients with bipolar disorder increased the probability of a transition to either mania or depression, but was nevertheless accompanied by high rates of false negatives/positives, sensitivity to estimation methods, and widespread variability across participants and affective states (F. M. Bos et al., 2022).

However, we were not able to determine whether early warning signals were present in our data. Twitter data had too much missingness at the individual-level to be useful for granular time-series analyses prior to onset of a depressive episode. Additionally, to acquire enough data for textual analysis, we needed to aggregate Tweets into daily bins. While this process increased the reliability of text features, it also reduced the sensitivity to detect short-term emotional changes. In the 8-week EMA study, it was challenging to identify robust evidence for state transitions within the relatively short timeframe. Even if we could identify states transitions in our data, there likely would not be enough data prior to the transition to estimate changes in variance and autocorrelation. Operationalising early warning signals remains difficult with significant challenges to overcome, nevertheless there is clear clinical value in the early detection of changes in mental health severity.

5.2.2 Clinical Usefulness and Ethical Considerations of Social Media

Social media has shown its potential to assess and monitor both physical and mental health at the population-level. Despite this, machine learning models cannot say much about the health of any given individual. Given the inability to produce meaningfully accurate individual predictions, is there a place for social media in the clinic? Social media data could potentially augment a person's therapeutic relationship with their therapist, by providing nuanced information about symptoms and to keep track of long-term changes (Yoo et al., 2021). Compared to self-reported data, social media's primary advantage is that it can be done without a patient's active involvement. Unobtrusive monitoring could also encourage people, who would otherwise not, to seek out treatment. As Chapter 2 demonstrated, social media data can only provide limited value in detecting mental health problems. Despite this limitation, social media could be the initial phase of a multi-stage public health initiative to reduce depression. By identifying individuals at-risk for depression via social media, people could be guided through additional checks, i.e., validated self-report questionnaires, that can more definitively establish mental health status. After that, people would then be directed to further resources to access care if needed.

The use of social media would also pose challenges to the workflow of clinicians. With constant access to a patient's social media, who is responsible for duty of care? For example, if a patient posts that they are suicidal on a Friday evening and subsequently dies by suicide later that evening, can their mental health team be held liable? Questions regarding clinical governance present significant barriers to practical adoption that need to be addressed in step with research on usefulness. Social media is a promising non-invasive tool to assess health at scale, yet not everyone is willing to provide this data due to privacy concerns. Qualitative semi-structured interviews of a general public focus group found broad support for using social media for research purpose, especially when it was seen to be beneficial to society (Yoo et al., 2021). Scepticism remains over using even publicly available and archived data because of the perceived harm. Yet, people routinely provide private companies with highly sensitive personal information, e.g., name, contact information, home address, banking details etc. The key difference then is not how sensitive the data is, but the breach of trust in how this data has been handled and the exact nature of data that is being shared (Zimmer,

2018). When users post on social media, they do not think that they are revealing anything more than what they explicitly state and do not expect their data to be used by third parties, e.g., researchers. Language use patterns are one way to reveal 'hidden' information about the user that they did not necessarily consent to sharing. In this thesis, we ensured that participants were explicitly made aware of how their data would be handled and only included people who consented to providing their Twitter data.

Social media's greatest potential perhaps lies in the ability to draw upon large samples in a naturalistic setting. One of the most ambitious applications of this is to quantify the happiness of all users posting in a particular language.

Researchers developed this application, called a Hedonometer, which assesses 10% of all Tweets per day (~50 million) and assigns an average happiness score (Dodds et al., 2011). This functions as a sort of global reading of mental health with the added benefit that the impact of major world events, e.g., natural disasters (Kryvasheyeu et al., 2015), can be observed in near real-time. Along with temporal information, differences in happiness based on geographic location, i.e., U.S. state, can be determined through Twitter (Mitchell et al., 2013). Moving beyond very broad population-level measurements, social media can also help uncover the emotional dynamics at much shorter timescales. Emotional intensity on Twitter, expressed through "I feel" statements, rapidly declines and within 1.5 hours is back to baseline levels of valence (Fan et al., 2019). This study would be challenging without already extant data, requiring large number of participants to report their emotions at the precise moment they occur with follow-up on the decline in intensity. Similarly, depressed Twitter users, compared to a control group, were found to post more during the evening and experience a rise in rumination and negative during through the early morning (Ten Thij et al., 2020). These studies provide examples of health-related questions that social media, in particular Twitter, can answer by employing novel sources of data. There are however, important confound to consider that can affect the validity of emotional states derived from social media. The use of sarcasm changes the valence of a statement, e.g., a negatively worded sentence implies a positive sentiment. Considering whether sarcasm is present is crucial for determining a person's true emotional state when writing. Despite some research into automatically identifying

sarcasm, classification performance remains low (González-Ibáñez et al., 2011). Another concern is how social desirability bias, i.e., the tendency for people to present themselves positively in order to be viewed favourably, affects what people are willing to share about their mental health via social media (Newman et al., 2011). Because of this, some people are less likely to post any health information on public social media platforms. As shown in Chapter 2, using keywords to assess mental health status leads to very different outcomes compared to validated questionnaires.

5.2.3 Treatment Applications of Network Theory

Network theory has generated much enthusiasm in both researchers and clinicians because it offers a compelling and intuitive theoretical model of depression. But perhaps even more so because of its potential to allow us to uncover the mechanisms by which existing treatments are effective and personalise treatments to improve their efficacy. In this thesis, we sought to evaluate key predictions of network theory regarding how increased connectivity manifests in depression (via severity vs. fluctuations). As demonstrated in Chapter 3, network connectivity is associated with greater fluctuations in depression severity over time. From this, we can infer that network connectivity relates to a person's emotional resilience. Within a more tightly connected network, i.e., a less resilient system, symptom activation easily propagates throughout the network which leads to enhanced sensitivity to external stressors. While less resilience is generally related to poorer outcomes, e.g., worse depression, it also signals that a system is able to change and incorporate novel information. Greater variability in depression has been shown to be associated with better treatment outcomes (Olthof et al., 2020; Schiepek et al., 2014; Tang & DeRubeis, 1999). Contrary to predictions, network connectivity increases, rather than decreases, after treatment. In light of some of the results presented in this thesis, together with this prior work, we posit that treatment may work to partially destabilise a person's depression network thereby facilitating the transition from a depressive state to a healthy one (Thelen & Smith, 1998). Treatments delivered during periods of relative instability may be more likely to enable change than those performed during stable times. Monitoring changes in network connectivity over time could

then be useful for identifying the best time to administer certain targeted interventions.

The other major promise of network theory is that it will lead to enhanced personalisation of treatment, moving away from a one size fits all approach. Although this makes intuitive sense, there have not yet been randomised trials to evaluate whether the use of a network informed approach improve clinical outcomes, compared to either randomly generated networks or care as usual (Bringmann, 2021). Additionally, network theory offers new avenues to explore mechanisms of treatment action. While studies have typically focused on connectivity as the primary characteristic to describe networks by, other network related factors also influence vulnerability to depression. One example, is the likelihood that a particular symptom within the network will become activated, independent of that symptom's relationship to other symptoms. Decreasing the probability that one symptom activates, reduces the likelihood that other symptoms within the depression network will then be activated.

5.2.4 Limitations and Future Directions

The studies described in this thesis were designed to be conservative and robust, incorporating replication and independent validation throughout. However, they were not without limitations. Several of these limitations concern the use of social media to estimate mental health status. Machine learning models were trained solely on LIWC text features, which groups words into predefined categories based on emotional content and grammatical characteristics. Algorithms trained on topics, i.e., groupings of words commonly used together, were shown to perform better in predicting psychological attributes since they can more effectively extract relevant patterns of language use. However, a set of topics is unique to a particular text and does not necessarily generalise well to other texts, because naturally these topics will change through time and in response to world events. While the LIWC may be at a disadvantage in predictive performance, it has the advantage of being reliable, interpretable and using psychologically validated categories (Chancellor & De Choudhury, 2020). The majority of people do not post frequently, for example on Twitter 80% of content is generated by 10% of users (Auxier & Anderson, 2021). Samples are social media, therefore, are

biased to certain groups with high engagement. People active on Twitter tend to be unrepresentative of the general population e.g., they are better educated, more politically liberal (Auxier & Anderson, 2021). Although the studies presented here used Twitter, some evidence suggests that Facebook data may be more predictive of mental health (Jaidka et al., 2018). Most text-based methods to assess the emotional content of language are based on English, and have not yet been extended to a wider range of languages. Languages are also known to have a strong positivity bias, i.e., positive words are more commonly used than negative ones (Dodds et al., 2015), which could lead to negative emotions being underrepresented in social media studies.

Furthermore, social media platforms have shifted away from being primarily text-based to primarily focused on images and video. This shift coincides with the emergence of new platforms that are radically different in function and form to Facebook, Twitter, and Reddit (the 3 main platforms mental health studies tend to use). Limited work exists on predicting mental health from either images (Birnbaum et al., 2020; Guntuku et al., 2019; Reece & Danforth, 2017) or video content (Ashraf et al., 2020), although these formats have the potential to improve over text-based models. Videos contain information not only about language but prosody, body language, and facial expressions all of which reflect unique aspects of mental health. Despite the expansion of social media into different formats, less research has been done on language used in private communication platforms, e.g., WhatsApp, and mental health (Glenn et al., 2020; Liu et al., 2022). Private messages are more likely to reflect an individual's true mental health status, without masking or image management, than public forums. While linguistic analysis of private messages may provide interesting research avenues, and a stronger signal to noise ratio than public data, it is unlikely to be accepted by the general public.

In Chapter 3, we measured the dynamical properties of emotions over time and related that to changes in depression symptom severity and fluctuations in depression over time. The primary limitation is that, like most EMA work, this was a naturalistic study. Consequently, any changes in emotional states that arose are due to uncontrolled daily events and stressors of varying duration, type, and magnitude. Treatment, either medication or therapy, likely directly affects network

connectivity, moderates the effect of stressors on the network, and causes issues with stationarity. Time-varying models are a possible solution that can account for systems with changing dynamic properties (Haslbeck & Waldorp, 2015). Our findings were restricted to the general population. While we attempted to replicate results in patients diagnosed with depression and anxiety, the sample size ($n = 45$) was underpowered to detect relevant associations with depression and the results are therefore equivocal.

While variability in depression symptoms is potentially an indicator of maladaptive emotional dysregulation (Bos et al., 2019), depending on the context, it could be an appropriate response. Therefore, the clinical significance of these fluctuations is not clear. Perhaps due to the short duration of the study, we did not find evidence for ‘critical transitions’, i.e., distinct shifts from depressed to non-depressed or vice versa; indeed, this precluded us from carrying out analyses pertaining to early warning signs. This limitation was partially addressed in Chapter 4, by using self-reported depressive episodes within the past year, but that does not address the core of the issue. Forcing individuals to binarize their experience of being depressed and non-depressed does not mean that the underlying reality was not a gradual transition at either end. Emotional states were assessed relatively infrequently, ranging from twice a day in Paid Students and Citizen Scientists to four times a day in the Clinical Sample. The coarse timescales mean that significant changes in emotions could have occurred, but would not necessarily show up in our data (Fisher & Newman, 2016; Fisher & Woodward, 2014). This could lead to mis-specified relationships among emotions and impede the recovery of a system’s true dynamics (Haslbeck & Ryan, 2021). As a partial solution, we focused our analyses on contemporaneous associations which are not only able to capture faster changes in affect, but are also considerably more stable than directed ones. Furthermore, vector autoregression – used here to estimate personalised networks – cannot estimate non-linear dynamics which is a characteristic of bistable systems, e.g., a healthy and depressed state.

One of the primary limitations of Chapter 4 is that the personalised networks were constructed using depression-relevant language features as proxies for depression symptoms. Because of this, it can be difficult to causally interpret

relationships between language features. Individual language features are known to correspond to particular aspects of depression symptomology. For example, increased usage of first-person singular pronouns is a marker of elevated self-rumination (Edwards & Holtzman, 2017). Zooming out to the network-level, however, connectivity among linguistic features can be a reasonable approximation of a person's affective dynamics. In an EMA study, participants complete a fixed set of items that are consistently asked each time they complete an assessment. For example, swear words occurred relatively infrequently, on average 30% of days, while articles were represented on 80% of days. Compared to EMA, each linguistic feature is not necessarily represented each day that a person tweets, given the relatively large possible feature space. Thus, the absence of a particular feature or emotion, e.g., negative emotions, on a particular day is not evidence of the absence of that feature. At the individual-level, personalised networks constructed from language are more subject to bias and may not accurately recapitulate emotional dynamics relevant to depression.

Participants were asked to self-report any depressive episodes in the past year based on periods where they experienced low mood and loss of interest/pleasure in activities for at least 2 weeks. These criteria were developed by us so that it could capture the key components of a depression diagnosis without overly burdening participants. This came at the expense of specificity, such that episodes could be reported even if they do not adequately meet the clinical threshold of a diagnosis (which requires more specific symptoms beyond just low mood and loss of interest). Also, because episodes are self-reported as a binary outcome – either present or not –, it is not possible to compare the severity of a depressive episode either across between individuals or within the same individual over time. Finally, we do not have any information regarding the reliability of participants to retrospectively report depressive episodes. Accurate recall is likely subject to a recency bias, which could lead to episodes being underreported if they occurred farther from the time when they completed self-report measures. In spite of these limitations, we were able to robustly demonstrate the network connectivity increased within naturally occurring depressive episodes at the within-individual level.

5.2.5 Conclusions

In summary, this thesis sought to test predictions of the network theory of depression and identify novel methods to predict its prevalence. Through this work, we confirmed a central prediction of network theory that depression is related to elevated network connectivity. In one study, we further interrogated this point to find that elevated network connectivity actually leads to larger fluctuations in depression, instead of simply greater symptom severity. In the process, we developed a novel way to construct personalised networks using language features from social media data, thereby expanding the capacity to evaluate longitudinal changes in mental health over time. While a promising methodological advancement, we also found that changes in language use are not specific to depression and largely unable to accurately measure an individual's mental health. This thesis used EMA and linguistic analysis to provide new mechanisms for understanding the emergence and maintenance of depression. Through early detection and personalisation, data science has tremendous potential to revolutionise our treatment and understanding of mental health.

6 References

- Aalbers, G., McNally, R. J., Heeren, A., De Wit, S., & Fried, E. I. (2019). Social media and depression symptoms: A network perspective. *Journal of Experimental Psychology: General*, 148(8), 1454.
- Abuse, S. (2019). Mental Health Services Administration.(2019). Key substance use and mental health indicators in the United States: Results from the 2018 National Survey on Drug Use and Health (HHS Publication No. PEP19-5068, NSDUH Series H-54). Rockville, MD: Center for Behavioral Health Statistics and Quality. *Substance Abuse and Mental Health Services Administration*. Retrieved from <https://www.samhsa.gov/data>.
- Aderka, I. M., Nickerson, A., Bøe, H. J., & Hofmann, S. G. (2012). Sudden gains during psychological treatments of anxiety and depression: a meta-analysis. *Journal of consulting and clinical psychology*, 80(1), 93.
- Afifi, T. O., Enns, M. W., Cox, B. J., Asmundson, G. J., Stein, M. B., & Sareen, J. (2008). Population attributable fractions of psychiatric disorders and suicide ideation and attempts associated with adverse childhood experiences. *American journal of public health*, 98(5), 946-952.
- Agrawal, A., Fu, W., & Menzies, T. (2016). What is wrong with topic modeling?(and how to fix it using search-based se). *arXiv preprint arXiv:1608.08176*.
- Al-Harbi, K. S. (2012). Treatment-resistant depression: therapeutic trends, challenges, and future directions. *Patient preference and adherence*, 6, 369.
- Al-Mosaiwi, M., & Johnstone, T. (2018). In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation. *Clin Psychol Sci*, 6(4), 529-542. <https://doi.org/10.1177/2167702617747074>

- Alim, T. N., Feder, A., Graves, R. E., Wang, Y., Weaver, J., Westphal, M., Alonso, A., Aigbogun, N. U., Smith, B. W., & Doucette, J. T. (2008). Trauma, resilience, and recovery in a high-risk African-American population. *American Journal of Psychiatry*, 165(12), 1566-1575.
- Altamura, A., Dell'Oso, B., Mundo, E., & Dell'Oso, L. (2007). Duration of untreated illness in major depressive disorder: a naturalistic study. *International Journal of Clinical Practice*, 61(10), 1697-1700.
- Althoff, T., Clark, K., & Leskovec, J. (2016). Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4, 463-476.
- Anderson, I., & Tomenson, B. (1995). Treatment discontinuation with selective serotonin reuptake inhibitors compared with tricyclic antidepressants: a meta-analysis. *Bmj*, 310(6992), 1433-1438.
- Andresen, E. M., Malmgren, J. A., Carter, W. B., & Patrick, D. L. (1994). Screening for depression in well older adults: Evaluation of a short form of the CES-D. *American journal of preventive medicine*, 10(2), 77-84.
- Ani, C., Bazargan, M., Hindman, D., Bell, D., Farooq, M. A., Akhanjee, L., Yemofio, F., Baker, R., & Rodriguez, M. (2008). Depression symptomatology and diagnosis: discordance between patients and physicians in primary care settings. *BMC family practice*, 9(1), 1-9.
- Ashraf, A., Gunawan, T. S., Rahman, F. D. A., Kartiwi, M., & Ismail, N. (2020). A summarization of the visual depression databases for depression detection. 2020 6th International Conference on Wireless and Telematics (ICWT),
- Association, A. P. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.
- Auxier, B., & Anderson, M. (2021). Social media use in 2021. *Pew research center*, 1, 1-4.
- Bak, M., Drukker, M., Hasmi, L., & van Os, J. (2016). An n=1 Clinical Network Analysis of Symptoms and Treatment in Psychosis. *PLoS One*, 11(9), e0162811.
<https://doi.org/10.1371/journal.pone.0162811>
- Baldwin, J. R., Caspi, A., Meehan, A. J., Ambler, A., Arseneault, L., Fisher, H. L., Harrington, H., Matthews, T., Odgers, C. L., & Poulton, R. (2021). Population vs Individual Prediction of Poor Health From Results of Adverse Childhood Experiences Screening. *JAMA pediatrics*.
- Bathina, K. C., Ten Thij, M., Lorenzo-Luaces, L., Rutter, L. A., & Bollen, J. (2021). Individuals with depression express more distorted thinking on social media. *Nature Human Behaviour*, 5(4), 458-466.
- Beard, C., Millner, A. J., Forgeard, M. J., Fried, E. I., Hsu, K. J., Treadway, M., Leonard, C. V., Kertz, S., & Björqvinnsson, T. (2016). Network analysis of depression and anxiety symptom relationships in a psychiatric sample. *Psychological medicine*, 46(16), 3359-3369.
- Beck, A. T. (1979). *Cognitive therapy of depression*. Guilford press.
- Benton, A., Mitchell, M., & Hovy, D. (2017). Multi-task learning for mental health using social media text. *arXiv preprint arXiv:1712.03538*.
- Berlim, M. T., Richard-Devantoy, S., Dos Santos, N. R., & Turecki, G. (2020). The network structure of core depressive symptom-domains in major depressive disorder following antidepressant treatment: a randomized clinical trial. *Psychol Med*, 1-15.
<https://doi.org/10.1017/S0033291720001002>
- Birnbaum, M. L., Ernala, S. K., Rizvi, A. F., De Choudhury, M., & Kane, J. M. (2017). A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals. *J Med Internet Res*, 19(8), e289.
<https://doi.org/10.2196/jmir.7956>
- Birnbaum, M. L., Norel, R., Van Meter, A., Ali, A. F., Arenare, E., Eyigoz, E., Agurto, C., Germano, N., Kane, J. M., & Cecchi, G. A. (2020). Identifying signals associated with psychiatric illness utilizing language and images posted to Facebook. *npj Schizophrenia*, 6(1), 1-10.

- Bjelland, I., Krokstad, S., Mykletun, A., Dahl, A. A., Tell, G. S., & Tambs, K. (2008). Does a higher educational level protect against anxiety and depression? The HUNT study. *Social science & medicine*, 66(6), 1334-1345.
- Boettiger, C., & Hastings, A. (2012). Quantifying limits to detection of early warning for critical transitions. *Journal of the Royal Society Interface*, 9(75), 2527-2539.
- Borgatti, S. P. (2005). Centrality and network flow. *Social networks*, 27(1), 55-71.
- Borsboom, D., & Cramer, A. O. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annu Rev Clin Psychol*, 9, 91-121. <https://doi.org/10.1146/annurev-clinpsy-050212-185608>
- Borsboom, D., Cramer, A. O., Schmittmann, V. D., Epskamp, S., & Waldorp, L. J. (2011). The small world of psychopathology. *PLoS One*, 6(11), e27407.
- Bos, E. H., de Jonge, P., & Cox, R. F. (2019). Affective variability in depression: Revisiting the inertia–instability paradox. *British Journal of Psychology*, 110(4), 814-827.
- Bos, F., von Klipstein, L., Emerencia, A. C., Veermans, E., Verhage, T., Snippe, E., Hadders-Prins, G., Wichers, M., & Riese, H. (2022). PErsonalized Treatment by Real-time assessment (PETRA): A web-application for personalized diaries in psychiatric care.
- Bos, F. M., Fried, E. I., Hollon, S. D., Bringmann, L. F., Dimidjian, S., DeRubeis, R. J., & Bockting, C. L. (2018). Cross-sectional networks of depressive symptoms before and after antidepressant medication treatment. *Social psychiatry and psychiatric epidemiology*, 53(6), 617-627.
- Bos, F. M., Fried, E. I., Hollon, S. D., Bringmann, L. F., Dimidjian, S., DeRubeis, R. J., & Bockting, C. L. H. (2018). Cross-sectional networks of depressive symptoms before and after antidepressant medication treatment. *Soc Psychiatry Psychiatr Epidemiol*, 53(6), 617-627. <https://doi.org/10.1007/s00127-018-1506-1>
- Bos, F. M., Schreuder, M. J., George, S. V., Doornbos, B., Bruggeman, R., van der Krieke, L., Haarman, B., Wichers, M., & Snippe, E. (2022). Anticipating manic and depressive transitions in patients with bipolar disorder using early warning signals. *International journal of bipolar disorders*, 10(1), 1-14.
- Bos, F. M., Snippe, E., Bruggeman, R., Doornbos, B., Wichers, M., & van der Krieke, L. (2020). Recommendations for the use of long-term experience sampling in bipolar disorder care: a qualitative study of patient and clinician experiences. *International journal of bipolar disorders*, 8(1), 1-14.
- Bos, F. M., Snippe, E., de Vos, S., Hartmann, J. A., Simons, C. J. P., van der Krieke, L., de Jonge, P., & Wichers, M. (2017). Can We Jump from Cross-Sectional to Dynamic Interpretations of Networks? Implications for the Network Perspective in Psychiatry. *Psychother Psychosom*, 86(3), 175-177. <https://doi.org/10.1159/000453583>
- Boschloo, L., van Borkulo, C. D., Rhemtulla, M., Keyes, K. M., Borsboom, D., & Schoevers, R. A. (2015). The Network Structure of Symptoms of the Diagnostic and Statistical Manual of Mental Disorders. *PLoS One*, 10(9), e0137621. <https://doi.org/10.1371/journal.pone.0137621>
- Bringmann, L. F. (2021). Person-specific networks in psychopathology: Past, present, and future. *Current opinion in psychology*, 41, 59-64.
- Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitsoff, T., & Filar, B. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.
- Bucci, W., & Freedman, N. (1981). The language of depression. *Bulletin of the Menninger Clinic*, 45(4), 334.
- Bukh, J. D., Bock, C., Vinberg, M., & Kessing, L. V. (2013). The effect of prolonged duration of untreated depression on antidepressant treatment outcome. *Journal of affective disorders*, 145(1), 42-48.

- Bulteel, K., Tuerlinckx, F., Brose, A., & Ceulemans, E. (2016). Using raw VAR regression coefficients to build networks can be misleading. *Multivariate behavioral research*, 51(2-3), 330-344.
- Burcusa, S. L., & Iacono, W. G. (2007). Risk for recurrence in depression. *Clinical psychology review*, 27(8), 959-985.
- Burger, J., Epskamp, S., Dablander, F., Schoevers, R. A., Fried, E. I., & Riese, H. (2021). A clinical PREMISE for personalized models: Towards a formal integration of case formulations and statistical networks.
- Cabrieto, J., Adolf, J., Tuerlinckx, F., Kuppens, P., & Ceulemans, E. (2018). Detecting long-lived autodependency changes in a multivariate system via change point detection and regime switching models. *Scientific reports*, 8(1), 1-15.
- Cacioppo, J. T., & Gardner, W. L. (1999). Emotion. *Annual review of psychology*, 50.
- Campbell, R. S., & Pennebaker, J. W. (2003). The secret life of pronouns: Flexibility in writing style and physical health. *Psychological science*, 14(1), 60-65.
- Capecelatro, M. R., Sacchet, M. D., Hitchcock, P. F., Miller, S. M., & Britton, W. B. (2013). Major depression duration reduces appetitive word use: an elaborated verbal recall of emotional photographs. *J Psychiatr Res*, 47(6), 809-815.
<https://doi.org/10.1016/j.jpsychires.2013.01.022>
- Carey, T. A. (2011). Exposure and reorganization: The what and how of effective psychotherapy. *Clinical psychology review*, 31(2), 236-248.
- Carpenter, S. R., & Brock, W. A. (2006). Rising variance: a leading indicator of ecological transition. *Ecology letters*, 9(3), 311-318.
- Carver, C. S. (2015). Control processes, priority management, and affective dynamics. *Emotion Review*, 7(4), 301-307.
- Chancellor, S., Birnbaum, M. L., Caine, E. D., Silenzio, V. M., & De Choudhury, M. (2019). A taxonomy of ethical tensions in inferring mental health states from social media. Proceedings of the conference on fairness, accountability, and transparency,
- Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1), 1-11.
- Chancellor, S., Lin, Z., Goodman, E. L., Zerwas, S., & De Choudhury, M. (2016). Quantifying and predicting mental illness severity in online pro-eating disorder communities. Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing,
- Chapman, D. P., Whitfield, C. L., Felitti, V. J., Dube, S. R., Edwards, V. J., & Anda, R. F. (2004). Adverse childhood experiences and the risk of depressive disorders in adulthood. *Journal of affective disorders*, 82(2), 217-225.
- Chekroud, S. R., Gueorguieva, R., Zheutlin, A. B., Paulus, M., Krumholz, H. M., Krystal, J. H., & Chekroud, A. M. (2018). Association between physical exercise and mental health in 1· 2 million individuals in the USA between 2011 and 2015: a cross-sectional study. *The Lancet Psychiatry*, 5(9), 739-746.
- Chen, L., Liu, R., Liu, Z. P., Li, M., & Aihara, K. (2012). Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Sci Rep*, 2, 342.
<https://doi.org/10.1038/srep00342>
- Chen, X., Sykora, M. D., Jackson, T. W., & Elayan, S. (2018). What about mood swings: Identifying depression on twitter with temporal measures of emotions. Companion Proceedings of the The Web Conference 2018,
- Cheng, Q., Li, T. M., Kwok, C. L., Zhu, T., & Yip, P. S. (2017). Assessing Suicide Risk and Emotional Distress in Chinese Social Media: A Text Mining and Machine Learning Study. *J Med Internet Res*, 19(7), e243. <https://doi.org/10.2196/jmir.7276>
- Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., & Goharian, N. (2018). SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. *arXiv preprint arXiv:1806.05258*.

- Coley, R. Y., Boggs, J. M., Beck, A., & Simon, G. E. (2021). Predicting outcomes of psychotherapy for depression with electronic health record data. *Journal of affective disorders reports*, 6, 100198.
- Collishaw, S., Pickles, A., Messer, J., Rutter, M., Shearer, C., & Maughan, B. (2007). Resilience to adult psychopathology following childhood maltreatment: Evidence from a community sample. *Child abuse & neglect*, 31(3), 211-229.
- Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality,
- Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015). From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality,
- Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). CLPsych 2015 shared task: Depression and PTSD on Twitter. Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality,
- Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomed Inform Insights*, 10, 1178222618792860. <https://doi.org/10.1177/1178222618792860>
- Coppersmith, G., Ngo, K., Leary, R., & Wood, A. (2016). Exploratory analysis of social media prior to a suicide attempt. Proceedings of the third workshop on computational linguistics and clinical psychology,
- Coyne, J. C., Thompson, R., Klinkman, M. S., & Nease Jr, D. E. (2002). Emotional disorders in primary care. *Journal of consulting and clinical psychology*, 70(3), 798.
- Craft, L. L., & Landers, D. M. (1998). The effect of exercise on clinical depression and depression resulting from mental illness: A meta-analysis. *Journal of Sport and Exercise Psychology*, 20(4), 339-357.
- Cramer, A. O., van Borkulo, C. D., Giltay, E. J., van der Maas, H. L., 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>
- Cuevas, A. C., Ots, C. V., Heeren, A. H., & Bentall, R. P. (2020). A temporal network approach to paranoia: A pilot study. *Frontiers in Psychology*, 11, 2359.
- Curtis, B., Giorgi, S., Buffone, A. E., Ungar, L. H., Ashford, R. D., Hemmons, J., Summers, D., Hamilton, C., & Schwartz, H. A. (2018). Can Twitter be used to predict county excessive alcohol consumption rates? *PLoS One*, 13(4), e0194290.
- Curtiss, J. E., Mischoulon, D., Fisher, L. B., Cusin, C., Fedor, S., Picard, R. W., & Pedrelli, P. (2021). Rising early warning signals in affect associated with future changes in depression: a dynamical systems approach. *Psychological medicine*, 1-9.
- Dablander, F., & Hinne, M. (2019). Node centrality measures are a poor substitute for causal inference. *Scientific reports*, 9(1), 1-13.
- Dakos, V., van Nes, E. H., Donangelo, R., Fort, H., & Scheffer, M. (2009). Spatial correlation as leading indicator of catastrophic shifts. *Theoretical Ecology*, 3(3), 163-174. <https://doi.org/10.1007/s12080-009-0060-6>
- Dakos, V., van Nes, E. H., Donangelo, R., Fort, H., & Scheffer, M. (2010). Spatial correlation as leading indicator of catastrophic shifts. *Theoretical Ecology*, 3(3), 163-174.
- De Choudhury, M. (2015). *Anorexia on Tumblr* Proceedings of the 5th International Conference on Digital Health 2015,
- De Choudhury, M., Counts, S., & Horvitz, E. (2013). Social media as a measurement tool of depression in populations. Proceedings of the 5th Annual ACM Web Science Conference,

- De Choudhury, M., Counts, S., Horvitz, E. J., & Hoff, A. (2014). Characterizing and predicting postpartum depression from shared facebook data. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*,
- De Choudhury, M., & De, S. (2014). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. *Eighth international AAAI conference on weblogs and social media*,
- De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *Seventh international AAAI conference on weblogs and social media*,
- De Choudhury, M., & Kiciman, E. (2017). The language of social support in social media and its effect on suicidal ideation risk. *Proceedings of the International AAAI Conference on Web and Social Media*,
- De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering shifts to suicidal ideation from mental health content in social media. *Proceedings of the 2016 CHI conference on human factors in computing systems*,
- De Choudhury, M., Sharma, S. S., Logar, T., Eekhout, W., & Nielsen, R. C. (2017). Gender and cross-cultural differences in social media disclosures of mental illness. *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*,
- Dejonckheere, E., Mestdagh, M., Houben, M., Rutten, I., Sels, L., Kuppens, P., & Tuerlinckx, F. (2019). Complex affect dynamics add limited information to the prediction of psychological well-being. *Nature Human Behaviour*, 3(5), 478-491.
- DEWAN, M. J., & Anand, V. S. (1999). Evaluating the tolerability of the newer antidepressants. *The Journal of nervous and mental disease*, 187(2), 96-101.
- DiMatteo, M. R., Lepper, H. S., & Croghan, T. W. (2000). Depression is a risk factor for noncompliance with medical treatment: meta-analysis of the effects of anxiety and depression on patient adherence. *Archives of internal medicine*, 160(14), 2101-2107.
- Dinga, R., Schmaal, L., Penninx, B. W., Veltman, D. J., & Marquand, A. F. (2020). Controlling for effects of confounding variables on machine learning predictions. *BioRxiv*.
- Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., Mitchell, L., Harris, K. D., Kloumann, I. M., & Bagrow, J. P. (2015). Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8), 2389-2394.
- Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS One*, 6(12), e26752.
- Dube, S. R., Anda, R. F., Felitti, V. J., Chapman, D. P., Williamson, D. F., & Giles, W. H. (2001). Childhood abuse, household dysfunction, and the risk of attempted suicide throughout the life span: findings from the Adverse Childhood Experiences Study. *Jama*, 286(24), 3089-3096.
- Dunn, A. L., Trivedi, M. H., Kampert, J. B., Clark, C. G., & Chambliss, H. O. (2005). Exercise treatment for depression: efficacy and dose response. *American journal of preventive medicine*, 28(1), 1-8.
- Dunning, D., Ahmed, S., Foulkes, L., Griffin, C., Griffiths, K., Leung, J., Parker, J., Pi-Sunyer, B. P., Sakhardande, A., & Bennett, M. (2022). The impact of mindfulness training in early adolescence on affective executive control, and on later mental health during the COVID-19 pandemic: a randomised controlled trial. *Evidence-based mental health*, 25(3), 110-116.
- Edwards, T. M., & Holtzman, N. S. (2017). A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality*, 68, 63-68.
<https://doi.org/10.1016/j.jrp.2017.02.005>
- Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., Jha, S., Agrawal, M., Dziurzynski, L. A., & Sap, M. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological science*, 26(2), 159-169.

- Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preotiuc-Pietro, D., Asch, D. A., & Schwartz, H. A. (2018). Facebook language predicts depression in medical records. *Proc Natl Acad Sci U S A*, 115(44), 11203-11208.
<https://doi.org/10.1073/pnas.1802331115>
- Elliott, H., Jones, P. J., & Schmidt, U. (2019). Central Symptoms Predict Posttreatment Outcomes and Clinical Impairment in Anorexia Nervosa: A Network Analysis. *Clinical Psychological Science*, 8(1), 139-154. <https://doi.org/10.1177/2167702619865958>
- Ellison, W. D. (2021). An initial study of practicing psychologists' views of the utility of ecological momentary assessment for difficult psychotherapy cases. *Administration and Policy in Mental Health and Mental Health Services Research*, 48(4), 597-607.
- Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behav Res Methods*, 50(1), 195-212.
<https://doi.org/10.3758/s13428-017-0862-1>
- Epskamp, S., van Borkulo, C. D., van der Veen, D. C., Servaas, M. N., Isvoranu, A. M., Riese, H., & Cramer, A. O. J. (2018). Personalized Network Modeling in Psychopathology: The Importance of Contemporaneous and Temporal Connections. *Clin Psychol Sci*, 6(3), 416-427. <https://doi.org/10.1177/2167702617744325>
- Epskamp, S., Waldorp, L. J., Möttus, R., & Borsboom, D. (2018). The Gaussian graphical model in cross-sectional and time-series data. *Multivariate behavioral research*, 53(4), 453-480.
- Ernala, S. K., Birnbaum, M. L., Candan, K. A., Rizvi, A. F., Sterling, W. A., Kane, J. M., & De Choudhury, M. (2019). Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. Proceedings of the 2019 chi conference on human factors in computing systems,
- Ernala, S. K., Rizvi, A. F., Birnbaum, M. L., Kane, J. M., & De Choudhury, M. (2017). Linguistic Markers Indicating Therapeutic Outcomes of Social Media Disclosures of Schizophrenia. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW), 1-27.
<https://doi.org/10.1145/3134678>
- Eronen, M. I. (2020). Causal discovery and the problem of psychological interventions. *New Ideas in Psychology*, 59, 100785.
- Esfahani, F. Z., Sayama, H., Visser, K. F., & Strauss, G. P. (2017). Sensitivity of the positive and negative syndrome scale (PANSS) in detecting treatment effects via network analysis. *Innovations in clinical neuroscience*, 14(11-12), 59.
- Ewbank, M. P., Cummins, R., Tablan, V., Bateup, S., Catarino, A., Martin, A. J., & Blackwell, A. D. (2020). Quantifying the association between psychotherapy content and clinical outcomes using deep learning. *JAMA psychiatry*, 77(1), 35-43.
- Fan, R., Varol, O., Varamesh, A., Barron, A., van de Leemput, I. A., Scheffer, M., & Bollen, J. (2019). The minute-scale dynamics of online emotions reveal the effects of affect labeling. *Nature Human Behaviour*, 3(1), 92-100.
- Fava, M. (2003). Diagnosis and definition of treatment-resistant depression. *Biological psychiatry*, 53(8), 649-659.
- Felitti, V. J., Anda, R. F., Nordenberg, D., Williamson, D. F., Spitz, A. M., Edwards, V., & Marks, J. S. (1998). Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults: The Adverse Childhood Experiences (ACE) Study. *American journal of preventive medicine*, 14(4), 245-258.
- Fiesler, C., & Proferes, N. (2018). "Participant" perceptions of Twitter research ethics. *Social Media+ Society*, 4(1), 2056305118763366.
- Fineberg, S., Leavitt, J., Deutsch-Link, S., Dealy, S., Landry, C., Pirruccio, K., Shea, S., Trent, S., Cecchi, G., & Corlett, P. (2016). Self-reference in psychosis and depression: a language marker of illness. *Psychological medicine*, 46(12), 2605.

- Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, 115(27), E6106-E6115.
- Fisher, A. J., & Newman, M. G. (2016). Reductions in the diurnal rigidity of anxiety predict treatment outcome in cognitive behavioral therapy for generalized anxiety disorder. *Behaviour research and therapy*, 79, 46-55.
- Fisher, A. J., Reeves, J. W., Lawyer, G., Medaglia, J. D., & Rubel, J. A. (2017). Exploring the idiographic dynamics of mood and anxiety via network analysis. *Journal of abnormal psychology*, 126(8), 1044.
- Fisher, A. J., & Woodward, S. H. (2014). Cardiac stability at differing levels of temporal analysis in panic disorder, post-traumatic stress disorder, and healthy controls. *Psychophysiology*, 51(1), 80-87.
- Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. M. (2002). The Obsessive-Compulsive Inventory: development and validation of a short version. *Psychological assessment*, 14(4), 485.
- Ford, E., Curlew, K., Wongkoblap, A., & Curcin, V. (2019). Public opinions on using social media content to identify users with depression and target mental health care advertising: mixed methods survey. *JMIR mental health*, 6(11), e12942.
- 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: a patient-level meta-analysis. *Jama*, 303(1), 47-53.
- Frewen, P. A., Allen, S. L., Lanius, R. A., & Neufeld, R. W. (2012). Perceived causal relations: novel methodology for assessing client attributions about causal associations between variables including symptoms and functional impairment. *Assessment*, 19(4), 480-493.
<https://doi.org/10.1177/1073191111418297>
- Frewen, P. A., Schmittmann, V. D., Bringmann, L. F., & Borsboom, D. (2013). Perceived causal relations between anxiety, posttraumatic stress and depression: extension to moderation, mediation, and network analysis. *Eur J Psychotraumatol*, 4.
<https://doi.org/10.3402/ejpt.v4i0.20656>
- 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. (2015). Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. *J Affect Disord*, 172, 96-102.
<https://doi.org/10.1016/j.jad.2014.10.010>
- 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), 1-11.
- Fried, E. I., van Borkulo, C. D., Cramer, A. O., Boschloo, L., Schoevers, R. A., & Borsboom, D. (2017). Mental disorders as networks of problems: a review of recent insights. *Soc Psychiatry Psychiatr Epidemiol*, 52(1), 1-10. <https://doi.org/10.1007/s00127-016-1319-z>
- 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.
- Froehlich, J., Chen, M. Y., Consolvo, S., Harrison, B., & Landay, J. A. (2007). MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones. Proceedings of the 5th international conference on Mobile systems, applications and services,

- Frumkin, M. R., Piccirillo, M. L., Beck, E. D., Grossman, J. T., & Rodebaugh, T. L. (2021). Feasibility and utility of idiographic models in the clinic: a pilot study. *Psychotherapy Research*, 31(4), 520-534.
- Garner, D., Olmsted, M., Bohr, Y., & Garfinkel, P. (1982). The eating attitudes test: Psychometric features. *Psychological medicine*, 12, 871-878.
- Gelo, O. C. G., & Salvatore, S. (2016). A dynamic systems approach to psychotherapy: A meta-theoretical framework for explaining psychotherapy change processes. *Journal of Counseling Psychology*, 63(4), 379.
- George, L. K., Blazer, D. G., Hughes, D. C., & Fowler, N. (1989). Social support and the outcome of major depression. *The British journal of psychiatry*, 154(4), 478-485.
- Ghio, L., Gotelli, S., Marcenaro, M., Amore, M., & Natta, W. (2014). Duration of untreated illness and outcomes in unipolar depression: a systematic review and meta-analysis. *Journal of affective disorders*, 152, 45-51.
- Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *Elife*, 5, e11305.
- Gkotsis, G., Oellrich, A., Hubbard, T., Dobson, R., Liakata, M., Velupillai, S., & Dutta, R. (2016). The language of mental health problems in social media. Proceedings of the third workshop on computational linguistics and clinical psychology,
- Glenn, J. J., Nobles, A. L., Barnes, L. E., & Teachman, B. A. (2020). Can text messages identify suicide risk in real time? a within-subjects pilot examination of temporally sensitive markers of suicide risk. *Clinical Psychological Science*, 8(4), 704-722.
- González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in twitter: a closer look. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,
- Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 424-438.
- Grant, J. E., & Chamberlain, S. R. (2018). Sleepiness and impulsivity: Findings in non-treatment seeking young adults. *Journal of behavioral addictions*, 7(3), 737-742.
- Greenberg, P. E., Fournier, A.-A., Sisitsky, T., Simes, M., Berman, R., Koenigsberg, S. H., & Kessler, R. C. (2021). The economic burden of adults with major depressive disorder in the United States (2010 and 2018). *Pharmacoeconomics*, 39(6), 653-665.
- Greene, D., O'Callaghan, D., & Cunningham, P. (2014). How many topics? stability analysis for topic models. Joint European conference on machine learning and knowledge discovery in databases,
- Guntuku, S. C., Preotiuc-Pietro, D., Eichstaedt, J. C., & Ungar, L. H. (2019). What twitter profile and posted images reveal about depression and anxiety. Proceedings of the international AAAI conference on web and social media,
- Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18, 43-49. <https://doi.org/10.1016/j.cobeha.2017.07.005>
- Hakulinen, C., Fried, E. I., Pulkki-Raback, L., Virtanen, M., Suvisaari, J., & Elovainio, M. (2020). Network structure of depression symptomology in participants with and without depressive disorder: the population-based Health 2000-2011 study. *Soc Psychiatry Psychiatr Epidemiol*. <https://doi.org/10.1007/s00127-020-01843-7>
- Hall, M., Scherner, P. V., Kreidel, Y., & Rubel, J. A. (2021). A systematic review of momentary assessment designs for mood and anxiety symptoms. *Frontiers in Psychology*, 12, 642044.
- Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science*, 26(1), 10-15.
- Hamilton, M. (1959). The assessment of anxiety states by rating. *British journal of medical psychology*.

- Hamilton, M. (1960). A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry*, 23(1), 56.
- Hammen, C., Henry, R., & Daley, S. E. (2000). Depression and sensitization to stressors among young women as a function of childhood adversity. *Journal of consulting and clinical psychology*, 68(5), 782.
- Hankin, B. L., Abramson, L. Y., Moffitt, T. E., Silva, P. A., McGee, R., & Angell, K. E. (1998). Development of depression from preadolescence to young adulthood: emerging gender differences in a 10-year longitudinal study. *Journal of abnormal psychology*, 107(1), 128.
- Haslam, N. (2003). Categorical versus dimensional models of mental disorder: The taxometric evidence. *Australian & New Zealand Journal of Psychiatry*, 37(6), 696-704.
- Haslbeck, J., & Waldorp, L. J. (2015). mgm: Estimating time-varying mixed graphical models in high-dimensional data. *arXiv preprint arXiv:1510.06871*.
- Haslbeck, J. M., & Ryan, O. (2021). Recovering within-person dynamics from psychological time series. *Multivariate behavioral research*, 1-32.
- Hayes, A. M., Laurenceau, J.-P., Feldman, G., Strauss, J. L., & Cardaciotto, L. (2007). Change is not always linear: The study of nonlinear and discontinuous patterns of change in psychotherapy. *Clinical psychology review*, 27(6), 715-723.
- Hayes, A. M., & Strauss, J. L. (1998). Dynamic systems theory as a paradigm for the study of change in psychotherapy: an application to cognitive therapy for depression. *Journal of consulting and clinical psychology*, 66(6), 939.
- Hayes, A. M., & Yasinski, C. (2015). Pattern destabilization and emotional processing in cognitive therapy for personality disorders. *Frontiers in Psychology*, 6, 107.
- Hayes, A. M., Yasinski, C., Barnes, J. B., & Bockting, C. L. (2015). Network destabilization and transition in depression: New methods for studying the dynamics of therapeutic change. *Clinical psychology review*, 41, 27-39.
- Heeren, A., & McNally, R. J. (2017). Social Anxiety Disorder as a Densely Interconnected Network of Fear and Avoidance for Social Situations. *Cognitive Therapy and Research*, 42(1), 103-113. <https://doi.org/10.1007/s10608-017-9876-3>
- Heinzel, S., Tominschek, I., & Schiepek, G. (2014). Dynamic patterns in psychotherapy-discontinuous changes and critical instabilities during the treatment of obsessive compulsive disorder. *Nonlinear Dynamics, Psychology, and Life Sciences*, 18(2), 155-176.
- Höller, I., Schreiber, D., Bos, F., Forkmann, T., Teismann, T., & Margraf, J. (2022). The Mereology of Depression—Networks of Depressive Symptoms during the Course of Psychotherapy. *International journal of environmental research and public health*, 19(12), 7131.
- Hosenfeld, B., Bos, E. H., Wardenaar, K. J., Conradi, H. J., van der Maas, H. L., Visser, I., & de Jonge, P. (2015). Major depressive disorder as a nonlinear dynamic system: bimodality in the frequency distribution of depressive symptoms over time. *BMC psychiatry*, 15(1), 1-9.
- Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological bulletin*, 141(4), 901.
- Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness, and patient progress. *American psychologist*, 51(10), 1059.
- Hswen, Y., Gopaluni, A., Brownstein, J. S., & Hawkins, J. B. (2019). Using Twitter to Detect Psychological Characteristics of Self-Identified Persons With Autism Spectrum Disorder: A Feasibility Study. *JMIR Mhealth Uhealth*, 7(2), e12264. <https://doi.org/10.2196/12264>
- Hung, C.-I., Liu, C.-Y., & Yang, C.-H. (2017). Untreated duration predicted the severity of depression at the two-year follow-up point. *PLoS One*, 12(9), e0185119.
- [Record #28 is using a reference type undefined in this output style.]

- Ireland, M., & Iserman, M. (2018). Within and between-person differences in language used across anxiety support and neutral reddit communities. *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*.
- Jacobson, N. C., & Chung, Y. J. (2020). Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. *Sensors*, 20(12), 3572.
- Jahng, S., Wood, P. K., & Trull, T. J. (2008). Analysis of affective instability in ecological momentary assessment: Indices using successive difference and group comparison via multilevel modeling. *Psychological Methods*, 13(4), 354.
- Jaidka, K., Guntuku, S., & Ungar, L. (2018). Facebook versus Twitter: Differences in Self-Disclosure and Trait Prediction. *Proceedings of the International AAAI Conference on Web and Social Media*.
- Jimeno, N., Gomez-Pilar, J., Poza, J., Hornero, R., Vogeley, K., Meisenzahl, E., Haidl, T., Rosen, M., Klosterkötter, J., & Schultze-Lutter, F. (2020). Main symptomatic treatment targets in suspected and early psychosis: new insights from network analysis. *Schizophrenia Bulletin*.
- Jollans, L., Boyle, R., Artiges, E., Banaschewski, T., Desrivieres, S., Grigis, A., Martinot, J.-L., Paus, T., Smolka, M. N., & Walter, H. (2019). Quantifying performance of machine learning methods for neuroimaging data. *NeuroImage*, 199, 351-365.
- Kashdan, T. B., & Rottenberg, J. (2010). Psychological flexibility as a fundamental aspect of health. *Clinical psychology review*, 30(7), 865-878.
- Katon, W., Rutter, C., Ludman, E. J., Von Korff, M., Lin, E., Simon, G., Bush, T., Walker, E., & Unützer, J. (2001). A randomized trial of relapse prevention of depression in primary care. *Archives of general psychiatry*, 58(3), 241-247.
- Kelley, S., & Gillan, C. (2020). Within-subject changes in network connectivity occur during an episode of depression: evidence from a longitudinal analysis of social media posts.
- Kelley, S., Gillan, C. (2021). Using linguistic features in social media posts to study the network dynamics of depression longitudinally. *Twitter_Depression*.
<https://doi.org/10.5281/zenodo.5745764>
- Kendler, K. S., Karkowski, L. M., & Prescott, C. A. (1998). Stressful life events and major depression: risk period, long-term contextual threat, and diagnostic specificity. *The Journal of nervous and mental disease*, 186(11), 661-669.
- Kendler, K. S., Karkowski, L. M., & Prescott, C. A. (1999). Causal relationship between stressful life events and the onset of major depression. *American Journal of Psychiatry*, 156(6), 837-841.
- Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Dziurzynski, L., Ungar, L. H., Stillwell, D. J., Kosinski, M., Ramones, S. M., & Seligman, M. E. (2014). The online social self: An open vocabulary approach to personality. *Assessment*, 21(2), 158-169.
- Kessler, R. C., Birnbaum, H. G., Shahly, V., Bromet, E., Hwang, I., McLaughlin, K. A., Sampson, N., Andrade, L. H., De Girolamo, G., & Demyttenaere, K. (2010). Age differences in the prevalence and co-morbidity of DSM-IV major depressive episodes: results from the WHO World Mental Health Survey Initiative. *Depression and anxiety*, 27(4), 351-364.
- Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry*, 62(6), 617-627.
- Kessler, R. C., Davis, C. G., & Kendler, K. S. (1997). Childhood adversity and adult psychiatric disorder in the US National Comorbidity Survey. *Psychological medicine*, 27(5), 1101-1119.
- Kessler, R. C., & Magee, W. J. (1993). Childhood adversities and adult depression: basic patterns of association in a US national survey. *Psychological medicine*, 23(3), 679-690.

- Kessler, R. C., McGonagle, K. A., Swartz, M., Blazer, D. G., & Nelson, C. B. (1993). Sex and depression in the National Comorbidity Survey I: Lifetime prevalence, chronicity and recurrence. *Journal of affective disorders*, 29(2-3), 85-96.
- Kessler, R. C., Nelson, C. B., McGonagle, K. A., Liu, J., Swartz, M., & Blazer, D. G. (1996). Comorbidity of DSM-III-R major depressive disorder in the general population: results from the US National Comorbidity Survey. *The British journal of psychiatry*, 168(S30), 17-30.
- Kessler, R. C., Petukhova, M., Sampson, N. A., Zaslavsky, A. M., & Wittchen, H. U. (2012). Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. *International journal of methods in psychiatric research*, 21(3), 169-184.
- Kim, D.-M., Wampold, B. E., & Bolt, D. M. (2006). Therapist effects in psychotherapy: A random-effects modeling of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. *Psychotherapy Research*, 16(02), 161-172.
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15), 5802-5805.
- Koval, P., Brose, A., Pe, M. L., Houben, M., Erbas, Y., Champagne, D., & Kuppens, P. (2015). Emotional inertia and external events: The roles of exposure, reactivity, and recovery. *Emotion*, 15(5), 625.
- Koval, P., Kuppens, P., Allen, N. B., & Sheeber, L. (2012). Getting stuck in depression: The roles of rumination and emotional inertia. *Cognition & Emotion*, 26(8), 1412-1427.
- Koval, P., Pe, M. L., Meers, K., & Kuppens, P. (2013). Affect dynamics in relation to depressive symptoms: variable, unstable or inert? *Emotion*, 13(6), 1132.
- Kramer, I., Simons, C. J., Hartmann, J. A., Menne-Lothmann, C., Viechtbauer, W., Peeters, F., Schruers, K., van Bemmelen, A. L., Myin-Germeys, I., & Delespaul, P. (2014). A therapeutic application of the experience sampling method in the treatment of depression: a randomized controlled trial. *World Psychiatry*, 13(1), 68-77.
- Krieke, L. V. D., Jeronimus, B. F., Blaauw, F. J., Wanders, R. B., Emerencia, A. C., Schenk, H. M., Vos, S. D., Snippe, E., Wichers, M., & Wigman, J. T. (2016). HowNutsAreTheDutch (HoeGekIsNL): A crowdsourcing study of mental symptoms and strengths. *International journal of methods in psychiatric research*, 25(2), 123-144.
- Kross, E., Bruehlman-Senecal, E., Park, J., Burson, A., Dougherty, A., Shablack, H., Bremner, R., Moser, J., & Ayduk, O. (2014). Self-talk as a regulatory mechanism: how you do it matters. *Journal of Personality and Social Psychology*, 106(2), 304.
- Kryvasheyev, Y., Chen, H., Moro, E., Van Hentenryck, P., & Cebrian, M. (2015). Performance of social network sensors during Hurricane Sandy. *PLoS One*, 10(2), e0117288.
- Kuppens, P., Sheeber, L. B., Yap, M. B., Whittle, S., Simmons, J. G., & Allen, N. B. (2012). Emotional inertia prospectively predicts the onset of depressive disorder in adolescence. *Emotion*, 12(2), 283.
- Kuranova, A., Booij, S. H., Menne-Lothmann, C., Decoster, J., van Winkel, R., Delespaul, P., De Hert, M., Derom, C., Thiery, E., & Rutten, B. P. (2020). Measuring resilience prospectively as the speed of affect recovery in daily life: a complex systems perspective on mental health. *BMC medicine*, 18(1), 1-11.
- Kuranova, A., Wigman, J. T., Menne-Lothmann, C., Decoster, J., van Winkel, R., Delespaul, P., Drukker, M., De Hert, M., Derom, C., & Thiery, E. (2021). Network dynamics of momentary affect states and future course of psychopathology in adolescents. *PLoS One*, 16(3), e0247458.
- Lee Pe, M., Kircanski, K., Thompson, R. J., Bringmann, L. F., Tuerlinckx, F., Mestdagh, M., Mata, J., Jaeggi, S. M., Buschkuhl, M., Jonides, J., Kuppens, P., & Gotlib, I. H. (2015). Emotion-

- Network Density in Major Depressive Disorder. *Clin Psychol Sci*, 3(2), 292-300.
<https://doi.org/10.1177/2167702614540645>
- Leis, A., Ronzano, F., Mayer, M. A., Furlong, L. I., & Sanz, F. (2019). Detecting Signs of Depression in Tweets in Spanish: Behavioral and Linguistic Analysis. *J Med Internet Res*, 21(6), e14199. <https://doi.org/10.2196/14199>
- Lépine, J.-P., Gastpar, M., Mendlewicz, J., & Tylee, A. (1997). Depression in the community: the first pan-European study DEPRES (Depression Research in European Society). *International clinical psychopharmacology*.
- Levine, F. M., Sandeen, E., & Murphy, C. M. (1992). The therapist's dilemma: Using nomothetic information to answer idiographic questions. *Psychotherapy: Theory, Research, Practice, Training*, 29(3), 410.
- Lichtwarck-Aschoff, A., Hasselman, F., Cox, R., Pepler, D., & Granic, I. (2012). A characteristic destabilization profile in parent-child interactions associated with treatment efficacy for aggressive children. *Nonlinear Dynamics-Psychology and Life Sciences*, 16(3), 353.
- Liebowitz, M. R. (1987). Social phobia. *Modern problems of pharmacopsychiatry*.
- Liu, T., Meyerhoff, J., Eichstaedt, J. C., Karr, C. J., Kaiser, S. M., Kording, K. P., Mohr, D. C., & Ungar, L. H. (2022). The relationship between text message sentiment and self-reported depression. *Journal of affective disorders*, 302, 7-14.
- Lorant, V., Delière, D., Eaton, W., Robert, A., Philippot, P., & Ansseau, M. (2003). Socioeconomic inequalities in depression: a meta-analysis. *American journal of epidemiology*, 157(2), 98-112.
- Lumontod III, R. Z. (2020). Seeing the invisible: Extracting signs of depression and suicidal ideation from college students' writing using LIWC a computerized text analysis. *International Journal of Research*, 9(4), 31-44.
- Lunansky, G., Hoekstra, R. H., & Blanken, T. (2021). Disentangling dynamic affect trajectories for distinct depression courses during the COVID-19 pandemic.
- Lunansky, G., Van Borkulo, C. D., Haslbeck, J. M., Van der Linden, M. A., Garay, C. J., Etchevers, M. J., & Borsboom, D. (2021). The mental health ecosystem: extending symptom networks with risk and protective factors. *Frontiers in psychiatry*, 12, 640658.
- Luppa, M., Sikorski, C., Luck, T., Ehreke, L., Konnopka, A., Wiese, B., Weyerer, S., König, H.-H., & Riedel-Heller, S. G. (2012). Age-and gender-specific prevalence of depression in latest-life—systematic review and meta-analysis. *Journal of affective disorders*, 136(3), 212-221.
- Lutz, W., Leon, S. C., Martinovich, Z., Lyons, J. S., & Stiles, W. B. (2007). Therapist effects in outpatient psychotherapy: A three-level growth curve approach. *Journal of Counseling Psychology*, 54(1), 32.
- Lydon-Staley, D. M., Xia, M., Mak, H. W., & Fosco, G. (2019). Adolescent emotion network dynamics in daily life and implications for depression. *Journal of abnormal child psychology*, 47(4), 717-729.
- Lyons, M., Aksayli, N. D., & Brewer, G. (2018). Mental distress and language use: Linguistic analysis of discussion forum posts. *Computers in Human Behavior*, 87, 207-211.
<https://doi.org/10.1016/j.chb.2018.05.035>
- MacGillivray, S., Arroll, B., Hatcher, S., Ogston, S., Reid, I., Sullivan, F., Williams, B., & Crombie, I. (2003). Efficacy and tolerability of selective serotonin reuptake inhibitors compared with tricyclic antidepressants in depression treated in primary care: systematic review and meta-analysis. *Bmj*, 326(7397), 1014.
- Maciejewski, D., van Roekel, E., Ha, T., De France, K., Lin, L., Lennarz, H., Trompetter, H., Meeus, W., Lichtwarck-Aschoff, A., & Branje, S. (2022). Beyond main effects? Affect level as a moderator in the relation between affect dynamics and depressive symptoms.
- Mansueto, A. C., Wiers, R., van Weert, J. C., Schouten, B. C., & Epskamp, S. (2020). Investigating the Feasibility of Idiographic Network Models.

- Marcus, S. C., & Olfson, M. (2010). National trends in the treatment for depression from 1998 to 2007. *Archives of general psychiatry*, 67(12), 1265-1273.
- Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., & Feczko, E. (2020). Towards reproducible brain-wide association studies. *BioRxiv*.
- Marin, R. S., Biedrzycki, R. C., & Firinciogullari, S. (1991). Reliability and validity of the Apathy Evaluation Scale. *Psychiatry research*, 38(2), 143-162.
- Markon, K. E., Chmielewski, M., & Miller, C. J. (2011). The reliability and validity of discrete and continuous measures of psychopathology: a quantitative review. *Psychological bulletin*, 137(5), 856.
- Mason, O., Linney, Y., & Claridge, G. (2005). Short scales for measuring schizotypy. *Schizophrenia research*, 78(2-3), 293-296.
- McElroy, E., Napoleone, E., Wolpert, M., & Patalay, P. (2019). Structure and Connectivity of Depressive Symptom Networks Corresponding to Early Treatment Response. *EClinicalMedicine*, 8, 29-36. <https://doi.org/10.1016/j.eclinm.2019.02.009>
- McHugh, R. K., Whitton, S. W., Peckham, A. D., Welge, J. A., & Otto, M. W. (2013). Patient preference for psychological vs pharmacologic treatment of psychiatric disorders: a meta-analytic review. *The Journal of clinical psychiatry*, 74(6), 13979.
- McLaughlin, K. A., & King, K. (2015). Developmental trajectories of anxiety and depression in early adolescence. *Journal of abnormal child psychology*, 43(2), 311-323.
- McLean, D. C., Nakamura, J., & Csikszentmihalyi, M. (2017). Explaining system missing: Missing data and experience sampling method. *Social Psychological and Personality Science*, 8(4), 434-441.
- McManus, K., Mallory, E. K., Goldfeder, R. L., Haynes, W. A., & Tatum, J. D. (2015). Mining Twitter data to improve detection of schizophrenia. *AMIA Summits on Translational Science Proceedings, 2015*, 122.
- Medagoda, N., Shanmuganathan, S., & Whalley, J. (2013). A comparative analysis of opinion mining and sentiment classification in non-english languages. 2013 International Conference on Advances in ICT for Emerging Regions (ICTer),
- Mellon, J., & Prosser, C. (2017a). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, 4(3), 2053168017720008.
- Mellon, J., & Prosser, C. (2017b). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, 4(3). <https://doi.org/10.1177/2053168017720008>
- Merrick, M. T., Ports, K. A., Ford, D. C., Afifi, T. O., Gershoff, E. T., & Grogan-Kaylor, A. (2017). Unpacking the impact of adverse childhood experiences on adult mental health. *Child abuse & neglect*, 69, 10-19.
- Miner, A. S., Fleming, S. L., Haque, A., Fries, J., Althoff, T., Wilfley, D. E., Agras, W. S., Milstein, A., Hancock, J., & Ash, S. M. (2022). Uncovering the linguistic characteristics of psychotherapy: a computational approach to measure therapist language timing, responsiveness, and consistency. *medRxiv*.
- Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS One*, 8(5), e64417.
- Mitchell, M., Hollingshead, K., & Coppersmith, G. (2015). Quantifying the language of schizophrenia in social media. Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality,
- Moitra, M., Santomauro, D., Collins, P. Y., Vos, T., Whiteford, H., Saxena, S., & Ferrari, A. J. (2022). The global gap in treatment coverage for major depressive disorder in 84 countries from

- 2000–2019: A systematic review and Bayesian meta-regression analysis. *PLoS medicine*, 19(2), e1003901.
- Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, 2(4), 201-218.
- Molendijk, M. L., Bamelis, L., van Emmerik, A. A., Arntz, A., Haringsma, R., & Spinhoven, P. (2010). Word use of outpatients with a personality disorder and concurrent or previous major depressive disorder. *Behav Res Ther*, 48(1), 44-51.
<https://doi.org/10.1016/j.brat.2009.09.007>
- Momen, N. C., Plana-Ripoll, O., Agerbo, E., Christensen, M. K., Iburg, K. M., Laursen, T. M., Mortensen, P. B., Pedersen, C. B., Prior, A., & Weyer, N. (2022). Mortality Associated With Mental Disorders and Comorbid General Medical Conditions. *JAMA psychiatry*.
- Morris, M. E., Kathawala, Q., Leen, T. K., Gorenstein, E. E., Guilak, F., DeLeeuw, W., & Labhard, M. (2010). Mobile therapy: case study evaluations of a cell phone application for emotional self-awareness. *Journal of Medical Internet Research*, 12(2), e10.
- Moussavi, S., Chatterji, S., Verdes, E., Tandon, A., Patel, V., & Ustun, B. (2007). Depression, chronic diseases, and decrements in health: results from the World Health Surveys. *The Lancet*, 370(9590), 851-858.
- Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of classification*, 31(3), 274-295.
- Musliner, K. L., Munk-Olsen, T., Eaton, W. W., & Zandi, P. P. (2016). Heterogeneity in long-term trajectories of depressive symptoms: Patterns, predictors and outcomes. *Journal of affective disorders*, 192, 199-211.
- Nakamura, T., Kubo, K., Usuda, Y., & Aramaki, E. (2014). Defining patients with depressive disorder by using textual information. 2014 AAAI Spring Symposium Series,
- Nemesure, M. D., Heinz, M. V., Huang, R., & Jacobson, N. C. (2021). Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Scientific reports*, 11(1), 1-9.
- Newman, M. W., Lauterbach, D., Munson, S. A., Resnick, P., & Morris, M. E. (2011). It's not that i don't have problems, i'm just not putting them on facebook: challenges and opportunities in using online social networks for health. Proceedings of the ACM 2011 conference on Computer supported cooperative work,
- Nobles, A. L., Glenn, J. J., Kowsari, K., Teachman, B. A., & Barnes, L. E. (2018). Identification of Imminent Suicide Risk Among Young Adults using Text Messages. *Proc SIGCHI Conf Hum Factor Comput Syst*, 2018. <https://doi.org/10.1145/3173574.3173987>
- ODea, B., Boonstra, T. W., Larsen, M. E., Nguyen, T., Venkatesh, S., & Christensen, H. (2018). The relationship between linguistic expression and symptoms of depression, anxiety, and suicidal thoughts: A longitudinal study of blog content. *arXiv preprint arXiv:1811.02750*.
- Oh, H. J., Lauckner, C., Boehmer, J., Fewins-Bliss, R., & Li, K. (2013). Facebooking for health: An examination into the solicitation and effects of health-related social support on social networking sites. *Computers in Human Behavior*, 29(5), 2072-2080.
- Olfson, M., & Marcus, S. C. (2009). National patterns in antidepressant medication treatment. *Archives of general psychiatry*, 66(8), 848-856.
- Olfson, M., Marcus, S. C., Druss, B., Elinson, L., Tanielian, T., & Pincus, H. A. (2002). National trends in the outpatient treatment of depression. *Jama*, 287(2), 203-209.
- Olthof, M., Hasselman, F., Strunk, G., Aas, B., Schiepek, G., & Lichtwarck-Aschoff, A. (2020). Destabilization in self-ratings of the psychotherapeutic process is associated with better treatment outcome in patients with mood disorders. *Psychotherapy Research*, 30(4), 520-531.
- Ong, A. D., Bergeman, C. S., Bisconti, T. L., & Wallace, K. A. (2006). Psychological resilience, positive emotions, and successful adaptation to stress in later life. *Journal of Personality and Social Psychology*, 91(4), 730.

- Ono, M., Schneider, S., Junghaenel, D. U., & Stone, A. A. (2019). What affects the completion of ecological momentary assessments in chronic pain research? An individual patient data meta-analysis. *Journal of Medical Internet Research*, 21(2), e11398.
- Ophir, Y., Sisso, I., Asterhan, C. S., Tikochinski, R., & Reichart, R. (2020). The turker blues: Hidden factors behind increased depression rates among Amazon's mechanical turkers. *Clinical Psychological Science*, 8(1), 65-83.
- Orabi, A. H., Buddhitha, P., Orabi, M. H., & Inkpen, D. (2018). Deep learning for depression detection of twitter users. Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic,
- Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt impulsiveness scale. *Journal of clinical psychology*, 51(6), 768-774.
- Paykel, E., Ramana, R., Cooper, Z., Hayhurst, H., Kerr, J., & Barocka, A. (1995). Residual symptoms after partial remission: an important outcome in depression. *Psychological medicine*, 25(6), 1171-1180.
- Pearson, S. D., Katzelnick, D. J., Simon, G. E., Manning, W. G., Helstad, C. P., & Henk, H. J. (1999). Depression among high utilizers of medical care. *Journal of general internal medicine*, 14(8), 461-468.
- Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological science*, 8(3), 162-166.
- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*.
- Perrin, A. (2015). Social media usage. *Pew research center*, 125, 52-68.
- Poole, J. C., Dobson, K. S., & Pusch, D. (2017). Childhood adversity and adult depression: The protective role of psychological resilience. *Child abuse & neglect*, 64, 89-100.
- Pratap, A., Atkins, D. C., Renn, B. N., Tanana, M. J., Mooney, S. D., Anguera, J. A., & Areán, P. A. (2019). The accuracy of passive phone sensors in predicting daily mood. *Depression and anxiety*, 36(1), 72-81.
- Prieto, V. M., Matos, S., Alvarez, M., CACHEDA, F., & Oliveira, J. L. (2014). Twitter: a good place to detect health conditions. *PLoS One*, 9(1), e86191.
- Reece, A. G., & Danforth, C. M. (2017). Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6(1). <https://doi.org/10.1140/epjds/s13688-017-0110-z>
- Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., & Langer, E. J. (2017). Forecasting the onset and course of mental illness with Twitter data. *Sci Rep*, 7(1), 13006. <https://doi.org/10.1038/s41598-017-12961-9>
- Regier, D. A., Goldberg, I. D., & Taube, C. A. (1978). The de facto US mental health services system: a public health perspective. *Archives of general psychiatry*, 35(6), 685-693.
- Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.-A., & Boyd-Graber, J. (2015). Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality,
- Riese, H., Von Klipstein, L., Schoevers, R., van der Veen, D., & Servaas, M. (2021). Personalized ESM monitoring and feedback to support psychological treatment for depression: a pragmatic randomized controlled trial (Therap-i). *BMC psychiatry*, 21(1), 1-11.
- Ringwald, W. R., & Wright, A. G. (2022). Overcoming the Confound of Means and Variability for Measuring Everyday Emotion Dynamics Related to Neuroticism.
- Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2019). Response compliance and predictors thereof in studies using the experience sampling method. *Psychological assessment*, 31(2), 226.
- Robinaugh, D. J., Hoekstra, R. H., Toner, E. R., & Borsboom, D. (2020). The network approach to psychopathology: a review of the literature 2008–2018 and an agenda for future research. *Psychological medicine*, 50(3), 353-366.

- Robinaugh, D. J., Millner, A. J., & McNally, R. J. (2016). Identifying highly influential nodes in the complicated grief network. *Journal of abnormal psychology, 125*(6), 747.
- Rodebaugh, T. L., Tonge, N. A., Piccirillo, M. L., Fried, E., Horenstein, A., Morrison, A. S., Goldin, P., Gross, J. J., Lim, M. H., Fernandez, K. C., Blanco, C., Schneier, F. R., Bogdan, R., Thompson, R. J., & Heimberg, R. G. (2018). Does centrality in a cross-sectional network suggest intervention targets for social anxiety disorder? *J Consult Clin Psychol, 86*(10), 831-844. <https://doi.org/10.1037/ccp0000336>
- Rodriguez, A. J., Holleran, S. E., & Mehl, M. R. (2010). Reading between the lines: the lay assessment of subclinical depression from written self-descriptions. *J Pers, 78*(2), 575-598. <https://doi.org/10.1111/j.1467-6494.2010.00627.x>
- Rost, K., Nutting, P., Smith, J. L., Elliott, C. E., & Dickinson, M. (2002). Managing depression as a chronic disease: a randomised trial of ongoing treatment in primary care. *Bmj, 325*(7370), 934.
- Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biological psychiatry, 84*(6), 443-451.
- Rude, S., Gortner, E.-M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion, 18*(8), 1121-1133.
- Rush, A. J., Carmody, T. J., Ibrahim, H. M., Trivedi, M. H., Biggs, M. M., Shores-Wilson, K., Crismon, M. L., Toprac, M. G., & Kashner, T. M. (2006). Comparison of self-report and clinician ratings on two inventories of depressive symptomatology. *Psychiatric Services, 57*(6), 829-837.
- Rush, A. J., Trivedi, M. H., Wisniewski, S. R., Nierenberg, A. A., Stewart, J. W., Warden, D., Niederehe, G., Thase, M. E., Lavori, P. W., & Lebowitz, B. D. (2006). Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR* D report. *American Journal of Psychiatry, 163*(11), 1905-1917.
- Saha, K., Sugar, B., Torous, J., Abrahao, B., Kiciman, E., & De Choudhury, M. (2019). A social media study on the effects of psychiatric medication use. Proceedings of the International AAAI Conference on Web and Social Media,
- Santos, H., Jr., Fried, E. I., Asafu-Adjei, J., & Ruiz, R. J. (2017). Network Structure of Perinatal Depressive Symptoms in Latinas: Relationship to Stress and Reproductive Biomarkers. *Res Nurs Health, 40*(3), 218-228. <https://doi.org/10.1002/nur.21784>
- Sap, M., Park, G., Eichstaedt, J., Kern, M., Stillwell, D., Kosinski, M., Ungar, L., & Schwartz, H. A. (2014). Developing age and gender predictive lexica over social media. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP),
- Sareen, J., Afifi, T. O., McMillan, K. A., & Asmundson, G. J. (2011). Relationship between household income and mental disorders: findings from a population-based longitudinal study. *Archives of general psychiatry, 68*(4), 419-427.
- Saunders, J. B., Aasland, O. G., Babor, T. F., De La Fuente, J. R., & Grant, M. (1993). Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption-II. *Addiction, 88*(6), 791-804.
- Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., van Nes, E. H., Rietkerk, M., & Sugihara, G. (2009). Early-warning signals for critical transitions. *Nature, 461*(7260), 53-59. <https://doi.org/10.1038/nature08227>
- Schiepek, G. K., Tominschek, I., & Heinzl, S. (2014). Self-organization in psychotherapy: testing the synergetic model of change processes. *Frontiers in Psychology, 5*, 1089.
- Schreuder, M. J., Wigman, J. T., Groen, R. N., Weinans, E., Wichers, M., & Hartman, C. A. (2022). Anticipating the direction of symptom progression using critical slowing down: a proof-of-concept study. *BMC psychiatry, 22*(1), 1-13.

- Schwartz, H. A., Eichstaedt, J., Kern, M., Park, G., Sap, M., Stillwell, D., Kosinski, M., & Ungar, L. (2014). Towards assessing changes in degree of depression through facebook. *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*,
- Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., & Seligman, M. E. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS One*, 8(9), e73791.
- Schweren, L., van Borkulo, C. D., Fried, E., & Goodyer, I. M. (2018). Assessment of symptom network density as a prognostic marker of treatment response in adolescent depression. *JAMA psychiatry*, 75(1), 98-100.
- Segal, A., Wald, I., Lubin, G., Fruchter, E., Ginat, K., Ben Yehuda, A., Pine, D. S., & Bar-Haim, Y. (2020). Changes in the dynamic network structure of PTSD symptoms pre-to-post combat. *Psychol Med*, 50(5), 746-753. <https://doi.org/10.1017/S0033291719000539>
- Seow, T. X., & Gillan, C. M. (2020). Transdiagnostic phenotyping reveals a host of metacognitive deficits implicated in compulsivity. *Scientific reports*, 10(1), 1-11.
- Shapiro, D. A., & Shapiro, D. (1982). Meta-analysis of comparative therapy outcome studies: a replication and refinement. *Psychological bulletin*, 92(3), 581.
- Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, 1(2), 213-220.
- Shen, Y.-C., Kuo, T.-T., Yeh, I.-N., Chen, T.-T., & Lin, S.-D. (2013). Exploiting temporal information in a two-stage classification framework for content-based depression detection. *Pacific-Asia conference on knowledge discovery and data mining*,
- Shin, K. E., Newman, M. G., & Jacobson, N. C. (2022). Emotion network density is a potential clinical marker for anxiety and depression: Comparison of ecological momentary assessment and daily diary. *British Journal of Clinical Psychology*, 61, 31-50.
- Simon, G. E., Von Korff, M., Ludman, E. J., Katon, W. J., Rutter, C., Unützer, J., Lin, E. H., Bush, T., & Walker, E. (2002). Cost-effectiveness of a program to prevent depression relapse in primary care. *Medical Care*, 941-950.
- Smith, K. E., Mason, T. B., Crosby, R. D., Cao, L., Leonard, R. C., Wetterneck, C. T., Smith, B. E. R., Farrell, N. R., Riemann, B. C., Wonderlich, S. A., & Moessner, M. (2019). A comparative network analysis of eating disorder psychopathology and co-occurring depression and anxiety symptoms before and after treatment. *Psychol Med*, 49(2), 314-324. <https://doi.org/10.1017/S0033291718000867>
- Smith, R., Alkozei, A., Killgore, W. D., & Lane, R. D. (2018). Nested positive feedback loops in the maintenance of major depression: An integration and extension of previous models. *Brain, behavior, and immunity*, 67, 374-397.
- Snijders, T. A. (2011). Statistical models for social networks. *Annual review of sociology*, 37, 131-153.
- Snippe, E., Viechtbauer, W., Geschwind, N., Klippel, A., de Jonge, P., & Wichers, M. (2017). The Impact of Treatments for Depression on the Dynamic Network Structure of Mental States: Two Randomized Controlled Trials. *Sci Rep*, 7, 46523. <https://doi.org/10.1038/srep46523>
- Solomon, D. A., Keller, M. B., Leon, A. C., Mueller, T. I., Lavori, P. W., Shea, M. T., Coryell, W., Warshaw, M., Turvey, C., & Maser, J. D. (2000). Multiple recurrences of major depressive disorder. *American Journal of Psychiatry*, 157(2), 229-233.
- Souery, D., Papakostas, G. I., & Trivedi, M. H. (2006). Treatment-resistant depression. *Journal of Clinical Psychiatry*, 67, 16.
- Spielberger, C. D. (1983). State-trait anxiety inventory for adults.
- Spiller, T. R., Levi, O., Neria, Y., Suarez-Jimenez, B., Bar-Haim, Y., & Lazarov, A. (2020). On the validity of the centrality hypothesis in cross-sectional between-subject networks of psychopathology. *BMC medicine*, 18(1), 1-14.

- Suls, J., Green, P., & Hillis, S. (1998). Emotional reactivity to everyday problems, affective inertia, and neuroticism. *Personality and Social Psychology Bulletin*, 24(2), 127-136.
- Sun, J., Schwartz, H. A., Son, Y., Kern, M. L., & Vazire, S. (2020). The language of well-being: Tracking fluctuations in emotion experience through everyday speech. *Journal of Personality and Social Psychology*, 118(2), 364.
- Sutin, A. R., Terracciano, A., Milaneschi, Y., An, Y., Ferrucci, L., & Zonderman, A. B. (2013). The trajectory of depressive symptoms across the adult life span. *JAMA psychiatry*, 70(8), 803-811.
- Tang, T. Z., & DeRubeis, R. J. (1999). Sudden gains and critical sessions in cognitive-behavioral therapy for depression. *Journal of consulting and clinical psychology*, 67(6), 894.
- Ten Thij, M., Bathina, K., Rutter, L. A., Lorenzo-Luaces, L., van de Leemput, I. A., Scheffer, M., & Bollen, J. (2020). Depression alters the circadian pattern of online activity. *Scientific reports*, 10(1), 1-10.
- Terluin, B., De Boer, M. R., & De Vet, H. C. (2016). Differences in connection strength between mental symptoms might be explained by differences in variance: Reanalysis of network data did not confirm staging. *PLoS One*, 11(11), e0155205.
- Thelen, E., & Smith, L. B. (1998). Dynamic systems theories.
- Torous, J., Kiang, M. V., Lorme, J., & Onnela, J.-P. (2016). New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. *JMIR mental health*, 3(2), e5165.
- Trull, T. J., Solhan, M. B., Tragesser, S. L., Jahng, S., Wood, P. K., Piasecki, T. M., & Watson, D. (2008). Affective instability: measuring a core feature of borderline personality disorder with ecological momentary assessment. *Journal of abnormal psychology*, 117(3), 647.
- Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., & Ohsaki, H. (2015). *Recognizing Depression from Twitter Activity* Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15,
- Tunnard, C., Rane, L. J., Wooderson, S. C., Markopoulou, K., Poon, L., Fekadu, A., Jurueña, M., & Cleare, A. J. (2014). The impact of childhood adversity on suicidality and clinical course in treatment-resistant depression. *Journal of affective disorders*, 152, 122-130.
- Turvey, C. L., Wallace, R. B., & Herzog, R. (1999). A revised CES-D measure of depressive symptoms and a DSM-based measure of major depressive episodes in the elderly. *International psychogeriatrics*, 11(2), 139-148.
- Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7), 456-464.
- van Borkulo, C., Borsboom, D., Nivard, M., & Cramer, A. (2011). NetLogo Symptom Spread Mode.
- van Borkulo, C., Boschloo, L., Borsboom, D., Penninx, B. W., Waldorp, L. J., & Schoevers, R. A. (2015). Association of Symptom Network Structure With the Course of [corrected] Depression. *JAMA psychiatry*, 72(12), 1219-1226. <https://doi.org/10.1001/jamapsychiatry.2015.2079>
- van Borkulo, C., Boschloo, L., Borsboom, D., Penninx, B. W., Waldorp, L. J., & Schoevers, R. A. (2015). Association of symptom network structure with the course of depression. *JAMA psychiatry*, 72(12), 1219-1226.
- van de Leemput, I. A., Wichers, M., Cramer, A. O., Borsboom, D., Tuerlinckx, F., Kuppens, P., van Nes, E. H., Viechtbauer, W., Giltay, E. J., Aggen, S. H., Derom, C., Jacobs, N., Kendler, K. S., van der Maas, H. L., Neale, M. C., Peeters, F., Thiery, E., Zachar, P., & Scheffer, M. (2014). Critical slowing down as early warning for the onset and termination of depression. *Proc Natl Acad Sci U S A*, 111(1), 87-92. <https://doi.org/10.1073/pnas.1312114110>
- van Genugten, C. R., Schuurmans, J., Lamers, F., Riese, H., Penninx, B. W., Schoevers, R. A., Riper, H. M., & Smit, J. H. (2020). Experienced burden of and adherence to smartphone-based

- ecological momentary assessment in persons with affective disorders. *Journal of clinical medicine*, 9(2), 322.
- van Os, J., Verhagen, S., Marsman, A., Peeters, F., Bak, M., Marcelis, M., Drukker, M., Reininghaus, U., Jacobs, N., & Lataster, T. (2017). The experience sampling method as an mHealth tool to support self-monitoring, self-insight, and personalized health care in clinical practice. *Depression and anxiety*, 34(6), 481-493.
- van Rooijen, G., Isvoranu, A. M., Kruijt, O. H., van Borkulo, C. D., Meijer, C. J., Wigman, J. T. W., Ruhe, H. G., de Haan, L., & investigators, G. (2018). A state-independent network of depressive, negative and positive symptoms in male patients with schizophrenia spectrum disorders. *Schizophr Res*, 193, 232-239.
<https://doi.org/10.1016/j.schres.2017.07.035>
- Van Veen, M., Karsten, J., & Lancel, M. (2017). Poor sleep and its relation to impulsivity in patients with antisocial or borderline personality disorders. *Behavioral Medicine*, 43(3), 218-226.
- Vos, T., Lim, S. S., Abbafati, C., Abbas, K. M., Abbasi, M., Abbasifard, M., Abbasi-Kangevari, M., Abbastabar, H., Abd-Allah, F., & Abdelalim, A. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, 396(10258), 1204-1222.
- Waller, G. (2009). Evidence-based treatment and therapist drift. *Behaviour research and therapy*, 47(2), 119-127.
- Waller, G., & Turner, H. (2016). Therapist drift redux: Why well-meaning clinicians fail to deliver evidence-based therapy, and how to get back on track. *Behaviour research and therapy*, 77, 129-137.
- Wang, J., Mann, F., Lloyd-Evans, B., Ma, R., & Johnson, S. (2018). Associations between loneliness and perceived social support and outcomes of mental health problems: a systematic review. *BMC psychiatry*, 18(1), 1-16.
- Wang, J. L., Schmitz, N., & Dewa, C. S. (2010). Socioeconomic status and the risk of major depression: the Canadian National Population Health Survey. *Journal of Epidemiology & Community Health*, 64(5), 447-452.
- Wang, T., Brede, M., Ianni, A., & Mentzakis, E. (2018). Social interactions in online eating disorder communities: A network perspective. *PLoS One*, 13(7), e0200800.
<https://doi.org/10.1371/journal.pone.0200800>
- Watson, D. (1988). The vicissitudes of mood measurement: effects of varying descriptors, time frames, and response formats on measures of positive and negative affect. *Journal of Personality and Social Psychology*, 55(1), 128.
- Whiteford, H. A., Degenhardt, L., Rehm, J., Baxter, A. J., Ferrari, A. J., Erskine, H. E., Charlson, F. J., Norman, R. E., Flaxman, A. D., & Johns, N. (2013). Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *The Lancet*, 382(9904), 1575-1586.
- Wichers, M., Groot, P. C., Psychosystems, E., & Group, E. (2016). Critical slowing down as a personalized early warning signal for depression. *Psychotherapy and psychosomatics*, 85(2), 114-116.
- Wichers, M., Smit, A. C., & Snippe, E. (2020). Early warning signals based on momentary affect dynamics can expose nearby transitions in depression: a confirmatory single-subject time-series study. *Journal for Person-Oriented Research*, 6(1), 1.
- Wigman, J. T., van Os, J., Borsboom, D., Wardenaar, K. J., Epskamp, S., Klippel, A., Merge, Viechtbauer, W., Myin-Germeys, I., & Wichers, M. (2015). Exploring the underlying structure of mental disorders: cross-diagnostic differences and similarities from a network perspective using both a top-down and a bottom-up approach. *Psychol Med*, 45(11), 2375-2387. <https://doi.org/10.1017/S0033291715000331>
- Wigman, J. T., van Os, J., Thiery, E., Derom, C., Collip, D., Jacobs, N., & Wichers, M. (2013). Psychiatric diagnosis revisited: towards a system of staging and profiling combining

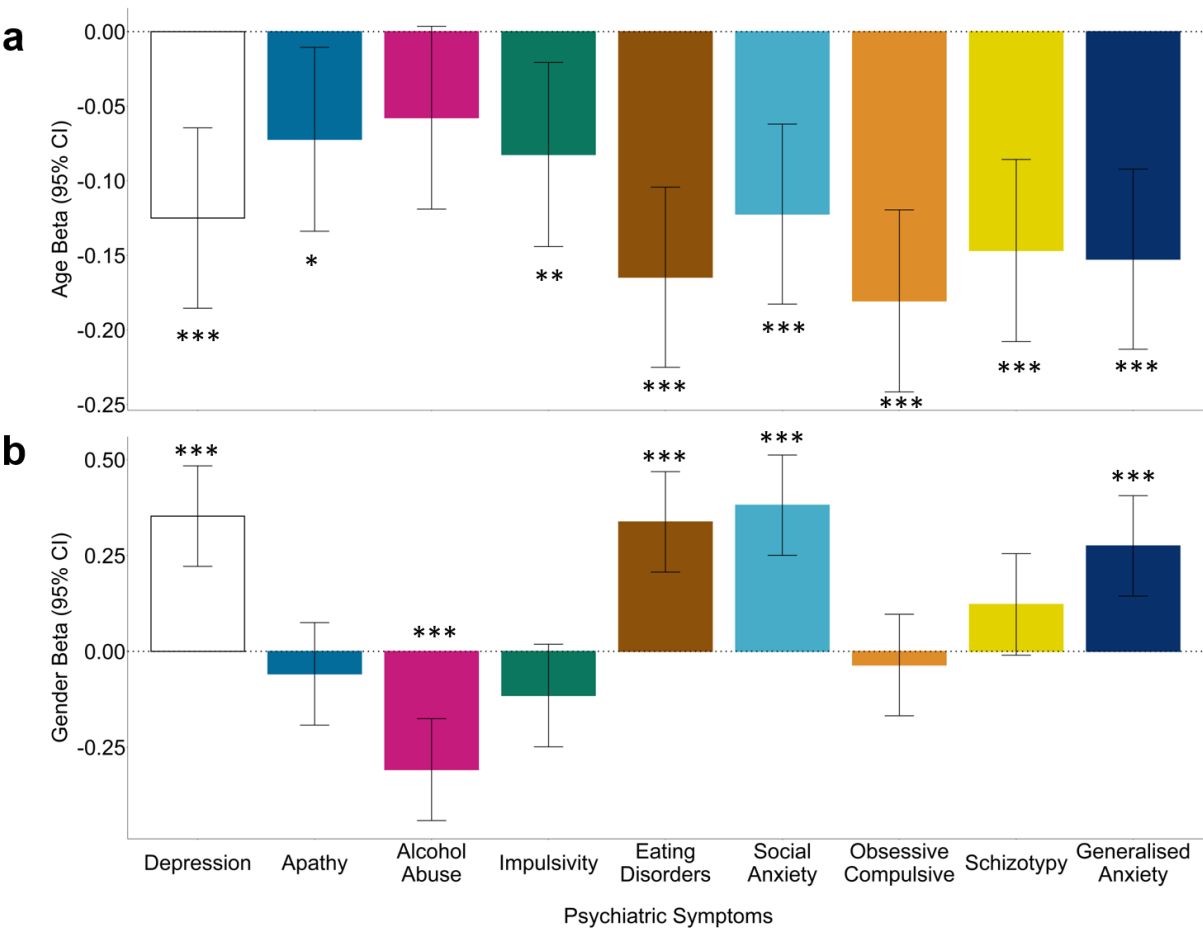
- nomothetic and idiographic parameters of momentary mental states. *PLoS One*, 8(3), e59559.
- Wild, B., Eichler, M., Friederich, H.-C., Hartmann, M., Zipfel, S., & Herzog, W. (2010). A graphical vector autoregressive modelling approach to the analysis of electronic diary data. *BMC medical research methodology*, 10(1), 1-13.
- Wingo, A. P., Wrenn, G., Pelletier, T., Gutman, A. R., Bradley, B., & Ressler, K. J. (2010). Moderating effects of resilience on depression in individuals with a history of childhood abuse or trauma exposure. *Journal of affective disorders*, 126(3), 411-414.
- Winter, N. R., Leenings, R., Ernsting, J., Sarink, K., Fisch, L., Emden, D., Blanke, J., Goltermann, J., Opel, N., & Barkhau, C. (2022). Quantifying Deviations of Brain Structure and Function in Major Depressive Disorder Across Neuroimaging Modalities. *JAMA psychiatry*.
- Wojcik, S., & Hughes, A. (2019). Sizing up Twitter users. *Washington, DC: Pew Research Center*.
- Wolf, M., Sedway, J., Bulik, C. M., & Kordy, H. (2007). Linguistic analyses of natural written language: unobtrusive assessment of cognitive style in eating disorders. *Int J Eat Disord*, 40(8), 711-717. <https://doi.org/10.1002/eat.20445>
- Wolf, M., Theis, F., & Kordy, H. (2013). Language Use in Eating Disorder Blogs. *Journal of Language and Social Psychology*, 32(2), 212-226. <https://doi.org/10.1177/0261927x12474278>
- Yang, J., Morris, M. R., Teevan, J., Adamic, L. A., & Ackerman, M. S. (2011). Culture matters: A survey study of social Q&A behavior. Fifth International AAAI Conference on Weblogs and Social Media,
- Yoo, D. W., Ernala, S. K., Saket, B., Weir, D., Arenare, E., Ali, A. F., Van Meter, A. R., Birnbaum, M. L., Abowd, G. D., & De Choudhury, M. (2021). Clinician Perspectives on Using Computational Mental Health Insights From Patients' Social Media Activities: Design and Qualitative Evaluation of a Prototype. *JMIR mental health*, 8(11), e25455.
- Zhang, R. (2017). The stress-buffering effect of self-disclosure on Facebook: An examination of stressful life events, social support, and mental health among college students. *Computers in Human Behavior*, 75, 527-537.
- Zimmer, M. (2018). Addressing conceptual gaps in big data research ethics: An application of contextual integrity. *Social Media+ Society*, 4(2), 2056305118768300.
- Zimmermann, J., Brockmeyer, T., Hunn, M., Schauenburg, H., & Wolf, M. (2017). First-person Pronoun Use in Spoken Language as a Predictor of Future Depressive Symptoms: Preliminary Evidence from a Clinical Sample of Depressed Patients. *Clin Psychol Psychother*, 24(2), 384-391. <https://doi.org/10.1002/cpp.2006>
- Zimmermann, J., Wolf, M., Bock, A., Peham, D., & Benecke, C. (2013). The way we refer to ourselves reflects how we relate to others: Associations between first-person pronoun use and interpersonal problems. *Journal of Research in Personality*, 47(3), 218-225.
- Zimmermann, J., Woods, W. C., Ritter, S., Happel, M., Masuhr, O., Jaeger, U., Spitzer, C., & Wright, A. G. (2019). Integrating structure and dynamics in personality assessment: First steps toward the development and validation of a personality dynamics diary. *Psychological assessment*, 31(4), 516.
- Zomick, J., Levitan, S. I., & Serper, M. (2019). Linguistic analysis of schizophrenia in reddit posts. Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology,
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2), 301-320.
- Zung, W. W. (1965). A self-rating depression scale. *Archives of general psychiatry*, 12(1), 63-70.

7 Appendices

7.1 Appendix I: Supplemental Information for Chapter 2

Minimum word count per user	Sample Size	Depression R ²	Depression r
5 days of Tweets/43 words	836	0.010	0.13
200	836	-0.001	0.07
400	836	0.044	0.22
500	836	0.034	0.21

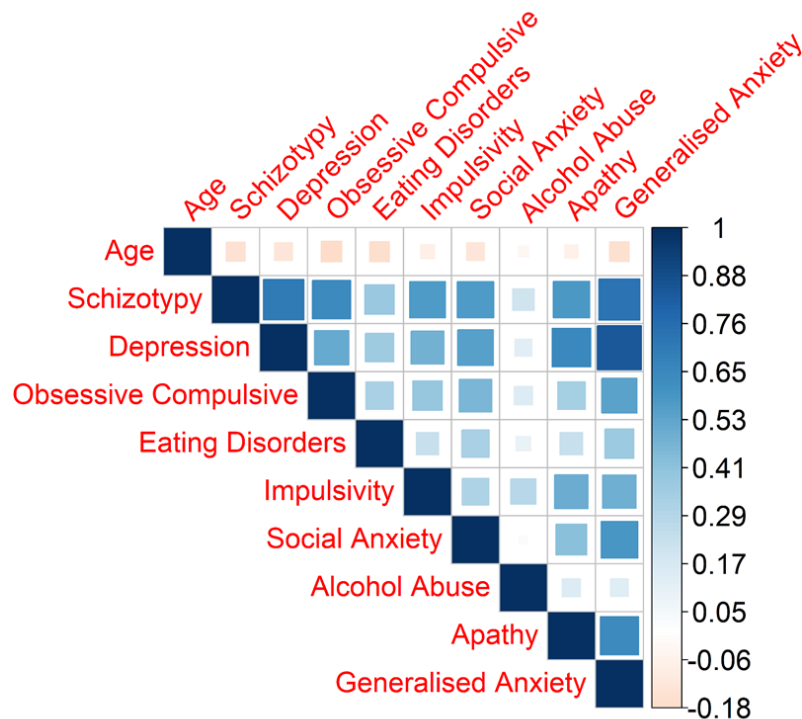
Supplementary Table 1: Effect of minimum word count per user on the depression trained model’s predictive performance, controlling for sample size. The minimum word count inclusion criterion used throughout the main text is ‘5 days of Tweets/43 words’.



Supplementary Figure 1: Association between 9 self-report questionnaires with age and gender

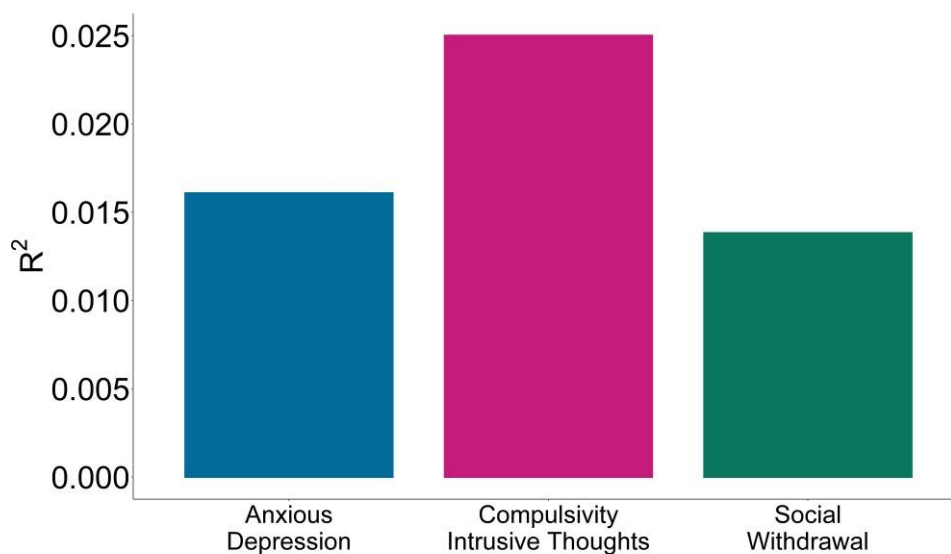
a) Associations between 9 psychiatric questionnaires and age. There were significant negative associations between all questionnaires and age except for alcohol abuse. b) Associations between questionnaires and gender. Female participants had significantly elevated levels of eating disorders, social anxiety, depression, and state anxiety compared to males. While male participants had significantly higher levels of alcohol abuse. Bars indicate a 95% confidence interval around the mean.

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



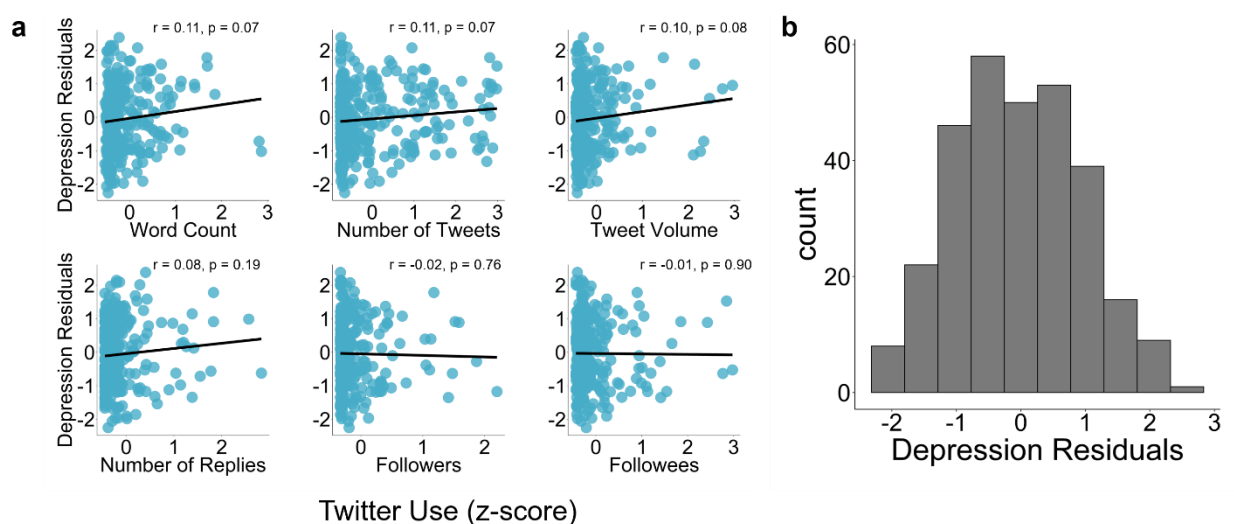
Supplementary Figure 2: Bivariate correlations among 9 psychiatric questionnaires and age

All psychiatric disorders are positively correlated with each other. Age is negatively associated with every disorder.



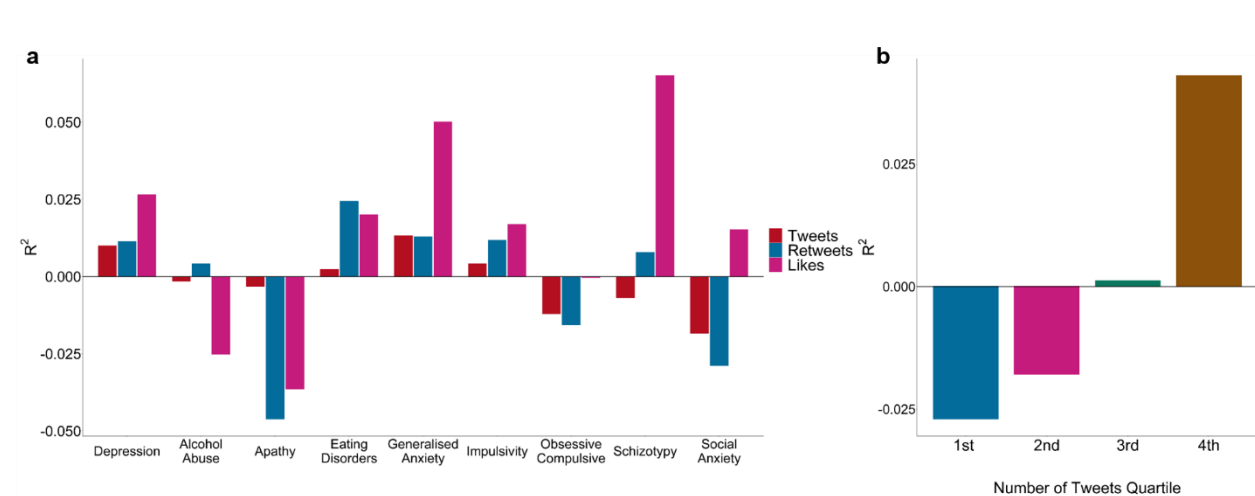
Supplementary Figure 3: Predictive performance of an anxious depression trained model tested on three transdiagnostic dimensions: anxious depression, compulsivity and intrusive thoughts, and social withdrawal

The anxious depression model performed best when tested on compulsivity and intrusive thoughts ($R^2 = 0.025$), but had above zero performance on all three dimensions.



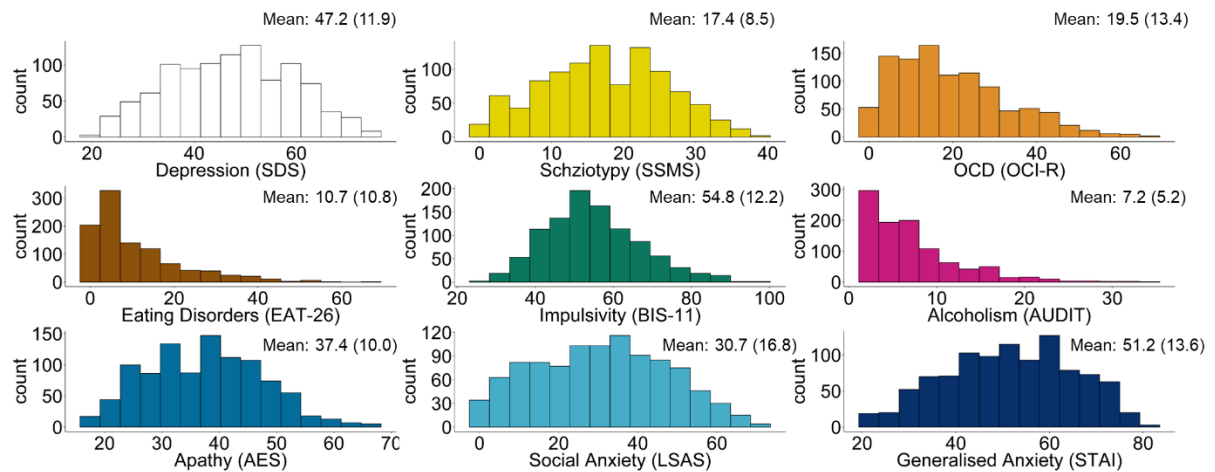
Supplementary Figure 4: No association between z-scored Twitter use and depression residuals derived from LIWC text feature model

a) Regression plots for depression residuals and z-scored Twitter use including: mean word count, total number of tweets, tweet volume, total number of replies, followers, and followees. There was no significant association between Twitter use and depression residuals (all $|\beta| > 0.02$, $p > 0.05$). b) Histogram of depression residuals were centered on zero (Mean = -0.05, $t = -0.84$ (df = 301), $p = 0.40$).

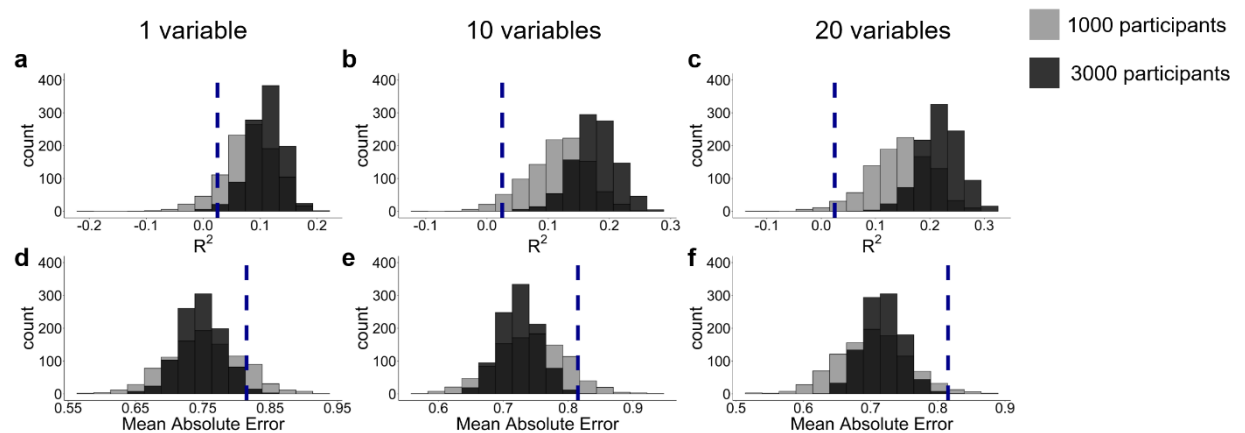


Supplementary Figure 5: Predictive performance of a depression model trained on subsets of Tweets

A) Predictive performance (R^2) of a depression model trained using only Tweets ($n = 756$), Retweets ($n = 637$), or Likes ($n = 902$). B) Depression model performance on four quartiles of text feature data: 1st quartile (mean Tweets = 23.5), 2nd quartile (mean Tweets = 159.2), 3rd quartile (mean Tweets = 867.2), and 4th quartile (mean Tweets = 3,625.8). Text features were divided into quartiles based on the total number of Tweets, Retweets, and Likes. Models trained on data from the 3rd and 4th quartiles had above zero performance compared to those trained on data from the lower two quartiles.



Supplementary Figure 6: Histograms of 9 psychiatric scales with means and standard deviations



Supplementary Figure 7: Power to detect an effect size double the observed depression predictive performance (i.e., $r = 0.32$) in simulated data

Simulation data of 3 types of datasets with 99 input features and 1 continuous target outcome with a sample size of either 1,000 or 3,000. The correlation between either 1, 10, or 20 features with the target outcome was set to $r = 0.32$. In datasets with more than 1 feature, multicollinearity was simulated among the relevant features by setting the correlation between those features to $r = 0.50$.

When only 1 variable was associated with the target outcome, 10.6% of simulated values had a worse performance than our observed value (blue dashed line; $R^2 = 0.025$, MAE = 0.815) (Figure S4a and S4a). However, this percentage dropped to only 2.3% when there were 20 variables associated with the target outcome (Figure S4c and S4f). Consequently, even in the worst-case scenario, where only 1 variable is truly associated with the target outcome, we would expect to report a larger R^2 value than what we observed in our study in approximately 90% of cases. Although increasing the sample size to 3,000 participants would further reduce the likelihood to 1%, it is already very unlikely to miss an effect size of this magnitude with a sample of 1,000. In our dataset, we have observed 51 variables significantly associated with depression severity. Thus, the likelihood of missing an effect size as large as $r = 0.32$ in our dataset is much smaller than 2.3%.

7.2 Appendix II: Supplemental Information for Chapter 3

7.2.1 Methods for Experiment 1: Paid Students and Citizen Scientists in Neureka

7.2.1.1 Procedure

Paid students were recruited through advertising at Fresher's week, societies, mailing lists, online and physical advertisements and paid €50 for participation. To assess changes in mental health treatment during the study, participants were asked weekly if in the past week they had: 1) seen a healthcare practitioner for their mental health (yes or no), 2) taken an antidepressant (yes or no), and 3) received talk/psychotherapy (yes or no). Participants in the paid student arm were paid €50 if they completed 90% of their assessments. They also completed one extra item each week: the number of college assignments they had in the past week (0 to 3+, where 3 indicated 3 or more assignments). Besides this additional item, the procedures for the two samples were identical. The Citizen Scientist arm was not paid for participation; they downloaded the public version of the app on the Apple App Store or Google Play Store and completed the assessments without supervision. For Citizen Scientists who attempted to complete Multi-Mood more than once, we included assessments from the most complete attempt. In March 2021, Neureka added a mood graph feature to provide visual feedback to users. In the final sample of N=194, N=126 (65%) had access to the mood graph.

7.2.1.2 Data Preparation

Positive items were first reverse coded to reflect negative affect, e.g., enthusiastic became 'not enthusiastic'. This was to standardise the directional interpretation of edges between nodes and overall network connectivity across networks that vary in proportion of positive vs negative nodes. EMA items were rated on a 7-point Likert scale. To ensure that there was sufficient variance for each network to be estimated, for every item in their time series, we added a small amount of random noise, -0.1 to + 0.1, to each EMA item. This meant that networks were estimable

even if a subject had rated a single item the same (i.e., with little to no variance) in their time series.

7.2.1.3 Sensitivity Analyses

Contemporaneous ($r > 0.51$, $p < 0.001$) and directed ($r > 0.25$, $p < 0.002$) connectivity were positively associated with network stability among Paid Students and Citizen Scientists (Figure S1). To understand if network stability might therefore bias our connectivity results in the two Neureka samples, we repeated our key analyses in these datasets with each participant's network stability as a covariate. Data missingness was low among our final samples (Table S1), and so we did not impute missing data. To confirm that this decision did not affect our results, we repeated our analysis of connectivity and baseline depression using imputed data. We did this using the *amelia* package, which performs multiple ($n=5$) imputations of missing data in a multivariate time-series and averages across them to produce an imputed timeseries.

7.2.1.4 Permutation Tests

To test if these distributions had greater correlations than expected by chance, we performed permutation tests. Specifically, in all 5-node network combinations, we randomised the timeseries of each participant's EMA data. We estimated personalised contemporaneous and directed networks based on this randomised data for each of the 4,368 network combinations and determined the correlation between (randomised) network connectivity and depression for each. We determined the statistical significance for each network by counting the number of values in the shuffled distribution that exceeded the observed value for a given true network, i.e., its p-value. We repeated this for all combinations of networks and then found the proportion of networks that reached $p < .05$. Finally, we tested if the valence of nodes in the networks affected the results by examining whether the number of negative (vs. positive) EMA items in a given network influenced the association between depression and network connectivity.

7.2.1.5 Network Stability Test

In order to estimate each network's stability, we successively dropped, within-individual, up to 85% of each participant's data in increments of 5% and calculated the within-individual correlation between network structure among the full dataset and the down sampled data. For each participant, we calculated the proportion of data that could be dropped to retain a correlation of 0.70 with their full timeseries, i.e., correlation stability coefficient (CS coefficient). Based on previously recommended guidelines (S. Epskamp, D. Borsboom, et al., 2018), the correlation stability coefficient should not be below 0.25 and ideally above 0.50. After examining internal reliability, we tested the generalisability of the structure of these networks across datasets by correlating the edge strengths (25 edges for the directed network, and 10 for the contemporaneous network) across the independent samples.

7.2.1.6 Observed Power from Experiment 1

A power analysis was completed in R using the pwr package with $\alpha=0.05$, $\text{power}=0.8$. The observed associations between network connectivity and baseline depression were relatively small and varied across our two samples. Before carrying out Experiment 1, we estimated power for detecting associations between the connectivity of both contemporaneous and directed networks and mean and variance of depression. The average Pearson R between baseline depression and mean contemporaneous network connectivity across Paid Students and Citizen Scientists was $r=.15$ (i.e., mean PS: $r=.07$ and CS: $r=.23$). At an alpha of $p=.05$, power of 0.8, a sample of $N=323$ is required to detect this. The same average for directed network connectivity was $r=.11$ (i.e., mean PS: $r=.03$ and CS: $r=.19$). A sample of $N=645$ would be required to detect this effect. In Experiment 2, the available sample size of our external dataset (HNATD) was $N=519$ and we were thus powered to test for contemporaneous network associations only. The effect size for the observed association between network connectivity and SD of depression was much stronger in both samples and for both network types (contemporaneous and directed). The average Pearson R between SD of depression and mean contemporaneous network connectivity across Paid Students and Citizen Scientists was $r=.345$ (i.e., mean PS: $r=.29$ and CS: $r=.40$). At an alpha of $p=.05$, power of 0.8, a sample of $N=63$ is required to detect this.

The same average for directed network connectivity was $r=.26$ (i.e., mean PS: $r=.21$ and CS: $r=.31$). At an alpha of $p=.05$, power of 0.8, a sample of $N=113$ is required to detect this. In Experiment 2, the sample size of the HNATD was therefore well-powered to detect associations between network connectivity and variance in depressive symptomology.

7.2.2 Methods for Experiment 2: Replication in a large and independent Citizen Science sample (HNATD)

7.2.2.1 Procedure

At baseline, participants completed the Quick Inventory of Depressive Symptoms (QIDS)(Rush, Carmody, et al., 2006). Participants rated their mood on a visual analogue scale between 0 (not at all) – 100 (very much). Among the 43 questions that participants answered, we selected a subset of 11 items (6 positive and 5 negative) from the Positive and Negative Affect Schedule (PANAS). Items selected were relaxed, energetic, anxious, enthusiastic, nervous, content, irritable, calm, dull, cheerful, and tired. There were 1,302 participants who completed at least 1 EMA assessment and 519 that completed at least 75% of assessments and were analysed here (Table S1). We retained the ‘gloomy’ for a separate analysis described below.

7.2.2.2 Data Preparation

As with the Experiment 1 Neureka datasets, positive items were first reverse coded to reflect negative affect, e.g., enthusiastic became ‘not enthusiastic’. The HNATD study used VAS scales from 0-100 and so variance in responding was not an issue for network estimation. Nonetheless, we added the same noise to the data as in Experiment 1 for consistency. More importantly perhaps, HNATD implemented a different EMA schedule to the Neureka studies; they had 3 assessments per day, resulting in unequal intervals between assessments overnight. To avoid this, we removed all overnight lags in our analyses, so the directed network edges uniformly reflect the 6-hour lag of one emotion onto another (and itself).

7.2.2.3 Data Analysis

Analysis was carried out in an identical manner to that described in Experiment 1. Because we do not have weekly depression scores, we used the standard deviation of the EMA item 'gloomy' as a proxy for the variance in depression and used it to test if network connectivity was best explained by variance in negative emotion, rather than its point-estimated severity per se. Note that the 'gloomy' item was not used in any network estimation analysis. For permutation analyses, as with the Neureka samples, participants classified as outliers in terms of their network connectivity scores were excluded on a per-network basis. The mean sample size in each network was therefore $N=515.3$, $\text{min}=512$, $\text{max}=516$.

7.2.3 Methods for Experiment 3: Comparison of Clinical and Non-Clinical Network Structure and Stability

7.2.3.1 Procedure

Data were gathered from 45 patients with a primary diagnosis of MDD or GAD who completed EMA for 30 days prior to undergoing therapy (Fisher et al., 2017). At baseline, participants completed the Hamilton Anxiety Rating Scale (HARS) (Hamilton, 1959) and the Hamilton Rating Scale for Depression (HRSD) (Hamilton, 1960). Participants were prompted to complete EMA assessments 4 times per day for a minimum of 30 days and rated items on a visual analogue scale from 0 (not at all) to 100 (as much as possible). For each assessment, participants answered 26 questions related to symptoms of MDD, GAD, positive affect, negative affect, rumination, behavioural avoidance, and reassurance seeking. We selected only the positive and negative affect items, and removed items that were either behavioural, e.g., procrastinated, or physical, e.g., experienced muscle fatigue. After removing these questions, we were left with 17 items including: energetic, enthusiastic, content, irritable, restless, worried, worthless or guilty, frightened or afraid, loss of interest or pleasure, angry, hopeless, positive, fatigued, difficulty concentrating, accepted or supported, threatened, and dwelled on the past (Table S2). The 'down or depressed' item was excluded from network estimation but was retained for a separate analysis detailed below.

7.2.3.2 Data Preparation

No exclusions were applied to this dataset, as all participants provided to us had completed the study requirement of 30 days with at least one assessment per day (Table S1 in 7.2.3.3 Supplementary Materials). As with in Experiment 1 and 2, positive items were reverse coded, and we added a small amount of noise to the data to ensure networks were estimable. Similar to HNATD, we removed all overnight lags in our analyses so that all directed edges referred to a 4-hour interval of one emotion onto another.

7.2.3.3 Data Analysis

Analysis was carried out in an identical manner to that described in Experiment 1 and 2. Because only baseline depression scores were available, we used the standard deviation of the ‘down or depressed’ item over the entire time-series to approximate depression variance, as a proxy for the weekly depression scores available in the Neureka samples. We then estimated the correlation between network connectivity and the variance of ‘down or depressed’.

For permutation analyses, participants classified as outliers in terms of their network connectivity scores were also excluded on a per-network basis. The mean sample size in each network was therefore $N=44.9$, $\text{min}=44$, $\text{max}=45$.

	Paid Students	Citizen Scientists	HNATD	Clinical Sample
Sample Characteristics				
Sample Size	155	194	519	40
Age (SD)[Range]	23 (4.7) [18, 41]	49.5 (13.6) [18, 82]	40.3 (13.6) [17, 73]	34.1 (13.2) [18, 60]
Gender				
% Female	74.8% (116)	64.9% (126)	83% (433)	65% (26)
% Male	23.9% (37)	32% (62)	17% (88)	35% (13)
% Other	1.3% (2)	3.1% (6)	NA	NA
% Depressed (N)	54.2% (84)	32.5% (63)	51.3% (266)	95% (38)
Study Design				
Depression Questionnaire	CES-D 20	CES-D 20	QIDS	HAM-D
Frequency	2x per day	2x per day	3x per day	4x per day
# PA items	7	7	6	5
# NA items	9	9	6	13
# Unique 5-Node Networks	4368	4368	462	6188
Response Option	Likert (-3 to +3)	Likert (-3 to +3)	VAS (0-100)	VAS (0-100)
Required # Assessments	84	84	68	30*
EMA Data				
Mean Num. Assessments	104.3 (5.4)	97.6 (7.4)	77.6 (5.5)	115.5 (13.6)
Min. Num. Assessments	84	84	68	87
Max Num. Assessments	112	112	90	151
Mean % Complete	93.2% (4.8%)	87.1% (6.6)	86.2% (6.1)	NA

Table S1: Comparison of demographic characteristics between Paid Students, Citizen Scientists, HNATD, and a clinical sample.

The “Other” gender category includes: non-binary, transgender male, transgender female, and did not disclose. Percent depressed refers to the percentage of a given sample on or above threshold for depression on the respective self-report instrument used in that study. For neureka datasets that was ≥ 20 on the CES-D, for HNATD > 5 on the QIDS, and for the clinical sample was > 7 on the HAM-D.

*Participants in Fisher et al., (2017) were required to complete a fixed number of assessments, but were asked to complete at least 1 assessment for 30 days.

Dataset	Positive Affect Items	Negative Affect Items
Paid students and Citizen Scientists	Enthusiastic, Cheerful, Strong, Happy, Energetic, Content, Talkative	Irritated, Down, Lonely, Anxious, Guilty, Indecisive, Restless, Agitated, Worried
HNATD	Relaxed, Energetic, Enthusiastic, Content, Calm, Cheerful	Gloomy, Anxious, Nervous, Irritable, Dull, Tired
Fisher	Energetic, Enthusiastic, Content, Positive, Accepted/Supported	Irritable, Restless, Worried, Worthless/Guilty, Frightened, Loss of interest/pleasure, Angry, Hopeless, Down/Depressed, Fatigued, Difficulty concentrating, Threatened/Judged/Intimated/Dwelled on Past

Table S2: Positive and negative affect items used to construct 5-node networks

7.2.4 Supplementary Results

Directed Network Connectivity and Depression in Experiment 1. Moving from contemporaneous to directed networks, we found that the exemplar network was again stable across samples, with the 25 edges highly correlated $r=.98$ (Figure S1A). In both datasets, the autoregressive coefficient of worried was the strongest edge, that is, exhibited the highest inertia. Unsurprisingly, directed networks had lower correlation stability compared to contemporaneous ones (Figure S1B). Paid Students and Citizen Scientists had a CS(0.7) of 0.35 and 0.30 respectively. This is expected, however, because correlation stability means something different in the directed case, where dropping assessments does not simply reduce power, it changes the *nature* of the lags (degrading them from 12-hour lags to 24, 36, 48-hour lags or more). Turning to our permuted networks, Citizen Scientists (Median $r = 0.13$) and Paid Students (Median $r = 0.01$) again had nominally positive distributions of the correlations between directed network connectivity with baseline depression scores (Figure S1C). Similar to the contemporaneous network analysis, we found that no networks in the Paid Student sample were significantly associated with baseline depression, while 50.25% were significant in the Citizen Scientists. Again, the more negative affect items in a given network, the more positive was the correlation with baseline depression. This was true of both Paid Students ($r=.54$, $p < 0.001$) and Citizen Scientists ($r=.36$, $p < 0.001$) (Figure S1D). On the subject-level, we found a significant association between mean directed network connectivity and baseline depression in Citizen Scientists ($r=.19$, $p=.01$), but not Paid Students ($r=.03$, $p=.73$) (Figure S1E). Extending these results to Week 8 depression, there was evidence of an overall positive association in both Paid Students (Median $r = 0.14$, IQR: 0.11, 0.18; 35.69% p -values below 0.05) and Citizen Scientists (Median $r = 0.14$, IQR: 0.09, 0.19; 51.05% p -values below 0.05) (Figure S1F). On the subject-level, however, Week 8 depression was only significantly associated with connectivity in Citizen Scientists ($r(189) = 0.25$, $p < 0.001$), but did not reach significance for Paid Students ($r = 0.15$, $p = 0.06$) (Figure S2B).

7.2.4.1 Directed Network Connectivity and Depression in Experiment 2.

The exemplar directed network was also very similar to the Neureka samples, with correlations of network edge strengths of $r = .87$ and $r = .89$ between HNATD and Paid Students and Citizen Scientists, respectively. Correlation stability was higher than Neureka samples at 0.5, likely due to the fact that EMA sampling was more frequent in this study (i.e., 6-hour vs 12-hour lags) (Figure S3B). Although slightly underpowered to examine the link between directed networks and baseline depression, we nonetheless estimated and examined these associations in an exploratory capacity. The median correlation was nominally positive, but very close to zero (Median $r = 0.03$). Just 3.25% of HNATD directed networks were significantly associated with baseline depression (Figure S4A), and 8.23 % were significantly associated with SD gloomy (Figure S4C). Mean directed connectivity was not significantly correlated with baseline depression ($r = .03$, $p = .57$) (Figure S4B) nor SD gloomy ($r = 0.07$, $p = 0.13$) (Figure S4D).

Dataset	Network	Model	Beta (SE)	p-value
Paid Students	Contemporaneous	Unadjusted	0.15 (0.08)	0.06
		Adjusted	0.01 (0.10)	0.92
	Directed	Unadjusted	0.19 (0.08)	0.02
		Adjusted	0.19 (0.08)	0.02
Citizen Scientists	Contemporaneous	Unadjusted	0.37 (0.07)	< 0.001
		Adjusted	0.31 (0.08)	< 0.001
	Directed	Unadjusted	0.23 (0.07)	0.002
		Adjusted	0.22 (0.07)	0.004
HNATD	Contemporaneous	Unadjusted	0.27 (0.04)	< 0.001
		Adjusted	0.27 (0.04)	< 0.001
	Directed	Unadjusted	0.04 (0.04)	0.32
		Adjusted	0.04 (0.04)	0.34
Clinical Sample	Contemporaneous	Unadjusted	-0.04 (0.15)	0.77
		Adjusted	-0.04 (0.16)	0.81
	Directed	Unadjusted	0.10 (0.15)	0.50
		Adjusted	0.10 (0.15)	0.53

Table S3: Association between network connectivity in the exemplar 5-node network with SD depression (Paid Students & Citizen Scientists), SD gloomy (HNATD), and SD down or depressed (Clinical Sample).

In the unadjusted model, symptom variability ~ network connectivity, while the adjusted model individual network stability is added as a covariate, i.e., symptom variability ~ network connectivity + network stability. Overall, adjusting for network stability had minimal effect on the relationship between network connectivity and symptom variability.

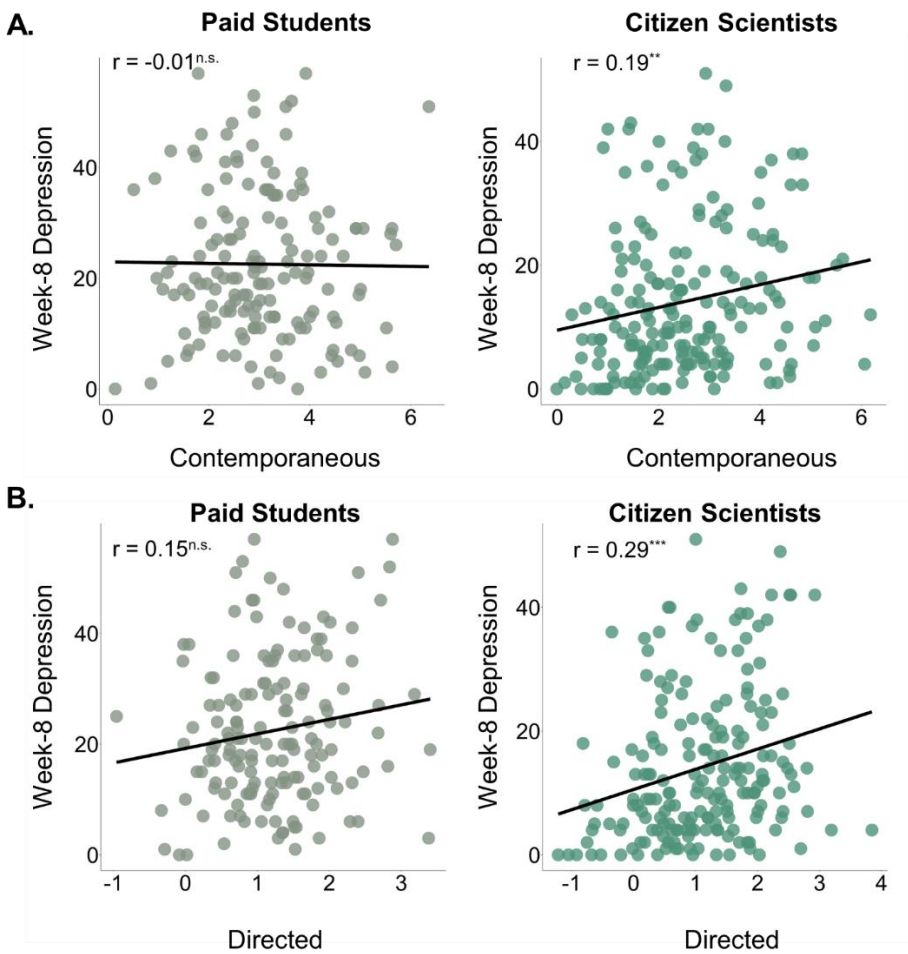


Figure S1: Association of per-participant network connectivity and Week-8 depression.

A) Contemporaneous network connectivity. Paid Students did not show a significant association ($r = -0.01$, $p = 0.89$) with Week-8 depression, but there was a significant association ($r = 0.19$, $p = 0.007$) for Citizen Scientists. B) Directed

network connectivity. Citizen Scientist network connectivity was significantly associated with Week-8 depression ($r = 0.25$, $p < 0.001$), with no significant association for Paid Students ($r = 0.15$, $p = 0.06$).

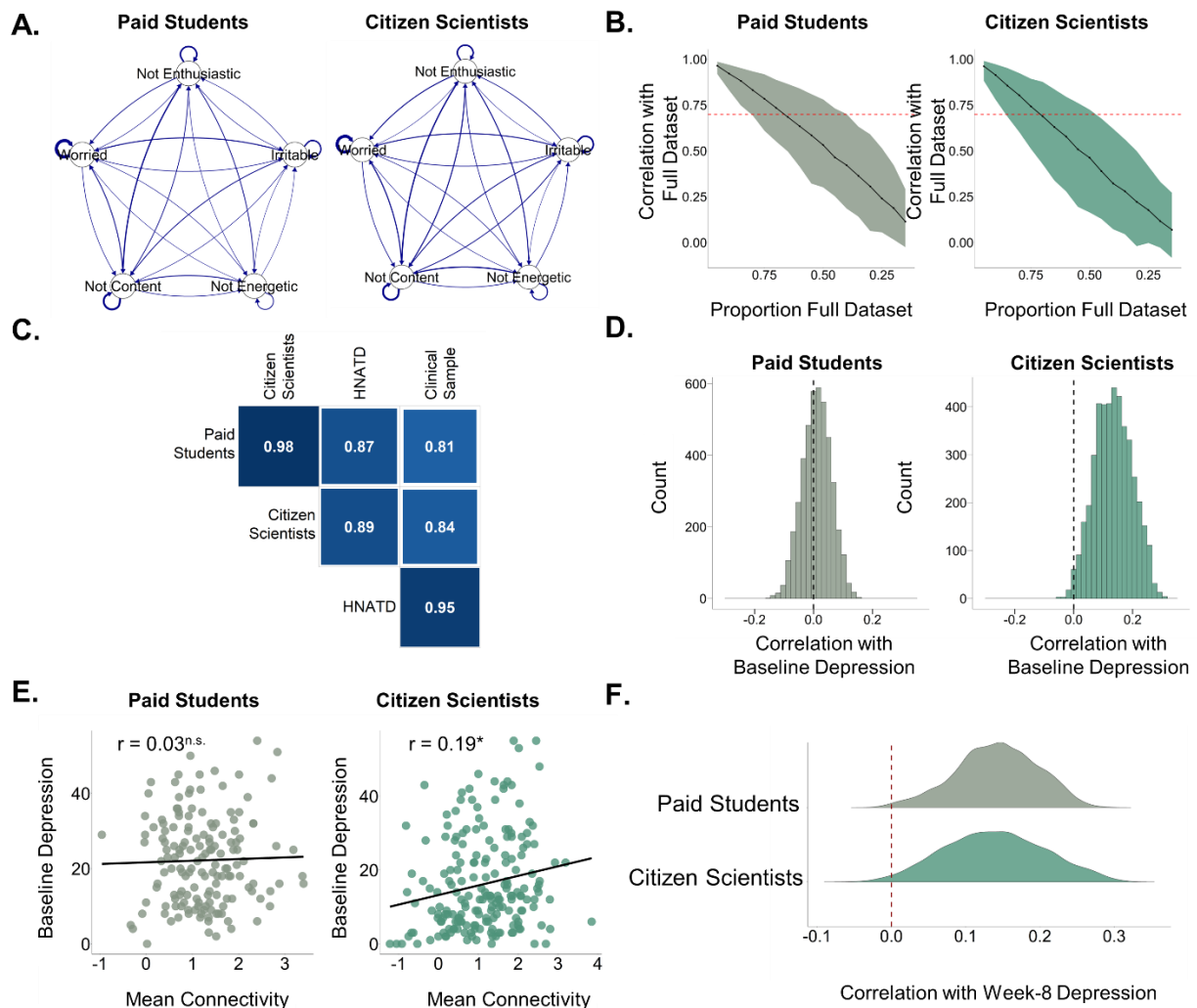


Figure S2. Directed Network Connectivity in Paid Students and Citizen Scientists.

A) Structure of the ‘exemplar’ network common to all datasets. Structure was highly consistent across samples (edges correlated at $r=.98$). B) Networks had low stability in both samples, with correlation stability values of 0.35 and 0.3 for Paid Students and Citizen Scientists respectively (where 0.5 is recommended). C) Correlation among edge strengths in the exemplar directed network across all 4 independent samples. D) Histograms of association between contemporaneous network connectivity and baseline depression for all possible combinations of 5-

node networks in Paid students and Citizen Scientists. Citizen Scientists, but not Paid Students, showed a positive association between network connectivity and depression. E) Association of per-participant mean network connectivity scores and baseline depression. Citizen scientists ($r=.19$), but not Paid Students ($r=.03$) showed a significant association with baseline depression. F) Histograms of association between directed network connectivity and Week 8 Depression.

Dataset	Network	Model	Beta (SE)	p-value
Paid Students	Contemporaneous	Unadjusted	0.29	<0.001
		Adjusted	0.17	0.05
	Directed	Unadjusted	0.21	0.008
		Adjusted	0.16	0.08
Citizen Scientists	Contemporaneous	Unadjusted	0.40	<0.001
		Adjusted	0.21	0.007
	Directed	Unadjusted	0.31	<0.001
		Adjusted	0.12	0.15
HNATD	Contemporaneous	Unadjusted	0.27	<0.001
		Adjusted	0.16	0.003
	Directed	Unadjusted	0.07	0.13
		Adjusted	0.12	0.04

Table S4: Association between depression variability and network connectivity either unadjusted or adjusted for mean EMA item variance.

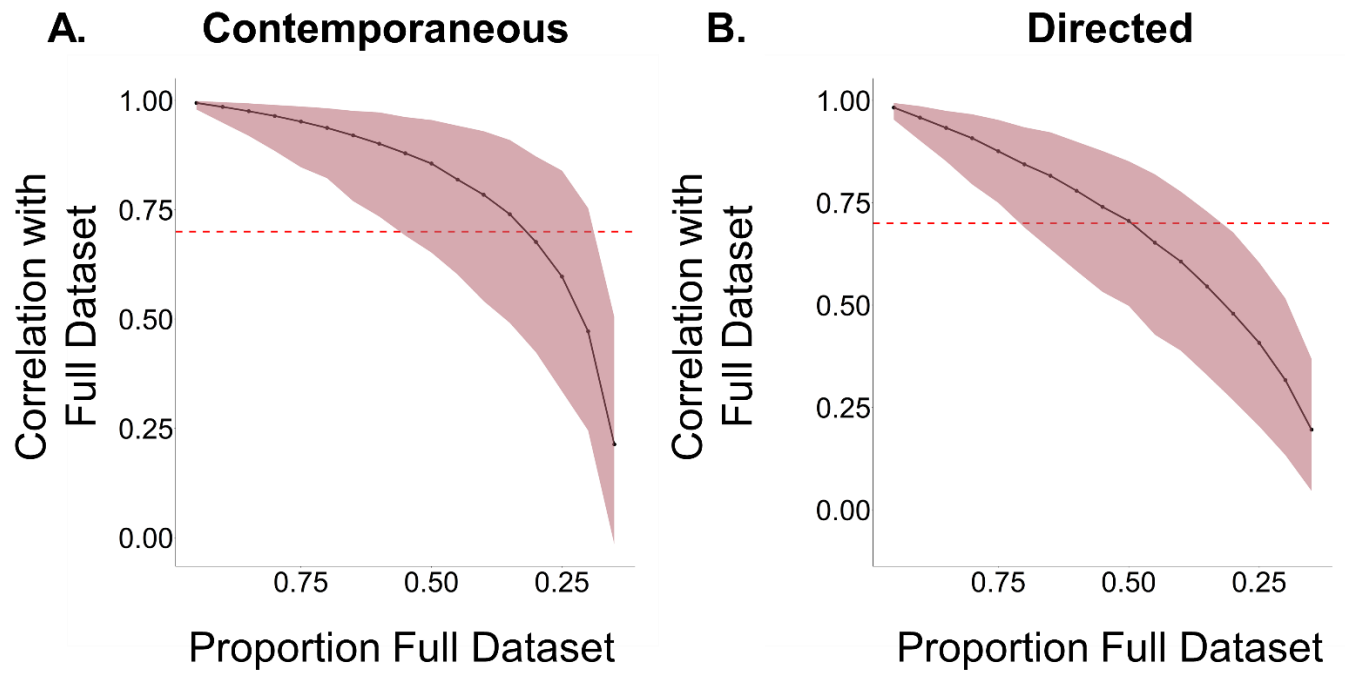


Figure S3: Exemplar network stability in a large independent sample (HNATD).

A) Contemporaneous exemplar network stability. The contemporaneous exemplar network had a correlation stability of 0.70. B) Directed exemplar network stability. The directed exemplar network had a correlation stability of 0.50.

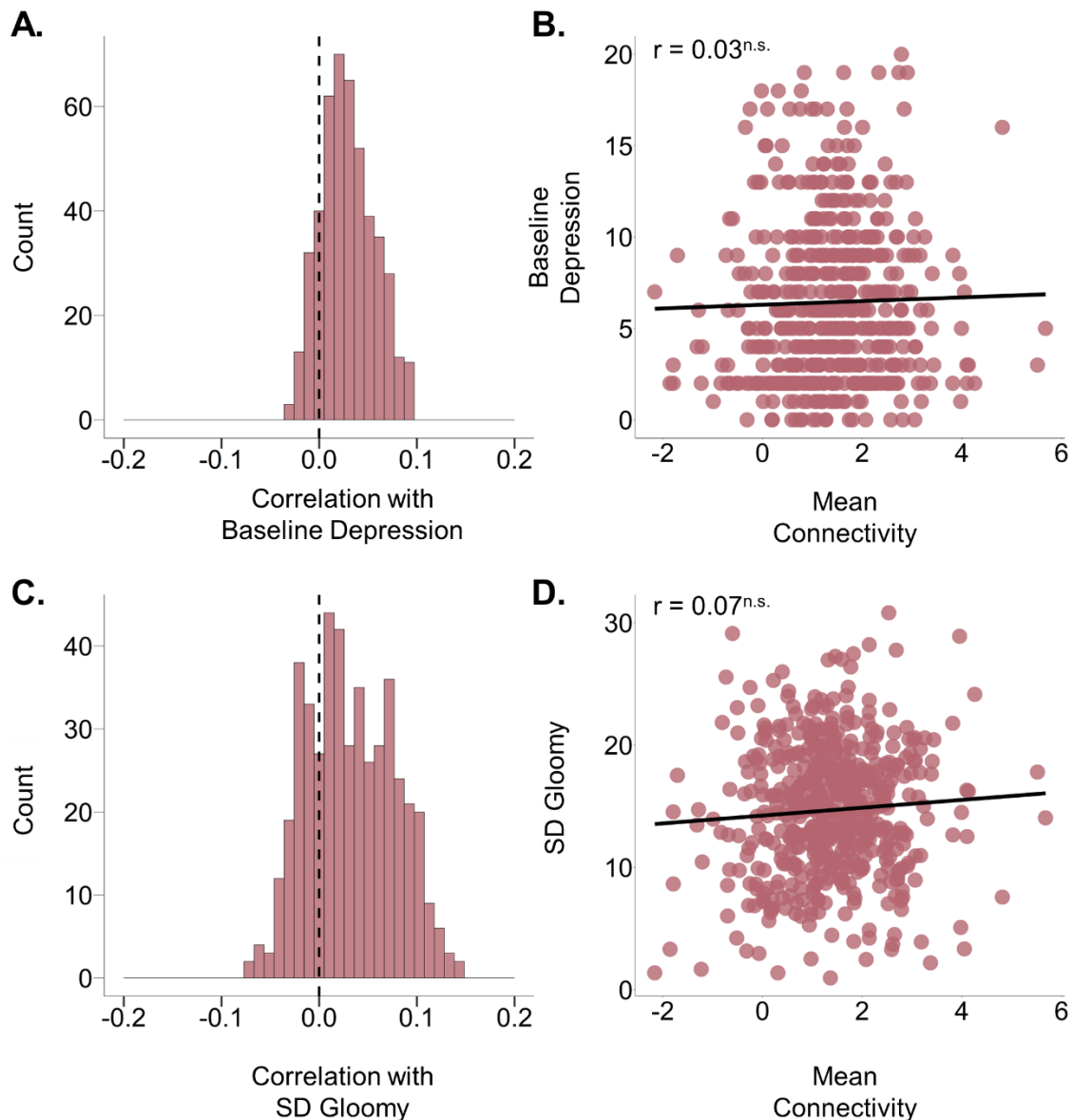


Figure S4: The association between directed network connectivity with baseline depression in a large and independent sample (HNATD).

A) Histograms of association between directed network connectivity and baseline depression for all possible combinations of 5-node networks (n = 462) in HNATD participants (proportion of p-values < 0.05: 3.25%). B) Association of per-participant mean directed network connectivity scores and baseline depression ($r=.03$, $p=.57$). C) Histograms of association between directed network connectivity and SD gloomy for all possible network combinations (proportion of p-values < 0.05: 8.23%). D) Association of per-participant mean directed network connectivity scores and baseline depression ($r=.07$, $p=.13$).

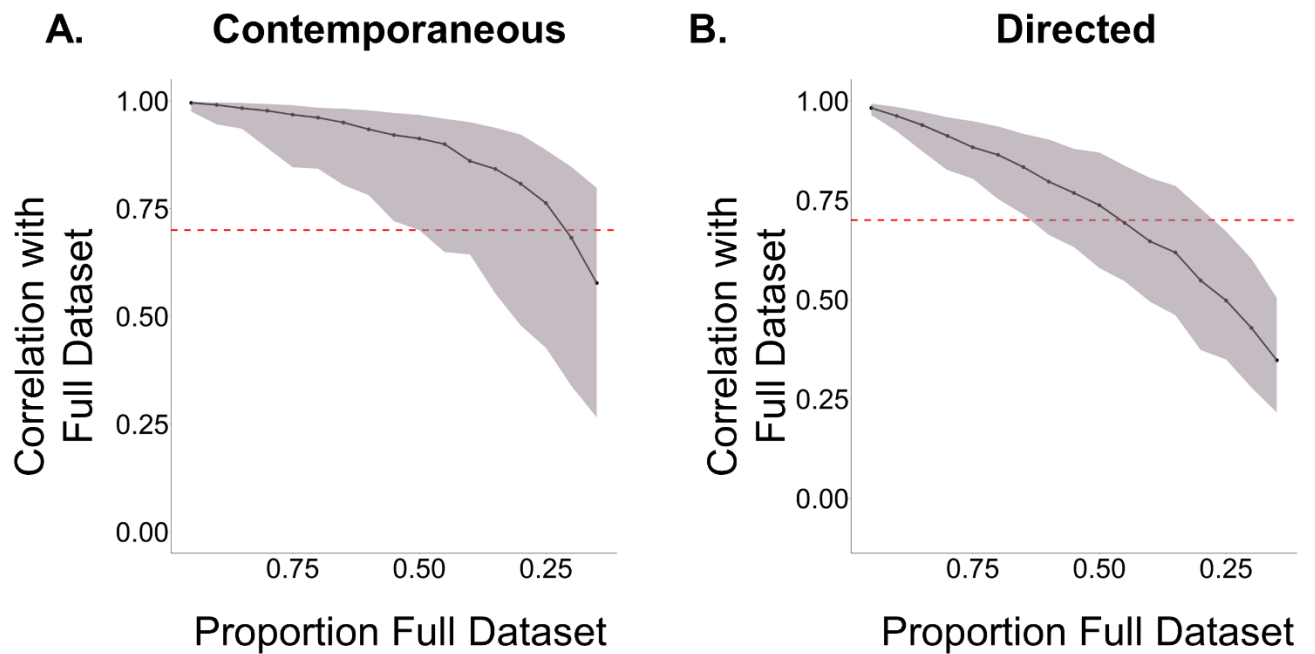


Figure S5: Exemplar network stability in a clinical sample

A) Contemporaneous exemplar network stability. The contemporaneous exemplar network had a correlation stability of 0.80. B) Directed exemplar network stability. The directed exemplar network had a correlation stability of 0.55.

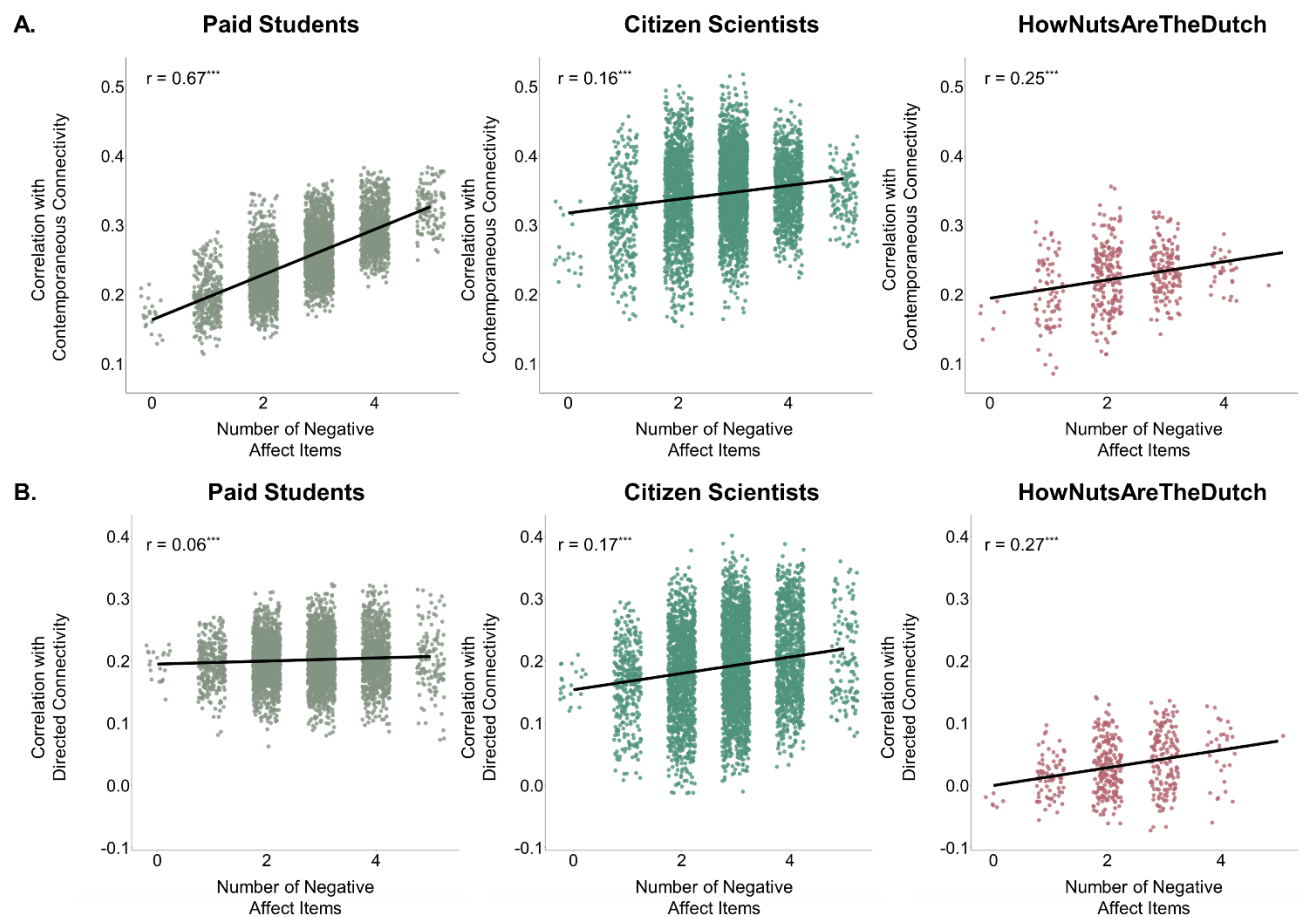


Figure S6: Impact of node valence on relationship between network connectivity and depression variability

Association between number of negative affect items and the correlation of depression variability and A) contemporaneous connectivity or B) directed connectivity. For Paid Students and Citizen Scientists, depression variability was operationalised as the SD of weekly depression scores. While in the HowNutsAreTheDutch sample, we used the SD of the 'I feel gloomy' item.

7.3 Appendix III: Supplemental Information for Chapter 4

7.3.1 Characteristics of participants recruited through paid vs free channels

	Paid N = 680	Free N = 266	p-value
Twitter Behaviour			
Tweets	294.8 (537.5)	512.6 (656.4)	1.9e-07***
Retweets	230.5 (473.8)	392.5 (525.7)	5.3e-06***
Likes	1033.9 (1090.3)	1260.9 (971.6)	0.003
Word count per day	117.2 (121.8)	158.1 (139.0)	1.2e-05***
Age (years)	29.6 (10.0)	29.5 (12.1)	0.92
Gender			
Male	216 (31.8%)	88 (33.1%)	0.77 ^a
Female	443 (65.1%)	174 (65.4%)	-
Transgender Male	5 (0.7%)	1 (0.4%)	-
Transgender Female	1 (0.1%)	0 (0%)	-
Non-Binary	13 (1.9%)	2 (0.8%)	-
Other	2 (0.3%)	1 (0.4%)	-
Country			
Ireland	19 (2.8%)	25 (9.4%)	3.2e-17 ^{a***}
United Kingdom	281(41.3%)	58 (21.8%)	-
United States	334 (49.1%)	146 (54.9%)	-
Canada	32 (4.7%)	10 (3.8%)	-
Australia	11 (1.6%)	4 (1.5%)	-
Other	3 (0.4%)	23 (8.6%)	-
Education			
Less than high school	11 (1.6%)	5 (1.9%)	5.5e-06 ^{a***}
High school	154 (22.6%)	36 (13.5%)	-
Some university	210 (30.9%)	107 (40.2%)	-
Bachelor's degree	221 (32.5%)	64 (24.1%)	-
Master's degree	70 (10.3%)	34 (12.8%)	-
Professional degree	6 (0.9%)	9 (3.4%)	-
Doctorate	8 (1.2%)	11 (4.1%)	-
Currently Employed (% Yes)	436 (64.1%)	22 (8.3%)	2.2e-53 ^{a***}
Physician diagnosed depression (% Yes)	292 (42.9%)	140 (52.6%)	0.01 ^{a*}

Table S1. Demographics and Twitter use characteristics of subjects recruited through paid (ClickWorker) vs free channels

Twitter and demographic characteristics of all participants along with differences between participants with and without a depressive episode. ^aChi-square test.

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

7.3.2 Reliability checks on a 9-node network

We conducted a split half reliability test to determine if the 36 unique edges in our network were reliable. We did this across subjects by splitting our sample into two equal halves, calculated personalised networks for all participants, and then correlated the mean edge strengths between the two halves. We found the estimates of edge strength to be highly reliable $r(36) = .99$, $p < .001$. The edge strength between Neg. Emo. and swear was much stronger than between other edges, reliability was $r(35) = 0.97$, $p < 0.001$ when we exclude this edge. Because individual networks tended to be sparse, the average of most edges tended

towards zero leading to a high correlation between split halves.

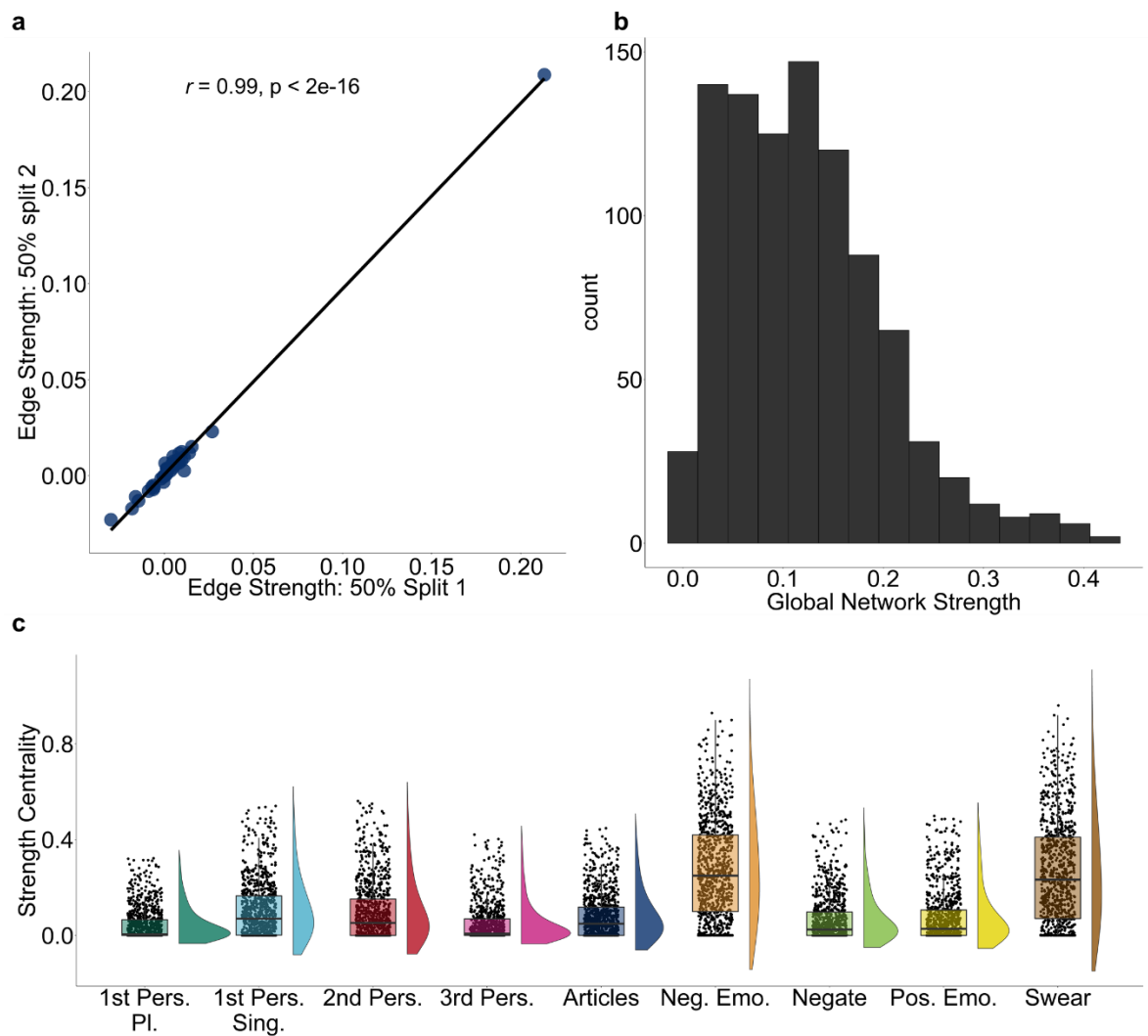


Figure S1. Split half reliability of primary 9-node network.

a): The edge strength among 36 unique connections between nodes in split halves ($n = 473$ for each half) of the sample were highly correlated with each other ($r(36) = 0.99$, $p < 2e-16$), indicating that the estimated edge strengths between nodes are highly reliable. b): Histogram of global network strength from personalised networks of all participants ($n = 946$). c) Variability of node strength centrality in the 9-node network ($n = 946$). Boxplots depict the median (centre line), upper and lower quartiles, i.e., interquartile range, whiskers, 1.5X interquartile range, and minimum and maximum values. ^aUnadjusted two-sided Pearson correlation.

7.3.3 Sensitivity of our analysis to the removal of ‘Supra’ categories in LIWC

In the LIWC library, certain categories are inclusive of multiple sub-categories. For example, within the personal pronoun category are: 1st person singular/plural, 2nd person, 3rd person, and impersonal pronouns. An increase in 1st person singular pronouns will then necessarily lead to an increase in the proportion of pronouns overall. To control for this potential, confound, the following supra-categories were removed from inclusion in the network: function words, impersonal pronouns, pronouns, affect, anxiety, anger, sad, social, cognitive processes, percept, biological, drives, relative, and informal. We then randomly selected 200 sets of 9 text features that were significantly (‘Depression Relevant’ x 100) or not-significantly (‘Depression Irrelevant’ x 100) associated with current depression severity. We then compared the change in within-episode connectivity between depression relevant and irrelevant networks. There was a significant increase in within-episode connectivity ($\beta = 0.01$, $SE = 0.001$, $p < 0.001$) even after removing the supra-categories within the LIWC. Thus, the increase in within-episode connectivity for depression relevant networks is not dependent on the inclusion of LIWC supra-categories.

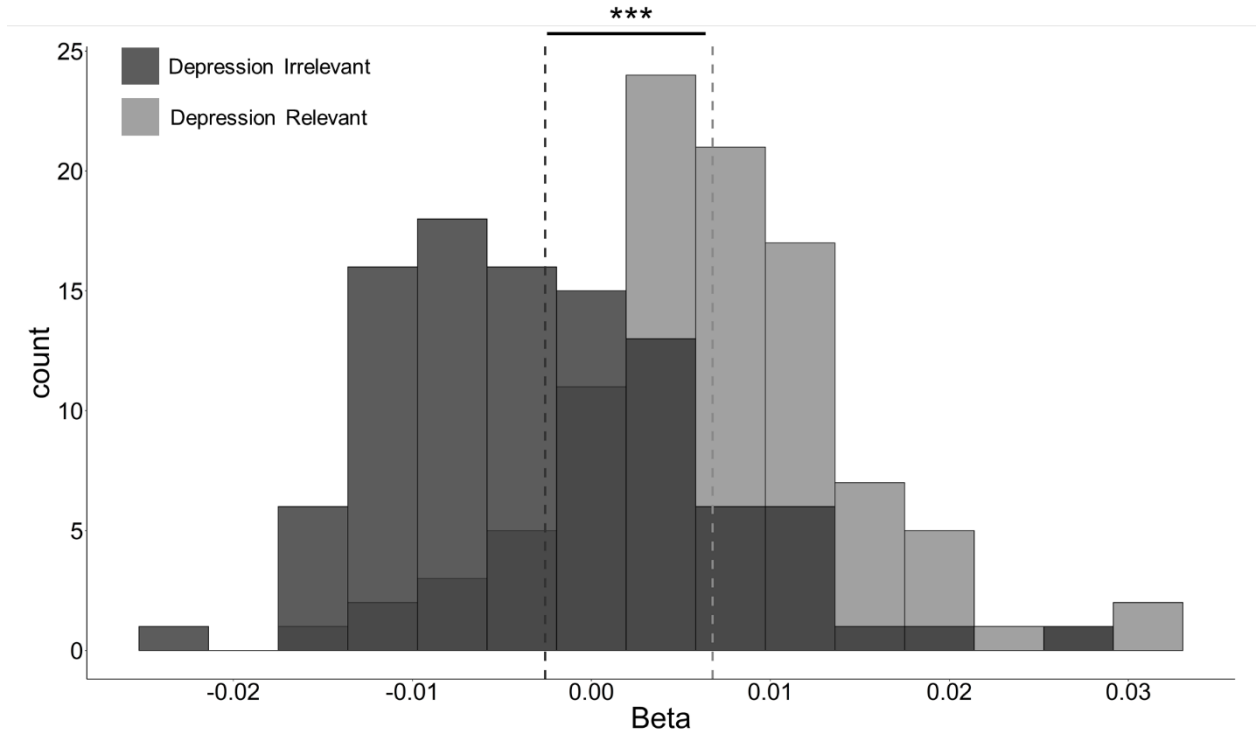


Figure S2: Random networks of 9 text features either significantly ('Depression Relevant') or not significantly ('Depression Irrelevant') associated with current depression excluding LIWC supra-categories.

There was a significant increase in within-episode network connectivity in depression relevant vs. depression irrelevant networks ($\beta = 0.01$, $SE = 0.001$, $p = 2.4e-13$). Results are from a general linear model with two-sided p-values without adjustment for multiple comparisons. Source data are provided as a Source Data file.

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

7.3.4 The effect of number of days on global network connectivity and associated analyses

We were concerned that between and within-subject differences in the number of days upon which we based our 9-node a priori network might affect the connectivity results. Indeed, we observed that subjects with more days had on average, more connected networks, $r=.19$, $p<.001$ (Figure S1a). Next, we checked if this might systemically bias our between-subject findings by testing for an association with depression symptom severity and the number of days with tweets. We found no association (Figure S1b). However, we found a significant within subject effect such that periods of time when subjects were in a depressed episode were significantly shorter than non-depressed periods of time (Figure S1c), $\beta = -90.8$, $SE = 6.2$, $p < 0.001$. We examined the network connectivity of within and between episode time periods (Figure S1d) and found a significant interaction ($\beta = 0.0005$, $SE = 0.0001$, $p < 0.001$) between the number of days and episode (within vs. outside).

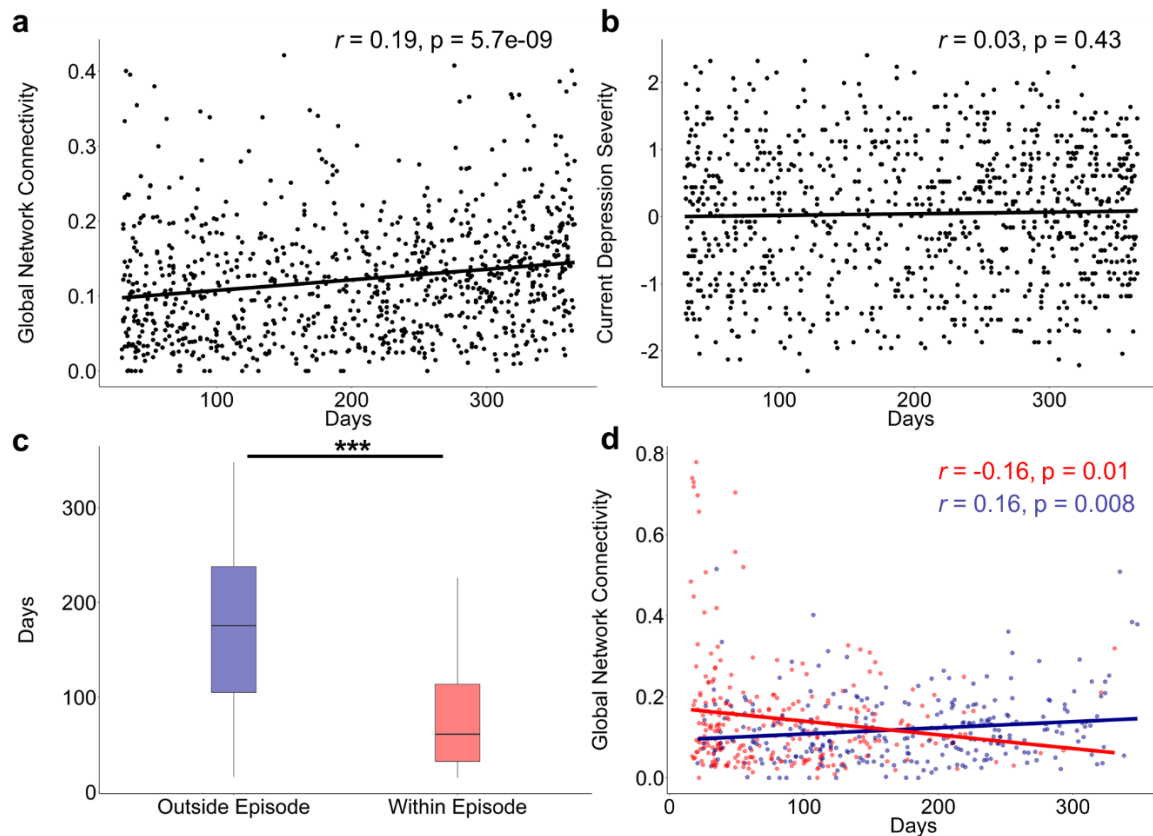


Figure S3. The effect of days on global network connectivity.

a) Among all participants ($N = 946$), there was a significant positive association between global network connectivity and number of days ($r(936) = 0.19$, $p = 5.7e-09$) in the full sample b). There was no significant association between current depression severity and number of days ($r(944) = 0.03$, $p = 0.43$). c) Within-episode periods have significantly fewer days on average than outside episode periods ($\beta = -90.8$, $SE = 6.2$, $p < 2e-16$). d) Association between number of days and within vs. outside episode network connectivity. There is significant interaction effect between the number of days and whether data is from a depressed vs non-depressed episode ($\beta = 0.0005$, $SE = 0.0001$, $p < 0.001$). For non-depressed periods of time (blue), the more days there were, the greater connectivity the connectivity, but for depressed episodes, the relationship was reversed. The boxplot depicts the median (centre line), upper and lower quartiles, i.e., interquartile range, and whiskers, i.e., 1.5x interquartile range. ^{a,b,d}Unadjusted two-sided Pearson correlation, ^{c,d}Two-sided general linear regression model.

As there was a credible confound from differences in days within subject, to control for the effect of the number of days on connectivity estimates, we repeated our analysis including days as a covariate (Table S2). This did not alter the significant positive association between current depression severity and negative emotions, swear words, articles, and global network connectivity in the full sample of participants (N = 946).

7.3.5 Frequency of occurrences of the 9 a priori text features in the 12-month time series

For each subject, we calculated the proportion of their days with tweets that contained each of the 9 text features. Swear was the least frequent, appearing on average in 30% of days with tweets, but with some individuals never swearing and others swearing on 100% of days with tweets. Articles were the most frequent, appearing on average in 80% of days with tweets, with the lower bound being 22% and upper 100%.

Text Feature	Proportion of Days with Non-Zero Text Features (Mean, SD) Range
1 st Person plural	0.38 (0.21) [0.02,1]
1 st Person singular pronouns	0.69 (0.20) [0.08,1]
2 nd Person singular pronouns	0.57 (0.20) [0.04,1]
3 rd Person pronouns	0.50 (0.22) [0.02,1]
Articles	0.80 (0.14) [0.22,1]
Negate	0.55 (0.22) [0.03,1]
Negative emotions	0.60 (0.22) [0.07,1]
Positive emotions	0.75 (0.16) [0.24,1]
Swear	0.30 (0.22) [0.00,1]

Table S2. Proportion of days with non-zero values for each of the 9 a priori text features

LIWC Text Feature	Unadjusted β	SE	p-value	Adjusted β	SE	p-value
Global Network Connectivity	0.008	0.003	0.002**	0.007	0.003	0.004**
Articles	0.01	0.003	<0.001***	0.01	0.003	<0.001***
1 st person singular pronouns	0.007	0.004	0.07	0.006	0.004	0.09
Negation words	0.003	0.003	0.36	0.003	0.003	0.36
Negative emotions	0.002	0.007	0.007**	0.02	0.007	0.009**
Positive emotions	0.006	0.004	0.12	0.006	0.004	0.12
3 rd person pronouns	0.005	0.003	0.08	0.005	0.003	0.08
Swear words	0.02	0.007	0.01*	0.02	0.007	0.01*
1 st person plural pronouns	0.0005	0.002	0.82	0.0004	0.002	0.85
2 nd person singular pronouns	0.006	0.004	0.12	0.006	0.004	0.13

Table S3: Controlling for days with tweets: adjusted vs unadjusted analyses for the association between depression severity and network connectivity

The association between network connectivity and depression severity is not affected by number of days ($n = 286$). Unadjusted within-subject linear model.

7.3.6 The association between network connectivity and depression severity is not affected by removal of 3rd person pronouns

Unlike the 8 other nodes selected a priori, there was no cross-sectional association with current depression severity and 3rd person pronouns. We therefore conducted a sensitivity analysis to test for the effect of removing 3rd person pronouns, she/he and they, on network structure. There was a significant positive association between overall network connectivity and current depression severity ($\beta = 0.008$, $SE = 0.003$, $p = 0.003$). Along with a significant increase in the node strength of articles (Article, $\beta = 0.01$, $SE = 0.003$, $p < 0.001$), swear words (Swear, $\beta = 0.02$, $SE = 0.007$, $p = 0.02$), and negative emotions (Neg. Emo., $\beta = 0.02$, $SE = 0.007$, $p = 0.02$). We found no significant changes to associations between network structure and depression severity caused by removing 3rd person pronouns.

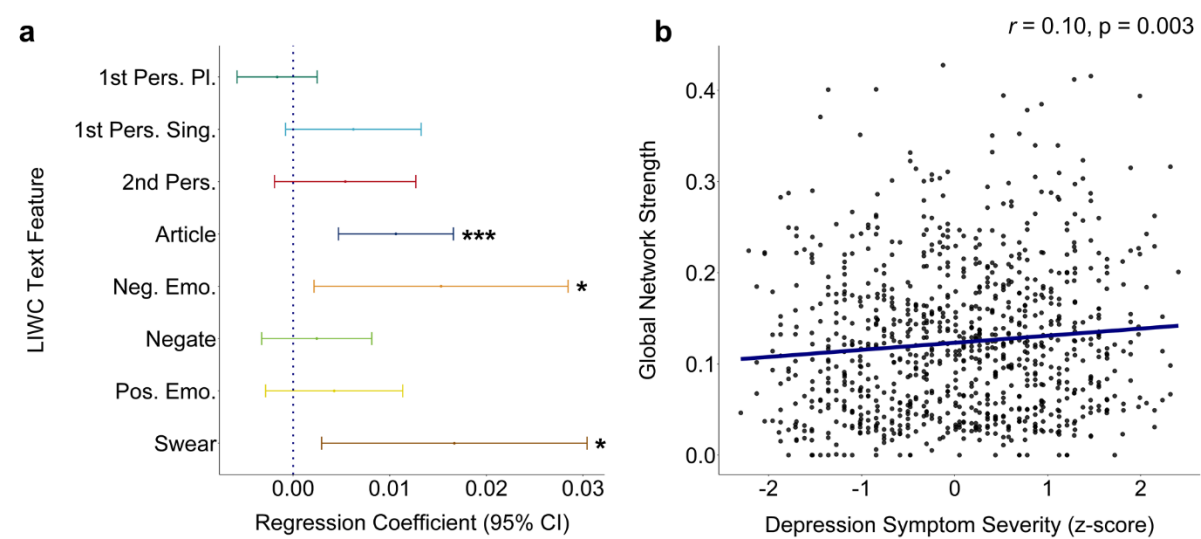


Figure S4. Tolerance of within subject analysis to the removal of 3rd person pronouns

Sensitivity analysis to test for the effect of removing 3rd person pronouns, she/he and they, on network structure due to lack of association with current depression severity. a) The association between individual network node strength and depression symptom severity with significant associations for the node strength of articles (Article, $\beta = 0.01$, SE = 0.003, $p = 0.0005$), swear words (Swear, $\beta = 0.02$, SE = 0.007, $p = 0.02$), and negative emotions (*Neg. Emo.*, $\beta = 0.02$, SE = 0.007, $p = 0.02$) ($n = 286$). b) The association between global network connectivity for personalised networks of all participants ($N = 946$) and depression symptom severity. ^aUnadjusted within-subject linear regression, ^bTwo-sided Pearson correlation unadjusted for multiple comparisons. Source data are provided as a Source Data file.

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

7.3.7 Bootstrapped control for unequal variances in within-episode network connectivity

To control for the unequal variances in the distribution of within vs. outside episode network connectivity, we subsampled 80% of global network connectivity from personalised networks of participants within and outside a depressive episode and re-ran the within-subject regression 1,000 times. We found that after bootstrapping the within-episode regression coefficient that the change in within-episode connectivity was still significant ($\beta = 0.03$, SE = 0.0001, $p < 0.001$).

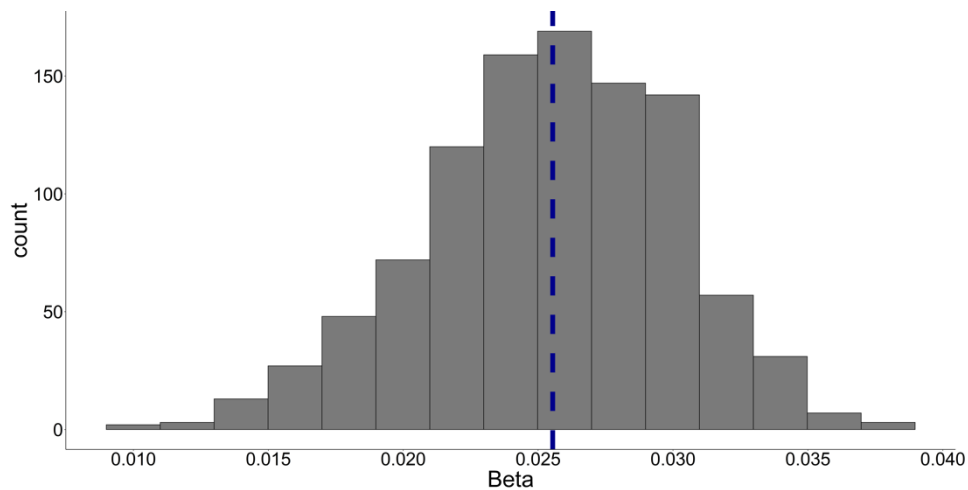


Figure S5: Bootstrapped regression coefficient of change in network connectivity within a depressive episode from 80% random sub-samples of the data repeated 1,000 times

7.3.8 Permutation test of within-episode identifier

We conducted a permutation test, randomising the indicator (within/outside episode) per subject 1000 times and comparing network strength within subject using this indicator. This amounts to comparing network connectivity for a random period of time (that has the same number of days as the real within episode period) to another random period of time (that has the same number of days as the real outside episode period). Networks were constructed using the same 9 LIWC text features as used in the main analysis and positive betas indicate that the fake ‘within episode’ network had more connectivity than the fake ‘outside episode’ one. These figures show that (A) 99.3% of betas were smaller than the true within/outside episode value in the unshuffled data, and (B) 11% of p-values were below 0.05. This indicates the discrepancy in number of days had an influence on global network connectivity differences, but this does not explain the findings within and outside episode.

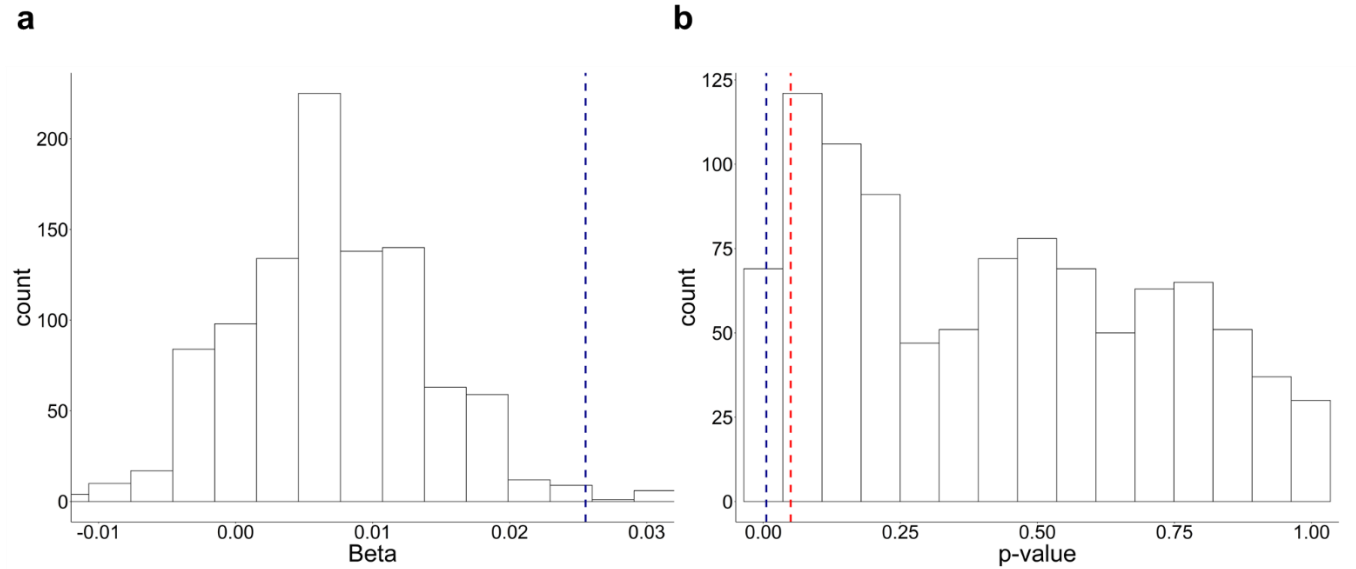


Figure S6. Permutation test randomizing within versus outside episode indicators.

Blue dashed lines indicate the observed beta and observed p-value for our a priori depression network, the red dashed line indicates the alpha level of 0.05. These figures show that (a) 99.3% of betas were smaller than the true within/outside episode value in the unshuffled data, and (b) 11% of p-values were below 0.05.

^aNumber of regression coefficients from 1,000 within-subject linear models that the observed beta value is below, unadjusted two-sided p-value, ^bNumber of p-values from the same 1,000 within-subject linear models that the observed p-value is below.

7.3.9 No mean increase in the use of 9 *a priori* text features outside vs within a depressive episode

In contrast to our principal findings that network connectivity increases when subjects are in a depressive episode, we found no significant change in any LIWC text feature within a depressive episode ($p > 0.05$).

LIWC Text Feature	β	SE	p-value
1 st person plural pronouns	-0.02	0.01	0.12
1 st person singular pronouns	-0.05	0.04	0.31
2 nd person singular pronouns	-0.03	0.03	0.32
3 rd person pronouns	0.01	0.01	0.38
Articles	-0.03	0.04	0.48
Negation words	-0.03	0.02	0.14
Negative emotions	-0.01	0.03	0.76
Positive emotions	-0.02	0.04	0.63
Swear words	0.01	0.01	0.69

Table S4. Results from regression analysis of the effect of episode (within, outside) on the mean use of the 9 *a priori* text features (N=286)

Mean difference in LIWC text features within a depressive episode compared to outside an episode from a within-subject regression (N = 286). Two-sided general linear regression unadjusted for multiple comparisons.

7.3.10 Association between all LIWC text features and current depression severity

LIWC Text Feature	<i>r</i>	95% CI	p-value
Conjunction	0.17	0.1, 0.23	< 0.001***
Negative emotions	0.17	0.11, 0.23	< 0.001***
Biological	0.17	0.1, 0.23	< 0.001***
Adverb	0.16	0.09, 0.22	< 0.001***
Sad	0.16	0.09, 0.22	< 0.001***
Analytic	-0.15	-0.21, -0.09	< 0.001***
Verb	0.15	0.09, 0.21	< 0.001***
Anxiety	0.15	0.08, 0.21	< 0.001***
Anger	0.15	0.08, 0.21	< 0.001***
Focus on present	0.15	0.09, 0.21	< 0.001***
Word Count	0.14	0.07, 0.2	< 0.001***
I (1 st person singular)	0.14	0.08, 0.20	< 0.001***
Sexual	0.14	0.07, 0.2	< 0.001***
Pronouns	0.13	0.07, 0.19	< 0.001***
Space	-0.13	-0.19, -0.07	< 0.001***
Words per sentence	0.12	0.05, 0.18	< 0.001***
Personal pronouns	0.12	0.06, 0.19	< 0.001***
Auxiliary verbs	0.12	0.05, 0.18	< 0.001***
Negation words	0.12	0.05, 0.18	< 0.001***
Cognitive processes	0.12	0.05, 0.18	< 0.001***
Differ	0.12	0.06, 0.18	< 0.001***
Achievement	-0.12	-0.19, -0.06	< 0.001***
Work	-0.12	-0.18, -0.06	< 0.001***
Informal	0.12	0.05, 0.18	< 0.001***
Function words	0.11	0.05, 0.17	0.001**
We (1 st person plural)	-0.11	-0.17, -0.05	0.001**
They (3 rd person singular)	0.11	0.05, 0.18	< 0.001***
Article	-0.11	-0.17, -0.05	0.001**
Discrepancy	0.11	0.05, 0.17	0.001**
Relative	-0.11	-0.17, -0.04	0.001**
Swear	0.11	0.05, 0.17	0.001**
Tone	-0.1	-0.17, -0.04	0.002**
Six letter words	-0.1	-0.16, -0.04	0.002**
Cause	0.1	0.04, 0.17	0.001**
Feel	0.1	0.04, 0.17	0.001**
Body	0.1	0.04, 0.16	0.002**
Health	0.1	0.04, 0.16	0.002**
Death	0.1	0.04, 0.17	0.001**
Clout	-0.09	-0.15, -0.02	0.008**
Drives	-0.09	-0.16, -0.03	0.004**
Reward	-0.09	-0.15, -0.03	0.006**
Dictionary words	0.08	0.01, 0.14	0.02*
You (2 nd person singular)	0.08	0.01, 0.14	0.02*
Number	-0.08	-0.14, -0.01	0.02
Time	-0.08	-0.14, -0.02	0.01*
Internet slang	0.08	0.01, 0.14	0.02*
Prepositions	-0.07	-0.13, 0	0.04*
Positive emotions	-0.07	-0.14, -0.01	0.03*
Female	0.07	0.01, 0.13	0.03*
Tentative	0.07	0, 0.13	0.04*
Power	-0.07	-0.13, -0.01	0.03*

3 rd person (she/he, they)	0.06	0, 0.13	0.06
Impersonal pronouns	0.06	0, 0.13	0.06
Insight	0.06	-0.01, 0.12	0.09
Ingest	0.06	-0.01, 0.12	0.08
Leisure	-0.06	-0.13, 0	0.06
Assent	0.06	0, 0.13	0.05
Affect	0.05	-0.02, 0.11	0.15
Friend	0.05	-0.01, 0.12	0.1
Percept	0.05	-0.01, 0.12	0.1
Authentic	0.04	-0.02, 0.10	0.21
Compare	-0.04	-0.11, 0.02	0.21
Interrogative	0.04	-0.02, 0.1	0.21
Quantifier	0.04	-0.02, 0.11	0.21
Affiliation	-0.04	-0.1, 0.02	0.23
Non-fluency words	0.04	-0.02, 0.11	0.19
Period	-0.04	-0.10, 0.03	0.27
Adjective	-0.03	-0.1, 0.03	0.31
Social	0.03	-0.03, 0.1	0.3
Risk	0.03	-0.04, 0.09	0.42
Focus on past	0.03	-0.03, 0.09	0.37
Focus on future	0.03	-0.03, 0.09	0.36
Motion	-0.03	-0.09, 0.04	0.37
Filler	0.03	-0.03, 0.09	0.35
All punctuation	-0.03	-0.09, 0.03	0.36
She/he (3 rd person singular)	0.02	-0.04, 0.09	0.45
Male	-0.02	-0.08, 0.05	0.58
Home	0.02	-0.05, 0.08	0.6
Exclamation point	0.02	-0.05, 0.08	0.6
Family	0.01	-0.05, 0.08	0.66
See	-0.01	-0.07, 0.06	0.86
Hear	0.01	-0.05, 0.08	0.69
Religion	-0.01	-0.07, 0.06	0.79
Question mark	0.01	-0.05, 0.08	0.66
Certain	0	-0.07, 0.06	0.95
Money	0	-0.06, 0.07	0.94
Apostrophe	0	-0.06, 0.07	0.94

Table S5: Bivariate correlations for all LIWC features and depression severity in full sample

Bivariate correlations between 87 LIWC text features and current depression severity with 95% confidence intervals (CIs). Approximately 59% of text features are correlated with depression severity at the $\alpha = 0.05$ level. Unadjusted two-sided Pearson correlation with 95% CIs. Source data are provided as a Source Data file.

