Doctoral Programme in Population Health
Department of Psychology and Logopedics,
Faculty of Medicine
University of Helsinki
Helsinki, Finland

DIGITAL INTERVENTIONS FOR DEPRESSION
PREDICTORS AND MODERATORS OF TREATMENT ADHERENCE AND OUTCOMES

Isaac Moshe

Doctoral dissertation, to be presented with the permission of the Faculty of Medicine of the University of Helsinki, for public examination in room 107, Athena building, on the 16th of December 2022, at 1 o’clock in the afternoon.

Helsinki 2022

Supervisors

Laura Pulkki-Råback, PhD
Docent, Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Helsinki, Finland

Giulio Jacucci, PhD
Professor, Department of Computer Science, University of Helsinki, Helsinki, Finland

Niklas Ravaja, PhD
Professor, Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Helsinki, Finland

Reviewers

Raimo Lappalainen, PhD
Professor, Department of Psychology, Faculty of Education and Psychology, University of Jyväskylä, Jyväskylä, Finland

Keegan Knittle, PhD
Assistant Professor, Faculty of Sport and Health Sciences, University of Jyväskylä, Jyväskylä, Finland

Opponent

Heleen Riper, PhD
Professor, Department of Clinical, Neuro and Developmental Psychology, Faculty of Behavioural and Movement Sciences, VU University Amsterdam, Amsterdam, The Netherlands
Visiting Honorable Professor, Department of Psychiatry, Faculty of Medicine, University of Turku, Turku, Finland

ISBN 978-951-51-8753-6 (paperback)
ISBN 978-951-51-8754-3 (PDF)

Unigrafia, Helsinki, 2022

May all beings have happiness and the causes of happiness.
May all beings be free from suffering and the causes of suffering.

From the Four Immeasurables, Buddhist Prayer

Table of Contents

Abstract

Background

Depression is a leading cause of disability worldwide. Although evidence-based treatments exist, less than one-in-five people in high-income countries and less than one-in-twenty-seven in low-income countries receive treatment, giving rise to a treatment gap in mental healthcare. Digital interventions have been proposed as a solution to address the treatment gap. As an increasing number of public and private healthcare providers adopt digital interventions to meet the growing demand for treatment, the current thesis set out to examine the latest evidence-base for digital depression interventions and the extent to which new technologies may be used to identify at-risk individuals.

Methods

Study 1 assessed the efficacy of digital interventions for the treatment of depressive symptoms based on the largest meta-analysis of digital depression interventions conducted to-date. Databases were searched for RCTs of a computer-, internet-, or smartphone-based intervention for depression versus an active or passive control condition. Participants were individuals with elevated symptoms of depression at baseline. Using a random-effects multilevel meta-regression model, we examined the effect size of treatment versus control (Hedges' *g*) and explored moderators of treatment outcome. Study II was a secondary analysis of data from two RCTs (N=253) of a digital intervention for the prevention and treatment of major depression. Using logistic regression, we first examined participant characteristics as potential predictors of intervention dropout. We then assessed to what extent dropout could be predicted following completion of the first module using a combination of participant characteristics and intervention usage data. Dropout was defined as completing less than six modules. Study III was an observational study of $N = 60$ adults (ages 24–68) who owned an Apple iPhone and Oura Ring. A smartphone app (*Delphi*) continuously monitored participants' location and smartphone usage behavior over a 4-week period. The Oura Ring provided measures of activity, sleep and heart rate variability (HRV). Participants were prompted to report their daily mood and self-reported measures of depression and anxiety were collected at baseline, midpoint and the end of the study using the DASS-21. Multilevel regression models were used to predict the association between smartphone and wearable data and mental health scores. Study IV was a secondary analysis of data from Study III in which we compared the accuracy of five supervised machine learning algorithms in the classification of individuals with normal versus above normal symptoms of depression, as defined by the DASS-21 cut-off scores.

Results

A systematic search of the literature in Study I identified 83 trials ($N = 15{,}530$). The overall effect size of digital interventions versus all controls was $g = .52$. Significantly lower effect sizes were found in studies conducted in real-world settings (effectiveness trials; *g* = .30) versus laboratory settings (efficacy trials; *g* = .59). Significantly higher effect sizes were found in interventions that involved human therapeutic guidance (*g* = .63) compared with unguided, self-help interventions (*g* = .34). Additionally, we found significant differences in effect size depending on the type of control used (WLC: *g* = .70; attention: *g* = .36; TAU: *g* = .31). No significant difference in outcomes was found between human-guided digital interventions and face-to-face therapy, although the number of studies was low. In Study II we found that lower level of education (OR=3.33) and both lower and higher age (a quadratic effect; age: OR=0.62, age^2: OR=1.55) were significantly associated with higher risk of dropout. In the analysis that aimed to predict dropout following completion of the first module, lower and higher age (age: OR=0.61, age^2: OR=1.58), medium versus high social support (OR=3.40) and a higher number of days to module completion (OR=1.05) predicted higher risk of dropout, whilst a self-reported negative event in the previous week was associated with lower risk of dropout (OR=0.22). In Study III, we found a significant negative association between the variability of locations visited and symptoms of depression (β = −0.21, *p* = 0.037) and significant positive associations between total sleep time and depression (β = 0.24, *p* = 0.023) and time in bed and depression (β = 0.26, *p* = 0.020). Additionally, we found that wake after sleep onset significantly predicted symptoms of anxiety (β = 0.23, *p* = 0.035). Study IV revealed that a Support Vector Machine using only sensor-based predictors had an accuracy of 75.90% and an Area Under the Curve (AUC) of 74.89%, whilst an XGBoost model that combined mood and sensor data as predictors classified participants as belonging to the group with normal or above normal levels of depressive symptoms with an accuracy of 81.43% and an AUC of 82.31%.

Conclusion

The current thesis provided evidence of the efficacy of digital interventions for the treatment of depression in a variety of populations. Importantly, we provided the first meta-analytic evidence that digital interventions are effective in routine healthcare settings, but only when accompanied by human guidance. Notwithstanding, adherence to digital interventions remains a major challenge with little more than 25% of patients completing the full intervention on average in real-world settings. Finally, we demonstrated that data from smartphone and wearable devices may provide valuable sources of data in predicting symptoms of depression, thereby helping to identify at-risk individuals.

Tiivistelmä

Tausta

Masennus on maailmanlaajuisesti keskeisimpiä toimintakykyä alentavia tekijöitä. Vaikka masennuksen hoitoon on kehitetty näyttöön perustuvia hoitomuotoja, hoidon tarjonta ei kohtaa kysyntää: korkean tulotason maissa vain viidennes hoitoa tarvitsevista saa hoitoa, ja matalan tulotason maissa hoitoa saavien osuus on vielä selkeästi alhaisempi. Hoidon saatavuusongelman ratkaisuksi on ehdotettu digitaalisia hoitomuotoja, ja digitaalisten masennushoitojen käyttö yleistyykin sekä julkisissa että yksityisissä hoitokonteksteissa. Tässä tutkimuksessa selvitettiin digitaalisten masennushoitojen tehokkuutta ja teknologiasovellusten käyttöä masennuksen riskiryhmien varhaisen tunnistamisen välineenä.

Menetelmät

Ensimmäinen osatutkimus tarkasteli masennusoireiden hoidossa käytettävien digitaalisten hoitojen tehokkuutta. Tutkimuksessa toteutettiin tähän mennessä kattavin meta-analyysi satunnaistettuihin koeasetelmiin perustuvista masennusinterventiotutkimuksista, joissa hoitomuotona oli digitaalinen ohjelma ja aktiivinen tai passiivinen kontrollitilanne. Digitaaliset hoidot olivat internetissä tai muulla digitaalisella alustalla toteutettuja hoitoja (esimerkiksi tietokone- tai älypuhelinperustaisia hoitoja). Analyyseissä käytettiin monitasometaregressiomallinnusta, joka estimoi efektikoon koeryhmälle verrattuna kontrolliryhmään (Hedgesin *g*). Lisäksi tarkasteltiin digitaalisten hoitojen tehokkuuteen mahdollisesti vaikuttavia muokkaavia tekijöitä. Toisessa osatutkimuksessa selvitettiin digitaalisen hoidon keskeyttämistä ennakoivia tekijöitä kahden satunnaistetun vertailututkimuksen aineistossa (*N*=253) logistisilla regressiomalleilla. Tutkimuksessa tarkasteltiin yksilöllisten ominaisuuksien yhteyttä digitaalisen hoidon keskeyttämistodennäköisyyteen ja lisäksi sitä, miten yksilölliset ominaisuudet ja osallistujan käyttäytyminen digitaalisella alustalla ennustivat keskeyttämistodennäköisyyttä osallistujien suoritettua hoidon ensimmäisen moduulin. Kolmannessa osatutkimuksessa selvitettiin, voidaanko älylaitteilla kerätyillä käyttäytymiseen ja hyvinvointiin liittyvillä tiedoilla ennustaa mielenterveysoireilua. Tutkimuksen aineistona 60 24–68-vuotiasta aikuista, joita seurattiin Applen iPhone-sovelluksen (*Delphi*) ja Oura-sormuksen avulla neljän viikon ajan. Kerätty aineisto sisälsi osallistujien sijaintia ja puhelimen käyttöä koskevat tiedot sekä aktiivisuuden, unen ja syketaajuuden vaihtelun mittaukset ja päivittäin raportoidun mielialan. Masennus-, ahdistus- ja stressioireet mitattiin osallistujilta tutkimuksen alussa, puolivälissä ja lopussa itseraportointikyselyillä (DASS-21). Kerätyn aineiston ja päivittäin raportoidun mielialan yhteyttä masennus-, ahdistus- ja stressioireiluun tutkittiin monitasoregressiomalleilla. Neljäs osatutkimus toteutettiin samassa aineistossa kuin kolmas osatutkimus. Siinä vertailtiin viiden ohjatun koneoppimisalgoritmin tarkkuutta luokitella osallistujat masentuneiden ja terveiden luokkiin.

Tulokset

Systemaattisen kirjallisuushaun perusteella ensimmäisen tutkimuksen meta-analyysiin sisällytettiin 83 tutkimusta (*N*=15,530). Digitaalisen intervention efektikoko kaikkiin kontrollitilanteisiin verrattuna oli *g* = 0.52. Kun digitaalinen hoito toteutettiin koeolosuhteiden ulkopuolella (ns. todellisessa elämässä), olivat efektikoot huomattavasti pienempiä (vaikuttavuus *g* = 0.30) kuin koeolosuhteissa havaitut (tehokkuus *g* = 0.59). Efektikoot olivat suurempia hoidoissa, joihin liittyi ohjaava ihmiskontakti (esimerkiksi terapeutti) (*g* = 0.63) verrattuna hoitoihin, joihin ei liittynyt ihmiskontaktia (*g* = 0.34). Efektikoot erosivat merkitsevästi myös kontrollitilanteesta riippuen (WLC: *g* = .70; attention: *g* = .36; TAU: *g* = .31). Ihmiskontaktin sisältävän digitaalisen hoidon havaittiin olevan yhtä tehokasta kuin kasvokkain tapahtuvan hoidon, joskin tutkimuksia tämän arvioimiseksi oli vain vähän. Toisessa osatutkimuksessa havaittiin, että matalampi koulutustaso (OR=3.33) sekä keskimääräistä(?) matalampi ja korkeampi ikä (kvadraattinen yhteys, ikä: OR=0.62, ikä^2: OR=1.55) ennustivat suurempaa todennäköisyyttä keskeyttää digitaalinen hoito. Niillä osallistujilla, jotka olivat suorittaneet hoidon ensimmäisen moduulin, keskeyttämistä ennustivat ikä (ikä: OR=0.61, ikä^2: OR=1.58), vähäisempi sosiaalinen tuki (OR=3.40) ja meneillään olevassa moduulissa jäljellä olevien päivien määrä (OR=1.05). Itseraportoitu ikävä tapahtuma edellisen viikon aikana oli puolestaan yhteydessä matalampaan keskeyttämistodennäköisyyteen (OR=0.22). Kolmannessa osatutkimuksessa havaittiin, että vähäisempi maantieteellinen liikkuvuus (β = −0.21, *p* = 0.037) ja suurempi unen (β = 0.24, *p* = 0.023) ja sängyssä vietetyn ajan määrä (β = 0.26, *p* = 0.020) olivat yhteydessä korkeampiin masennusoirepisteisiin. Lisäksi havaittiin yhteys nukahtamisen jälkeisen heräämisen ja ahdistuneisuusoireiden välillä (β = 0.23, *p* = 0.035). Neljäs tutkimus osoitti, että sensoripohjaisia ennustajia käyttävistä algoritmeista Support Vector Machine luokitteli ihmiset masennusoirepistemäärän perusteella oikein masentuneisiin ja terveisiin 75.90% tarkkuudella (käyrän alle jäävä pinta-ala (AUC) = 74.89%). Sensoripohjaisia ja päivittäisiin mielialamittauksiin perustuvia ennustajia yhdistävän XGBoost-algoritmin tarkkuus oli 81.43% (AUC = 82.31%).

Johtopäätös

Tämä väitöskirjatutkimus tuotti uutta tietoa digitaalisten masennushoitojen tehokkuudesta. Tutkimuksessa esitettiin ensimmäinen kattava meta-analyysi, joka osoitti, että digitaaliset hoidot voivat olla tehokkaita psykiatrisen hoidon välineitä, mikäli digitaaliseen hoitoon sisältyy ohjaava ihmiskontakti. Digitaalisten hoitojen laajamittaisen käytön suurin haaste liittyy yhä hoitoon sitoutumiseen; keskimäärin vain joka neljäs potilas suorittaa hoidon loppuun. Tutkimustulosten mukaan kannettavien ja puettavien älylaitteiden avulla voidaan kerätä arvokasta tietoa, jonka avulla ennakoida masennusoireilua ja siten varhain tunnistaa erityisessä riskissä olevat.

Acknowledgements

There is an old saying, "*if you want to go fast, go alone. If you want to go far, go together*". Nothing is more true of this thesis. During my doctoral journey, I have had the honor of working with some incredible people from whom I have learned a tremendous amount. Beyond that, I have been supported and encouraged throughout by a wonderful team of family, friends, and colleagues. This thesis is dedicated to you all.

I would like to begin by thanking my supervisors, Docent Laura Pulkki-Råback, Professor Giulio Jacucci, and Professor Niklas Ravaja. Laura, thank you for taking me under your wing and for all the guidance and support you have provided over the past years. Your meticulous attention to detail, coupled with your ability to frame the research agenda within a wider scientific and societal context is rare and have been invaluable in informing this work. Thank you also for trusting me to forge my own path and giving me the space to experiment, make mistakes and learn. We have had many great conversations and shared many laughs over the years, and I look forward to many more. Giulio, thank you for championing my cause and for your boundless positivity. Your interdisciplinary expertise gave me the confidence to bring together the worlds of clinical psychology and computer science in novel ways. Niklas, thank you for your unwavering support from the very beginning, for always being there when I have needed help, and for bringing together a superb supervisory team to help guide me throughout my doctoral journey.

To my thesis advisory committee, Research Professor Timo Partonen and Associate Professor Benjamin Cowley, thank you for helping me navigate my doctoral path from beginning to end and for your valuable input on my research. I would like to thank the official reviewers of this thesis, Professor Raimo Lappalainen and Assistant Professor Keegan Knittle for their positive and insightful comments on the manuscript. I sincerely thank Professor Heleen Riper for kindly agreeing to be the opponent at my doctoral defense; it is an honor.

I would like to express profound gratitude to my co-authors and collaborators, without whom this work would simply not have been possible. Special appreciation goes to my two *amigos*, Lasse Sander and Yannik Terhorst. Lasse, you welcomed me into this exciting research field unconditionally and your guidance has been invaluable. I cannot express how grateful I am for how generous you have been in sharing your time and expertise with me throughout this journey. Yannik, I owe you a tremendous debt for the methodological and statistical expertise you have brought to this work and for your instruction. You are a teacher *par excellence*; kind and patient. I would like to extend my thanks to my other co-authors, Paula Phillipi, Denzil Ferreira, Kennedy Opoku Asare, Matthias Domhardt, Assistant Professor Ioana Cristea, Professor David Ebert, Professor David Mohr, Professor Harald Baumeister, and Professor Pim Cuijpers. I have learned a great deal from you all and am proud of the work we have produced together. Additionally, I would like to thank the Psychosocial Factors and Health Research Group, Professor Marko Elovanio, Christian Hakulinen, Kaisla Komulainen, Professor Markus Jokela, and my colleagues at the Department of Psychology for the wonderful camaraderie.

I am grateful to the Doctoral School in Population Health at the University of Helsinki, the Finnish Psychiatric Foundation, and the Finnish Cultural Foundation for the funding to produce this work.

Finally, I would like to thank my dear family and friends for their support and encouragement throughout this work and over the past years. Mum, thank you for providing the foundation that has enabled me to pursue this journey and for your unwavering love and support throughout the years. Phil, you are a rock that we all lean on, and I am deeply appreciative of all that you do. To my father-in-law, Emeritus Professor Mohamed M. Ahmed, thank you for sage advice and the many diverse and wonderful conversations we have shared over the years. I look forward to many more. To my mother-in-law, Dr. Marjaliisa Puukko-Ahmed, thank you for the sincere interest you have taken in my work from the start and for all the support you have provided in enabling me to get it done. Your ability to marry the worlds of eastern and western philosophy is rare and inspirational, and I have cherished our discussions.

Most of all, I would like to thank my beautiful wife, Noora, and our amazing children, Safiya and Laila. Noora, you have been by my side every step of this journey as it evolved from unstructured thoughts and a deep yearning to do something important and meaningful to published articles and a doctoral dissertation. Thank you for helping to create the space I have needed to explore these topics in depth and the practical, intellectual, and emotional support you have given me throughout. You are my best friend and I feel the deepest gratitude for what we have. Safiya and Laila, you two illuminate my world. Thank you for the joy you bring to each and every day, and for your patience and understanding during the times I needed to focus. You have given me a sense of purpose and meaning that is unparalleled and I love you both more than words can express.

Helsinki, November 2022

Isaac Moshe

Isaac Moshe

List of Original Publications

I Moshe, I., Terhorst, Y., Philippi, P., Domhardt, M., Cuijpers, P., Cristea, I., Pulkki-Råback, L., Baumeister, H., & Sander, L. B. (2021). Digital interventions for the treatment of depression: A meta-analytic review. *Psychological Bulletin*, *147*(8), 749–786. https://doi.org/10.1037/bul0000334

II Moshe, I., Terhorst, Y., Paganini, S., Schlicker, S., Pulkki-Råback, L., Baumeister, H., Sander, L. B., & Ebert, D. D. (2022). Predictors of Dropout in a Digital Intervention for the Prevention and Treatment of Depression in Patients With Chronic Back Pain: Secondary Analysis of Two Randomized Controlled Trials. *Journal of Medical Internet Research*, *24*(8), e38261. https://doi.org/10.2196/38261

III Moshe, I., Terhorst, Y., Opoku Asare, K., Sander, L., Ferreira, D., Baumeister, H., Mohr, D. C., & Pulkki-Råback, L. (2021). Predicting Symptoms of Depression and Anxiety Using Smartphone and Wearable Data. *Frontiers in Psychiatry*, *12*, 625247. https://doi.org/10.3389/fpsyt.2021.625247

IV Opoku K, Moshe I, Terhorst Y, Vega J, Hosio S, Baumeister H, Pulkki-råback L, Ferreira D. (2022). Mood ratings and digital biomarkers from smartphone and wearable data differentiates and predicts depression status: A longitudinal data analysis. *Pervasive and Mobile Computing*. Elsevier B.V.; 2022;83:101621. doi: 10.1016/j.pmcj.2022.101621

The original publications have been reproduced with the permission of the copyright holders.

Abbreviations

AIC	Akaike information criterion
AUC / AUROC	Area under the receiver operating curve
BIC	Bayesian information criterion
CBP	Chronic back pain
CBT	Cognitive behavioral therapy
CI	Confidence interval
COVID-19	Coronavirus disease 2019
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders, 4th Edition
EEG	Electroencephalogram
ES	Effect size
GPS	Global positioning system
HRV	Heart rate variability
iCBT	Internet-based cognitive behavioral therapy
IPDMA	Individual participant data meta-analysis
LRT	Life review therapy
ML	Machine learning
OR	Odds ratio
PDT	Psychodynamic therapy
PSG	Polysomnography
PST	Problem solving therapy
RoB	Risk of bias
RCT	Randomized controlled trial
SD	Standard deviation
SE	Standard error
SOL	Sleep onset latency
TAU	Treatment as usual
TIB	Time in bed
TST	Total sleep time
WASO	Wake after sleep onset
WLC	Waitlist control

Introduction

Each year, approximately 300 million people are affected with depression, making it a leading cause of disability worldwide (World Health Organization, 2017). Across the lifespan, one in five women and one in eight men will experience at least one episode of major depression (Bromet et al., 2011; Kessler & Bromet, 2013). What is more, these numbers are likely to underestimate the true extent of the burden as many more people suffer from symptoms of depression but do not meet the threshold required for a formal clinical diagnosis: so-called "subthreshold" or "subsyndromal" depression (Kessler & Bromet, 2013). Depression is evenly distributed across the age groups with peaks in prevalence during the second and third decades of life and again during the fifth and six decades (Hirschfeld, 2012). Although sociocultural factors such as socioeconomic status are known to influence the course of depression, similar prevalence rates are found between high- and low-and-middle income countries (De Aquino et al., 2018; Whiteford et al., 2013). Depression has been identified as a risk factor for several chronic health conditions. It is associated with lower quality of life (Malhi & Mann, 2018), significant social impact and increased morbidity and mortality (Erskine et al., 2015; Ormel et al., 1994; Vos et al., 2017). Besides the personal suffering experienced by patients and their families, mental disorders are also responsible for considerable economic and societal damage. The global cost of mental health conditions in 2010 was estimated at US$ 2.5 trillion and the cost is projected to surge to US$ 6.0 trillion by 2030 (Bloom, et al., 2011).

There is a large body of evidence demonstrating the efficacy of psychotherapy for the treatment of depression across all age groups and levels of depression severity (David et al., 2018). Psychotherapy has been demonstrated to be as effective as pharmacotherapy in the short-term (Cuijpers, Noma, et al., 2020), has less unwanted side-effects and may be more effective than antidepressants in the long-term (Karyotaki et al., 2016). What is more, the majority of patients prefer psychotherapy as their first-line of treatment over medication (Lee et al., 2020; McHugh et al., 2013). As such, most national clinical guidelines now recommend psychotherapy as the first line of treatment for mild-to-moderate depression (American Psychiatric Association, 2000; NICE, 2017). However, despite the demonstrated efficacy of existing treatments, it is estimated that only one in five (20%) of people in high-income countries and one in twenty-seven (3.7%) in low-and-middle income countries receive appropriate care, giving rise to a "treatment gap" in mental healthcare: the difference between the proportion of people who need care and the proportion that receive it (Evans-Lacko et al., 2018; Kohn et al., 2004).

Although there are many barriers to accessing care, one of the most significant is the dominant model of delivering psychotherapy (Kazdin, 2017). The dominant model is characterized by one-to-one, in-person treatment delivered by a highly trained mental health professional and held within a clinical setting. This model places major limitations on the degree to which treatment can be extended to reach the large number of people currently in need

(Fairburn & Patel, 2014; Kazdin & Blase, 2011). First, there are simply too few trained mental health professionals to meet the demand for treatment - and this is projected to get worse in the coming years (KFF, 2021). Furthermore, the majority of mental health professionals are trained to provide care for adults with mild-to-moderate disorders. Yet, there is a currently a major unmet demand for the treatment of children and adolescents, the elderly, minority groups, people with physical and intellectual disabilities and those living in rural areas (Andreas et al., 2017; Eden et al., 2012).

Digital interventions have been proposed as an alternative method of delivering mental healthcare that may address several of the limitations of the dominant model and thus help meet the growing demand for treatment (Kazdin, 2017). Although digital interventions vary considerably in their form and content, they typically package up the core components of psychotherapy (e.g., cognitive restructuring and behavioral activation in CBT or brainstorming possible solutions in PST) into digital media such as videos, interactive exercises, text and imagery and deliver them as structured programs, most commonly via the internet or a smartphone application. Interventions usually last between 6 to 12 weeks and are designed to be self-guided or accompanied by some form of brief human support (e.g., weekly emails or phone calls) (Andersson & Carlbring, 2021). The purported benefits of digital interventions include reducing waiting list times, enabling access for rural or hard-to-reach communities and lessening the stigma associated with seeking support (many interventions are anonymous).

As an increasing number of digital interventions are being deployed within public and private healthcare settings to meet the growing demand for treatment (Clark, 2018; Titov et al., 2018) – driven in part by the acceleration of digital healthcare due to the coronavirus pandemic (Torous et al., 2020) – a comprehensive analysis of the extent to which these interventions are effective in the treatment of depression, together with a robust exploration of their limitations, is both timely and important. In particular, whilst numerous studies have been conducted assessing the efficacy of digital mental health interventions in highly controlled settings, less well known is how effective these interventions are in real-world healthcare settings where participants are far more representative of the general population. Equally important is understanding how the outcomes of digital interventions compare with traditional, face-to-face psychotherapy as many practitioners have expressed concerns with the lack of in-person contact in digital interventions (Topooco et al., 2017). As digital interventions start to take their place as part of the wider treatment offering within public healthcare (Perera-Delcourt & Sharkey, 2019), clinicians and policy makers would also benefit from a more in-depth understanding of which patient groups digital interventions may (and may not) be effective for. That is to say, are there specific patient characteristics such as age, depression severity level or physical comorbidity that moderate treatment outcomes?

Another major question relates to the acceptability of digital interventions. Qualitative analyses of patient experiences with digital interventions have revealed that a number of patients find the digital format challenging (Knowles et al., 2015) and many studies report high dropout rates for digital interventions (Cuijpers, Noma, et al., 2019; Eysenbach, 2005). A more thorough understanding of what factors are associated with higher levels of treatment dropout at both the participant and intervention level could help inform both treatment allocation and intervention design, thereby maximizing the overall efficacy of digital interventions.

Even if we are able to close the treatment gap and provide evidence-based treatments to all those in need, modeling studies suggest that over 60% of the global burden from depression would remain due to the limited efficacy of existing solutions (Andrews et al., 2004). On average, 40-50% of patients fail to respond to initial treatment (Cuijpers, Sijbrandij, et al., 2013). Even when patients do respond, 80% of those who remit will experience at least one further episode during their lifetime (Spijker et al., 2002). We are thus faced not only with a treatment gap, but also an "efficacy gap".

One of the most promising ways of reducing the burden of depression is the early identification of individuals who may be at-risk of developing the disorder. Over 20% of people are estimated to suffer from subclinical depression each year (Cuijpers et al., 2004; Fergusson et al., 2005) and these individuals are four times more likely to develop Major Depression than those who do not display any symptoms (Cuijpers et al., 2004). Yet, despite this, less than half of the cases of subclinical depression are identified (McQuaid et al., 1999; Wells et al., 1988). Successfully identifying these individuals and ensuring they receive proper treatment could prevent more than 70% of morbidity and mortality caused by depression (Docherty, 1997).

Digital interventions may also have a valuable role to play here. Over the past decade, the adoption of smartphone and wearable devices has exploded, with over 6 billion people now having access to a smartphone device (Statista, 2022). These devices continuously produce an exhaust of passive data which have been shown to correlate with and predict symptoms related to a number of mental disorders (Rohani et al., 2018). Being able to leverage this data to identify individuals exhibiting early warning signs and known risk factors associated with depression may provide clinicians with the opportunity to intervene early in the progression of the disease to prevent it from developing into a full-blown disorder. The field here is still young, however, and many results are conflicting. As such, research would benefit from more studies assessing whether previous findings replicate as well as trials exploring the role of new sources of data that may be used to improve predictive models.

The current investigation set out to explore the efficacy, effectiveness, and moderators of treatment outcome in digital interventions for depression in a meta-analysis of randomized controlled trials comparing digital interventions to active and inactive controls. Additionally, we explored what factors may predict treatment dropout in a digital intervention for depression in a secondary analysis of two large-scale RCTs. Finally, we assessed to what extent data from a smartphone and novel consumer wearable device may be used to predict symptoms of depression in a longitudinal observation study.

It is the core thesis of the current study that new technologies have the potential to play a major role in addressing the growing and unmet demand for treatment in depression. As we shall explore in the closing sections of this study, they may also provide us with an opportunity to go beyond the existing limits to efficacy that have been a hallmark of the field for the past 50 years (Cristea et al., 2017; Johnsen & Friborg, 2015). In the studies herein we provide a robust assessment of the efficacy and effectiveness of these solutions at the same time as exposing their current limitations. We explore possible ways of addressing these limitations and close by providing clear directions for the future of the field in order to realize the full potential digital interventions may hold.

Let us begin…

2 Review of the Literature

2.1 What are digital interventions?

Emerging in the 1980s, digital interventions started out as little more than psychotherapy treatment manuals delivered via CD-ROM. Over time, they have come to be delivered over the internet, first via computers and, more recently, via smartphone devices. With their close approximation to treatment manuals, digital interventions mirror many of the aspects of face-to-face psychotherapy. They often begin with some form of assessment, for example the self-report Patient Health Questionnaire (PHQ-9) for depression, following which patients are presented with an overview of the program and an explanation of how it has been designed to help alleviate their symptoms. Programs are typically structured into a series of lessons or modules based on the therapeutic approach used in the intervention. Modules are delivered as a mixture of text, video, audio clips and interactive exercises and patients are often given homework to complete at the end of each module to help practice and consolidate what they have learned. Whilst the duration of interventions varies, the majority are designed to be completed in approximately 6-12 weeks.

An example of one of the most widely researched interventions, Deprexis, can be found in **Appendix 1**. Deprexis consists of 10 modules covering a variety of content broadly consistent with a cognitive behavioral-approach but also drawing from other psychotherapeutic approaches (Berger et al., 2011; Meyer et al., 2009). The content of the modules relate to: (1) psychoeducation (Beck, 1979). (2) cognitive restructuring (Beck, 1979), (3) behavioral activation (Jacobson et al., 1996), (4) mindfulness and acceptance (Hayes et al., 1999), (5) relaxation, physical exercise and lifestyle modification (Pollock, 2001), (6) problem-solving (Nezu, 1986), (7) interpersonal skills (Weissman & Klerman, 1990), (8) expressive writing and forgiveness (Pennebaker, 2004), (9) positive psychology interventions (Seligman & Csikszentmihalyi, 2000) and (10) dreamwork and emotion-focus interventions (Hill, 1996). Each module combines a theoretical overview of the relevant therapeutic component(s) together with specific recommendations of how patients can apply the theory practically in their own lives to help alleviate symptoms. For example, the Deprexis module focused on cognitive restructuring first explains the theory behind automatic thoughts and how these thoughts are related to emotions, behavior, and events. It then asks the user to identify the cognitive distortions in their own automatic thoughts and provides them with simple techniques to challenge those automatic thoughts. Digital interventions may be offered as “unguided”, self-help interventions, where the onus is entirely on the individual to complete the program or as “guided interventions”, where they are accompanied by some form of minimal human support such as brief weekly emails or telephone calls. In some cases, interventions are provided as an adjunct to traditional face-to-face therapy, so called “blended therapy”.

Digital interventions may provide a number of benefits over face-to-face therapy, including saving therapist time, and thus reducing waitlists, improving access for people in hard-to-reach areas or with disabilities, and allowing patients to access therapy at a time and place that is convenient to them (Ebert, Van Daele, et al., 2018). Furthermore, the relative anonymity enabled by digital interventions may reduce the stigma associated with seeking therapy (Thomas et al., 2015). As we shall explore further in the Discussion, digital interventions may also provide a number of advantages for psychotherapy researchers, including the opportunity to recruit trial participants more efficiently and cost-effectively, and the ability to use more standardized treatment protocols than trials of face-to-face psychotherapy (Holmes et al., 2018).

However, digital interventions are not without disadvantages and challenges. Among these are the negative attitudes of some patients towards the digital medium, either due to lack of in-person interaction with the therapist or the burden of completing tasks online (Knowles et al., 2015). The highly structured nature of interventions are also felt by some clinicians and patients as being too rigid to enable the responsiveness that is a cornerstone of psychotherapy practice (Stiles et al., 1998). Related to this, many clinicians question whether a proper therapeutic alliance can develop between a patient and clinician within a digital intervention (Sucala et al., 2012). This is important as the therapeutic alliance has been found to be one of the most robust factors predicting treatment outcomes in psychotherapy (Flückiger et al., 2018). Finally, the rate of dropout in digital interventions is often much higher than in face-to-face therapy, especially when delivered within routine healthcare settings (Van Ballegooijen et al., 2014). As many people who access digital therapy do so anonymously, this is especially concerning as it is often not possible to follow-up with these patients and assess whether their symptoms have deteriorated.

2.2 The efficacy of digital interventions for the treatment of depression and factors moderating outcomes

The first study published assessing the efficacy of a digital intervention for depression was by Selmi and colleagues in 1990 (Selmi et al., 1990). The three-arm RCT compared a computer-based CBT program with weekly face-to-face therapy and a waitlist control (WLC) over a six-week period. Participants were individuals diagnosed with mild-to-moderate depression. The study found high effect size (ES) superiority for both the computer-based CBT program (g = 0.88) and face-to-face therapy (g = 0.74) compared to WLC. Moreover, no significant difference was found between the two treatment arms, leading the authors to conclude that the computer-based intervention was as effective in the treatment of depression as face-to-face therapy.

During the three decades since that seminal study, the field has evolved rapidly. There are now hundreds of RCTs and dozens of meta-analyses assessing the efficacy of digital interventions for the treatment of depression for a variety of populations (Andersson & Cuijpers, 2009; Barak et al., 2008; Carlbring et al., 2018; Firth, Torous, Nicholas, Carney, Rosenbaum, et al., 2017; Karyotaki et al., 2017, 2018; Konigbauer et al., 2017; Linardon et al., 2019). Previous meta-analytic results suggest small-to-moderate effect size superiority of between cohen's *d* = .32 (Spek, Cuijpers, et al., 2007) and *d* = .90 (Konigbauer et al., 2017) for digital interventions versus control conditions. However, outcomes vary considerably between studies, with some studies finding effect sizes as large as *d* = 1.60 (Titov et al., 2010) and others finding no effect of digital interventions at all (e.g., Boeschoten et al., 2017; G. Clarke et al., 2002). The considerable range of reported effect sizes may be explained by several factors that have been demonstrated to influence outcomes in digital interventions, including control type, the role of human guidance, participant characteristics, study setting and design, and study quality. In the following sections we shall explore each of these in turn.

2.2.1 The influence of control type on effect size

As in trials of face-to-face therapy (Cuijpers et al., 2010), the type of control used in digital intervention studies can have a significant influence on reported effect sizes. The majority of studies conducted to-date on digital mental health interventions have used waitlist control groups as the comparator (Webb et al., 2017). However, it is well demonstrated that waitlist control groups can undermine the internal validity of a trial and often lead to an overestimation of treatment effects (Cuijpers, Karyotaki, et al., 2019; Furukawa et al., 2014; Gold et al., 2017). For example, a meta-analysis on digital interventions for the treatment of depression and anxiety revealed a significantly higher mean effect size for studies using a waitlist control (*d* = 0.90) compared to studies using care-as-usual (*d* = 0.38) as the control condition. One reason for the difference in outcomes may be that participants allocated to a waitlist are more "motivated to remain depressed" so they can receive the treatment they originally desired, whilst those allocated to care-as-usual may actively seek alternative treatments (Furukawa et al., 2014; Kiluk et al., 2011). Another explanation might be that waitlist-based trials are often more "convenient trials", with a lower trial budget and thus lower quality standards (Cuijpers et al., 2010). Regardless of the underlying reasons, it is important to understand how control types may influence reported effect sizes across studies and, within that, clearly establish the incremental value of digital interventions above alternatives such as usual care or active psychoeducational controls.

Another important question related to comparator conditions is how digital interventions compare to face-to-face therapy. A survey assessing the acceptability of digital treatments for depression by mental health stakeholders across Europe (the E-COMPARED study), revealed a number of concerns with eliminating the face-to-face interaction between patient and therapist as most digital interventions do (Topooco et al., 2017). Specifically, stakeholders were concerned that digital interventions are "impersonal, [with] no direct eye contact", that "the personal relationship with the therapist is lost, this is only possible via face-to-face", and the treatment format "does not adequately address comorbidity/crisis/suicide risk" (Topooco et al., 2017, p.

5). Indeed, the dominant theoretical models of psychotherapy seem to support the contention that a face-to-face therapist is required for large treatment outcomes (Wampold, 2001). As the strength of that relationship, the "therapeutic alliance," is one of the largest predictors of outcomes in psychotherapy (Wampold, 2015), this raises questions as to whether one can develop a therapeutic alliance in digital interventions and, if not, does this influence outcomes?

Despite these concerns, a meta-analysis by Carlbring and colleagues (2018) comparing iCBT with face-to-face therapy for depression found no significant difference in outcomes between the two, causing the authors to conclude that the two treatment formats are equally effective in treating depressive symptoms. However, it is important to highlight a number of limitations with the meta-analysis by Carlbring and colleagues. First, the meta-analysis had high levels of heterogeneity, pooling together studies on individual face-to-face and group-based psychotherapy as well as studies using different control types. Second, sample sizes of many of the studies included in the analysis may have been too small to detect the differences typically found in noninferiority or comparative trials (only one trial had more than 40 participants in each arm). Finally, the included trials were conducted in highly controlled laboratory settings, typically in a clinician's office, which are not representative of real-world conditions. As we shall explore later in this study, the type of setting in which digital interventions are delivered may have a major influence on outcomes. Given the substantial difference in time required for therapists to deliver face-to-face therapy compared with digital interventions (7.8 times the amount, according to one meta-analysis (Andrews et al., 2018)), and the potential for digital interventions to scale the delivery of treatment outside of in-person settings, robust studies assessing whether digital interventions are indeed non-inferior to face-to-face therapy are paramount before making clinical and policy decisions.

2.2.2 The influence of human guidance

Digital interventions typically come in two forms: those where the module content is accompanied by human guidance (guided interventions) and those without any human support (self-guided or unguided interventions). A number of studies and meta-analyses have demonstrated that guided interventions lead to superior outcomes compared to unguided interventions (Baumeister et al., 2014; Musiat et al., 2022; Spek, Nyklicek, et al., 2007). For example, a meta-analysis by Andersson and Cuijpers (2009) found an average ES of $d = 0.61$for guided interventions but a significantly lower ES of $d = .25$ for unguided interventions. More recently, however, researchers have argued that the difference in effect size between the two may be diminishing. A meta-analysis by Koenigbauer and colleagues (2017) found no significant differences between guided and unguided interventions, although the subgroup analyses consisted of only three trials. A review of digital interventions for depression and anxiety by Shim and colleagues (2017) found mixed evidence when comparing the two treatment formats, proposing that recent developments in technology affecting the quality and navigability of digital interventions may reduce the need for human guidance.

Another important factor related to guidance is that not all types of guidance are the same. One of the most clinically meaningful distinctions can be made between technical guidance and therapeutic guidance. In interventions with technical guidance, the content of the guidance is limited to answering technical questions regarding the intervention (e.g., trouble logging in or accessing content) and providing motivational support to maximize intervention adherence. The person providing the technical guidance is usually not a trained clinician but rather a support person, explicitly informed not to provide advice of a clinical or therapeutic nature. By contrast, in interventions with therapeutic support, the person providing the guidance is a trained clinician (e.g., psychotherapist or psychiatrist) and the content of the support is therapeutic in nature, including helping the patient set goals, discussing strategies related to overcoming challenges and providing feedback on the content of homework. In one of the first RCTs to directly compare technical versus therapeutic guidance in a digital intervention for depression, Titov and colleagues (2010), found that both treatment arms led to large effect sizes compared to a waitlist control group (d = 1.60 and d = 1.54, respectively), but there was no significant difference between the two.

Within therapeutic guidance, the qualification and experience of the person providing support may also differ. In a meta-analysis examining the influence of therapist qualification on effect size, Baumeister et al (2014) found no difference in outcomes when support was provided by experienced clinicians with several years of experience compared to when it was provided by students of clinical psychology.

Finally, the question of just how much guidance is needed is also an important consideration. Similar to pharmacological treatment, understanding when and how much of a specific dose - in this case, guidance time - we should apply to achieve optimal outcomes has both clinical and pragmatic implications. For example, is the dose-response relationship between guidance time and outcomes linear? Is there a cut-off point after which a higher dose has no impact on outcomes, as Titov (2011) has hypothesized? In one of the few trials to examine the relationship between amount of guidance and outcomes in a digital mental health intervention, Klein and colleagues (2009) found that there was no difference in outcomes when an intervention targeting panic disorder was accompanied by frequent (three emails per week) or infrequent (one email per week) support. However, to-date, no studies have addressed this question in the field of digital interventions for depression.

All else being equal, unguided interventions are the most scalable form of digital intervention, being less dependent on financial resources, therapist time and availability. But whether unguided interventions are effective for all individuals, or only a subset, is unclear. As digital interventions make their way into real world healthcare settings, and commissioners are required to make important decisions regarding their implementation, an up-to-date and in-depth understanding of the role of human guidance on outcomes is needed. This should include the influence of guidance type on outcomes, the qualifications of the person providing guidance and the dose-response relationship.

2.2.3 Participant characteristics moderating outcomes

Identifying which participant characteristics might predict or moderate differential treatment outcomes is critical if we are to match patients with the right treatment formats and increase the overall efficacy of digital interventions (Andersson et al., 2008; Donker, Batterham, et al., 2013; Warmerdam et al., 2013). Predictors are pre-treatment variables that predict outcome in all treatment groups (Kraemer et al., 2002, 2006), whilst moderators are pre-treatment variables that identify which individuals are more likely to benefit from a specific treatment or treatment component (Kazdin, 2009). As in face-to-face psychotherapy, a number of studies have identified specific patient characteristics as predictors of treatment outcome in digital interventions (Donker, Batterham, et al., 2013; Ebert et al., 2013; Warmerdam et al., 2013), although, as has been the case in face-to-face therapy, results have often been conflicting or inconclusive.

In an RCT of a guided iCBT intervention for sub-threshold depression, Spek and colleagues (2008) found that female gender was associated with superior outcomes, whilst a meta-analysis using individual participant data (IPDMA of guided iCBT interventions for depression by Karyotaki and colleagues (2018) found that there was no influence of gender. Education level has also been shown to predict differential outcomes in digital interventions. In a study comparing two internet-based interventions for the treatment of depressive symptoms – one based on CBT and the other problem-solving therapy (PST) - Warmerdam and colleagues (2013) found that high education level increased the likelihood of improvement by 2.41 when compared with a low or medium level of education. In addition, the study found an interaction effect between theoretical orientation and education: individuals with lower levels of education benefited significantly more from the intervention based on PST than the intervention based on CBT, highlighting the complex relationship between predictors, moderators and outcomes that may be present in digital interventions.

Whether digital interventions are acceptable and efficacious for people of all ages groups is also an important consideration. Whilst meta-analyses comparing outcomes of face-to-face psychotherapy for depression across different age groups have shown lower effect sizes for children (g = .35) than for adults (g = .77) (Cuijpers, Karyotaki, et al., 2020), results for digital interventions may differ. A meta-analysis of internet- and computer-based CBT for depression in young people by Ebert and colleagues (2015) reported significant medium-to-large effect sizes over all control types (d = .76) leading the study authors to conclude that digital interventions may be a promising treatment alternative when evidence based face-to-face treatment is not feasible. However, it is important to note that the review included young people with a wide age variation (13-25 years of age), blurring the distinctions between children, adolescents, and young adults. As discussed previously, the type of control used in these studies may also influence outcomes. A more recent meta-analysis of digital mental health interventions for young people by Garrido and colleagues (2019) found a small (d = .33) pooled effect size superiority of digital interventions compared to no intervention controls but no significant difference in outcomes when the digital intervention was compared to active control conditions. Again, the study also included participants up to the age of 25. Thus, whether digital interventions for depression are indeed effective for children (<18 years of age) remains an open question.

Depression in older adults is a significant problem due to its association with reduced quality of life, increased disability and increased utilization of medical services (Bunce et al., 2012; Unützer et al., 2002). Compounding the problem, depression is underdiagnosed in older adults (Byers et al., 2010) and only a small proportion of older people seek and receive evidence-based treatments (MacKenzie et al., 2012; Trollor et al., 2007). Although digital interventions hold promise in reaching this population, natural questions arise as to whether challenges using new technology formats such as the internet and smartphone devices may prevent them from experiencing the benefits. Despite the concerns, initial evidence suggests that digital interventions may be highly effective for older people. An RCT assessing the clinical and cost-effectiveness of guided iCBT for older adults (>60 years) with symptoms of depression revealed very large clinical improvements ($d = 2.08$) compared to a waitlist control group. Moreover, outcomes were maintained at 12-month follow-up. Unguided, self-help interventions may also be effective for older adults. A large RCT comparing guided and unguided iCBT in over 400 older adults (aged 60-78) with symptoms of depression and comorbid anxiety found large within-group effect sizes for both guided ($d = 1.45$) and unguided ($d = 1.44$) interventions, with no significant difference between the two treatment formats (Titov et al., 2016). However, whilst results from these two trials are promising, other studies involving older adults have not found any effect size superiority of digital intervention versus controls (O'Moore et al., 2018), prompting the need for a meta-analysis to synthesize findings across studies.

Whether digital interventions are effective and feasible for individuals with severe depression is unclear. The majority of clinical guidelines do not recommend digital interventions as first-line treatment for patients with severe depression (American Psychiatric Association, 2000; NICE, 2017). Indeed, many trials explicitly exclude participants with very high levels of depression severity or suicidality (Bailey et al., 2020; Pearson et al., 2001; Sander, Gerhardinger, et al., 2020). Furthermore, the current perception held by clinicians and policymakers is that digital interventions are only acceptable for patients with mild depression (Topooco et al., 2017). Despite this, meta-analyses of studies that have included more severely depressed patients have actually found the largest mean effect sizes in participants with the highest pre-treatment depression severity (J. H. Wright et al., 2019). How the intervention is delivered may be an important moderator here, however. An IPDMA comparing the effect of guided versus unguided interventions across individuals with different severity levels found that digital interventions were only effective for individuals with severe depression levels when accompanied by human support. On the other hand, both guided and unguided interventions were demonstrated to be effective for individuals with mild symptom severity. Moderator analyses that explore the interaction between participant characteristics as predictors and intervention components as moderators of treatment outcomes are thus extremely important in informing clinical considerations when it comes to the question of whether to offer digital interventions to patients with severe mental illness. It also carries pragmatic implications for how groups may be stratified to receive different treatment formats to tackle the population-level burden more cost-effectively.

A final consideration when it comes to participant characteristics relates to physical comorbidity. Up to 50% of adults diagnosed with a chronic physical disease experience symptoms of depression (Clark & Currie, 2009). Depression has been identified as both a risk factor and a negative prognostic factor for diabetes, back pain, cardiovascular disease, arthritis and hypertension (Steffen et al., 2020). Comorbid depression is related to poor quality of life, poorer outcomes, increased healthcare utilization and increased mortality compared to the presence of depression or the physical disease alone (Barnett et al., 2012; Goldberg, 2010). Due to the fragmented nature of most healthcare systems, where physical and psychological symptoms are treated differently and up to 75% of patients say they have difficulty accessing psychological treatment (Mohr et al., 2010), digital interventions may provide a particularly valuable treatment format for patients with comorbid somatic conditions. In addition, digital interventions could provide an alternative to antidepressant medication for these patients, where there are often considerable risks associated with combining medication types. In a study assessing the efficacy of a guided intervention for the treatment of depressive symptoms in patients with Type 1 or 2 diabetes, Ebert and colleagues (2017) found a large effect size superiority ($d = .83$) for the digital intervention compared to TAU plus psychoeducation. An RCT by O'Moore and colleagues (2018) assessing the effects of an iCBT intervention for depression in older adults with osteoarthritis found that participants who received the intervention reported significantly fewer depressive symptoms compared to participants who received the standard treatment for osteoarthritis. Moreover, participants in the intervention group also reported lower pain disability and improved physical function after completing the intervention, compared with participants receiving usual care. Given the potential of digital interventions for this population, a meta-analytic review assessing the efficacy of digital interventions in the treatment of depression for individuals experiencing a coexisting somatic disorder - and whether there are specific disorders or severity levels that might moderate outcomes - is thus another important area of research that has not been addressed until now.

2.2.4 Study design & quality

As digital interventions become increasingly adopted within public and private healthcare settings, understanding whether they are effective outside of highly controlled laboratory settings is a critical question. Here we may distinguish between *efficacy* and *effectiveness* trial designs. Broadly speaking, efficacy studies are designed to assess whether the intervention produces the expected result under ideal conditions, whilst effectiveness studies measure the effect of the intervention within "real world" clinical settings (Godwin et al., 2003). Efficacy trials are typically designed to be high in internal validity, whilst effectiveness trials focus on maximizing external validity, i.e., will the results of the study generalize to other populations and other healthcare settings. Distinguishing between the two is particularly important when informing public policy as there is a large body of evidence from across medicine demonstrating that effect sizes found in efficacy trials are often significantly higher than those found in effectiveness trials (Eichler et al., 2011; Pagoto & Lemon, 2013).

This difference may also hold true for studies on digital interventions. In contrast to the promising results found in many efficacy studies, one of the largest pragmatic trials conducted within a public healthcare setting (the REEACT trial in the UK National Health Service) found no difference between two internet-based interventions and care-as-usual. Consequently, the study authors concluded that the benefits of computer-based interventions may not transfer to clinical settings (Gilbody et al., 2015a). The study was met with strong criticism, however (Dowrick, 2015; Gilbody et al., 2015b). Notably, the amount of guidance provided in the trial was extremely low (on average 6 minutes per participant) which may have led to the low levels of adherence found (less than 20% of participants completed the entire intervention). Indeed, a follow-up study that provided weekly guidance of 10-20 minutes over the telephone found significantly greater outcomes for participants receiving the additional guidance compared to those receiving the original intervention protocol (Gilbody et al., 2017). A more recent study by Richards and colleagues (Richards et al., 2020) assessing effectiveness in the same setting and in a similar population but using a different guided intervention reported a significant benefit in favor of the intervention, with superior outcomes maintained at 12-month follow up.

Another important consideration in effectiveness trials is the comparator. Both the Gilbody and Richardson studies used a waitlist control condition. Yet, as we have seen, WLC conditions tend to overinflate effect sizes compared to treatment as usual. Treatment as usual is arguably a more robust comparator in effectiveness trials given that the central research question is "do these interventions offer benefits over and above the usual care a patient would receive otherwise?" (Gilbody et al., 2015a). To-date, however, no research has been conducted comparing real-world outcomes across control types, leaving the question unanswered until now.

The quality of the study may also influence reported outcomes. Significant associations have been found between the reported effect sizes in face-to-face psychotherapy and quality criteria such as concealment of allocation and intention to treat analyses (Gellatly et al., 2007). A meta-analysis of face-to-face psychotherapy by Cuijpers and colleagues (2010) found strong evidence that high quality studies of psychotherapy for adult depression resulted in smaller effect sizes than low quality studies, concluding that the effects of psychotherapy for adult depression have previously been overestimated. As meta-analyses are particularly susceptible to the issue of "garbage in and garbage out" (Egger et al., 2001), a thorough assessment of study quality across trials and how it may influence reported effect sizes is thus critical before any robust conclusions can be drawn about the efficacy and effectiveness of digital interventions.

2.3 Adherence in digital interventions

2.3.1 The law of attrition in digital interventions

One of the main challenges associated with digital interventions is the high level of treatment dropout found in many studies (Batterham et al., 2008; Donkin et al., 2011; Musiat et al., 2022). In a meta-analysis comparing adherence in internet-based CBT and face-to-face CBT for depression, van Ballegooijen and colleagues (2014) reported that, on average, only 65.1% of participants completed the iCBT intervention. In comparison, 84.7% of participants completed treatment in face-to-face therapy. Similar findings have been found across studies , giving rise to what Eysenbach (2005, p. 2) has referred to as the "law of attrition", *"the phenomenon of participants stopping usage and/or being lost to follow-up, as one of the fundamental characteristics and methodological challenges in the evaluation of eHealth applications".*

Dropout may be exacerbated further when interventions are delivered outside of highly controlled laboratory settings or when there is no human guidance accompanying the intervention. For example, in one of the largest community-based studies of a self-guided iCBT intervention involving more than 80,000 users, 90% of participants failed to proceed beyond the first module (Batterham et al., 2008). These findings are important as there is a large body of evidence demonstrating a strong dose-response relationship between treatment adherence and outcomes in digital interventions (Donkin et al., 2011). As with guidance, determining the amount, frequency, and intensity of engagement with digital interventions has both practical and clinical implications. In the same way that pharmacological studies seek to establish the optimal dose of a medical substance in order to ensure that patients do not receive too little or too much of the medication, so too is there likely to be an optimal dose-response for digital interventions (Cuijpers, Huibers, et al., 2013). In face-to-face psychotherapy, research has shown that a larger number of sessions leads to greater improvement, with the optimal number of sessions around nine to ten (Clark, 2018) and recent research suggests a similar dose-response may apply to the number of modules in digital interventions. In a machine-learning analysis of engagement patterns in an iCBT intervention for depression and anxiety, Chien and colleagues (2020) identified that the maximal efficacious dose was to complete the 7 modules during an 8-week period at the pace of 1 module per week. In practice, as in pharmacotherapy, the optimal dose-response relationship in digital interventions will likely differ across participants depending on a number of factors, not least baseline depression severity level (Barkham et al., 2006; Reynolds et al., 1996). A meta-analytic examination of the dose-response relationship – and factors that may moderate the relationship – may thus help inform intervention developers and clinical guidelines when designing and prescribing interventions for specific populations.

2.3.2 Factors moderating adherence in digital interventions

Why some participants drop out of treatment and others do not is unclear. Several factors have been shown to predict adherence and dropout in digital interventions across sociodemographic variables, psychological comorbidity, intervention-related components and the therapeutic alliance (Christensen et al., 2009; Donkin et al., 2011; Kok et al., 2017). With regards to sociodemographic variables, male gender, lower education status, and older age have all been associated with lower adherence to digital interventions. Differences in gender may reflect the fact that females generally present with a higher effort to cope with depression compared to males (Karyotaki et al., 2015). At the same time, males are more likely to engage with health-risk behaviors which may also have a negative influence on adherence (Babwah et al., 2006). Lower educational status has been proposed to increase dropout due to greater difficulties in understanding the intervention content and/or limited abilities in using information technology. These factors may also account for the lower adherence found in studies involved in older participants (Christensen et al., 2009), although there are conflicting findings here (Karyotaki et al., 2015). Regarding comorbidity, both comorbid psychological disorders (Karyotaki et al., 2015; Kok et al., 2017) and higher baseline depression severity (Christensen et al., 2009) have been associated with lower adherence. One explanation for this may be the greater degree of impaired functioning found in these groups, making it more difficult for individuals to focus on the perceived demands of digital interventions (Johansson et al., 2015; Klein Hofmeijer-Sevink et al., 2012).

Whether the presence of a comorbid somatic condition - or the severity of that somatic condition - impacts the likelihood of dropout is unknown. As discussed previously, digital interventions may have a particularly valuable role to play in supporting individuals with co-morbid depression and a physical illness such as chronic pain. However, chronic pain and depression are often associated with reduced motivation (Felger & Treadway, 2016; Jonsson et al., 2011). As the treatment schedules of patients with multi-morbidities are often already demanding, higher levels of pain disability or lower levels of pain self-efficacy (the confidence in carrying out activities whilst in pain), may impact an individual's ability to engage with a digital intervention, thereby increasing the likelihood of dropout.

Knowing whether there may be certain participant characteristics associated adherence and dropout would help establish whether digital interventions are acceptable for all patients or just a subset (e.g., those experiencing low levels of pain). Such information could then help inform treatment allocation, reducing aggregate dropout levels and maximizing overall treatment outcomes.

2.3.3 Using intervention usage data to improve prediction models

Once a patient has started treatment, there is also a wealth of data from the intervention itself that can be used to improve the identification of those at high risk. Most digital interventions exhibit a so-called "attrition phase" early on in the intervention - typically the first one or two modules -

during which the majority of participants who fail to complete the intervention will dropout (Eysenbach, 2005; Farvolden et al., 2005; Wallert et al., 2018). Being able to identify which patients are at high risk of dropping out before this attrition phase may provide valuable prompts for clinicians to intervene early and, ideally, prevent dropout before it occurs. Digital interventions capture an array of passive and active data related to how users are interacting with the intervention. For example, many of them track when and for how long patients log in, the content they consume and responses to self-report questionnaires, interactive exercises, and homework. In cases where the intervention is guided, they may also capture the amount of guidance time provided (and requested) and the content of those interactions if guidance is delivered via email or chat (Bateup et al., 2020; Chien et al., 2020).

A growing body of evidence is now emerging demonstrating how such data may be used in predicting how likely an individual user is to drop out of an intervention once they have begun treatment. For example, in an iCBT intervention for the treatment of depression and anxiety following myocardial infarction, Wallert and colleagues (2018) found that the best predictors of dropout were the baseline characteristics of cardiac-related fear and gender, but also a novel set of linguistic predictors derived from the patients' written homework assignments. These included the number of words written, which the authors contended may be a proxy for patient effort during therapy, and the number of mutual words, which they proposed might have been a proxy for therapeutic alliance. Of relevance to the current study, a model that combined the baseline participant characteristics with the intervention-usage variables achieved the highest accuracy in predicting dropout, demonstrating the incremental value of intervention-usage data in identifying individuals at-risk of dropout.

A study by Bremer and colleagues (2020) of an iCBT intervention for insomnia also demonstrated the value of intervention usage data in improving dropout predictions. Using a combination of baseline characteristics (e.g., self-reported stress levels) and data from the intervention (e.g., number of days to complete each module), the authors were able to predict treatment dropout with an Area Under the Receiver Operating Curve (AUC) of 0.719. Moreover, the model was able to identify individuals at risk of dropout early on in the intervention (after they completed the introductory module), highlighting the potential of these models to provide valuable signals to supporting clinicians or in tailoring the intervention itself (e.g., via timely automatic reminders).

Finally, as an increasing number of regulations surrounding the acquisition and usage of personal data are now being put in place to safeguard patient privacy and protection, there may be cases where intervention providers do not have access to any participant characteristics. The problem may be further compounded when the interventions are accessed anonymously. To assess the feasibility of predicting dropout using a minimally data-sensitive model, Cote-Allard and colleagues (2022) developed a deep learning mode to predict dropout based on anonymized user IDs and login/logout timestamps only. The best-fitting model was able to identify participants at risk of dropout with a balanced accuracy of 70%.

Although the above findings are promising, it is important to realize that they are still exploratory in nature and the predictive power of most models remains moderate. As such, future research would benefit from additional studies that can identify novel predictors of dropout as well as establishing a consistent set of variables that may generalize across interventions and population groups, especially when implementing interventions in real-world healthcare settings.

2.4 Beyond the treatment gap: using digital technologies for identifying at-risk individuals

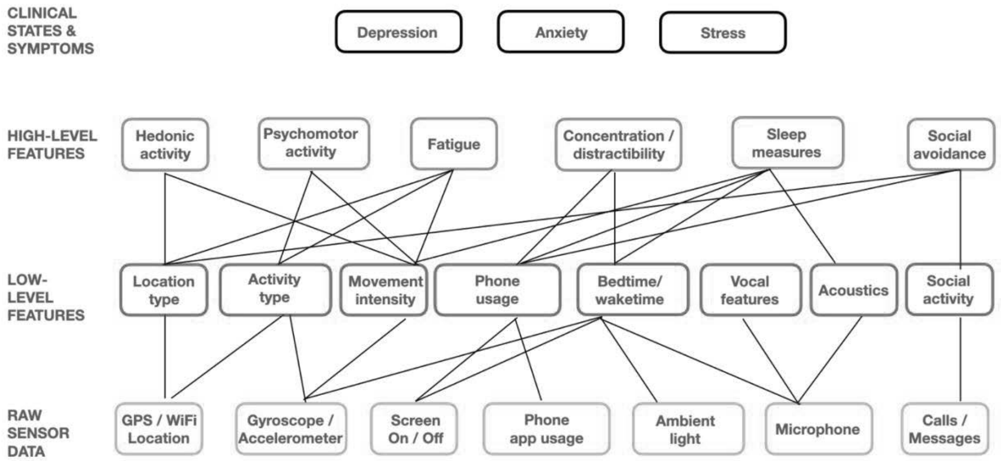
2.4.1 Digital phenotyping and its potential within mental health

Given the small-to-moderate effect sizes of existing treatments, one of the most promising methods of reducing the high burden of disease in depression is prevention (Campion et al., 2012; Cuijpers et al., 2012; Ormel, Kessler, et al., 2019). Until recently, understanding and predicting the onset and relapse of mental disorders has primarily been based on prospective population-based studies. Yet, whilst these broad insights are undoubtedly useful, the data often fails to capture the differences between individuals as well as the fine-grained temporal relationship between the causes and symptoms of mental ill-health (Torous et al., 2021). This issue is particularly relevant in the case of depression, which is a highly heterogeneous disorder and symptom profiles vary considerably (Goldberg, 2011). Smartphone devices may have a valuable role to play here as they provide longitudinal, multimodal, and temporally rich data that could generate insights into how disorders unfold at the individual level. Such data may be used to help identify individuals exhibiting symptoms or risk-factors for developing a disorder and thus enable more timely intervention.

Enter the relatively new field of digital phenotyping. Digital phenotyping has been defined as the moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices (Torous et al., 2016; Wisniewski et al., 2019). Digital phenotypes (also referred to as *digital biomarkers*) are of particular interest in clinical psychology and psychiatry where consistent biological markers are lacking, and existing diagnostic tools rely on self-report measures. Armed with an array of sensors, smartphone devices are able to automatically capture data related to movement (using the accelerometer), location (using the global positioning system, GPS), vocal biomarkers (using the microphone) and facial expression (using the camera). Furthermore, all of this data may be captured passively, reducing the burden of self-report often found with traditional active measures.

A simplified framework of how digital phenotyping data may be used to detect symptoms of depression is displayed in **Figure 1**. The bottom layer of the hierarchy is the raw sensor data from the smartphone which form the inputs to the digital phenotyping platform. These may include GPS data, gyroscope and accelerometer, microphone recordings or call logs. As raw sensor data alone is insufficient for making clinical inferences so first need to be transformed into features, the second layer of the platform. Features can be understood as higher-level abstractions of the raw sensor data and their construction is arguably the most important step in the process (Bengio et al., 2013). Features can be engineered using hand-crafted or automated approaches. In hand-crafted approaches, features are constructed “top down” on the basis of human expertise or knowledge and thus often represent pre-existing constructs. For example, raw data from the accelerometer, gyroscope and GPS sensors may be combined to create a feature called “activity type” (walking, running or driving); raw data from phone usage may be transferred into meaningful features such as the number of incoming and outgoing calls, call times and call duration and the ratios between them (Mohr et al., 2017). Alternatively, features can be engineered using machine learning techniques such as autoencoders (Lopez Pinaya et al., 2020) that are able to identify novel representations of the raw data. These representations may not map onto pre-existing constructs but may yield higher precision when attempting to predict certain higher-level markers or states (Z. S. Chen et al., 2022).

Figure 1. Example of a layered digital phenotyping sensemaking framework



Note. Abbreviations: GPS, global positioning system; SMS, short message service.
Adapted from “Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning”, by Mohr, D. C., Zhang, M., & Schueller, S. M. (2017) *Annual Review of Clinical Psychology*, *13*, 23–47. Copyright 2017 by “Annual Reviews”.

From lower-level features and sensor data, we may then derive higher-level features: cognitive and behavioral markers that represent thoughts, behaviors, and emotions. For example, a behavioral marker for social avoidance may include features such as in-phone social activity or location type. These higher-level markers are often developed using machine learning models that identify feature sets based on their ability to classify the relevant cognitive and behavioral markers, for example, sleep disruption based on features derived from the microphone (noise), phone camera (light) and gyroscope (picking up the phone) (Z. Chen et al., 2013).

Finally, clinical states, such as depression and anxiety may be predicted from a combination of high-level behavioral markers and low-level features. For example, depression may be identified from disruptions to sleep, social avoidance, activity levels and movement. One of the potential applications of digital phenotyping is the ability to identify specific symptoms associated with a disorder and thus identify individuals who are subsyndromal but nonetheless at risk of developing a full-blown disorder. In so doing, it also holds the potential to identify possible targets for those sub-clinical preventative interventions, for example, symptoms related to sleep disturbance or fatigue.

2.4.2 Digital phenotyping in depression research

A growing body of research is now emerging demonstrating the role of digital phenotyping data in identifying individuals with symptoms of mental disorders. One of the most widely studied sensors is the GPS sensor, from which a number of meaningful features related to movement and location can be derived. In a meta-analysis of studies assessing the relationship between digital phenotyping data and affective disorders, Rohani and colleagues (2018) found that the positive association between time spent at home and depressive symptom severity was the most consistent finding across studies. In a study of smartphone GPS and phone usage features that correlated with depression, Saeb and colleagues (2015) found that the regularity of participants' 24-hour movement patterns ($r = -0.63$), the variance of locations they visited ($r = -0.58$) and the proportion of time they spent at home ($r = 0.49$) were all associated with self-reported depression severity. Similar relationships were found by Farhan and colleagues (2016) in a study of university students, although the correlations were a lot lower.

Several studies have also found relationships between symptoms of depression and features derived from smartphone usage (Rohani et al., 2018). In an exploratory study of 25 participants, Mehrotra and colleagues (2016) found that phone usage time, the number of applications used, the number of times the phone was unlocked, the number of notifications received and the speed of responding to those notifications were all significantly correlated with self-reported symptoms of depression using the PHQ-8. Furthermore, correlations were moderate-to-high and highly significant based on only two weeks' worth of digital phenotyping data. In a clinical sample of 29 patients diagnosed with bipolar disorder, Faurholt-Jepsen and colleagues (2016) found that higher depressive symptom severity was associated with a higher number of incoming calls per day but a lower number of outgoing calls and fewer answered incoming calls.

However, not all findings have been consistent. In a year-long study of 13 patients diagnosed with bipolar affective disorder, Beiwinkel and colleagues (2016) found a significant negative correlation between the number of outgoing SMS text messages and self-reported symptoms of depression, whilst the study by Faurholt-Jepsen and colleagues (2016) found a significant positive correlation between the two. A study by Asselbergs and colleagues (2016) that captured data from 26 students over a 6-week period, found a negative correlation between mobile phone usage frequency and depressive symptoms, whilst the studies by Saeb and colleagues (2015) and Mehrotra and colleagues (2016) found a positive correlation.

Thus, although the initial findings related to the potential of smartphone data to predict symptoms of depression may be promising, the field is still young: most studies are observational and exploratory, sample sizes are often small, and a number of findings are conflicting. For the field to progress, new studies are needed aimed at replicating existing findings using larger sample sizes and providing a deeper understanding of under what conditions certain relationships may vary. For example, relationships between smartphone data and symptoms of mental health have been shown to differ across gender (Cho et al., 2016), cultures (Hernández et al., 2015), personality types (DeMasi et al., 2016) and between clinical- and non-clinical samples (Rohani et al., 2018).

2.4.3 Consumer wearables: a novel source of digital phenotyping data?

Over the past decade, the number of consumer wearable devices has increased dramatically, with over a billion connected wearable devices projected to be in use by the end of 2022 (Statisa, 2020). Wearable devices, such as the Apple Watch or Fitbit, capture a wide range of behavioral and physiological data, including step count, activity types, energy expenditure, heart rate, and sleep measures, all of which may be used to enrich digital phenotyping models. Moreover, unlike smartphones, these devices are often worn continuously by the individual, including at night, so may capture more accurate data (Piccinini et al., 2020).

Perhaps the most widely available sensors in consumer wearable devices are the accelerometer and gyroscope. Together, these sensors are able to track both the type and duration of physical activity. A large body of research already exists on the relationship between physical exercise and depression (Cooney et al., 2013; Schuch et al., 2018) and a number of large scale studies using laboratory-grade actigraph device data have demonstrated how accelerometer data may be used to identify individuals with symptoms of depression. A meta-analysis of studies using actigraphy data in patients with diagnosed depression found significantly less daytime activity in patients with depression compared with non-depressed individuals (SMD=-0.76) and an increase in daytime activity and reduction in nighttime activity over the course of treatment (Burton et al., 2013). In a study of patients with depression or bipolar disorder, Jacobson and colleagues were able to distinguish clinically diagnosed patients from healthy controls with 89% precision using just 2 weeks of actigraph data. Moreover, the actigraphy data was able to significantly predict symptom change during the two-week period, highlighting the potential of digital phenotyping data to identify clinical deterioration as it unfolds.

Sleep is another behavioral measure that is strongly associated with mental health, with decades of research identifying sleep disturbance as a common symptom of a number of disorders (Harvey et al., 2011; Sivertsen et al., 2009; Taylor et al., 2005). Disturbance in sleep can be detected by the timing of sleep (when a person falls asleep and when they wake up), the duration (how long an individual sleeps) and its quality (how long it takes for a person to fall asleep and how many times they wake up following sleep onset). The majority of studies to-date assessing the relationship between sleep and mental health have used polysomnography (PSG) to measure sleep. PSG is a multi-parametric test using a combination of EEG, blood oxygen levels, heart rate and eye and leg movements, typically carried out during the night and within the laboratory. Studies using PSG have found several variables associated with depression, including total sleep time, sleep latency (the time taken to fall asleep), number of awakenings and the duration of awakenings after sleep onset (Ilanković et al., 2014).

A number of consumer wearable devices are now commercially available that capture data related to the main measures of sleep, offering the potential to use these devices in larger populations. One example is the Oura Ring device (www.ouraring.com). Launched in 2013, the Oura Ring is the first consumer wearable device that has demonstrated validity in measuring common measures of sleep and high agreement with PSG measures (de Zambotti et al., 2019). However, to-date no research has been carried out assessing whether sleep data derived from a validated consumer wearable device may be used to predict symptoms of mental health. Given the potential for these devices to collect data within larger populations, in naturalistic settings, over longer periods of time and at considerably less cost, this a promising area for future research that may provide the foundation for consumer wearable data to eventually play a role in supporting clinical decisions within traditional healthcare.

2.4.4 Using machine learning to increasing model accuracy

For predictive models to have clinical utility they need to be able to make generalizable predictions about individuals. Generalizability can be defined as how accurately a statistical model generated in one group performs in a new group of individuals (Dwyer et al., 2018). In recent years, a number of researchers have demonstrated the limitations of classical statistical approaches in providing generalizable predictions that translate into clinical practice (Yarkoni & Westfall, 2017). One of the issues is the assumption of linearity in many classical approaches such as linear and logistic regression, which rarely represent the true underlying relationships in the data (Ernst & Albers, 2017). Another is the reliance on human expertise or *a priori* theoretical constructs for predictor selection that may fail to identify important predictors or interactions and thus limit the model performance (Chowdhury & Turin, 2020). Finally, the majority of classical statistical models are evaluated based on their “goodness of fit” between the statistical model and the sample data and/or the extent to which the size and direction of the regression coefficient align with the proposed theoretical model. However, such an approach often leads to the problem of *overfitting*, where the model describes features that arise from the sampling or measurement error in the data rather than the underlying population distribution (Yarkoni & Westfall, 2017). As a result, the model often produces overly optimistic results and fails to perform well when applied to other samples drawn from the same population.

Over the past decade, machine learning methods have become increasingly used within clinical decision-making to identify at-risk individuals and individual response to treatment

(Chekroud et al., 2016; Lee et al., 2018; Shatte et al., 2019). Developed from the study of pattern recognition and computational learning, machine learning uses algorithms to learn the relationships between complex variables by minimizing the error between the predicted and observed outcomes (Dreiseitl & Ohno-Machado, 2002). Machine learning methods may offer an alternative approach to classical statistical methods that address some of their limitations (Weng et al., 2017). For example, whilst traditional statistical methods analyze the relationship between variables sequentially, machine learning methods are able to iteratively and simultaneously analyze multiple interacting relationships between variables (Orrù et al., 2012). In so doing, they may be better placed to identify non-linear and higher-dimensional patterns in the data, especially in large "big data" sets. Machine learning may also improve the identification of which variables in a dataset are relevant and irrelevant for predicting specific outcomes and reveal patterns in the data that do not necessarily map to preexisting theoretical frameworks (Bzdok & Meyer-Lindenberg, 2018).

To minimize prediction error, i.e., predict future observations as accurately as possible, machine learning models typically use a combination of *cross validation* to assess a model's performance and how it might be improved, and *regularization* to prevent overfitting (i.e., maximize generalizability). Cross-validation involves training and testing a model on different samples of data (Browne, 2000). Although the most robust way to assess the performance of a model is often to assess whether findings replicate in a completely independent dataset, this is not always possible. Cross-validation works by randomly splitting a given dataset into two sets: a *training* set and a *test* set. The training set is used to fit the model and the training set is then used to quantify the prediction error of the trained model. By ensuring that none of the data points in the test set are used to fit the model in the training set, the problem of overfitting is reduced. However, it comes at the cost of potentially *underfitting*. To address this, the most common approach is k-fold cross validation where datasets can be split into k-number of "folds" (typically 3-10) (Dwyer et al., 2018). One by one, each fold is selected as the test set and the other sets are combined into the training set. The process is then repeated until all folds have been used as the training set. The overall model performance is then assessed by averaging the test performance scores of the folds.

Cross-validation provides a simple but powerful method of estimating how well a model will generalize to new data that addresses the problem of overfitting caused when the same data is used to both train and test a model. To then *prevent* the model from overfitting, machine learning practitioners typically use a method called *regularization*. Regularization involves "penalizing" the cost function of a model as it grows more complex. One of the most commonly employed regularization techniques is lasso regression (Tibshirani, 1996). Unlike classical linear or logistic regression models which will often produce a nonzero coefficient for every variable in the model due to the existence of some statistical association between each predictor and the outcome variable, lasso regression identifies small coefficients with low relative contribution and *shrinks* them to zero. To avoid being penalized, lasso regression ensures that a coefficient is only retained if the incremental predictive utility outweighs the penalty being applied (where the size of the penalty parameter can be modified). The largest benefits from regularization are often obtained when the number of potential predictors is large relative to the sample size. Here, classical approaches based on ordinary least squares will often overfit the data, whilst lasso regression will prioritize stronger associations in the data and ignore small variations and thus be more likely to generalize to out of sample datasets.

Due to the large, complex and high dimensional feature space, machine learning models may lend themselves particularly well to predicting class membership (e.g., whether someone is depressed or at risk of depression) based on digital phenotyping data. A number of different machine learning algorithms exist, including Support Vector Machines (SVM), Random Forest (RF), XGBoost (XGB), K-Nearest Neighbors (KNN) and Neural Networks (NN). Whilst it is beyond the scope of the current thesis to explore these models in detail, a growing body of evidence is emerging of the ability of machine learning models to identify individuals with depression. For example, Farhan and colleagues (2016) were able to predict clinically-diagnosed depression status from iPhone GPS data with a precision of 0.93, recall of 0.73 and specificity of 0.97 using SVM and RF classifiers. Furthermore, models that combined digital phenotyping data with PHQ-9 scores outperformed models using PHQ-9 scores alone, suggesting that digital phenotyping data may have captured relevant features of the DSM-IV diagnosis that were not reflected in the PHQ-9 scores.

To-date, no research has assessed the extent to which machine learning models applied to digital phenotyping data from a smartphone and a scientifically validated consumer wearable device may be used to identify individuals at-risk of depression. The ability to develop simple classification tools from digital phenotyping data (e.g., individuals with above normal levels of depressive symptoms) could provide another critical component in the translation of digital phenotyping data into real-world clinical practice.

3 Aims

As digital interventions are becoming increasingly adopted within healthcare systems across the globe, an in-depth understanding of the efficacy and limitations of digital mental health interventions is both timely and important. The current study begins with the largest meta-analytic review of digital interventions for the treatment of depression conducted to-date (Study I). Here, we assess the efficacy and effectiveness of digital interventions as well as which participant, intervention and study-related factors influence outcomes. We then proceed to address one of the major challenges in digital interventions, namely treatment dropout (Study II). Here, we assess which participant variables may predict dropout in a digital intervention for depression in patients with comorbid physical illness and whether intervention usage variables may improve the prediction of dropout early on in the intervention when it matters the most. Finally, in an attempt to look forward to the future of the field, we assess to what extent digital phenotyping data from smartphone and wearable devices may be used to predict symptoms of depression. Here, we examine the role of novel features derived from wearable devices and assess the accuracy of machine learning models in the classification of at-risk individuals (Studies III and IV).

We address the following research questions:

The Efficacy & Effectiveness of Digital Interventions for Depression (Study I)

1.1 Are digital interventions effective in reducing symptoms of depression?

1.2 Is there a difference in outcomes between control types?

1.3 What factors moderate outcomes across participant characteristics, intervention-related variables, study design and setting?

Predicting Dropout in a Digital Intervention for Depression (Study II)

2.1 What participant factors predict dropout in a digital intervention for depression?

2.2 Can dropout be predicted from intervention usage data?

2.3 Which set of predictors leads to the highest predictive accuracy?

Using Digital Phenotyping to Predict Depression Symptom Severity (Study III + Study IV)

3.1 Can features derived from smartphone and consumer wearable device data be used to predict symptoms of depression?

3.3 Which digital phenotyping features have the strongest predictive power?

3.3 How accurately can machine learning models classify individuals with *normal* and *above normal* levels of depressive symptoms using smartphone and wearable data? Which features are the most important in distinguishing the two groups?

4 Methods

4.1 Study I: Meta-analysis on the efficacy of digital interventions for the treatment of depressive symptoms

4.1.1 Identification and selection of studies

Studies were identified in a three-step procedure: First, we searched the Cochrane Central Register of Controlled trials (CENTRAL), PsycINFO, EMBASE and MEDLINE for relevant articles. The search was conducted on October 13, 2020. Second, we checked the reference lists of relevant existing systematic reviews and meta-analyses (Andersson & Cuijpers, 2009; Barak et al., 2008; Baumeister et al., 2014; Carlbring et al., 2018; Cuijpers et al., 2008, 2011; Firth, Torous, Nicholas, Carney, Pratap, et al., 2017; Karyotaki et al., 2017, 2018; Königbauer et al., 2017; Linardon et al., 2019; Spek, Cuijpers, et al., 2007; Weisel et al., 2019). Third, we conducted backward searches in all included articles. The full texts of all relevant articles were then obtained.

Studies were included if (a) they included participants of any age, (b) participants had elevated symptoms of depression at baseline, as measured by validated self-reported or clinician-rated depression scales (i.e., studies included individuals with and without a clinical diagnosis) (c) treatment was provided via a computer offline (e.g., via CD-ROM) or online over the internet, or via a smartphone device (including both via the internet and via native apps), (d) the study was an RCT with an inactive control condition (i.e., waitlist control or no treatment) or active comparison condition (treatment as usual (TAU), attention control, face- to-face psychotherapy). The selection process was conducted by two independent reviewers. Disagreements were resolved by a discussion among the reviewers. If needed, a third reviewer was consulted. The agreement between the reviewers was good in both the title and abstract screening (88.5%, κ = .61) and full-text assessment (96.5%, κ = .72).

The systematic review was registered with the International Prospective Register of Systematic Reviews (PROSPERO CRD42019136554). Further details of the methodology and analysis are provided in the study protocol published in advance (Moshe et al., 2020).

4.1.2 Data extraction

The following data were extracted and coded by two independent reviewers. Any disagreements were solved in discussion. If not indicated otherwise, perfect agreement was reached between the reviewers.

Participant Characteristics

We extracted (a) age, (b) gender (percentage of females), (c) target population, (d) presence of somatic comorbidities, and (e) baseline severity levels. Baseline severity was analyzed as a continuous variable to avoid bias resulting from categorization (healthy, mild, severe depression). As the PHQ-9 was the most frequently used scale, we used the PHQ-9 scores where available. For all other studies, the information was recoded to PHQ-9.

Depression Outcome measures

All outcome measurements were extracted (e.g., self-report and clinician ratings). We included outcomes at different assessment times and details of instruments used. The mean and standard deviation for all intervention and control conditions within a study were coded for the calculation of the effect size.

Intervention Components

To account for the different types of interventions used in the study, we extracted (a) the number of intervention modules, and (b) the theoretical orientation of the intervention: third-wave, cognitive behavioral therapy, psychodynamic therapy (PDT), problem-solving therapy (PST), life-review therapy (LRT), other.

Guidance

We extracted (a) the type of guidance: unguided, therapeutic guidance and technical guidance, (b) qualification of the guiding personnel: "high" (MSc or Diploma degree in psychology, or professional psychotherapist, psychiatrist or psychotherapist in training), or "low" (BSc, other qualifications, or mixed coding), (c) communication mode (synchronous or asynchronous), (d) average guidance time for each participant in minutes. In unguided interventions no human support was involved, i.e., interventions were completely self-guided. In guided interventions with technical guidance, the support was restricted to solving technical problems and motivating patients to adhere to the intervention. In interventions with therapeutic guidance, the support was extended to include content and processes related to the treatment and was of a genuine therapeutic nature.

Adherence Outcome Measures

Adherence was extracted for intervention adherence and assessment completion. The extracted variables were the proportion of participants completing assessments, the proportion of participants completing the first intervention module, the average number of intervention modules completed and the proportion of participants completing all modules.

Design and Study Features

We extracted the following design and study features: (a) year of publication, (b) type of control, (c) sample size, (d) region (Asia, Europe, North America, Oceania, multiple, other), and (e) efficacy or effectiveness trial setting.

Risk of Bias Assessment
We assessed Risk of Bias as “low,” “unclear,” or “high” separately across six domains: (a) “selection bias”; (b) “performance bias”; (c) “detection bias”; (d) “attrition bias”; (e) “reporting bias”; and (f) “other bias.”

4.1.3 Meta-analytic procedure

Effect size (Hedges’ *g*) was calculated as the post-test difference between the mean of the intervention condition and the mean of the control group divided by their pooled standard deviation, adjusted for sample size (Hedges, 1981). Intention to treat (ITT) data was used in the analysis.

As many studies contained multiple outcomes for depression, multiple assessment time points and multiple comparator conditions, we calculated separate effect sizes for each. To account for the dependencies within a study, we used a three-level meta-regression model with random effects (Assink & Wibbelink, 2016; Harris Cooper & Hedges, 2009; Pastor & Lazowski, 2018). The three-level model avoids biases caused by the pooling of different effect sizes within a study, where the correlations between outcomes are not reported. The three levels of the model represented the three different variance components: sampling variance of the extracted effect sizes at level one; variance between the extracted effect sizes from the same study at level two; and variance between studies at level three (Assink & Wibbelink, 2016).

The average ES of digital interventions was calculated using an intercept-only model. To control for potential differences in baseline depression severity, for example introduced by post-randomization attrition, baseline depression level was included as a covariate in all analyses unless indicated otherwise. Subgroup analyses and meta-regression were used to assess the influence of predictors. Profile likelihood plots were used to check for overparameterization and identifiability.

To assess the influence of control type on outcomes, subsets were created for each control type (as well as for active and inactive control conditions). Meta-regression was used to test for significant differences between control type(s). Similarly, we calculated the ES for each delivery modality and tested for significant differences. To answer the question of long-term efficacy and effectiveness of digital interventions, meta-regression with assessment time as the predictor was conducted, testing for linear, quadratic, and cubic change in ES over assessment time.

To examine the influence of guidance on outcomes, we calculated the ES for the subsets of studies containing unguided interventions, interventions with technical guidance and interventions with therapeutic guidance. Using meta-regression, we then tested for significant differences between technical guidance versus unguided interventions and interventions with therapeutic guidance versus unguided interventions. To assess whether the qualification of the person providing guidance influenced outcomes, we tested for an interaction effect between the qualification level and therapeutic guidance on ES. Similarly, we tested for an interaction effect between guidance time and therapeutic guidance on outcomes to determine whether a dose-response relationship existed between guidance time and outcomes. To assess whether there was a difference in outcomes between efficacy and effectiveness trials, we calculated the pooled ES separately for each study design and then tested for significant differences using meta-regression. Similar procedures were used to assess the influence of the following moderators on reported ES: (a) pretreatment depression severity, (b) somatic comorbidities, (c) gender, (d) age, (e) therapeutic approach, (f) number of modules, (g) study quality (item-wise RoB), and (h) year of publication.

Finally, we explored intervention adherence using a three-level random effects meta-regression model. Intervention adherence was defined in two ways: (a) the percentage of participants that completed all modules (= completer rate) and (b) the average percentage of modules completed by a participant (= module completion rate). A subset for each of these was created and meta-regression was used to test for a significant influence of adherence on outcomes. Only studies reporting information on intervention adherence were included.

Study Heterogeneity and Variance Components

Heterogeneity was calculated using the I2 statistic (Borenstein et al., 2017). Profile likelihood confidence intervals were also calculated (Borenstein et al., 2017; Jackson et al., 2014). We expected high heterogeneity (above 75% (Ioannidis et al., 2007)) based on the findings of previous meta-analyses and the extended time frame in the present study.

Small Study Effects and Publication Bias

We used a funnel plot to detect potential biasing effects by small studies. Here, we plotted the ES against the precision (standard error). Asymmetry would indicate bias. Asymmetry was tested using the Egger’s test adapted to the three level structure of the present meta-analysis (Egger et al., 1997): the influence of precision as a predictor on effect size was tested in the three-level meta-regression model. However, instead of the standard error, we used the weight as a predictor in the meta-regression model, as using SE tends to over-reject due to artifactual correlations with ES (Pustejovsky & Rodgers, 2019).

4.2 Study II. Predictors of dropout in a digital intervention for the treatment of depression: secondary analysis of two RCTs

4.2.1 Study design

This was a secondary analysis of data from two trials assessing the efficacy of a therapist-guided internet-based intervention for the treatment (Baumeister et al., 2020) and prevention (Sander, Paganini, et al., 2020) of depressive symptoms in patients with comorbid chronic back pain

(CBP). Both trials were observer-masked, multicenter, pragmatic randomized controlled trials with a parallel design. The trials were conducted simultaneously using the same intervention, procedures and research setting but targeted individuals with different levels of depressive symptomatology at intake. In Baumeister et al., 2020, participants were diagnosed with depressive disorder and chronic back pain; in Sander et al., 2020, participants had subclinical mild levels of depressive symptoms. For the purposes of the current study the trial data was therefore combined. All procedures were approved by the ethics committee of the Albert-Ludwigs University of Freiburg, Germany.

4.2.2 Participants

All participants (N=253) assigned to the intervention arms of the primary studies were included in the current analysis. The inclusion criteria of the primary studies were: (1) age 18 years and older, (2) depressive symptoms, either meeting DSM-IV criteria for a mild-to-moderate depressive episode or persistent depressive disorder (Baumeister et al., 2020) or reported persistent subthreshold depressive symptoms in the past 3 months (Sander et al., 2020), (3) diagnosed back pain chronicity of at least 6 months, (4) German language skills, and (5) internet and PC access. The exclusion criteria were: (1) having ongoing or planned psychotherapy within the forthcoming 3 months, (2) being currently suicidal or having had suicidal attempts within the past 5 years, or (3) having had a severe depressive episode within the past 6 months. In the primary studies, participants were recruited during or following discharge from one of 82 orthopedic clinics across Germany. They were recruited personally via a clinician or online using flyer and information letters distributed by the clinic.

4.2.3. Intervention

The intervention is a guided internet- and mobile-based intervention for the treatment (eSano BackCare-D (Lin et al., 2017)) or prevention (eSano BackCare-DP (Sander et al., 2017)) of depression in patients with comorbid CBP. The content of the intervention is based on cognitive behavioral therapy for depression and includes elements of psychoeducation, social skills, problem solving, behavioral activation, relaxation, motivation for physical exercises and psychological pain intervention elements. Modules consist of information provided by text, video, audio and interactive exercises and include a homework assignment. At the start of each module, participants report their perceived stress level at the time and whether they had experienced any negative events in the previous 7 days. There are six regular modules and three optional modules focusing on sleep, partnership/sexuality, and work. Participants were advised to complete one session per week. During the intervention, participants were guided by trained and supervised psychologists (e-coaches) who provided written feedback within 48 hours of each completed module and by answering any queries.

4.2.4. Measures

Baseline measures

For the current study, eight baseline characteristic variables were assessed as potential predictors of dropout. Variables were chosen on the basis of previous research pointing to demonstrated or hypothetical relationships between the predictor variables and intervention adherence or dropout (Bremer et al., 2020; Christensen et al., 2009; Karyotaki et al., 2015; Kok et al., 2017; Schmidt et al., 2019). Demographic characteristics included age, gender (male or female), education level (based on the International Standard of Education low: level 1-2, medium: level 3-4, high: level 5+, UNESCO, 2017)), marital status (single, in a relationship, divorced/widowed) and social support (low, medium, high). Clinical characteristics included depression, as measured by the Hamilton Depression Rating Scale (HAM-D) (Hamilton, 1960), pain disability as measured by the Oswestry Disability Index (ODI) (Mannion et al., 2006) and pain self-efficacy as measured by the Pain Self-Efficacy Questionnaire (PSEQ). The Process variables included internet affinity, as measured by an Internet Affinity Scale (IAS). More details on all measurements are provided in the original study protocols (Lin et al., 2017; Sander et al., 2017).

Intervention Usage Measures

Intervention usage measures included both active and passive measures. The active measures were the stress level reported by the participant at the start of each module ("Burden") and the occurrence of any negative events experienced in the past seven days, self-reported by the participant at the start of each module ("Negative Events"). Burden was assessed using a Likert scale of 0-10, where 0="Not burdened at all" and 10="Extremely burdened". Negative events was dummy coded as 0="no negative event in the past week", 1="at least one negative event in the past week". For passive measures, we included the number of days taken to complete each module ("N Days to complete module") and the number of minutes spent online completing each module ("Time spent online completing module").

Dropout

Dropout was defined as completing less than 6 of the intervention modules, in accordance with the intervention developers (Lin et al., 2017; Sander et al., 2017). It was operationalized as a binary outcome (dropped out / did not drop out).

4.2.5 Statistical analysis

Predicting dropout from participant baseline characteristics

To assess whether participant baseline characteristics could predict dropout, analyses were conducted using logistic regression in 3 steps. First, we conducted a series of bivariate analyses to assess the Odds Ratios (ORs) of each baseline variable at a time (the "bivariate model"). Second, we repeated the analyses with all baseline variables simultaneously entered into the binomial model (the "complete model"). Finally, we built a "parsimonious model" in which we excluded non-significant predictors with no incremental predictive power from the complete model in a stepwise procedure.

Akaike information criterion (AIC) and Bayesian information criterion (BIC) were used as measures of model fit and for model comparison. For nested models, Likelihood Ratio Tests (LRTs) were used to directly compare whether two models were significantly different from one another (Hosmer & Lemeshow, 2000). Collinearity was assessed using Variance Inflation Factors (VIF) and Tolerance (1/VIF). The assumption of linearity of the logit (a linear relationship between the predictors and dropout) was assessed for all continuous predictor variables and any variables found violating the assumption were transformed based on visual inspection of the plot.

As this was an exploratory study we did not adjust for multiple testing. The study was not powered for confirmatory analysis of the predictors and alpha adjustment may have increased the likelihood of Type II errors.

Predicting dropout early on in the intervention

To assess whether we could identify people at risk of dropout early on in intervention, we first created a subset of the data available up until the point of module 1 completion (i.e., baseline assessment data and intervention data captured until participants had completed the first module). We then compared the performance of 3 separate logistic regression models using the constrained dataset: (1) a model based on participant baseline characteristics only: the "baseline characteristics model" (2) a model based on intervention-usage variables only: the "intervention usage model", (3) a model combining all baseline characteristics and intervention-usage variables: the "combined model". The quality of the models was assessed using the Area Under the Receiver Operating Characteristic curve (AUROC) and related measures of sensitivity and specificity (Hosmer & Lemeshow, 2000). The optimal threshold for the AUROC was determined using Youden's J statistic (Youden, 1950).

Sensitivity analyses were conducted to assess whether predictors differed in the prevention and treatment studies. Here, study was included as a dummy-coded variable (0=PROD-BP, 1=WARD/BP) in all parsimonious models, first as an additional predictor to assess for a main effect of study type on dropout and then as an interaction term with other predictors in the model to assess whether the effect of a predictor differed across studies.

To assess whether the number of modules completed influenced the relative risk of dropout, we conducted sensitivity analyses using Cox proportional hazards regression (Cox, 1972). Here, we assessed whether significant predictors of dropout differed between the two methods. Analyses were conducted according to procedures outlined in Eysenbach (2005). The number of completed modules was used as a proxy for time. Models were built using the same 3-step procedure outlined above for logistic regression.

Principle for handling missing data

Missingness occurred in 111 (3.6%) data points and was assumed to be missing at random (MAR), meaning that missingness depended on observed data (Enders, 2010). To avoid bias introduced by missingness, missing data was imputed using multiple imputation by chained equations (MICE) (Van Buuren, 2018; van Buuren & Groothuis-Oudshoorn, 2011). Predictors for missing values were selected based on (1) model induced predictors, (2) predictors based on bivariate correlation and (3) bivariate correlation with missingness, according to the procedures outlined by van Buuren & Groothuis-Oudshoorn (2011). Predictive mean matching was used as the imputation method. The number of imputed data sets was set to 20 and the number of iterations to 10. Convergence was visually assessed and confirmed. Regression analysis was performed on each imputed data set and results were pooled according to Rubin's rule (Rubin, 1996). Sensitivity analyses were conducted using observed (non-imputed) data to compare with the results of the complete models using imputed data

4.3 Study III + IV: Predicting symptoms of depression and anxiety from smartphone and wearable data

4.3.1 Study design

This was an observational study with repeated measurements conducted over 30 days. Measurements consisted of daily digital phenotyping features extracted from smartphone and wearable data, daily mood reports and questionnaires conducted at baseline, midpoint, and the end of the study.

4.3.2 Participants

We recruited 60 adult participants in April 2020, using posts on online communities and social media sites. Participants were eligible if they were (a) at least 18 years old, (b) able to speak and read English, (c) owned an iPhone with access to the internet and (d) owned an Oura Ring. Participants signed an online consent form agreeing with the study protocol, data collection and analysis. The study was exempt from formal ethical committee approval according to the local ethical guidelines in the conduct of research (*Guidelines for Ethical Review in Human Sciences | Tutkimuseettinen Neuvottelukunta*, 2020). As compensation for participation, participants received a mental health and well-being report at the end of the study.

After completing the consent form, participants received an email link to download the *Delphi* iOS smartphone app for Apple smartphones. The Delphi app was used to collect all study data including questionnaire data, mental health outcomes, mood assessments and digital phenotyping data. Participants were requested to uninstall the app at the end of the study.

4.3.3 Measures

Mental Health Outcomes

Symptoms of depression and anxiety were assessed at baseline (T0), midpoint (T1; 16 days) and at the end of the study (T2; 31 days) using the Depression Anxiety Stress Scales (DASS-21). The DASS-21 is a 21-item short form of the DASS (Lovibond & Lovibond, 1995). It measures depressive mood, anxiety, and chronic tension/stress during the past week (e.g., "I was aware of dryness of my mouth"; "I couldn't seem to experience any positive feeling at all."). All items are rated on a 4-point Likert scale ranging from 0 ("did not apply to me at all") to 3 ("applied to me very much or most of the time"). The subscores range from 0 to 21, with higher subscores indicating more severe symptoms. The DASS-21 has shown high internal consistency for all subscales in previous administrations (Antony et al., 1998). To provide categorical definitions of depression severity, we used the cut-off values provided by the DASS-21. A score of 0–4 was categorized as "normal", 5–6 as "mild", 7–10 as "moderate", 11–13 as "severe", and 14+ as "extremely severe (Antony et al., 1998; Lovibond & Lovibond, 1995).

Ecological Momentary Assessment (EMA) of Mood

Notifications were sent by the Delphi app 3 times per day asking participants to report their mood. Notifications were randomized within a 30-minute window during the morning, afternoon and evening (i.e., ~09:00, 14:30, and 20:00). Mood was assessed using the circumplex model of affect (Russell, 1980). The model conceptualizes mood as a two-dimensional construct comprising of valence (positive/negative) and arousal (low/high). A single item question ("How are you feeling right now?") was used to assess mood. To answer the questions, participants were provided with two 9-point response scales from −4 to 4 (low to high) representing the two dimensions of mood. The scales were set to zero by default.

Smartphone Sensor Data

Delphi uses the AWARE open source framework (Ferreira et al., 2015; Nishiyama et al., 2020) to collect raw data from smartphone sensors. In the current study we collected data from the Battery, GPS, Screen and Timezone. Additionally, we used the ESM Scheduler plugin to deliver the EMAs. The data collected by the app is first stored locally on the participant's device. The data is then uploaded onto a secure cloud server when the device is connected to the internet via Wi-Fi. To ensure data protection and participant privacy we used application permissions, certificates, secure network connections and user authentication. Additionally, AWARE obfuscates and encrypts all data using one-way hashing of logged personal identifiers. For further details, please refer to Ferreira and colleagues (2015) and Nishiyama and colleagues (2020).

Wearable Data: Sleep, Activity & HRV

The Oura Ring was used to measure sleep, activity and heart rate variability (HRV). To assess sleep, we measured participants' total sleep time (TST), total time in bed (TIB), sleep onset latency (SOL) and wake after sleep onset (WASO). The Oura Ring has been demonstrated to have high agreement with polysomnography (the gold-standard for measuring sleep) in the aforementioned sleep variables (de Zambotti et al., 2019). To assess activity, we measured the participant's step count (the number of steps as determined by the Oura Ring's 3D accelerometer) and metabolic equivalent for task (MET). MET is a standardized measurement of the amount of energy used by the body during physical activity, as compared to resting metabolism (Jetté et al., 1990). One MET is defined as the energy the body uses at rest. The Oura Ring was also used to provide measures of average night-time HRV. The device calculates HRV using the root mean square of successive differences between heartbeats (RMSSD) and has been shown to have high agreement with ECG, the gold standard for measuring heart rate variability (Kinnunen et al., 2020).

Location Features

As sensor data is recorded using the UNIX timestamps, we converted each sensor data entry into a local date using the participant's timezone data. Data was then aggregated at the day level. To maximize accuracy of location data, all duplicate entries were removed, as well as GPS coordinates with latitude 0.0 and longitude 0.0. Preprocessing and extraction of the location features were computed according to the procedures detailed in Saeb and colleagues (2015). Prior to extracting the features, each GPS location data sample was coded as representing a stationary state (e.g., at home) or transition state (e.g., moving outside). This was calculated using the movement speed at each location, determined by its time derivative. A transition state was defined as a speed more than 1 km/hour. K-means clustering (Arthur & Vassilvitskii, 2007) was then applied to the stationary state data to identify the locations where participants spent the majority of their time (location clusters). Five location features were then extracted from the GPS data: *Total Distance* (the total number of kilometers traveled by the participant during the specified time period), *Location Variance* (the variability in participants' GPS locations), *Location Entropy* (the variability of the time participants spent at the location clusters), *Normalized Location Entropy* (a measure of entropy that is invariant to the number of clusters a participant spent time at), and *Time at Home* (the proportion of time a participant spent at home relative to other location clusters). Details on how each feature was calculated can be found in the Methods section of Study III.

Phone Usage Features

We extracted two features related to phone usage: *Phone Usage Frequency* (the number of times a participant interacted with their phone during the specified time period), and *Phone Usage Duration* (the total number of minutes a participant interacted with their phone during the specified time period). Interactions were calculated based on a screen unlocking event. Usage session duration was calculated as the time from when the phone was unlocked until it was locked.

4.3.5 Statistical analyses

Predicting Depression Symptom Severity as a Continuous Outcome

Prior to the analyses, all independent variables (smartphone and wearable digital phenotyping features and EMA mood data) and dependent variables (DASS-21 scores for depression and anxiety) were synchronized to the relevant study day. To ensure that all variables in the analysis reflected the same time period, each independent variable was then pooled for the first two weeks of the study and the second two weeks of the study, thereby aligning with the timing of the DASS-21 midpoint and endpoint measurements.

Predicting Mental Health Symptom Severity From Smartphone and Wearable Data

Multilevel regression models (MLM) were used to assess the association between digital phenotyping features and the scores on the DASS-21 depression and anxiety subscales (H. Goldstein, 1995; MacCallum et al., 1997; Nezlek, 2012). MLMs were used to take into account that the data was nested within persons, i.e., the observations are not independent (Nezlek, 2001) and reduce the likelihood of Type I errors (Musca et al., 2011). In the current study, the repeated measures (level 1) are nested within the person (level 2). Intraclass correlations (ICC) were used to confirm the necessity of multi-level models (all ICC > 0.05).

MLMs with random intercepts and random slopes were applied separately to four sets of independent variables: GPS location features (total distance, location variance, entropy, normalized entropy, and time at home); smartphone usage features (usage duration and usage frequency); wearable device data (step count, MET, TST, SOL, WASO, TIB, and HRV); and EMA mood data (valence and arousal). All variables were z-standardized.

The intercept represents the average depression and anxiety scores across the study and the beta coefficient (slope) models the relationship between smartphone and wearable data and mental health scores. A 2-tailed p value < 0.05 was considered statistically significant.

Analyses were conducted using a three-step procedure. First, we conducted a series of bivariate analyses for each predictor to investigate its association with depression and anxiety scores. Second, we built multivariate "combined" models using predictors that were significant in the bivariate analyses. Third, we conducted likelihood ratio tests (LRTs) to compare whether models that contained more predictors were significantly superior to those with less predictors. All models were fitted using maximum likelihood (Enders, 2010; Singer & Willett, 2009; van Buuren & Groothuis-Oudshoorn, 2011).

Missing Data Handling

Multiple imputation was used to handle missing data. Missingness occurred in 10% of the DASS-21 assessment and 9.1% of the sensor data. Data was assumed to be missing at random (MAR) (Enders, 2010; Revelle, 2020). The imputation model followed guidelines for multilevel multiple imputations (van Buuren & Groothuis-Oudshoorn, 2011). Predictive mean matching for multilevel was used as the imputation method. The number of imputed data sets was set to 20 and the number of iterations to 15. Convergence was visually assessed and confirmed (van

Buuren & Groothuis-Oudshoorn, 2011). Regression analysis was performed on each imputed data set and the results pooled using the Rubin's rule (Rubin, 1996).

Predicting Depression Symptom Class with Machine Learning

For the machine learning analyses, participants were divided into two groups: "normal" and "above normal" based on their levels of depressive symptom severity at baseline (T0). Participants in the normal group had a score of 0-9 on the DASS-21 depression subscale, whilst participants in the "above normal" group had a score greater than nine. The scores correspond to the DASS-21 cut-offs (Lovibond & Lovibond, 1995).

52 day-level (24 hours from midnight to midnight) features were computed from the smartphone and wearable data using the processes described above. Details of all features can be found in supplemental materials of Study IV. To clean the data, all features with zero variance and more than 15% of missing data were first excluded. Subsequently, participants with less than 15 days of data were removed as well as days with more than 30% of missing data. After the data cleaning process, there were 54 participants with 1,556 days (mean 28.81, range 21-30 per participant) and 49 digital phenotyping features. On average, participants were missing 7.56% (range 0% -12.98%) of data values.

For the predictive analyses, participant group (normal with label 0 and above normal with label 1) was modeled as a function of mood ratings (Valence and Arousal), Demographics (Gender and Age) and digital phenotyping features (GPS location, Phone usage, Sleep, and Physical activity). Each participant's daily digital phenotyping features and mood ratings as well as their age and gender were labeled with their depressive symptom group status. Population-based models leveraging five supervised machine learning (ML) algorithms were then created: Support Vector Machines (SVM), Random Forest (RF), XGBoost (XGB), K-Nearest Neighbors (KNN), and Logistic Regression (LR). These ML algorithms were selected based on their common usage in supervised classification and the fact that they are easy to train and interpretable (Garcia-Ceja et al., 2018).

A robust machine learning model training approach using nested cross-validation was used to reduce the chances of model overfitting A time-series aware leave-one-participant-day-out and stratified three-fold cross-validation for the outer and inner cross-validation, respectively, was also employed. For each iteration of the nested cross-validation, one participant's day was designated as the test set, and the rest of the participants' dataset were designated as the training set. For time-series awareness, all training set samples recorded after the test set were removed from the training set. This ensured that future dataset was not used to predict the past. The inner cross-validation was for missing data imputation, feature scaling, feature selection and hyperparameter optimization of the classifiers. The hyperparameters of the classifiers were optimized using grid search over a predefined set of parameters (see Supplementary Table 3 in Study IV for more details). Missing data imputation was carried out separately for each nested cross-validation iteration. Missing data was imputed in the train set (participants' data for training the model) separately per participant. For training set imputation, a Bayesian Ridge Regression iterative feature imputation process was used (Pedregosa et al., 2011). The method uses all other features with no missing data as predictors. Missing values in the test set (i.e., one record of a participant's day) were imputed with the mean of the corresponding feature in the training set. Similar imputation of the test set and training set can be found in Low and colleagues (2021) and Poulos and colleagues (2018).

All features were scaled using min-max scaling. Feature scaling on the test set was applied using the min-max parameters of the training set. To mitigate biases in the output of the machine learning models, the imbalanced training set was handled by oversampling the minority class with the synthetic minority over-sampling technique (SMOTE). The 45 best features were then selected based on the mutual information between the features and the target (participants' depression status). The SHAP (SHapley Additive exPlanations) method (Lundberg et al., 2020) was used to compute feature importance.

The predictive performance of the ML models was evaluated using the area under the receiver operating characteristic curve (AUC), F1, F1 Macro, Accuracy, Recall, and Precision metrics. Three baseline classifiers were used as a benchmark for the performance of the ML algorithms: (1) a naive classifier that predicts only the Majority Class (MC), (2) a Decision Tree (DT) classifier trained (same training approach as ML classifiers) with only the demographic dataset, and (3) a Random Weighted Classifier (RWC), that is, ten thousand randomly generated predictions according to the multinomial distribution of the normal and above normal group labels. F1, Precision, Recall metrics are reported for above normal (F11, Precision1, Recall1) and normal groups (F10, Precision0, Recall 0).

5 Results

5.1 The efficacy & effectiveness of digital interventions for depression (Study I)

5.1.1 Study selection

A total of 14,513 articles were examined from the following databases: CENTRAL (N = 3,711), Embase (N=4,333), Medline (N=3,764), PsycINFO (N=2,705). After removing duplicates, we screened 7,651 studies by title and abstract. In total, 88 articles covering 83 unique studies met the inclusion criteria and were included in the analysis. **Figure 2** shows the PRISMA flow diagram.

5.1.2 Characteristics of included studies

Table 1 displays a descriptive summary of the characteristics of the studies included in the meta-analysis (k = 83). A total of k = 79 (95%) of studies focused on adult populations. The mean age across all studies was M = 41.33 (SD = 9.68). Overall, 69.5% of participants were women. The majority of participants were reported to have mild-to-moderate depression symptom severity (recoded PHQ-9: M = 12.91, SD = 2.95). The majority of studies were conducted in Europe (k = 51, 61%), followed by North America (k = 15, 18%), and Australia and New Zealand (k = 13, 16%). Only three studies were conducted in Asia (4%) and one study in South America (1%). We did not find any studies conducted in Africa.

CBT was the most common theoretical orientation (k = 67, 74.4%), followed by third-wave (k = 9, 10.0%), PST (k = 7, 7.8%), PDT (k = 1, 1.1%), LRT (k = 1, 1.1%), and other (k = 5, 5.6%; e.g., combined approaches). A total of k = 72 (80%) interventions were guided, of which k = 47 (52.2%), provided therapeutic guidance and k = 25 (27.8%) provided technical guidance; k = 18 interventions (20%) were unguided. The average number of intervention modules was M = 7.4 (SD = 2.1). In most studies, the intervention was delivered via the Internet (k = 75, 90.36%). Only four studies (4.82%) reported on the ES of computer-based interventions: two on smartphone-based apps and Internet combined interventions (2.41%), and two on interventions using smartphone-based apps exclusively (2.41%). Of the k = 83 studies, k = 62 (74.7%) studies were conducted in efficacy settings compared with k = 21 (25.3%) in effectiveness settings.

WLC was the most common comparator used across studies (k = 43, 46.7%), followed by TAU (k = 24, 26.1%) attention control (k = 19, 20.7%), individual F2F (k = 3, 3.2%), group F2F (k = 2, 2.2%) and other (k = 1, 1.1%).

Figure 2. PRISMA flowchart

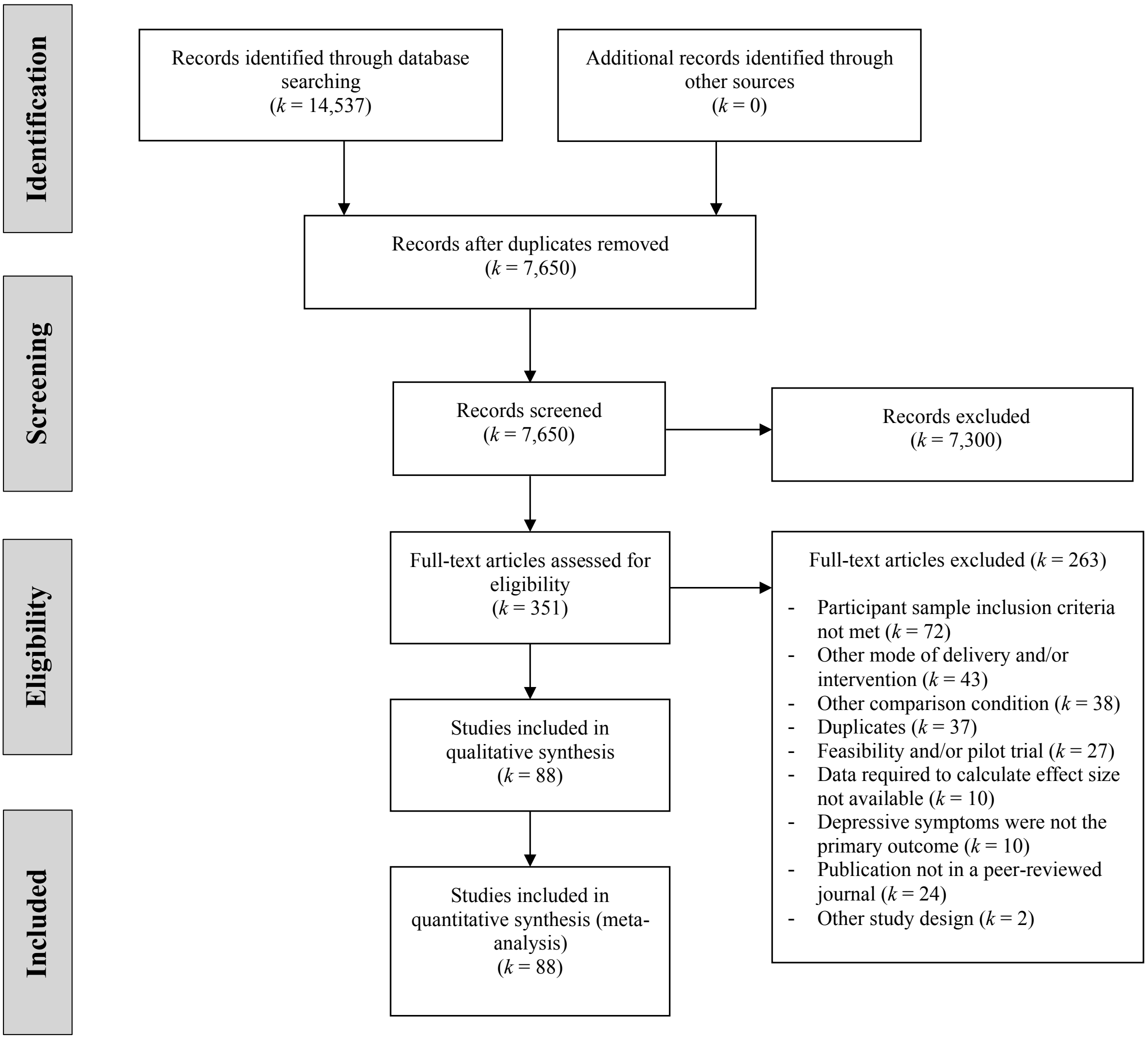

Identification
Screening
Eligibility
Included
Records identified through database searching (k = 14,537)
Additional records identified through other sources (k = 0)
Records after duplicates removed (k = 7,650)
Records screened (k = 7,650)
Records excluded (k = 7,300)
Full-text articles assessed for eligibility (k = 351)
Full-text articles excluded (k = 263)
- Participant sample inclusion criteria not met (k = 72)
- Other mode of delivery and/or intervention (k = 43)
- Other comparison condition (k = 38)
- Duplicates (k = 37)
- Feasibility and/or pilot trial (k = 27)
- Data required to calculate effect size not available (k = 10)
- Depressive symptoms were not the primary outcome (k = 10)
- Publication not in a peer-reviewed journal (k = 24)
- Other study design (k = 2)
Studies included in qualitative synthesis (k = 88)
Studies included in quantitative synthesis (meta-analysis) (k = 88)

Table 1. Summary of the Characteristics of the Studies included in the Meta-Analysis (k=83)

Name	**Total**
Number of studies	83
Participant characteristics	
Age mean (*SD*)	41.3 (9.7)
Females	69.5%
Target populations	
Children and adolescents	5 (6%)
Adults	76 (91.6%)
Older adults (>50 years)	2 (2.4%)
Baseline severity (PHQ-9)	12.9 (2.9)
Comorbid diseases	15 (18.1%)
Intervention characteristics	
Guidance	
Therapeutic	47 (52.2%)
Technical	25 (27.8%)
Unguided	18 (20.0%)
Number of modules	7.3 (2.2)
Theoretical orientation	
3rd-Wave	9 (10.0%)
CBT	67 (74.4%)
LRT	1 (1.1%)
DYN	1 (1.1%)
PST	7 (7.8%)
Other	5 (5.6%)
Study design	
Passive control	
Waitlist control	43 (46.7%)
Active control conditions	
Treatment as usual	24 (26.1%))
Attention control	19 (20.7%)
Face-to-face	3 (3.2%)
Group face-to-face	2 (2.2%)
Other[a]	1 (1.1%)
Setting	
Efficacy	62 (74.7%)
Effectiveness	21 (25.3%)
Sample size	
Total N	15530
Mean (*SD*)	173.4 (148.0)
Location	
Europe	51 (61%)
Australia & New Zealand	13 (16%)
North America	15 (18%)
Asia	3 (4%)
Africa	0 (0%)
South America	1 (1%)

Note. Abbreviations: CBT: cognitive behavioral therapy; LRT: life review therapy; DYN: psychodynamic therapy; PST: problem solving therapy. [a] Psychoeducation with weekly guidance (Johansson 2012a).

Overall risk of bias was low across all items other than the blinding of participants and outcome assessors, which were coded as having high overall risk of bias (RoB). However, it should be noted that in psychotherapy research masking of participants is generally not feasible (Munder & Barth, 2018) and thus self-report ratings might be susceptible to bias. In assessing whether RoB influenced reported ES, we found no significant difference in outcomes between low- and high-risk studies across all RoB items (all *p* > .05).

5.1.3 Overall effect size: improvements in depression severity compared with control groups

Selected characteristics of the 83 trials included in the meta-analysis are displayed in **Appendix 2**, grouped by number of participants, mean age, control type, primary outcome measures, delivery method, guidance type, effect size of intervention compared to control condition (Hedges *g*), intervention completion rate, percentage of completers, and country in which the study was conducted. Unless otherwise stated, all findings report outcomes adjusted for baseline differences.

The overall effect size superiority of digital interventions over control groups was *g* = .52, 95% CI [.43, .60], *p* < 001. Heterogeneity was high: I2 = 84, 95% CI [57, 100]. Effect size varied considerably across studies depending on the comparator used. **Table 2** displays the pooled ES for digital interventions broken down by control type. In studies with WLC comparisons, the between-group ES was medium-to-large with a pooled ES of *g* = .70, 95% CI [.58, .83] *p* < .001. In studies using attention control conditions, the average ES for digital interventions was small to moderate: *g* = .36, 95% CI [.19, .54], *p* < .001. In studies using TAU control conditions, the average ES was also small to moderate: *g* = .31, 95% CI [.21, .41], *p* < .001.

We found no significant differences in overall outcomes in studies comparing digital interventions with individual face-to-face therapy (*g* = .01, 95% CI [2.73, 2.72], *p* = .982). However, we found only three studies comparing digital interventions and individual face-to-face therapy. Moreover, all of these involved interventions with human guidance. For group face-to-face therapy we identified only two studies, which provided a total of three data points at post. Again, there was no significant difference in outcomes between the two conditions (*g* = .17, 95% CI [2.91, 3.26], *p* = .609).

Table 2. Pooled Effect Size (Hedges g) of Digital Interventions Versus Type of Control Condition

Control type	***g***	**95%-CI**	***p*-value**	***I^2***
All control conditions	0.52	0.43 to 0.60	<.001	84 (57 to 100)
Passive control conditions				
WLC	0.70	0.58 to 0.83	<.001	79 (43 to 100)
Active control conditions				
TAU	0.31	0.21 to 0.41	<.001	60 (0 to 100)
Attention	0.36	0.19 to 0.54	<.001	84 (41 to 100)
Face-to-face	-0.01	-2.73 to 2.72	.982	<.001 (0.00 to 100)

Note. Hedges g according to the random-effects model. Abbreviations: WLC: waitlist control, TAU: treatment-as-usual. I^2: heterogeneity

5.1.4 Factors moderating effect size

Participant characteristics

There was a significant influence of baseline depression severity on outcomes (z-standardized: β = .12, 95% CI [.04, .20], *p* = .005), indicating that individuals with higher depression severity benefit more from digital interventions compared to individuals with lower baseline symptom severity. Although there was no influence of age on outcomes using z-standardized age as a predictor in meta-regression models, we found only sparse evidence assessing the efficacy of digital interventions for children and adolescents (k = 4). A subset analysis on these four studies yielded a nonsignificant effect of digital interventions (*g* = .15, 95% CI [1.34, 1.63], *p* = .708). We found no influence of gender on outcomes (β = .03, 95% CI [.06, .11], *p* = .517). The existence of a comorbid somatic condition did not have an impact on outcomes either (β = .05, 95% CI [.13, .23], *p* = .551).

Delivery Method

Subset analyses comparing different delivery methods against all control conditions revealed an average effect size of *g* = .53, 95% CI [.43, .62], *p* < 001 for Internet-based interventions, *g* = .45, 95%CI [.42,1.31], *p* =.151 for computer-based interventions and *g* = .39, 95% CI [.27, 1.06], *p* = .122 for smartphone-based interventions. Despite the difference in point estimates, meta-regression found no significant effect of delivery method on overall effect size (*p* > 0.05). However, it is worth noting that only one trial (Guo et al., 2020) demonstrated the efficacy of a stand-alone smartphone app intervention on the reduction of depressive symptoms.

Adherence

Adherence was operationalized as (a) the percentage of participants that completed the full intervention (completer rate), and (b) the average percentage of modules completed by a participant (module completion rate). A total of 49 studies contained information on the percentage of completers and 36 studies on the module completion rate. The intercept-only model yielded an average completer rate of 53.49%, 95% CI [44.62%, 62.37%], *p* < .001 and an average module completion rate of 67.85%, 95% CI [59.00%, 76.07%], *p* < .001. In the completer subset, the overall ES of digital interventions was *g* = .65, 95% CI [.53, .78], *p* < .001. Meta-regression found a strong influence of adherence on outcomes: the estimated increase in ES if all participants were completers was estimated at β = .57, 95% CI [.04, 1.10], *p* = .037.

Given the strong influence of adherence on ES, additional analyses were conducted to identify moderators of adherence. The average module completion rate in unguided interventions was 53.67%, 95% CI [34.00%, 73.35%], *p* < .001, compared with 60.90%, 95% CI [40.67%, 81.14%], *p* < .001 in interventions with technical guidance and 76.31%, 95% CI [65.76%, 86.85%], *p* < .001 in interventions with therapeutic guidance. A comparison of module completion in guided vs. unguided interventions showed a significant effect of β = 22.81%, 95% CI [5.18%, 40.43%], *p* = .016, and a nonsignificant effect for technical guidance compared with unguided interventions, β = 7.73%, 95% CI [15.26%, 30.72%], *p* = .474.

Similar differences across guidance formats were found for the percentage of completers. In unguided interventions, the percentage of completers was 38.11%, 95% CI [7.87%, 68.35%], *p* = .022. In interventions with technical guidance, the percentage of completers was 50.30%, 95% CI [26.68%, 73.91%], *p* = .001 and in interventions with therapeutic guidance 56.36% of participants completed the full intervention (95% CI [47.95%, 64.76%], *p* <.001).

There was also a significant difference in adherence between efficacy and effectiveness trials. The percentage of completers in effectiveness trials was estimated at 25.22%, 95% CI [10.95%, 39.48%], *p* = .004, compared to 60.89% in efficacy trials. The percentage of module completion in effectiveness trials was estimated at 53.61%, 95% CI [41.70%, 65.53%], *p* < .001, compared to 74.62% in efficacy trials. Neither age nor gender significantly influenced adherence (*p* <.05). Temporal analyses found no significant change in intervention adherence in the two decades between 2000 and 2020, irrespective of the operationalization used.

Publication Year

In assessing whether the reported outcomes of digital interventions have changed over time, meta-regression using publication year as a predictor found no significant change in effect size over the past two decades (*p*<0.05).

5.1.5 Efficacy versus effectiveness settings

The overall ES of digital interventions compared to all controls in effectiveness studies was *g* = .30, 95% CI [.15, .45], *p* < .001 compared with *g* = .59, 95% CI [.50, .69], *p* < .001 in efficacy trials. Overall ES was significantly lower in effectiveness trials compared with efficacy trials (β = .30, 95% CI [.11, .48], *p* = .002) explaining 13.5% of the between-study variance.

Subset analyses assessing differences between control conditions, revealed that studies using WLC conditions reported the highest average effect size (*g* = .81, 95% CI [111, 2.74], *p* = .161). TAU control conditions yielded a significant overall ES superiority of *g* = .30, 95% CI [.19, .41], *p* < .001. Interestingly, effectiveness trials using attention control conditions showed a null finding of *g* = .01, 95% CI [.51, .49], *p* = .939. Finally, in the only effectiveness trial comparing a digital intervention against a face-to-face control (Wagner et al., 2014), the difference was nonsignificant, *g* = .01, 95% CI [.79, .76] .

5.1.6 The influence of human guidance on outcomes

Subset analyses revealed that the average ES of unguided interventions compared with all control conditions was *g* = .34, 95% CI [.24, .45], *p* < .001. For interventions with technical guidance, the pooled ES was *g* = .46, 95% CI [.29, .62], *p* < .001. For interventions with human therapeutic guidance the pooled ES was *g* = .63, 95% CI [.50, .76], *p* < .001. There was a significant difference in outcomes between interventions with human therapeutic guidance and interventions with technical guidance (β = .22, 95% CI [.03, .41], *p* = .024).

In interventions where human therapeutic support was provided, the qualification level of the person providing guidance (high versus low) did not have a significant influence on outcomes (β = .17, 95% CI [12, .46], *p* = 254). Nor did we find an interaction between time and guidance on effect size (β = .00, 95% CI [.01, .01], p = .316), revealing no significant dose-response relationship between the amount of guidance provided and outcomes.

Finally, in assessing whether the influence of guidance on outcomes differed between efficacy and effectiveness settings, we found a significant interaction (β = .10, 95% CI [.03, .17], *p* = .014) between study setting (efficacy vs. effectiveness) and guidance (unguided versus guided interventions), suggesting that guidance may be especially important to achieving outcomes in real-world settings.

Publication bias

A funnel plot of the ES of included studies was created to assess for publication bias and small study effects (see **Figure 3**). The funnel plot shows an asymmetrical distribution of reported effect sizes: studies with larger sample sizes (and thus higher precision) tended to report lower effect sizes, suggesting that studies with smaller sample sizes finding low or negative outcomes may not have been published. This visual finding was further corroborated by the modified Egger's regression model which revealed a significant negative effect of precision on ES (β=-

0.29, 95%-CI: -0.07 to -0.51, p=.016). As effectiveness trials tend to have higher sample sizes, we tested the modified Egger's regression model separately on the subsets of efficacy and effectiveness trials. Here we found that the Egger's test was only significant in the efficacy trials subset (β=-0.31, 95%-CI: -0.57 to -0.05, p=.029) and not in the effectiveness trials subset (β=-0.12, 95%-CI: -0.47 to 0.22, p=.272), indicating that bias may only be present in studies conducted in efficacy settings, but not those carried out in effectiveness settings.

Figure 3a. Funnel Plot to Assess for Publication Bias by Relating Effect Sizes to Standard Errors

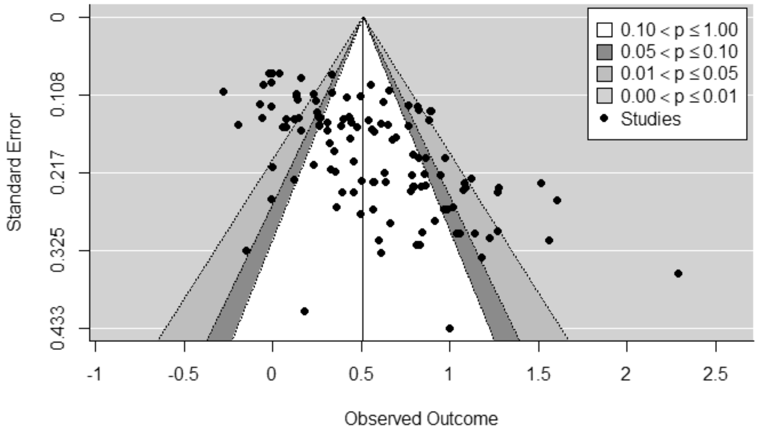


Figure 3b. Efficacy: Funnel Plot to Assess for Publication Bias by Relating Effect Sizes to Standard Errors

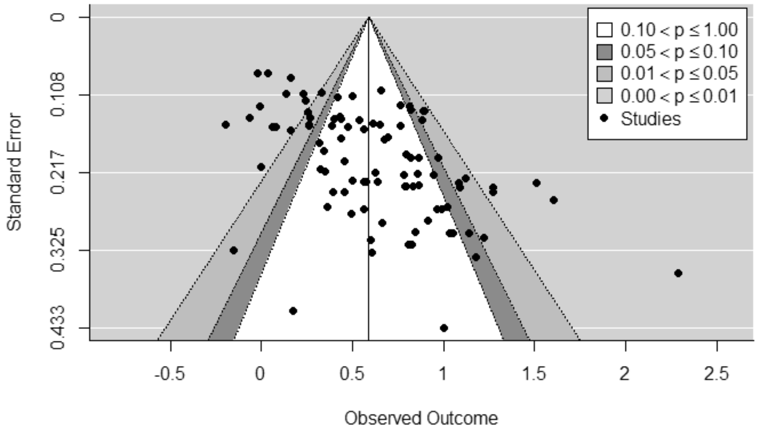
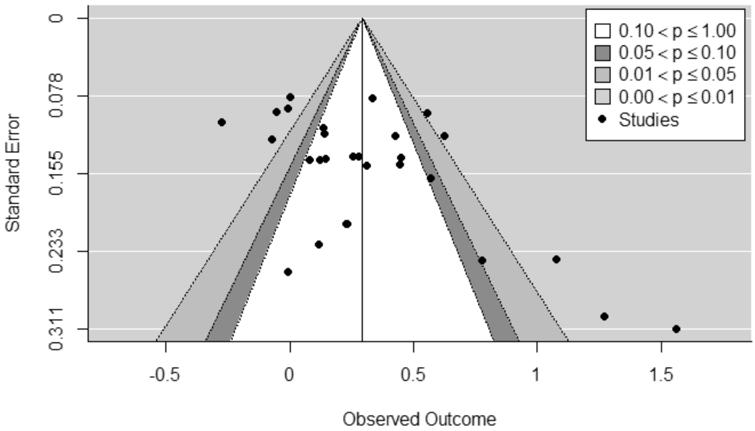


Figure 3c. Effectiveness: Funnel Plot to Assess for Publication Bias by Relating Effect Sizes to Standard Errors



5.2 Predicting dropout in a digital intervention for depression (Study II)

5.2.1 Descriptive characteristics

Among the 253 participants, 149 (58.9%) were female and 104 (41.1%) were male. Age ranged from 24 to 78, with a mean age of 51.1 (SD 8.88). 171 (67.6%) participants reported a low level of education. 34 (13.4%) were single, 180 (71.1%) were in a relationship or married and 39 (15.4%) were divorced or separated. The mean depression severity at baseline was 10.3 (SD = 5.93) as measured by the HAM-D and 9.94 (SD = 4.41) as measured by the PHQ-9. The mean level of pain disability was 31.3 (SD=14.7) as measured by the ODI and the mean level of pain self-efficacy was 34.9 (SD = 13.0), as measured by the PSEQ. **Table 3** provides a detailed summary of the demographic and clinical characteristics of the sample.

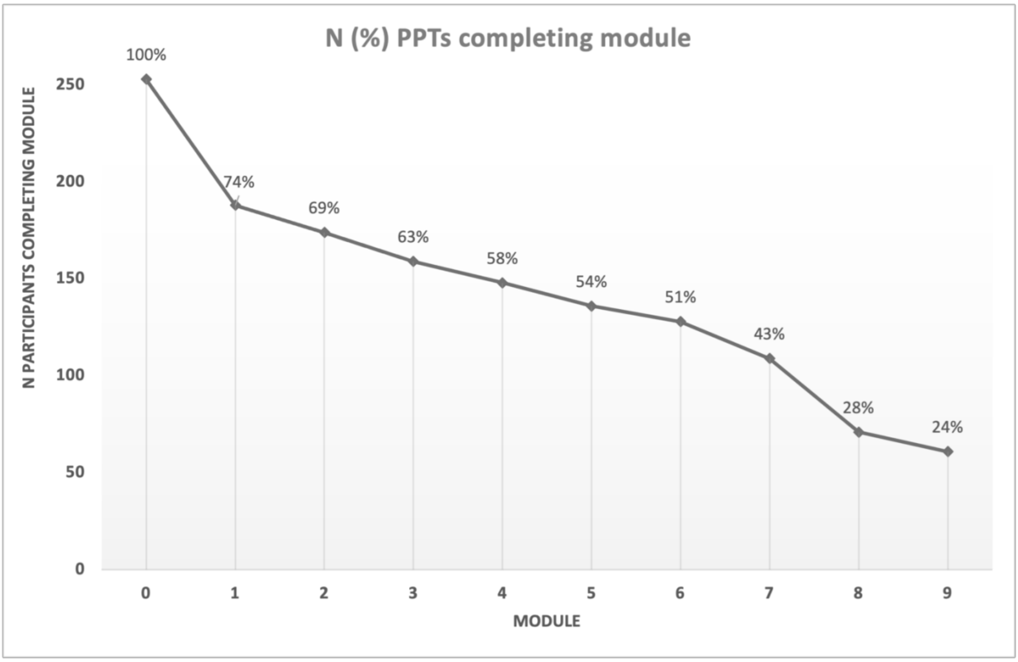
On average, participants completed 4.65 out of the six regular and three optional modules (SD=3.48). Participants took an average of 17.64 days (SD=19.55) to complete each module and the mean time online taken to complete a module was 80.26 minutes (SD=136.96). The mean self-reported burden was 4.55 (SD=1.97) and the mean number of self-reported negative events across the intervention was 0.80 (SD=1.33). **Figure 4** shows that a total of 114 out of 253 participants (45.1%) dropped out of the intervention prior to completing at least 6 modules and that the number of participants completing the modules decreased steadily as the intervention progressed in time.

Table 3. Demographic and clinical characteristics of sample (N=253)

Variable	N (%) / Mean (SD)
Age	
Mean (SD)	51.1 (8.88)
Gender	
male	104 (41.1%)
female	149 (58.9%)
Education Level	
low	171 (67.6%)
medium	45 (17.8%)
high	37 (14.6%)
Marital Status	
single	34 (13.4%)
in a relationship (incl married)	180 (71.1%)
divorced or separated	39 (15.4%)
Children	
yes	200 (79.1%)
no	53 (20.9%)
Social Support	
None	9 (3.6%)
Low	67 (26.5%)
Sufficient	81 (32.0%)
High	73 (28.9%)
Very high	23 (9.1%)
IAS	
Mean (SD)	9.33 (4.00)
HAMD-D	
Mean (SD)	10.3 (5.93)
PHQ-9	
Mean (SD)	9.94 (4.41)
Pain disability (ODI)	
Mean (SD)	31.3 (14.7)
Pain self-efficacy (PSEQ)	
Mean (SD)	34.9 (13.0)
Dropout	
No	139 (54.9%)
Yes	114 (45.1%)

Note. Values are based on observed data

Figure 4. Module completion rates for the eSano BackCare-D/DP digital intervention



5.2.2 Participant characteristics as predictors of dropout

Table 4 displays the performance of the models used to predict dropout based on participant baseline characteristics. The results of the bivariate analysis indicated that having a lower level of education was significantly associated with higher risk of dropout (OR=2.43, 95% CI:1.19 to 4.97, *p*=.018), whilst higher age predicted lower risk of dropout (OR=0.97, 95% CI: 0.94 to 0.99, *p*=.015). None of the other potential predictors (gender, social support, internet affinity, baseline depression severity and baseline pain intensity) were statistically significant at the level of *p*<0.05 in the bivariate analysis.

In the complete model, being single was found to be an additionally significant predictor of dropout (OR=2.53, 95% CI: 1.09 to 5.88, *p*=.031). When age was added as a quadratic term (age^2) to the model to account for the non-linear relationship between age and dropout, we found that both age (OR=0.63, 95% CI: 0.47 to 0.84, *p*=0.002) and age^2 (OR=1.55, 95% CI: 1.17 to 2.05, *p*=0.003) were significant predictors, such that both lower and higher age were associated with increased risk of dropout.

In the parsimonious model, where predictors were stepwise reduced to relevant predictors only, low education (OR=3.33, 95% CI: 1.51 to 7.32, p=.003) and age (OR=0.62, 95% CI: 0.47

to 0.82, p= 0.001; age^2: OR=1.55, 95% CI: 1.18 to 2.04, p=0.002) remained as significant predictors of dropout. Marital status, internet affinity, baseline depression severity and baseline pain intensity were found to be non-significant after controlling for the other predictors.

Table 4. Predictors of dropout from participant baseline characteristics

Predictors	**Bivariate Model**			**Complete Model**			**Parsimonious Model**		
	OR	**95% CI**	***p***	**OR**	**95% CI**	***p***	**ORs**	**95% CI**	***p***
Age	0.97	0.94-0.99	0.02*	0.63	0.47-0.84	0.00**	0.62	0.47-0.82	0.00***
Age^2				1.55	1.17-2.05	0.00**	1.55	1.18-2.04	0.00**
Gender (male)	1.60	0.96-2.66	0.07.	1.68	0.96-2.94	0.07.			
Marital status:									
Single vs in a relationship	1.97	0.93-4.20	0.08.	2.54	1.09-5.90	0.03*			
Divorced/widowed vs in a relationship	0.54	0.26-1.14	0.11	0.62	0.27-1.42	0.25			
Education:									
Low vs medium	2.43	1.19-4.97	0.01*	3.77	1.68-8.49	0.00**	3.33	1.51-7.32	0.00**
High vs medium	1.88	0.75-4.71	0.18	2.08	0.74-5.83	0.16			
Social support:									
Low vs high	0.83	0.45-1.53	0.55	0.83	0.41-1.69	0.60			
Medium vs high	1.60	0.88-2.90	0.13	1.64	0.86-3.14	0.13			
IAS	1.02	0.96-1.09	0.53	1.02	0.95-1.10	0.58			
HAMD	0.99	0.95-1.03	0.67	0.98	0.93-1.03	0.36			
Pain Disability	1.00	0.98-1.02	0.85	1.00	0.97-1.02	0.69			

Note. OR = odds ratio, CI = confidence interval, IAS = Internet Affinity Score, HAMD = Hamilton Depression Rating Scale

5.2.3 Predicting dropout early on in the intervention

In the second set of analyses, we assessed whether participant baseline characteristics and intervention usage variables could be used to predict dropout following completion of the first module. In the parsimonious model using only participant baseline characteristics, higher and lower age (OR=0.61, 95% CI: 0.43 to 0.87, *p*= 0.006; age^2: OR=1.58, 95% CI: 1.12 to 2.24, *p*=0.01) and low education (OR=3.60, 95% CI: 1.19 to 10.88, *p*=.023) were significant predictors of dropout. The AUROC for the model was 0.70, the sensitivity was 68% and the specificity was 62%.

In the parsimonious model using only intervention usage data, a higher number of days to module completion predicted higher risk of dropout (OR=1.04, 95% CI:1.01 to 1.07, *p*=0.005) whilst a self-reported negative event in the previous week was associated with lower risk of dropout (OR=0.30, 95% CI: 0.11 to 0.81, *p*=0.018). The AUROC for the model was 0.61, the sensitivity was 56% and the specificity was 54%.

In the parsimonious model that combined participant baseline characteristics and intervention usage variables as predictors, higher and lower age (OR=0.61, 95% CI: 0.43 to 0.87, *p*=0.006; age^2: OR=1.58, 95% CI: 1.12 to 2.24, *p*=0.01), medium versus high social support (OR=3.40, 95% CI: 1.33 to 8.64, *p*=.011) and a higher number of days to module completion (OR=1.05, 95% CI:1.02 to 1.08, *p*=0.004) all predicted higher risk of dropout, whilst a self-reported negative event in the previous week was associated with lower risk of dropout (OR=0.22, 95% CI: 0.07 to 0.68, *p*=0.009). The AUROC for the model was 0.72, the sensitivity was 76% and the specificity was 59%.

A comparison of the parsimonious models based on participant baseline characteristics and intervention usage variables can be found in **Table 5**. The model that combined baseline and intervention usage variables was the most accurate in predicting dropout (AIC=187.51, BIC=210.16) and significantly more accurate than the model using participant baseline characteristics only (AIC= 207.77, BIC=223.95), $\chi^2(181)=5.32$, *p*=.006.

Sensitivity analyses

Sensitivity analyses assessing whether findings differed between the treatment (WARD-BP) and prevention (PROD-BP) studies found no significant difference between the two, either in terms of main effect or interaction effects with other predictors. Sensitivity analyses assessing whether findings differed when using Cox proportional hazards regression versus logistic regression found no difference in the significant predictors. Sensitivity analyses assessing whether the results differed between models using observed data and those using imputed data revealed no difference in the predictors found to be significant. Results from the sensitivity analyses can be found in the Supplemental Materials of Study II.

Table 5. Predictors of dropout following completion of Module 1: model comparison

Model	AIC	BIC	AUROC	Sensitivity	Specificity
Model 1: Baseline Variables	212.57	254.65	0.70	68%	62%
Model 2: Intervention Variables	207.48	223.66	0.61	56%	54%
Model 3: Baseline + Intervention Variables	198.88	253.9	0.72	76%	59%

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion, AUROC = Area under the receiver operating curve

5.3 Using digital phenotyping to predict depression symptom severity (Study III + IV)

5.3.1 Descriptive characteristics

Of the 60 participants recruited to the study and who completed the baseline questionnaire, 1 participant (1.7%) dropped out due to concerns over privacy, 2 participants (3.4%) dropped out due to burden of self-report and 2 participants (3.4%) dropped out for unknown reasons. Of the remaining 55 participants, 47 (85.5%) completed the midpoint questionnaire and 54 (98.2%) completed the endpoint questionnaire.

In total, 30 out of the 55 included participants (54.5%) were female and 25 (45.5%) were male. Age ranged from 24 to 68 with a mean of 42.8 (SD 11.6). The majority of participants (80%) had a bachelor's or higher degree, 17% reported high school as their highest education and 2% reported having no secondary education. The mean depression severity was M = 3.78, SD = 3.48 (normal: 67.3%, mild-moderate: 25.4%, severe: 7.3%), the mean anxiety severity was M = 2.73, SD = 2.68 (normal: 70.9%, mild-moderate: 23.7%, severe: 5.5%). **Table 6** provides a detailed summary of all participants included in the final analysis.

Table 6. Participant characteristics

Variable	% / M (SD)	n
Age	42.8 (11.6)	55
Gender		
Female	54.5	30
Male	45.5	25
Ethnicity		
Asian / Pacific Islander	1.8	1
Hispanic or Latino	3.6	2
White	92.7	51
Other	1.8	1
Education		
Less than a high school diploma	1.9	1
High school degree or equivalent	16.7	9
Bachelor's degree (e.g. BA, BS)	44.4	24
Master's degree (e.g. MA, MS)	33.3	18
Doctorate (e.g. PhD, EdD)	1.9	1
Prefer not to say	1.9	1
Employment		
Employed full-time (35+ hours a week)	49.1	27
Employed part-time (less than 35 hours a week)	7.3	4
Unemployed (currently looking for work)	1.8	1
Unemployed (not currently looking for work)	3.6	2
Student	5.5	3
Retired	1.8	1
Self-employed	21.8	12
Unable to work	7.3	4
Prefer not to say	1.8	1
Marital Status		
Single (never married)	25.5	14
Married	49.1	27
In a domestic partnership	20.0	11
Divorced	5.5	3
Mental Health Status at Baseline		
Depression	3.78 (3.48)	-
Anxiety	2.73 (2.68)	-
Stress	6.00 (3.82)	-

Note: Values are based on observed data

5.3.2 Predicting depression symptom severity from smartphone data

From the GPS-derived location features, we found that location variance was negatively associated with subsequent depressive symptom severity (β = −0.21, SE = 0.10, t(81) = −2.13, *p* = 0.037). None of the other GPS-derived features (total distance, location entropy, normalized location entropy and time at home) were found to be significantly associated with symptoms of depression. We found no significant association between smartphone usage duration or usage frequency and symptoms of depression.

5.3.3 Predicting depression symptom severity from wearable data

From the Oura Ring, we found a significant relationship between total sleep time and symptoms of depression (β = 0.24, SE = 0.11, t(73) = 2.33, *p* = 0.023) and time in bed and symptoms of depression (β = 0.26, SE = 0.11, t(59) = 2.39, *p* = 0.020). Additionally, we found a significant association between WASO and anxiety β= 0.23, SE = 0.11, t(90) = 2.13, *p* = 0.035].[1] No significant association was found between the physical activity measures (MET and steps) and symptoms of depression. From the EMA mood data, we found that valence (positive or negative) was significantly related to depression [β= −0.39, SE = 0.11, t(55) = −3.43, *p* = 0.001].

5.3.4 Combining passive and active data for superior model performance

A comparison of models derived from active EMA data and passive smartphone and wearable data revealed that the model based on EMA data was more accurate at predicting depressive symptoms than the model based on smartphone and wearable data, but the combination of the two yielded the best fit. See **Table 7** for a comparison of all models.

5.3.5 Predicting normal vs above normal depression severity status using ML models

The predictive performance of ML classifiers trained with demographics, mood ratings and digital phenpotyping features and the performance of the three baseline models can be found in **Table 8.**

[1] Although anxiety disorders are not the focus of the current thesis, given the high rates of comorbidity between anxiety and depression and the fact that anxiety is a risk-factor for depression, the finding is worth highlighting.

Table 7. A comparison of MLM model performance on the prediction of depression

Model	Fixed effects		Goodness of fit and Comparison		
	Estimate	**p-value**	**AIC**	**BIC**	**P value**
Baseline model					
Intercept	0	>.999	271.12	279.22	-
EMA model					
Intercept	0	>.999	258.25	269.05	.001[a]
Arousal	-0.39	.001			
GPS model					
Intercept	0	>.999	267.59	278.39	.033[a]
Variance	-0.21	.035			
Extended digital phenotyping model: GPS and wearable data					
Intercept	0	>.999	261.57	275.07	.024[b]
Variance	-0.21	.033			
TIB	0.25	.018			
Combined model: EMA and digital phenotyping model					
Intercept	0	>.999	247.46	263.66	.001[c]
Variance	-0.21	.023			.005[d]
TIB	0.24	.014			
Arousal	-0.38	.001			

Note. [a] Comparison against baseline, [b] Comparison against GPS, [c] Comparison against extended sensing model, [d] Comparison against EMA model

Table 8. Performance of population models classifying normal versus above normal depressive symptom status.

Models	Accuracy %	AUC %	F1 Macro %	Precision1 %	Recall1 %	F11 %	Precision0 %	Recall0 %	F10 %
Baseline1: MC	74.07	50.00	42.55	0.00	0.00	0.00	74.07	100.00	85.11
Baseline2: DT	59.26	46.96	46.96	21.43	21.43	21.43	72.50	72.50	72.50
Baseline3: RWC	61.68	50.13	49.88	26.07	26.14	25.75	74.15	74.12	74.01
Models with Demographics, Digital phenotyping data and Mood ratings as features									
Logistic Regression	64.91	67.26	60.30	38.71	59.11	46.78	82.26	66.96	73.83
Random Forest	70.82	68.08	62.11	44.06	43.84	43.95	80.21	80.35	80.28
K-Nearest Neighbors	68.25	69.35	64.12	42.93	65.76	51.95	85.12	69.13	76.30
Support Vector	75.90	74.89	67.67	54.25	48.77	51.36	82.54	85.48	83.98
XGBoost	81.43	82.31	73.34	69.97	50.49	58.66	84.09	92.35	88.02
Models with Demographics and Digital phenotyping features									
Logistic Regression	63.50	65.74	58.58	36.81	55.67	44.31	80.89	66.26	72.85
Random Forest	69.60	67.07	61.41	42.26	45.07	43.62	80.14	78.26	79.19
K-Nearest Neighbors	69.22	64.80	63.25	43.02	55.42	48.44	82.48	74.09	78.06
Support Vector	77.06	74.25	68.67	57.10	48.52	52.46	82.74	87.13	84.88
XGBoost	79.31	80.71	70.33	64.29	46.55	54.00	82.81	90.87	86.65

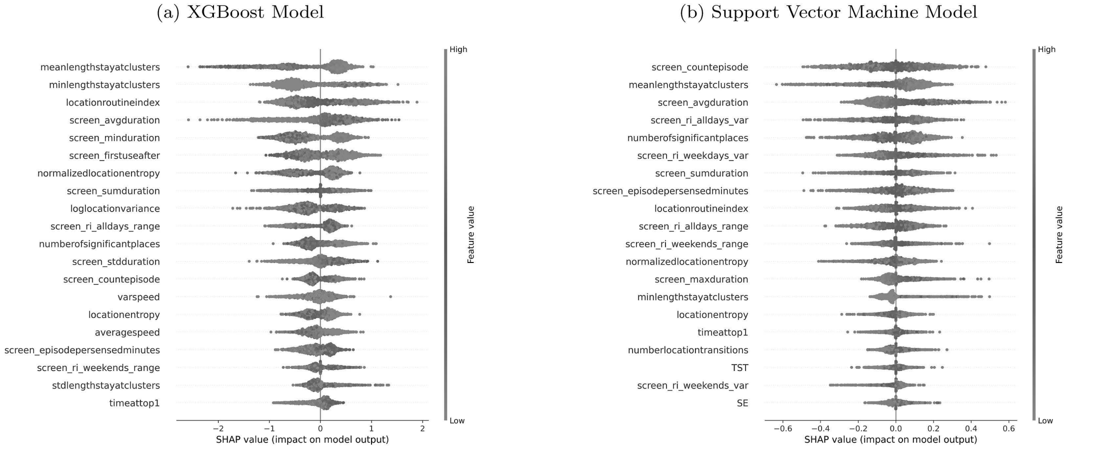
Note. Abbreviations: MC=Majority Class, RWC= Random Weighted Classifier, DT=Decision Tree

Only the XGB and SVM classifiers outperformed all three baseline models. The XGB was the best performing classifier. XGB predicted whether a participant belonged to the group with above normal symptoms of depression with 81.43% accuracy (AUC=82.31%, Precision1=69.97%, Recall1=50.49%, F11=58.66%, Precision0=84.09%, Recall0=92.35%, F10=88.02%). SVM predicted whether a participant belonged to the group with above normal symptoms of depression with an accuracy of 75.90% (AUC=74.89%, Precision1=69.97%, Recall1=48.77%, F11=51.36%, Precision0=82.54%, Recall0=85.48%, F10=83.98%).

The twenty most important digital phenotyping features for XGB and SVM classifiers in the models with no mood ratings are illustrated in **Figure 5**. The most important features included phone usage (sum, average, standard deviation of screen unlock duration, and count of screen unlocks), GPS mobility (mean length of stay at significant places, number of significant places, location routine index, and normalized location entropy), and Sleep (TST, Sleep Efficiency). **Figure 5** lists the digital phenotyping features on the y-axis in descending order of importance on the model's output. Each dot represents the SHAP value of one participant's feature, with blue and red colors representing low and high values of that feature, respectively.

SHAP dependence plots revealed interactions between important digital phenotyping features, enhancing the understanding of how the features impacted the model's output. For example, for most participants, a lower number of significant places combined with an increased average length of stay in significant places increased the likelihood for the classifiers to classify the participant under the group with above normal levels of depressive symptoms. When it came to phone usage features, an increased number of screen unlocks per day coupled with a lower average unlock duration per unlock increased the likelihood of the participant being classified within the above normal group. Conversely, participants who unlocked their phones less in a day and spent more screen time per unlock were more likely to be classified within the normal group.

Figure 5. Density scatter plot of SHAP values for (a) XGBoost and (b) SVM models, illustrating feature importance and impact on model output. Features are listed in descending order of importance.



(a) XGBoost Model

(b) Support Vector Machine Model

6 Discussion

This thesis investigated the role of digital technologies for the treatment of depression and identification of at-risk individuals. A systematic review of RCTs assessing the efficacy of digital interventions for the treatment of depression revealed that digital interventions are moderately effective in reducing symptoms of depression compared to all controls. Moreover, we provided the first meta-analytic evidence that guided digital interventions are more effective than treatment as usual in routine healthcare settings. In contrast, we found that unguided interventions (interventions without human support) may not be significantly more effective than treatment as usual in routine healthcare settings. Adherence to digital interventions remains a major challenge, with only 25% of participants completing the full intervention in routine healthcare settings. However, a secondary analysis of two large-scale RCTs of a digital intervention for depression in patients with chronic back pain demonstrated that dropout can be predicted from participant baseline variables (higher and lower age and lower education level) and that the inclusion of intervention usage variables (number of days to module completion and self-reported negative events) may improve the prediction of dropout early on in treatment. Finally, a longitudinal observational study assessing the role of digital phenotyping data in the prediction of depressive symptoms demonstrated that features derived from smartphone and wearable devices (GPS location data and sleep measures) may be used to predict depression symptom severity. Although the sample size was small and the study was conducted during the first wave of the COVID-19 pandemic, the findings contribute to a growing body of evidence of the potential of this new research field to contribute to early identification of at-risk individuals, thereby enabling more timely preventative interventions.

6.1 The efficacy and effectiveness of digital interventions

The thesis began with the largest and most comprehensive meta-analysis of the efficacy and effectiveness of digital interventions for the treatment of depression conducted to-date. Across 83 RCTs and 15,530 participants, we found a moderate effect size of $g = .52$ for digital interventions compared to all control conditions. Furthermore, effect size superiority was sustained at follow-up.

6.1.1 The difference in effect size across comparators

As has been reported in studies on face-to-face psychotherapy (Cuijpers, Karyotaki, et al., 2020), we found a significant difference in the reported effect size of digital interventions across control conditions. Subgroup analyses revealed an average ES of $g = .70$ for studies using WLC conditions, $g = .36$ for studies using attention control conditions and $g = .31$ for studies using TAU as a control. The appropriate choice of control condition for digital interventions is a matter of debate. Research has demonstrated that WLC controls tend to overinflate the “real” effect size, and may act as a *nocebo* condition, bringing negative psychological expectations whilst patients wait for active treatment (Furukawa et al., 2014). As such, TAU may offer a more realistic estimate of what an intervention can offer beyond usual care.

One of the most pertinent questions for digital interventions is how outcomes compare to traditional, face-to-face psychotherapy. Study I found no significant differences in the average effect size of digital interventions and face-to-face therapy, corroborating the findings from Carlbring and colleagues (2018). However, unlike the meta-analysis by Carlbring and colleagues, which pooled the results from both individual and group-based psychotherapy, the meta-analysis in Study I was limited to individual therapy alone. Notwithstanding, it is worthy of note that we only found three RCTs that directly compared digital interventions with face-to-face therapy. Moreover, these were all efficacy trials with small, highly selective subsamples of depressed patients who self-referred to treatment.

An alternative way of comparing the outcomes of digital interventions versus face-to-face therapy may be found by benchmarking the effect sizes against those reported in meta-analyses on psychotherapy. In the largest meta-analysis of trials on face-to-face psychotherapy for depression, Cuijpers and colleagues (2020) identified a moderate-to-large overall effect size of $g = .75$ compared to all control conditions, a large effect size of $g = .91$ when compared to WLC conditions and a moderate effect size of $g = .61$ when compared to TAU – all of which are considerably larger than the effect sizes of digital interventions found in Study I. Based on these data points and the lack of high-quality studies providing a direct comparison between digital interventions and face-to-face therapy, we believe that there is insufficient evidence to claim that digital interventions are as effective as face-to-face therapy, despite what is currently being proposed across the field (e.g., Andersson et al., 2019; Holmes et al., 2018).

6.1.2 For whom do digital interventions work?

Whether digital interventions are equally effective for all individuals - or whether certain participant characteristics are associated with differential outcomes - is important to know if we are to allocate patients to the appropriate treatment. Study I assessed a number of participant characteristics as potential moderators of effect size across age, gender, baseline depression severity and the presence of a comorbid somatic illness. In contrast to previous studies (eg., Donker et al., 2013), we found no significant impact of gender on outcomes, suggesting that males and females may benefit equally from digital interventions. This is important as males are significantly less likely than women to access mental health services (Gagné et al., 2014; Wang et al., 2005), most likely due to the increased social stigma associated with emotional vulnerability and seeking mental health treatment (Courtenay, 2000; Heifner, 1997; Möller-Leimkühler, 2002). At the same time, men are far more likely to commit suicide or suffer from substance use disorders, both of which depression is a known risk factor for (OECD, 2022). As such, the relative anonymity enabled by digital interventions may provide men with a "safer" alternative to access care where they may not otherwise.

Subgroup analyses comparing the efficacy of digital interventions across age groups found a significant difference in effect size between children and adolescents (*g* = .15) and adults (*g* = .53). Our findings are lower than what has been reported in previous meta-analyses on digital interventions for youth. Ebert and colleagues (2015) and Garrido and colleagues (2019) found overall effect sizes of *g* = .72 and *d* = .33 respectively for digital interventions compared to all controls. However, both of these reviews included young adults up to the age of 25 years of age, whilst the meta-analysis in Study I was limited to children up to the age of 18, thereby providing a more accurate estimate of effect size for this population.

Smaller effect sizes in children compared to adults have also been found in meta-analyses of face-to-face psychotherapy for depression (Cuijpers, Karyotaki, et al., 2020). Although it is unclear exactly why outcomes are different between the groups, one reason may be that the underlying therapies commonly used with children and adolescents are often adapted from therapies that were originally designed for adults. As the same is true for most studies on digital interventions, we would benefit from a better understanding of the mechanisms and mediators of change specific to children and how interventions may be better targeted to this subgroup based on the hypothesized mechanisms (Cuijpers, Karyotaki, et al., 2020; Domhardt et al., 2020). Indeed, despite the comprehensive search criteria used in Study I, we found only four studies in 30 years targeting children and adolescents (Gladstone et al., 2018; Ip et al., 2016; Smith et al., 2015; B. Wright et al., 2017). Given the growing prevalence of depression in young people (Racine et al., 2021) and the lack of robust evidence for the efficacy of pharmacological interventions for children, we mark this out as an important area for future research.

Another understudied subpopulation in digital interventions are individuals with severe mental illness (Sander et al., 2016). Despite the fact that clinical guidelines do not recommend digital interventions as first line treatment for patients severe depression (American Psychiatric Association, 2000; NICE, 2017), Study I found that outcomes were actually greater for individuals with higher baseline depression severity than those with lower baseline severity. The finding may reflect the fact that these individuals are more motivated to engage in treatment (Karyotaki et al., 2018), or simply that these participants have greater room for improvement. Nonetheless, only 1% of studies in the meta-analysis involved participants with severe pre-treatment depression levels.

There is a common tendency for trials on digital interventions to exclude individuals with severe depression or at risk of suicide, owing to a range of ethical and practical challenges (Bailey et al., 2020; Sander, Gerhardinger, et al., 2020). Furthermore, there is a strong feeling amongst the majority of clinicians that digital interventions should be limited to individuals with milder depression (Topooco et al., 2017). However, a growing body of evidence is emerging that the negative effects of digital interventions may be no different to what is experienced in usual healthcare settings (Ebert et al., 2016; Nielssen et al., 2022). Given the large evidence base demonstrating the efficacy of digital interventions for mild-to-moderate depression, it is now time for the field to conduct a more thorough exploration of the efficacy of digital interventions for individuals with high levels of depression severity. This should include the assessment of potential adverse conditions such as hospitalization and suicide for which these individuals are at high risk, as well as an examination of participant factors that may increase individual risk of deterioration/negative outcomes for digital interventions specifically.

Study I provided the first evidence that the presence of a comorbid physical illness does not influence the efficacy of digital interventions for depression. Combined with findings from Study II that patients with higher levels of pain disability were no more likely to drop out of the intervention than patients with lower levels of pain, the current thesis thus provides strong support for the efficacy, effectiveness, and acceptability of digital interventions in individuals with a comorbid somatic illness. This is promising for a number of reasons: first, comorbid depression is highly prevalent in individuals with chronic physical illness, often leading to increased use of pain medication, lower treatment adherence, poorer treatment outcomes, higher healthcare utilization and greater medical complications (DiMatteo et al., 2000; Rayner et al., 2016; Wing et al., 2002). Second, despite the high prevalence, only a small proportion of patients with a chronic physical illness access mental health treatment. Finally, as antidepressant treatment may lead to adverse drug interactions, digital interventions provide a high scalable alternative treatment format. In so doing, they may also help bridge the current divide between physical and mental health treatment, thereby enabling new forms of collaborative care for which there is a major need (Foster et al., 2018; Ormel, Cuijpers, et al., 2019).

6.1.3 The influence of human guidance on outcomes

Whilst a number of recent studies have suggested that there is no difference in outcomes between guided and unguided interventions (Bennett et al., 2019; Karyotaki et al., 2021; Konigbauer et al., 2017), Study I found that interventions that provided therapeutic guidance had a significantly higher overall effect size (*g* = .63) than unguided interventions (*g* = .34) and interventions with technical guidance (*g* = .46). The superior outcomes of guided interventions may be due to the increase in adherence found when human support is provided. According to the Supportive Accountability Model (Mohr et al., 2011) guidance increases adherence through accountability to a person who is seen as trustworthy, benevolent, and having expertise. Indeed, findings from Study I would appear to support this: across all intervention types, participants completed 67.9% of the intervention, whilst participants who received therapeutic guidance completed 76.3% of the intervention, participants who received technical guidance completed 60.9% of the intervention and participants who received no human guidance completed little more than half of the intervention (53.7%).

Our finding that therapeutic guidance led to higher effect sizes than technical guidance suggests that human support may be doing more than simply facilitating adherence as has been proposed until now (Ebert, Buntrock, et al., 2018; Musiat & Tarrier, 2014). Just as therapist behaviors may significantly influence treatment outcomes in face-to-face psychotherapy (e.g., through affirmation, encouragement, and self-disclosure), so too may similar mechanisms operate within digital interventions (Holländare et al., 2016). Future research would thus benefit from a better understanding of the relationship between guidance factors that contribute to adherence, including ways that some of these factors may be integrated within the designs of the intervention themselves (Kelders et al., 2012; Musiat & Tarrier, 2014).

Other important directions for future research that have both clinical and practical implications are understanding the dose-response relationship between guidance and outcomes and the role of clinical experience in the individual providing guidance. Study I found no significant dose-response relationship between the amount of guidance provided and outcomes. Thus, the question of "how much guidance is enough" - and whether that may differ depending on participant characteristics (e.g., baseline severity, education, or motivation) - remains unanswered. Importantly, we also found no significant difference in outcomes when guidance was provided by individuals with lower levels of qualification and experience and when it was provided by highly experienced clinicians. This may be due to the highly standardized nature of digital interventions, where the role of the therapist is to provide clarification and reinforcement of the therapeutic content delivered through the modules. Whatever the case, these findings have significant implications on the ability to scale digital interventions without relying on the already reduced pool of highly qualified clinicians available to deliver treatment, thereby potentially allocating them to more complex or severe cases, as is currently being explored in some public healthcare settings (Clark, 2018).

6.1.4 Digital interventions outside the lab: findings from effectiveness trials

Whether findings from efficacy trials translate into real-world healthcare settings is a critical question if digital interventions are to be considered as a viable treatment option within private and public healthcare settings. Study I found a small-to-moderate effect size superiority of $g = .30$ for digital interventions in effectiveness trials when compared to all controls. Moreover, sub-group analyses found that the effect size superiority was maintained when digital interventions were compared to treatment as usual ($g = .30$), putting to rest the contentious question of whether digital interventions offer benefits over and above the usual care an individual receives in real-world healthcare settings.

Notwithstanding, the pooled effect size for digital interventions was significantly lower in effectiveness (real-world) trials than in efficacy (laboratory-based) trials, mirroring what has been found in trials in face-to-face psychotherapy and antidepressant medication (Pigott et al., 2010; Singal et al., 2014). One reason for this is likely to be the lower adherence found in effectiveness settings compared to efficacy settings. On average, participants in effectiveness trials completed 53.6% of the intervention (compared to 74.6% in efficacy trials) and 25.2% completed the full intervention (compared to 60.9% in efficacy trials). Although it was not directly explored in the current studies, findings from other studies suggest that the lower adherence and lower effect sizes may be a result of the different sample populations and recruitment strategies employed in two trial designs (Streiner, 2002). In efficacy trials, participants often self-refer, are more highly motivated, better educated and more internet-savvy. In addition, they may receive remuneration for participating and adhering to the trial which, as was demonstrated in the current analysis, can have a significant influence on outcomes. By contrast, individuals in effectiveness studies are more representative of the general population: they are more likely to present with comorbid psychological and somatic conditions (often an exclusion criterion in efficacy trials), may be less well-educated and are less willing to accept psychological therapy without face-to-face contact (Knowles et al 2015).

With the effectiveness of digital interventions for depression now established in controlled trials, it is time to turn our attentions to the science of implementation and address the research-to-practice gap that befalls all new evidence-based treatments. Implementation frameworks are now emerging for digital mental health interventions to successfully guide the transition from research-to-practice (Graham et al., 2019). Frameworks such as the Nonadoption, Abandonment, Scale-up, Spread, and Sustainability framework (NASS) (Greenhalgh et al., 2017) can be used to identify the common barriers and facilitators for implementing digital interventions in health care settings that have been documented in published reviews. From these, appropriate implementation strategies can then be devised. Barriers that need to be addressed will range from the stigma associated with mental health to patient preferences for traditional face-to-face delivery (Knowles et al., 2015), privacy issues, and practitioner concerns about the acceptability and efficacy of digital interventions (Topooco et al., 2017).

These barriers are not insurmountable, however, and when properly addressed the results can be transformative. One such case in point is the UK's Improving Access to Psychological Therapies (IAPT) program. Started in 2008, the IAPT program is now accessed by over one million people a year with symptoms of anxiety or depression (*Adult Improving Access to Psychological Therapies Programme*, 2022). Digital interventions have a significant role to play within IAPT services. As part of a "stepped care" offering, IAPT first offers short-term digital interventions to individuals with mild-to-moderate depression and/or anxiety ("low intensity" interventions), whilst longer face-to-face therapy ("high-intensity" interventions) are offered to those with more severe or complex symptomatology. With outcome monitoring after every session, care pathways can be updated during treatment to allocate patients to the right level of care, thereby optimizing service capacity and maximizing overall treatment outcomes (Clark, 2018). On average, half of people who have a course of treatment in IAPT recover and two-thirds show worthwhile improvements in their mental health (National Collaborating Centre for Mental Health (Great Britain), 2021).

6.2 The challenge of adherence

A critical component of ensuring the successful implementation of digital interventions within real world settings is adherence. Study I found a strong dose-response relationship between the number of modules completed and effect size, supporting the common finding in digital mental health interventions that “more is better” (Donkin et al., 2011). Furthermore, we found that completing the full intervention had the largest influence on outcomes of all the moderators assessed in the meta-analysis. Nonetheless, we found that only 53.5% of participants across all trials completed the full intervention and only 25.2% in effectiveness trials. This stands in stark contrast with a meta-analysis on face-to-face psychotherapy which found that 84.7% of participants completed the full treatment on average (Van Ballegooijen et al., 2014).

6.2.1 Predicting dropout from participant characteristics

The ability to identify which participant characteristics predict a higher likelihood of dropout from digital interventions could help inform intervention design (e.g., tailoring interventions to specific sub-populations) or support clinical decisions when selecting appropriate care pathways. In a secondary analysis of data from two large scale RCTs, Study II examined what factors predicted dropout in a digital intervention for the prevention and treatment of depression in patients with comorbid chronic back pain. Overall, we found that lower education level and lower and higher age (a quadratic effect) significantly predicted a higher risk of dropout.

The relationship between lower education levels and higher treatment dropout aligns with findings across other digital interventions as well as in face-to-face psychotherapy and medication adherence generally (AL-Asadi et al., 2014; Batterham et al., 2008; Jarrett et al., 2013). The association may reflect the fact that these individuals find it harder to understand the intervention material and/or the digital format and are thus unable to engage with the intervention as intended. Indeed, research suggests that the language in a typical CBT intervention has a reading age of 17 years, which could pose a significant challenge for many individuals in clinical settings (Williams & Garland, 2002).

That both younger and older age were associated with higher risk of dropout suggests that the relationship between age and dropout may be more complex that has been considered until now and may reflect the influence of different mechanisms. For example, Donker and colleagues (2013) found that outcomes in a iCBT intervention were lower than outcomes in a intervention based on IPT, suggesting that the content of IPT (e.g, interpersonal conflicts and role transitions) may make IPT more appropriate for this group. At the same time, research has demonstrated that older people may struggle with the technical skills required to engage with a digital intervention, thus explaining the higher rates of dropout for this group. Future research would therefore benefit from meta-analyses based on individual participant data to understand the mechanisms and moderators of adherence in digital interventions. Findings from these could then be used to inform trials assessing the most effective methods of tailoring an intervention – either through design or more relevant content – to improve adherence.

Our finding that baseline depression severity did not influence dropout provides further evidence that digital interventions may be both acceptable and efficacious for severely depressed individuals (notwithstanding the points addressed earlier in this thesis). Study II also provided the first evidence that pain disability was not significantly associated with risk of dropout suggesting that digital interventions are acceptable for individuals of varying levels of pain intensity and thus may provide an effective, scalable approach for integrating psychological treatment within pain management routines in clinical settings.

6.2.2 Improving dropout prediction with intervention usage data

As it is not always possible to identify individuals in advance of treatment who may dropout, Study II also explored to what extent dropout could be predicted once an individual had started treatment. Here we assessed whether dropout could be predicted early on in the intervention (following completion of the first module) and whether intervention usage data could be used to improve the accuracy of predictions beyond participant baseline characteristics alone. We found several variables derived from intervention usage data that were associated with higher or lower risk of dropout.

An increasing number of days taken to complete the first module significantly predicted higher risk of dropout. This may reflect challenges interacting with the intervention, low motivation, lack of available time or low perceived value of the intervention – all of which have been found to impact adherence in qualitative analyses assessing reasons for dropout (Beatty & Binnion, 2016; Johansson et al., 2015; Knowles et al., 2015). We also found that participants who reported the occurrence of a negative event in the previous week had a lower risk of dropout. This may be due to the fact that experiencing a negative event provided patients with a focus for the module and thus greater intrinsic motivation to complete the intervention. In contrast, that absence of a negative effect predicted a higher likelihood of dropout is also consistent with research demonstrating that some participants drop out of interventions because they no longer feel they need it (Eysenbach, 2005; Knowles et al., 2015; Waller & Gilbody, 2009). Irrespective of the underlying mechanisms, this is the first study to demonstrate the value of a 1-item self-report questionnaire to improve the prediction of dropout during a digital intervention and thus highlights the potential of incorporating such assessments within digital interventions in the future.

Finally, when comparing performance across models, we found that the model that combined baseline characteristics and intervention usage data had the highest predictive accuracy and was significantly more accurate that models based on participant baseline characteristic variables only or intervention usage variables only. The AUROC of the model was 0.72, with a sensitivity of 76% which exceeded the accuracy threshold of 65-70% at which clinicians reportedly become willing to act on predictions (Eisenberg & Hershey, 1983). Taken together, these findings demonstrate the potential of “early warning systems” such as these to alert supporting clinicians when an individual is at high-risk of dropout enabling them to intervene before dropout occurs and, ideally, prevent it.

Notwithstanding the above, it is important to note that these findings are still exploratory and limited to one intervention and one sample population. Moreover, predictive accuracy remains modest. Future research would thus benefit from examining the role of additional sources of data in improving predictive performance and how these may differ across interventions, settings, target disorders and populations. New sources of data should include both passive data (e.g., the number and timing of logins, guidance used and interaction with specific components of the modules such as homework (Chien et al., 2020) as well as active, self-report data such as measures of therapeutic alliance in the case of guided interventions (Knowles et al., 2015; Lawler et al., 2021). Exploring the potential of novel machine learning methods that are capable of representing high dimensional, non-linear relationships in this data may also provide valuable insights and improve the identification of at-risk individuals.

6.3 The potential of digital phenotyping data for identifying at-risk individuals

6.3.1 Predicting symptoms of depression using data from a consumer wearable device

The ability to identify individuals with symptoms of depression before progressing into a full-blown disorder has strong potential to reduce the overall burden of disease. In Study III, we assessed to what extent digital phenotyping data from smartphone and wearable devices could be used to predict symptoms of depression in a 4-week observational study. The study provided the first evidence that a validated consumer wearable device – the Oura Ring – could significantly predict symptoms of depression from common measures of sleep. Specifically, we found that increased time in bed and total sleep time both predicted greater depression symptom severity as measured by the self-report DASS-21. One explanation for this may be the lack of motivation and fatigue commonly seen in individuals suffering from depression. Indeed, several of the items on the DASS-21 depression subscale measure this, e.g., "I found it difficult to work up the initiative to do things", "I was unable to become enthusiastic about anything". In addition, we found that longer periods of wakefulness after falling asleep (WASO) significantly predicted symptoms of anxiety, which is commonly comorbid with depression. This is likely due to the hypervigilance or hyperarousal common in anxiety disorders that cause individuals to wake up more frequently during sleep (Saletu-Zyhlarz et al., 1997).

Although similar findings to these have been reported in studies using polysomnography in highly controlled laboratory settings, it is worthy of note that the current findings were obtained from naturalistic settings using consumer devices. Given that sleep disturbances are a common risk factor for a number of mental disorders, our findings highlight the potential of consumer wearables to provide highly scalable methods of identifying early warning signs of mental illness, thereby enabling more timely intervention and prevention.

6.3.2 Predicting symptoms of depression from smartphone data

In addition to wearable device data, we also found that GPS-location data derived from the smartphone significantly predicted symptoms of depression. Specifically, the more varied the locations that a person visited each day during the 4-week period (the *location variance* feature), the lower their depressive symptoms. The finding supports previous research demonstrating that people who move about more through geographic space are less depressed (Farhan et al., 2016; Saeb et al., 2015). However, in Study III, we did not find any relationship between the other features derived from GPS data and symptoms of depression: total distance traveled, location entropy and time at home. There may be a number of reasons for this. First, the study was conducted during the onset of the COVID-19 pandemic where restrictions on social movement are likely to have influenced movement patterns and thus the relationships between GPS and symptoms of depression found previously. Second, the sample size of the current study was likely underpowered to find statistically significant results for a number of predictors exhibiting small effect sizes. Finally, previous findings themselves have been conflicting, suggesting that the relationship between GPS data and mental health remains unclear and is likely to differ between participants. These last two points also apply to the relationship between smartphone usage and symptoms of depression; in Study III, we found no significant association between the duration or frequency of daily smartphone usage and symptoms of depression in the current study.

As small sample sizes are common in the field of digital phenotyping (Benoit et al., 2020), future research is needed to assess whether these findings replicate in larger sample sizes. The field would also benefit from examining the relationship between additional digital phenotyping variables and outcomes across different mental disorders as well as within both clinical and non-clinical populations (Rohani et al., 2018). Such features may be derived from vocal biomarkers (Fagherazzi et al., 2021), keyboard interactions (Mastoras et al., 2019), physiological measures (Coutts et al., 2020) and the use of specific smartphone apps (Rozgonjuk et al., 2020).

6.3.3 Using ML models to identify individuals with above normal levels of depressive symptoms

It is unlikely that any one specific digital phenotyping variable will account for a large proportion of variance in symptoms at the population level. As we saw in Study III, a model that combined self-report EMA data with both smartphone and wearable data was most accurate at predicting depressive symptoms, demonstrating the value of combining data from different sources. As relationships are likely to be complex at both the inter- and intra-individual level, machine learning methods may have a valuable role to play here, too.

In Study IV, we assessed the extent to which machine learning models could differentiate between individuals with normal levels of depressive symptoms (a score of 0-9 on the DASS depression subscale) versus those that had above normal levels (a score of >9). Using digital phenotyping data alone, the best performing model (the XGBoost) was able to predict which class an individual belonged to with an accuracy of 79.31% and AUC of 80.71%.

In addition, SHAP values revealed a number of features that contributed to model performance that were not found in Study III. In particular, features derived from phone usage such as the duration and number of screen unlocks were found to be among the top 10 most important features in classifying normal vs above normal levels of depressive symptoms in both the XGBoost and SVM models. In contrast, the linear regression analyses in Study III did not find any significant relationship between depressive symptoms and any of the features derived from smartphone usage data. This may be due to the greater precision provided by the features in Study III (e.g., Study IV included the sum, average and standard deviation of screen unlock duration, whilst Study IV only included the sum), or the complex, non-linear relationship between these features and depressive symptom severity.

Further evidence of the complexity of the relationship between smartphone usage data and depressive symptoms was revealed by the interactions shown in the SHAP dependence plots. For example, an increased number of screen unlocks per day combined with a lower average unlock duration per unlock (what may be defined as "excessive checking") was associated with higher likelihood of being classified as having above normal levels of depressive symptoms. Conversely, participants who unlocked their phones less throughout the day but spent more screen time per unlock were more likely to be classified as having normal levels of depressive symptoms. These findings align with research demonstrating that addictive mobile phone usage may be due to the need for reassurance caused by anxiety, poor self-esteem, or increased emotional instability (Billieux et al., 2015). Study IV provided the first evidence that this behavior and the purported relationship with depressive symptoms can be detected using digital phenotyping data. It also highlights the potential of machine learning models to identify complex, high dimensional relationships in the data. As the field of psychotherapy evolves and moves away from a reliance on traditional diagnostic categories to more complex, multi-dimensional models such as proposed by the Research Domain Criteria project (RDoC) (Insel et al., 2010), digital phenotyping data may thus have a valuable role to play in elucidating some of the underlying physiological and behavioral domains that inform future clinical diagnoses.

6.4 Limitations

There are a number of limitations of the current thesis. First, the majority of participants were white individuals from relatively high-income settings. 95% of studies included in the meta-analysis were conducted across Europe, Australasia, and the US, whilst participants in Study II were from Germany and the majority of participants in Study III were from Finland. We only found one study published in the last 30 years assessing the efficacy of a digital intervention for the treatment of depression conducted in South America and none in Africa. As a result, the generalizability of our findings may not extend to these continents where populations, health care systems and access to technology can differ considerably. As depression is a global public health problem with similar prevalence rates between low-income and high-income countries, yet significantly lower rates of adequate treatment in low income settings (Cuijpers, Quero, et al., 2019), we thus mark this out as a critical area for future research.

Another limitation relates to the methodologies used. Study I was conducted with meta-regression using aggregate data at the study level. Although a novel strength of the analysis was the multi-level approached used to account for the nested structure of the data (i.e., dependencies between outcomes measures within a study), a meta-analysis based on individual participant data (IPDMA) would have been a superior method. It is well known that meta-analyses are subject to publication bias and IPDMA is one step toward overcoming that limitation. An IPDMA approach would also have enabled the inclusion of more studies in the analysis (e.g., those with incomplete outcome data that had to be excluded in the data extraction phase), provided more standardized classifications of participant characteristics related to in/exclusion criteria, enabled a consistent unit of analysis for individual outcomes across studies and accounted for missing data at the patient level. As such, more robust measures of effect size would have been derived at the IPD level. In addition, there would have been greater statistical power for assessing potential moderators and interaction effects. The complex relationships found in the current thesis between individual moderators and mediators influencing outcomes in digital interventions – and the inconsistency of findings related to these - underscores this. Related to this, the small sample size in Study III was likely underpowered to find statistically significant relationships between digital phenotyping data and symptoms of depression, especially in cases where effect size was small.

Finally, whilst we found a number of significant relationships in the current analyses, none of the findings can be interpreted as causal in nature. For example, although we found that a higher number of modules completed was associated with superior outcomes in both the Study I and Study II, it may be that another, third, variable accounts for the relationship between the two. The same is true for the participant characteristics that moderated effect size and dropout, and the relationships between digital phenotyping data and depressive symptoms. Although we tried to control for potential confounders, there is the possibility that other variables that were not captured in the analysis were responsible for the relationship found, or at least influenced it. As we shall explore in the Conclusion, understanding these causal relationships is critical if we are to move the field forward.

7 Conclusion

***"The question towards which all outcome research should ultimately be directed is the following: What treatment, by whom, is most effective for this individual with that specific problem, and under which set of circumstances?"* Gordon Paul, 1967**

The current thesis provided evidence of the efficacy and acceptability of digital interventions for the treatment of depression for a variety of populations and across a number of settings. Notably, we provided the first evidence of the effectiveness of digital interventions compared to usual care in real-world healthcare settings. Given the limitations of the dominant model of delivering psychotherapy and its inherent constraints in addressing the growing treatment gap in mental healthcare, digital interventions provide a promising alternative treatment format.

However, if digital interventions are to take their place within the wider treatment mix, attention now needs to focus on bridging the research-to-practice gap and towards the science of implementation. Addressing the problem of adherence is one critical topic here. As Study I demonstrated, only one in four individuals complete the full intervention on average when delivered in real world settings. At the same time, completing the full intervention was associated with the largest reductions in symptoms of all moderators studied in the meta-analysis. Yet, we also provided evidence that the challenge of adherence is not insurmountable. As we demonstrated in Study II, intervention dropout could be predicted from a combination of participant baseline characteristics and intervention-usage data. What is more, at-risk individuals can be identified early on in treatment, thereby providing an opportunity for timely intervention. Future research will need to assess whether these findings can be extended into practice and assess their potential to reduce dropout within appropriately designed trials.

An equally important component of implementation studies is maximizing external validity. As we saw in our analysis, the majority of trials to-date have involved WEIRD samples (western, education, industrialized, rich and democratic). Depression permeates these divides. It is a pervasive disorder with a global prevalence and future research will need to reflect that. In practice, that means a greater research focus on the efficacy, effectiveness, and feasibility of digital interventions in low-and-middle income countries, across all age groups, across cultures and for individuals of all levels of depression severity.

As technology continues to develop in sophistication, digital interventions may also provide the opportunity to go beyond increasing treatment *access* to improving treatment *outcomes*. Studies III and IV contributed further evidence towards the potential of digital phenotyping data from smartphone and wearable devices for the identification of individuals with symptoms of common mental disorders. Moreover, for the first time, we demonstrated the potential of consumer-grade wearable devices capturing common measures of sleep to significantly predict symptoms of depression and anxiety. Such data may help identify individuals at risk of developing a disorder and thus enable more timely, preventative interventions which will be critical if we are to reduce the overall burden of disease associated with depression. Although the field is still young and substantial scientific and systemic developments will be required before such data is used within routine clinical practice, it is not unrealistic to imagine a future where such data is used to support clinical decision making.

7.1 Future directions

If we step back and contextualize the role of digital interventions within the field of psychotherapy as a whole, we begin to see exciting potential for these new technologies to inform the field more generally. On average, half of people who receive evidence-based treatments for depression will fail to respond. Of those that do respond, many remain at significant risk of future relapse (Malhi & Mann, 2018). Indeed, even when patients present with similar symptoms and are treated by the same therapist using similar methods, they respond differently (Kiesler, 1997). Some patients will demonstrate large improvements whilst others will show no reduction in symptoms at all (Bromet et al., 2011). Most problematic of all, we currently do not know who will benefit and who will not, nor which treatment format is likely to lead to best improvements for a specific patient (Verduijn et al., 2017). It is thus perhaps not surprising that outcomes of even the most widely studied psychological treatments such as CBT have not improved in the 40 years since its introduction (Johnsen & Friborg, 2015).

If this is to change in the years that follow, we need to move away from a research agenda focused on demonstrating *that* psychotherapy works to one that demonstrates *how* it works (Cuijpers, 2016). The majority of studies to-date have compared outcomes across two groups of individuals in different conditions, proposing underlying mechanisms of change based on the differences observed. However, these types of studies can only provide correlational evidence for the hypothesized mechanisms, not causal, and are thus limited in their explanatory power. To understand how psychotherapy works, new experimental designs are needed that are able to identify the moderators, mediators and mechanisms of change. A mechanism of change is the intermediate causal process through which a therapeutic ingredient (or some independent variable) produces a change in the outcome, whilst a mediator is a variable that may account statistically for the relationship between the independent variable and the outcome, and a moderator is a characteristic that may influence the direction of that relationship (e.g., gender).

In order for us to establish that a proposed mechanism is likely to be a causal factor in the psychological change of a patient, a number of methodological criteria have to be met (Kazdin, 2007). These include temporal precedence, plausibility, experimental manipulation, consistency, association, the dose-response relationship, and specificity. Meeting many of these requirements has long been a challenge in traditional, face-to-face psychotherapy research (Domhardt et al., 2021; Lemmens et al., 2016). Therapeutic ingredients and the change processes that follow them are rarely disentangled in consistent ways, most questionnaires are limited in their ability to capture process changes and sample sizes are often too small to identify potential mediators of change (Huibers et al., 2021).

Digital interventions may provide a powerful new paradigm that overcomes many of these challenges. In particular, they are highly standardized (consistent), provide the opportunity to deliver modular treatment components (experimental manipulation and specificity) and enable granular data collection related to the timing (temporal precedence), dose (dose-response relationship) and proximal outcomes (association) of those intervention components. Moreover, digital interventions can be scaled far more efficiently to the large sample sizes required to identify potential moderators or mediators of change.

With a more robust understanding of how therapy works we will then be better placed to develop new therapies and improve existing ones. It will also allow us to personalize therapies to the specific needs of the individual patient based on predicting which components are likely to be the most important for a patient based on their principal symptoms and concerns, as well as predicting the optimal order in which they should be delivered. Although the field is only just starting out here, there are already promising developments towards this end (e.g., Weisz et al., 2012).

With the addition of digital phenotyping data, digital interventions may also help provide a more detailed understanding of individual symptom networks and how disorders evolve at the intra-individual level. For example, intensive longitudinal measurement of mood data ahead of treatment using ecological momentary assessment has been used to reveal individual symptom profiles. These profiles have then been used to generate personalized treatment plans that bring about large effect sizes (Fisher et al., 2019). Realtime data such as this may also be used to deliver just-in-time-adaptive-interventions (JITAI) where specific treatment components are delivered at the optimal time for each individual based on their presenting symptoms, whether those symptoms are captured via self-report or proxy measures from passive digital phenotyping data (Nahum-Shani et al., 2018). Although the field is still young here, Study II in the current thesis contributes to an emerging body of evidence demonstrating its potential (Bae et al., 2018; S. P. Goldstein et al., 2017).

To deliver on that potential will require us to embrace a number of changes to the way we conduct research going forward. These include methodological requirements related to statistical power, generalizability, model specification, representative samples and prospective tests. Beyond that, are the issues related to implementation. As we have seen in the current thesis, we are only beginning to understand what is required for digital interventions to successfully make their way into healthcare practice. As that unfolds, we will need to address similar issues related to implementing precision medicine within psychotherapy. For machine learning models and digital phenotyping data to become integrated into regular clinical practice, concerns related to potential bias in these algorithms and to data privacy will need to be addressed.

These changes will require true interdisciplinary collaboration, both within academia and outside of it; with industry, government, policy makers and healthcare systems. This will take time and tremendous efforts, but the potential to contribute to our understanding of mental illness and bring about large and meaningful reductions to the growing burden of disease make it imperative.

References

Adult Improving Access to Psychological Therapies programme. (2022). https://www.england.nhs.uk/mental-health/adults/iapt/

AL-Asadi, A. M., Klein, B., & Meyer, D. (2014). Pretreatment attrition and formal withdrawal during treatment and their predictors: An exploratory study of the anxiety online data. *Journal of Medical Internet Research*, *16*(6), e2989. https://doi.org/10.2196/jmir.2989

American Psychiatric Association. (2000). Practice guideline for the treatment of patients with major depressive disorder (revision). *American Journal of Psychiatry*, *157*(4 SUPPL.), 1–45. https://doi.org/10.4088/jcp.v63n0416a

Andersson, G., & Carlbring, P. (2021). Cognitive behavioral therapy delivered using the internet. *Handbook of Cognitive Behavioral Therapy: Applications (Vol. 2).*, 607–631. https://doi.org/10.1037/0000219-019

Andersson, G., Carlbring, P., & Grimlund, A. (2008). Predicting treatment outcome in internet versus face to face treatment of panic disorder. *Computers in Human Behavior*, *24*(5), 1790–1801. https://doi.org/10.1016/J.CHB.2008.02.003

Andersson, G., & Cuijpers, P. (2009). Internet-based and other computerized psychological treatments for adult depression: a meta-analysis. *Cognitive Behaviour Therapy*, *38*(4), 196–205. https://doi.org/10.1080/16506070903318960

Andersson, G., Titov, N., Dear, B. F., Rozental, A., & Carlbring, P. (2019). Internet-delivered psychological treatments: from innovation to implementation. *World Psychiatry*, *18*(1). https://doi.org/10.1002/wps.20610

Andreas, S., Schulz, H., Volkert, J., Dehoust, M., Sehner, S., Suling, A., Ausín, B., Canuto, A., Crawford, M., Da Ronch, C., Grassi, L., Hershkovitz, Y., Muñoz, M., Quirk, A., Rotenstein, O., Santos-Olmo, A. B., Shalev, A., Strehle, J., Weber, K., … Härter, M. (2017). Prevalence of mental disorders in elderly people: The European MentDis_ICF65+ study. *The British Journal of Psychiatry*, *210*(2), 125–131. https://doi.org/10.1192/BJP.BP.115.180463

Andrews, G., Cuijpers, P., Craske, M. G., McEvoy, P., Titov, N., Basu, A., English, C. L., & Newby, J. M. (2018). Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: An updated meta-analysis. *Journal of Anxiety Disorders*, *55*(February), 70–78. https://doi.org/10.1016/j.janxdis.2018.01.001

Andrews, G., Issakidis, C., Sanderson, K., Corry, J., & Lapsley, H. (2004). Utilising survey data to inform public policy: Comparison of the cost-effectiveness of treatment of ten mental disorders. *British Journal of Psychiatry*, *184*(JUNE), 526–533. https://doi.org/10.1192/bjp.184.6.526

Antony, M. M., Cox, B. J., Enns, M. W., Bieling, P. J., & Swinson, R. P. (1998). Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample. *Psychological Assessment*, *10*(2), 176–181. https://doi.org/10.1037/1040-3590.10.2.176

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, *07-09-January-*

2007, 1027–1035.

Asselbergs, J., Ruwaard, J., Ejdys, M., Schrader, N., Sijbrandij, M., & Riper, H. (2016). Mobile phone-based unobtrusive ecological momentary assessment of day-to-day mood: An explorative study. *Journal of Medical Internet Research*, *18*(3), e5505. https://doi.org/10.2196/JMIR.5505

Assink, M., & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *The Quantitative Methods for Psychology*, *12*. https://doi.org/https://doi.org/10.20982/tqmp.12.3.p154

Babwah, F., Baksh, S., Blake, L., Cupid-Thuesday, J., Hosein, I., Sookhai, A., Poon-King, C., & Hutchinson, G. (2006). The role of gender in compliance and attendance at an outpatient clinic for type 2 diabetes mellitus in Trinidad. *Revista Panamericana de Salud Publica/Pan American Journal of Public Health*, *19*(2), 79–84. https://doi.org/10.1590/S1020-49892006000200002

Bae, S., Chung, T., Ferreira, D., Dey, A. K., & Suffoletto, B. (2018). Mobile phone sensors and supervised machine learning to identify alcohol use events in young adults: Implications for just-in-time adaptive interventions. *Addictive Behaviors*, *83*, 42–47. https://doi.org/10.1016/J.ADDBEH.2017.11.039

Bailey, E., Mühlmann, C., Rice, S., Nedeljkovic, M., Alvarez-Jimenez, M., Sander, L., Calear, A. L., Batterham, P. J., & Robinson, J. (2020). Ethical issues and practical barriers in internet-based suicide prevention research: A review and investigator survey. *BMC Medical Ethics*, *21*(1), 1–16. https://doi.org/10.1186/s12910-020-00479-1

Barak, A., Hen, L., Boniel-Nissim, M., & Shapira, N. (2008). A Comprehensive Review and a Meta-Analysis of the Effectiveness of Internet-Based Psychotherapeutic Interventions. *Journal of Technology in Human Services*, *26*(2–4), 109–160. https://doi.org/10.1080/15228830802094429

Barkham, M., Connell, J., Miles, J. N. V., Evans, C., Stiles, W. B., Margison, F., & Mellor-Clark, J. (2006). Dose-effect relations and responsive regulation of treatment duration: The good enough level. *Journal of Consulting and Clinical Psychology*, *74*(1), 160–167. https://doi.org/10.1037/0022-006X.74.1.160

Barnett, K., Mercer, S. W., Norbury, M., Watt, G., Wyke, S., & Guthrie, B. (2012). Epidemiology of multimorbidity and implications for health care, research, and medical education: A cross-sectional study. *The Lancet*, *380*(9836), 37–43. https://doi.org/10.1016/S0140-6736(12)60240-2

Bateup, S. E., Palmer, C. R., & Catarino, A. (2020). Using technology to understand how therapist variables are associated with clinical outcomes in IAPT. *Cognitive Behaviour Therapist*, *13*(August). https://doi.org/10.1017/S1754470X20000252

Batterham, P. J., Neil, A. L., Bennett, K., Griffiths, K. M., & Christensen, H. (2008). Predictors of adherence among community users of a cognitive behavior therapy website. *Patient Preference and Adherence*, *2*, 97–105. http://www.ncbi.nlm.nih.gov/pubmed/19920949

Baumeister, H., Paganini, S., Sander, L. B., Lin, J., Schlicker, S., Terhorst, Y., Moshagen, M., Bengel, J., Lehr, D., & Ebert, D. (2020). Effectiveness of a guided internet- and mobile-based intervention for patients with chronic back pain and depression (WARD-BP): A multicenter, pragmatic randomized controlled trial. *Psychotherapy and Psychosomatics*, 1–14. https://doi.org/10.1159/000511881

Baumeister, H., Reichler, L., Munzinger, M., & Lin, J. (2014). The impact of guidance on Internet-based mental health interventions — A systematic review. *Internet Interventions*, *1*(4), 205–215. https://doi.org/10.1016/j.invent.2014.08.003

Beatty, L., & Binnion, C. (2016). A Systematic Review of Predictors of, and Reasons for, Adherence to Online Psychological Interventions. *International Journal of Behavioral Medicine*, *23*(6), 776–794. https://doi.org/10.1007/s12529-016-9556-9

Beck, A. T. (1979). Cognitive Therapy of Depression. In A. T. Beck (Ed.), *Guilford Press*.

Beiwinkel, T., Kindermann, S., Maier, A., Kerl, C., Moock, J., Barbian, G., & Rössler, W. (2016). Using Smartphones to Monitor Bipolar Disorder Symptoms: A Pilot Study. *JMIR Mental Health*, *3*(1), e2. https://doi.org/10.2196/mental.4560

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Bennett, S. D., Cuijpers, P., Ebert, D. D., McKenzie Smith, M., Coughtrey, A. E., Heyman, I., Manzotti, G., & Shafran, R. (2019). Practitioner Review: Unguided and guided self-help interventions for common mental health disorders in children and adolescents: a systematic review and meta-analysis. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *60*(8), 828–847. https://doi.org/10.1111/jcpp.13010

Benoit, J., Onyeaka, H., Keshavan, M., & Torous, J. (2020). Systematic Review of Digital Phenotyping and Machine Learning in Psychosis Spectrum Illnesses. *Harvard Review of Psychiatry*, *28*(5), 296–304. https://doi.org/10.1097/HRP.0000000000000268

Berger, T., Hämmerli, K., Gubser, N., Andersson, G., & Caspar, F. (2011). Internet-Based Treatment of Depression: A Randomized Controlled Trial Comparing Guided with Unguided Self-Help. *Cognitive Behaviour Therapy*, *40*(4), 251–266. https://doi.org/10.1080/16506073.2011.616531

Billieux, J., Maurage, P., Lopez-Fernandez, O., Kuss, D. J., & Griffiths, M. D. (2015). Can Disordered Mobile Phone Use Be Considered a Behavioral Addiction? An Update on Current Evidence and a Comprehensive Model for Future Research. In *Current Addiction Reports* (Vol. 2, Issue 2, pp. 156–162). Springer. https://doi.org/10.1007/s40429-015-0054-y

Bloom, D.E., Cafiero, E.T., Jané-Llopis, E., Abrahams-Gessel, S., Bloom, L.R., Fathima, S., Feigl, A.B., Gaziano, T., Mowafi, M., Pandya, A., Prettner, K., Rosenberg, L., Seligman, B., Stein, A.Z., & Weinstein, C. (2011). *The Global Economic Burden of Noncommunicable Diseases*.

Boeschoten, R. E., Dekker, J., Uitdehaag, B. M. J., Beekman, A. T. F., Hoogendoorn, A. W., Collette, E. H., Cuijpers, P., Nieuwenhuis, M. M., & Van Oppen, P. (2017). Internet-based treatment for depression in multiple sclerosis: A randomized controlled trial. *Multiple Sclerosis*, *23*(8), 1112–1122. https://doi.org/10.1177/1352458516671820

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

Bremer, V., Chow, P. I., Funk, B., Thorndike, F. P., & Ritterband, L. M. (2020). Developing a Process for the Analysis of User Journeys and the Prediction of Dropout in Digital Health Interventions: Machine Learning Approach. *Journal of Medical Internet Research*, *22*(10). https://doi.org/10.2196/17738

Bromet, E., Andrade, L. H., Hwang, I., Sampson, N. A., Alonso, J., de Girolamo, G., de Graaf, R., Demyttenaere, K., Hu, C., Iwata, N., Karam, A. N., Kaur, J., Kostyuchenko, S., Lépine, J.-P., Levinson, D., Matschinger, H., Mora, M. E. M., Browne, M. O., Posada-Villa, J., … Kessler, R. C. (2011). Cross-national epidemiology of DSM-IV major depressive episode. *BMC Medicine*, *9*(1), 90. https://doi.org/10.1186/1741-7015-9-90

Browne, M. W. (2000). Cross-Validation Methods. *Journal of Mathematical Psychology*, *44*(1), 108–132. https://doi.org/10.1006/JMPS.1999.1279

Bunce, D., Batterham, P. J., Mackinnon, A. J., & Christensen, H. (2012). Depression, anxiety and cognition in community-dwelling adults aged 70 years and over. *Journal of Psychiatric Research*, *46*(12), 1662–1666. https://doi.org/10.1016/J.JPSYCHIRES.2012.08.023

Burton, C., McKinstry, B., Szentagotai Tătar, A., Serrano-Blanco, A., Pagliari, C., & Wolters, M. (2013). Activity monitoring in patients with depression: A systematic review. *Journal of Affective Disorders*, *145*(1), 21–28. https://doi.org/10.1016/J.JAD.2012.07.001

Byers, A. L., Yaffe, K., Covinsky, K. E., Friedman, M. B., & Bruce, M. L. (2010). High Occurrence of Mood and Anxiety Disorders Among Older Adults: The National Comorbidity Survey Replication. *Archives of General Psychiatry*, *67*(5), 489–496. https://doi.org/10.1001/ARCHGENPSYCHIATRY.2010.35

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(3), 223–230. https://doi.org/10.1016/J.BPSC.2017.11.007

Campion, J., Bhui, K., & Bhugra, D. (2012). European Psychiatric Association (EPA) guidance on prevention of mental disorders. *European Psychiatry : The Journal of the Association of European Psychiatrists*, *27*(2), 68–80. https://doi.org/10.1016/J.EURPSY.2011.10.004

Carlbring, P., Andersson, G., Cuijpers, P., Riper, H., & Hedman-Lagerlöf, E. (2018). Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. *Cognitive Behaviour Therapy*, *47*(1), 1–18. https://doi.org/10.1080/16506073.2017.1401115

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., & Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry*, *3*(3), 243–250. https://doi.org/10.1016/S2215-0366(15)00471-X

Chen, Z., Lin, M., Chen, F., Lane, N. D., Cardone, G., Wang, R., Li, T., Chen, Y., Choudhury, T., & Campbell, A. T. (2013). *Unobtrusive Sleep Monitoring Using Smartphones*. 145–152. https://doi.org/10.4108/icst.pervasivehealth.2013.252148

Chen, Z. S., Prathamesh, Kulkarni, Galatzer-Levy, I. R., Bigio, B., Nasca, C., & Zhang, Y.

(2022). Modern Views of Machine Learning for Precision Psychiatry. *ArXiv Preprint ArXiv:2204.01607*. https://doi.org/10.48550/arxiv.2204.01607

Chien, I., Enrique, A., Palacios, J., Regan, T., Keegan, D., Carter, D., Tschiatschek, S., Nori, A., Thieme, A., Richards, D., Doherty, G., & Belgrave, D. (2020). A Machine Learning Approach to Understanding Patterns of Engagement With Internet-Delivered Mental Health Interventions. *JAMA Network Open*, *3*(7), e2010791. https://doi.org/10.1001/jamanetworkopen.2020.10791

Cho, Y. M., Lim, H. J., Jang, H., Kim, K., Choi, J. W., Shin, C., Lee, S. K., Kwon, J. H., & Kim, N. (2016). A cross-sectional study of the association between mobile phone use and symptoms of ill health. *Environmental Health and Toxicology*, *31*, e2016022. https://doi.org/10.5620/EHT.E2016022

Chowdhury, M. Z. I., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, *8*(1), 262. https://doi.org/10.1136/FMCH-2019-000262

Christensen, H., Griffiths, K. M., & Farrer, L. (2009). Adherence in internet interventions for anxiety and depression. *Journal of Medical Internet Research*, *11*(2), 1–16. https://doi.org/10.2196/jmir.1194

Clark, D. M. (2018). Realizing the Mass Public Benefit of Evidence-Based Psychological Therapies: The IAPT Program. *Annual Review of Clinical Psychology*, *14*(January), 159–183. https://doi.org/10.1146/annurev-clinpsy-050817-084833

Clark, D. M., & Currie, K. C. (2009). Depression, anxiety and their relationship with chronic diseases: a review of the epidemiology, risk and treatment evidence. *Medical Journal of Australia*, *190*(7 SUPPL.), S54–S60. https://doi.org/10.5694/J.1326-5377.2009.TB02471.X

Clarke, G., Reid, E., Eubanks, D., O'Connor, E., DeBar, L. L., Kelleher, C., Lynch, F., & Nunley, S. (2002). Overcoming Depression on the Internet (ODIN): A randomized controlledtrialof an Internet depression skills intervention program. *Journal of Medical Internet Research*, *4*(3), 5–17. https://doi.org/10.2196/jmir.4.3.e14

Cooney, G. M., Dwan, K., Greig, C. A., Lawlor, D. A., Rimer, J., Waugh, F. R., McMurdo, M., & Mead, G. E. (2013). Exercise for depression: Some benefits but better trials are needed. *Saudi Medical Journal*, *34*(11), 1203. https://doi.org/10.1002/14651858.CD004366.pub6

Cooper, Harris, & Hedges, L. V. (2009). Research synthesis as a scientific process. In Harr Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* (2nd ed., pp. 3–17). Russell Sage Foundation.

Côté-Allard, U., Pham, M. H., Schultz, A. K., Nordgreen, T., & Torresen, J. (2022). *Adherence Forecasting for Guided Internet-Delivered Cognitive Behavioral Therapy: A Minimally Data-Sensitive Approach*. 1–9. http://arxiv.org/abs/2201.04967

Courtenay, W. H. (2000). Constructions of masculinity and their influence on men's well-being: a theory of gender and health. *Social Science & Medicine*, *50*(10), 1385–1401. https://doi.org/10.1016/S0277-9536(99)00390-1

Coutts, L. V., Plans, D., Brown, A. W., & Collomosse, J. (2020). Deep learning with wearable based heart rate variability for prediction of mental and general health. *Journal of*

Biomedical Informatics, *112*, 103610. https://doi.org/10.1016/J.JBI.2020.103610

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202. https://doi.org/10.1111/J.2517-6161.1972.TB00899.X

Cristea, I., Stefan, S., Karyotaki, E., David, D., Hollon, S. D., & Cuijpers, P. (2017). The effects of cognitive behavioral therapy are not systematically falling: A revision of Johnsen and Friborg (2015). *Psychological Bulletin*, *143*(3), 326–340. https://doi.org/10.1037/bul0000062

Cuijpers, P. (2016). The future of psychotherapy research: stop the waste and focus on issues that matter. *Epidemiology and Psychiatric Sciences*, *25*(4), 291–294. https://doi.org/10.1017/S2045796015000785

Cuijpers, P., Beekman, A. T. F., & Reynolds, C. F. (2012). Preventing Depression: A Global Priority. *JAMA*, *307*(10), 1033–1034. https://doi.org/10.1001/JAMA.2012.271

Cuijpers, P., De Graaf, R., & Van Dorsselaer, S. (2004). Minor depression: risk profiles, functional disability, health care use and risk of developing major depression. *Journal of Affective Disorders*, *79*(1–3), 71–79. https://doi.org/10.1016/S0165-0327(02)00348-8

Cuijpers, P., Geraedts, A. S., van Oppen, P., Andersson, G., Markowitz, J. C., & van Straten, A. (2011). Interpersonal psychotherapy for depression: a meta-analysis. *The American Journal of Psychiatry*, *168*(6), 581–592. https://doi.org/10.1176/appi.ajp.2010.10101411

Cuijpers, P., Huibers, M., Daniel Ebert, D., Koole, S. L., & Andersson, G. (2013). How much psychotherapy is needed to treat depression? A metaregression analysis. *Journal of Affective Disorders*, *149*(1–3), 1–13. https://doi.org/10.1016/j.jad.2013.02.030

Cuijpers, P., Karyotaki, E., Eckshtain, D., Ng, M. Y., Corteselli, K. A., Noma, H., Quero, S., & Weisz, J. R. (2020). Psychotherapy for Depression across Different Age Groups: A Systematic Review and Meta-analysis. *JAMA Psychiatry*, *77*(7), 1–9. https://doi.org/10.1001/jamapsychiatry.2020.0164

Cuijpers, P., Karyotaki, E., Reijnders, M., & Ebert, D. (2019). Was Eysenck right after all? A reassessment of the effects of psychotherapy for adult depression. *Epidemiology and Psychiatric Sciences*, *28*(1), 21–30. https://doi.org/10.1017/S2045796018000057

Cuijpers, P., Noma, H., Karyotaki, E., Cipriani, A., & Furukawa, T. A. (2019). Effectiveness and Acceptability of Cognitive Behavior Therapy Delivery Formats in Adults with Depression: A Network Meta-analysis. *JAMA Psychiatry*, *76*(7), 700–707. https://doi.org/10.1001/jamapsychiatry.2019.0268

Cuijpers, P., Noma, H., Karyotaki, E., Vinkers, C. H., Cipriani, Andrea, & Furukawa, T. A. (2020). A network meta-analysis of the effects of psychotherapies, pharmacotherapies and their combination in the treatment of adult depression. *World Psychiatry*, *19*(1), 92–107. https://doi.org/10.1002/wps.20701

Cuijpers, P., Quero, S., Papola, D., Cristea, I., & Karyotaki, E. (2019). Care-as-usual control groups across different settings in randomized trials on psychotherapy for adult depression: A meta-analysis. *Psychological Medicine*, *51(4)*, 634–644. https://doi.org/10.1017/S0033291719003581

Cuijpers, P., Sijbrandij, M., Koole, S. L., Andersson, G., Beekman, A. T., & Reynolds, C. F. (2013). The efficacy of psychotherapy and pharmacotherapy in treating depressive and anxiety disorders: A meta-analysis of direct comparisons. *World Psychiatry*, *12*(2), 137–148. https://doi.org/10.1002/wps.20038

Cuijpers, P., Van Straten, A., Bohlmeijer, E., Hollon, S. D., & Andersson, G. (2010). The effects of psychotherapy for adult depression are overestimated: A meta-analysis of study quality and effect size. *Psychological Medicine*, *40*(2), 211–223. https://doi.org/10.1017/S0033291709006114

Cuijpers, P., van Straten, A., Warmerdam, L., & Andersson, G. (2008). Psychological treatment of depression: A meta-analytic database of randomized studies. *BMC Psychiatry*, *8*(1), 36. https://doi.org/10.1186/1471-244X-8-36

David, D., Cristea, I., & Hofmann, S. G. (2018). Why Cognitive Behavioral Therapy Is the Current Gold Standard of Psychotherapy. *Frontiers in Psychiatry*, *9*(JAN). https://doi.org/10.3389/FPSYT.2018.00004

De Aquino, J. P., Londono, A., & Carvalho, A. F. (2018). An Update on the Epidemiology of Major Depressive Disorder Across Cultures. *Understanding Depression*, *1*, 309–315. https://doi.org/10.1007/978-981-10-6580-4_25

de Zambotti, M., Rosas, L., Colrain, I. M., & Baker, F. C. (2019). The Sleep of the Ring: Comparison of the ŌURA Sleep Tracker Against Polysomnography. *Behavioral Sleep Medicine*, *17*(2), 124–136. https://doi.org/10.1080/15402002.2017.1300587

DeMasi, O., Aguilera, A., & Recht, B. (2016). Detecting change in depressive symptoms from daily wellbeing questions, personality, and activity. *2016 IEEE Wireless Health, WH 2016*, 22–29. https://doi.org/10.1109/WH.2016.7764552

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. https://doi.org/10.1001/ARCHINTE.160.14.2101

Docherty, J. P. (1997). Barriers to the diagnosis of depression in primary care. . *The Journal of Clinical Psychiatry*, *58(Suppl 1)*, 5–10. https://psycnet.apa.org/record/1997-08036-001

Domhardt, M., Cuijpers, P., Ebert, D., & Baumeister, H. (2021). More Light? Opportunities and Pitfalls in Digitalized Psychotherapy Process Research. *Frontiers in Psychology*, *12*, 863. https://doi.org/10.3389/fpsyg.2021.544129

Domhardt, M., Steubl, L., & Baumeister, H. (2020). Internet- and Mobile-Based Interventions for Mental and Somatic Conditions in Children and Adolescents. *Zeitschrift Für Kinder- Und Jugendpsychiatrie Und Psychotherapie*, *48*(1), 33–46. https://doi.org/10.1024/1422-4917/a000625

Donker, T., Batterham, P. J., Warmerdam, L., Bennett, K., Bennett, A., Cuijpers, P., Griffiths, K. M., & Christensen, H. (2013). Predictors and moderators of response to internet-delivered Interpersonal Psychotherapy and Cognitive Behavior Therapy for depression. *Journal of Affective Disorders*, *151*(1), 343–351. https://doi.org/10.1016/j.jad.2013.06.020

Donker, T., Bennett, K., Bennett, A., Mackinnon, A., Van Straten, A., Cuijpers, P., Christensen, H., & Griffiths, K. M. (2013). Internet-delivered interpersonal psychotherapy versus internet-delivered cognitive behavioral therapy for adults with depressive symptoms: Randomized controlled noninferiority trial. *Journal of Medical Internet Research*, *15*(5), 1–18. https://doi.org/10.2196/jmir.2307

Donkin, L., Christensen, H., Naismith, S. L., Neal, B., Hickie, I. B., & Glozier, N. (2011). A Systematic Review of the Impact of Adherence on the Effectiveness of e-Therapies. *Journal of Medical Internet Research*, *13*(3), e52. https://doi.org/10.2196/jmir.1772

Dowrick, C. (2015). Computerised self help for depression in primary care. *BMJ (Online)*, *351*(November), 10–11. https://doi.org/10.1136/bmj.h5942

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, *35*(5–6), 352–359. https://doi.org/10.1016/S1532-0464(03)00034-0

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology*, *14*(1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Ebert, D., Buntrock, C., Lehr, D., Smit, F., Riper, H., Baumeister, H., Cuijpers, P., & Berking, M. (2018). Effectiveness of Web- and Mobile-Based Treatment of Subthreshold Depression With Adherence-Focused Guidance: A Single-Blind Randomized Controlled Trial. *Behavior Therapy*, *49*(1), 71–83. https://doi.org/10.1016/J.BETH.2017.05.004

Ebert, D., Donkin, L., Andersson, G., Andrews, G., Berger, T., Carlbring, P., Rozenthal, A., Choi, I., Laferton, J. A. C., Johansson, R., Kleiboer, A., Lange, A., Lehr, D., Reins, J. A., Funk, B., Newby, J., Perini, S., Riper, H., Ruwaard, J., … Cuijpers, P. (2016). Does Internet-based guided-self-help for depression cause harm? An individual participant data meta-analysis on deterioration rates and its moderators in randomized controlled trials. *Psychological Medicine*, *46*(13). https://doi.org/10.1017/S0033291716001562

Ebert, D., Gollwitzer, M., Riper, H., Cuijpers, P., Baumeister, H., & Berking, M. (2013). For whom does it work? Moderators of outcome on the effect of a transdiagnostic Internet-based maintenance treatment after inpatient psychotherapy: Randomized controlled trial. *Journal of Medical Internet Research*, *15*(10), 1–17. https://doi.org/10.2196/jmir.2511

Ebert, D., Nobis, S., Lehr, D., Baumeister, H., Riper, H., Auerbach, R. P., Snoek, F., Cuijpers, P., & Berking, M. (2017). The 6-month effectiveness of Internet-based guided self-help for depression in adults with Type 1 and 2 diabetes mellitus. *Diabetic Medicine*, *34*(1), 99–107. https://doi.org/10.1111/dme.13173

Ebert, D., Van Daele, T., Nordgreen, T., Karekla, M., Compare, A., Zarbo, C., Brugnera, A., Øverland, S., Trebbi, G., Jensen, K. L., Kaehlke, F., Baumeister, H., & Taylor, J. (2018). Internet- and Mobile-Based Psychological Interventions: Applications, Efficacy, and Potential for Improving Mental Health: A Report of the EFPA E-Health Taskforce. *European Psychologist*, *23*(2), 167–187. https://doi.org/10.1027/1016-9040/a000318

Ebert, D., Zarski, A. C., Christensen, H., Stikkelbroek, Y., Cuijpers, P., Berking, M., & Riper, H. (2015). Internet and computer-based cognitive behavioral therapy for anxiety and depression in youth: A meta-analysis of randomized controlled outcome trials. *PLoS ONE*, *10*(3), e0119895. https://doi.org/10.1371/journal.pone.0119895

Eden, J., Maslow, K., Le, M., Blazer, D., Populations, C. on the M. H. W. for G., Services, B. on H. C., & Medicine, I. of. (2012). The Mental Health and Substance Use Workforce for Older Adults. *The Mental Health and Substance Use Workforce for Older Adults: In Whose Hands?*, 1–396. https://doi.org/10.17226/13400

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ (Clinical Research Ed.)*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Egger, M., Smith, G. D., & Sterne, J. A. C. (2001). Uses and abuses of meta-analysis. *Clinical Medicine*, *1*(6), 478–484. https://doi.org/10.7861/CLINMEDICINE.1-6-478

Eichler, H. G., Abadie, E., Breckenridge, A., Flamion, B., Gustafsson, L. L., Leufkens, H., Rowland, M., Schneider, C. K., & Bloechl-Daum, B. (2011). Bridging the efficacy–effectiveness gap: a regulator's perspective on addressing variability of drug response. *Nature Reviews Drug Discovery 2011 10:7*, *10*(7), 495–506. https://doi.org/10.1038/nrd3501

Eisenberg, J. M., & Hershey, J. C. (1983). Derived Thresholds: Determining the Diagnostic Probabilities at Which Clinicians Initiate Testing and Treatment. *Medical Decision Making*, *3*(2), 155–168. https://doi.org/10.1177/0272989X8300300203

Enders, C. K. (2010). *Applied missing data analysis*. https://www.guilford.com/books/Applied-Missing-Data-Analysis/Craig-Enders/9781606236390

Ernst, A. F., & Albers, C. J. (2017). Regression assumptions in clinical psychology research practice-a systematic review of common misconceptions. *PeerJ*, *2017*(5). https://doi.org/10.7717/peerj.3323

Erskine, H. E., Moffitt, T. E., Copeland, W. E., Costello, E. J., Ferrari, A. J., Patton, G., Degenhardt, L., Vos, T., Whiteford, H. A., & Scott, J. G. (2015). A heavy burden on young minds: the global burden of mental and substance use disorders in children and youth. *Psychological Medicine*, *45*(7), 1551–1563. https://doi.org/10.1017/S0033291714002888

Evans-Lacko, S., Aguilar-Gaxiola, S., Al-Hamzawi, A., Alonso, J., Benjet, C., Bruffaerts, R., Chiu, W. T., Florescu, S., De Girolamo, G., Gureje, O., Haro, J. M., He, Y., Hu, C., Karam, E. G., Kawakami, N., Lee, S., Lund, C., Kovess-Masfety, V., Levinson, D., … Wojtyniak, B. (2018). Socio-economic variations in the mental health treatment gap forpeople with anxiety, mood, and substance use disorders: Results from the WHOWorld Mental Health (WMH) Surveys. *Psychological Medicine*, *48*(9), 1560. https://doi.org/10.1017/S0033291717003336

Eysenbach, G. (2005). The law of attrition. *Journal of Medical Internet Research*, *7*(1), 1–9. https://doi.org/10.2196/jmir.7.1.e11

Fagherazzi, G., Fischer, A., Ismael, M., & Despotovic, V. (2021). Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice. *Digital Biomarkers*, *5*(1), 78. https://doi.org/10.1159/000515346

Fairburn, C. G., & Patel, V. (2014). The global dissemination of psychological treatments: A road map for research and practice. *American Journal of Psychiatry*, *171*(5), 495–498. https://doi.org/10.1176/appi.ajp.2013.13111546

Farhan, A. A., Yue, C., Morillo, R., Ware, S., Lu, J., Bi, J., Kamath, J., Russell, A., Bamis, A., & Wang, B. (2016). Behavior vs. introspection: Refining prediction of clinical depression via smartphone sensing data. *2016 IEEE Wireless Health, WH 2016*, 30–37. https://doi.org/10.1109/WH.2016.7764553

Farvolden, P., Denisoff, E., Selby, P., Bagby, R. M., & Rudy, L. (2005). Usage and longitudinal effectiveness of a web-based self-help cognitive behavioral therapy program for panic disorder. *Journal of Medical Internet Research*, *7*(1), e129. https://doi.org/10.2196/jmir.7.1.e7

Faurholt-Jepsen, M., Vinberg, M., Frost, M., Debel, S., Margrethe Christensen, E., Bardram, J. E., & Kessing, L. V. (2016). Behavioral activities collected through smartphones and the association with illness activity in bipolar disorder. *International Journal of Methods in Psychiatric Research*, *25*(4), 309–323. https://doi.org/10.1002/MPR.1502

Felger, J. C., & Treadway, M. T. (2016). Inflammation Effects on Motivation and Motor Activity: Role of Dopamine. *Neuropsychopharmacology 2017 42:1*, *42*(1), 216–241. https://doi.org/10.1038/npp.2016.143

Fergusson, D. M., Horwood, L. J., Ridder, E. M., & Beautrais, A. L. (2005). Subthreshold Depression in Adolescence and Mental Health Outcomes in Adulthood. *Archives of General Psychiatry*, *62*(1), 66–72. https://doi.org/10.1001/ARCHPSYC.62.1.66

Ferreira, D., Kostakos, V., & Dey, A. K. (2015). AWARE: Mobile Context Instrumentation Framework. *Frontiers in ICT*, *2*(April), 1–9. https://doi.org/10.3389/fict.2015.00006

Firth, J., Torous, J., Nicholas, J., Carney, R., Pratap, A., Rosenbaum, S., & Sarris, J. (2017). The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. *World Psychiatry*, *16*(3), 287–298. https://doi.org/10.1002/wps.20472

Firth, J., Torous, J., Nicholas, J., Carney, R., Rosenbaum, S., & Sarris, J. (2017). Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. *Journal of Affective Disorders*, *218*, 15–22. https://doi.org/10.1016/j.jad.2017.04.046

Fisher, A. J., Bosley, H. G., Fernandez, K. C., Reeves, J. W., Soyster, P. D., Diamond, A. E., & Barkin, J. (2019). Open trial of a personalized modular treatment for mood and anxiety. *Behaviour Research and Therapy*, *116*, 69–79. https://doi.org/10.1016/J.BRAT.2019.01.010

Flückiger, C., Del, A. C., Wampold, B. E., & Horvath, A. O. (2018). The Alliance in Adult Psychotherapy: A Meta-Analytic Synthesis. *Psychotherapy*, *55*(4), 316–340. https://doi.org/10.1037/PST0000172

Foster, N. E., Anema, J. R., Cherkin, D., Chou, R., Cohen, S. P., Gross, D. P., Ferreira, P. H., Fritz, J. M., Koes, B. W., Peul, W., Turner, J. A., Maher, C. G., Buchbinder, R., Hartvigsen, J., Underwood, M., van Tulder, M., Menezes Costa, L., Croft, P., Ferreira, M., … Woolf, A. (2018). Prevention and treatment of low back pain: evidence, challenges, and promising directions. *Lancet (London, England)*, *391*(10137), 2368–2383. https://doi.org/10.1016/S0140-6736(18)30489-6

Furukawa, T. A., Noma, H., Caldwell, D. M., Honyashiki, M., Shinohara, K., Imai, H., Chen, P., Hunot, V., & Churchill, R. (2014). Waiting list may be a nocebo condition in psychotherapy trials: A contribution from network meta-analysis. *Acta Psychiatrica Scandinavica*, *130*(3), 181–192. https://doi.org/10.1111/acps.12275

Gagné, S., Vasiliadis, H. M., & Préville, M. (2014). Gender differences in general and specialty outpatient mental health service use for depression. *BMC Psychiatry*, *14*(1), 1–11. https://doi.org/10.1186/1471-244X-14-135

Garcia-Ceja, E., Riegler, M., Nordgreen, T., Jakobsen, P., Oedegaard, K. J., & Tørresen, J. (2018). Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive and Mobile Computing*, *51*, 1–26. https://doi.org/10.1016/J.PMCJ.2018.09.003

Garrido, S., Millington, C., Cheers, D., Boydell, K., Schubert, E., Meade, T., & Nguyen, Q. V. (2019). What Works and What Doesn't Work? A Systematic Review of Digital Mental Health Interventions for Depression and Anxiety in Young People. *Frontiers in Psychiatry*, *10*, 759. https://doi.org/10.3389/fpsyt.2019.00759

Gellatly, J., Bower, P., Hennessy, S., Richards, D., Gilbody, S., & Lovell, K. (2007). What makes self-help interventions effective in the management of depressive symptoms? Meta-analysis and meta-regression. *Psychological Medicine*, *37*(9), 1217–1228. https://doi.org/10.1017/S0033291707000062

Gilbody, S., Brabyn, S., Lovell, K., Kessler, D., Devlin, T., Smith, L., Araya, R., Barkham, M., Bower, P., Cooper, C., Knowles, S., Littlewood, E., Richards, D. A., Tallon, D., White, D., & Worthy, G. (2017). Telephone-supported computerised cognitive-behavioural therapy: REEACT-2 large-scale pragmatic randomised controlled trial. *British Journal of Psychiatry*, *210*(5), 362–367. https://doi.org/10.1192/bjp.bp.116.192435

Gilbody, S., Littlewood, E., Hewitt, C., Brierley, G., Tharmanathan, P., Araya, R., Barkham, M., Bower, P., Cooper, C., Gask, L., Kessler, D., Lester, H., Lovell, K., Parry, G., Richards, D. A., Andersen, P., Brabyn, S., Knowles, S., Shepherd, C., … White, D. (2015a). Computerised cognitive behaviour therapy (cCBT) as treatment for depression in primary care (REEACT trial): Large scale pragmatic randomised controlled trial. *The BMJ*, *351*, 2–5. https://doi.org/10.1136/bmj.h5627

Gilbody, S., Littlewood, E., Hewitt, C., Brierley, G., Tharmanathan, P., Araya, R., Barkham, M., Bower, P., Cooper, C., Gask, L., Kessler, D., Lester, H., Lovell, K., Parry, G., Richards, D. A., Andersen, P., Brabyn, S., Knowles, S., Shepherd, C., … White, D. (2015b). Computerised cognitive behaviour therapy (cCBT) as treatment for depression in primary care (REEACT trial): Large scale pragmatic randomised controlled trial- Rapid Responses. *The BMJ*, *351*, 2–5. https://doi.org/10.1136/bmj.h5627

Gladstone, T., Terrizzi, D., Stinson, A., Nidetz, J., Canel, J., Ching, E., Berry, A., Cantorna, J., Fogel, J., Eder, M., Bolotin, M., Thomann, L. O., Griffith, K., Ip, P., Aaby, D. A., Brown, C. H., Beardslee, W., Bell, C., Crawford, T. J., … Van Voorhees, B. W. (2018). Effect of Internet-based Cognitive Behavioral Humanistic and Interpersonal Training vs. Internet-based General Health Education on Adolescent Depression in Primary Care: A Randomized Clinical Trial. *JAMA Network Open*, *1*(7), 1–15. https://doi.org/10.1001/jamanetworkopen.2018.4278

Godwin, M., Ruhland, L., Casson, I., MacDonald, S., Delva, D., Birtwhistle, R., Lam, M., & Seguin, R. (2003). Pragmatic controlled clinical trials in primary care: The struggle between external and internal validity. *BMC Medical Research Methodology*, *3*, 1–7.

https://doi.org/10.1186/1471-2288-3-28

Gold, S. M., Enck, P., Hasselmann, H., Friede, T., Hegerl, U., Mohr, D. C., & Otte, C. (2017). Control conditions for randomised trials of behavioural interventions in psychiatry: a decision framework. In *The Lancet Psychiatry* (Vol. 4, Issue 9, pp. 725–732). Elsevier Ltd. https://doi.org/10.1016/S2215-0366(17)30153-0

Goldberg, D. (2010). The detection and treatment of depression in the physically ill. *World Psychiatry*, *9*(1), 16–20. https://doi.org/10.1002/j.2051-5545.2010.tb00256.x

Goldberg, D. (2011). The heterogeneity of "major depression." *World Psychiatry*, *10*(3), 226. https://doi.org/10.1002/J.2051-5545.2011.TB00061.X

Goldstein, H. (1995). Hierarchical Data Modeling in the Social Sciences. *Journal of Educational and Behavioral Statistics*, *20*(2), 201–204. https://doi.org/10.3102/10769986020002201

Goldstein, S. P., Evans, B. C., Flack, D., Juarascio, A., Manasse, S., Zhang, F., & Forman, E. M. (2017). Return of the JITAI: Applying a Just-in-Time Adaptive Intervention Framework to the Development of m-Health Solutions for Addictive Behaviors. *International Journal of Behavioral Medicine*, *24*(5), 673–682. https://doi.org/10.1007/s12529-016-9627-y

Graham, A. K., Lattie, E. G., & Mohr, D. C. (2019). Experimental Therapeutics for Digital Mental Health. *JAMA Psychiatry*, *76*(12). https://doi.org/10.1001/jamapsychiatry.2019.2075

Greenhalgh, T., Wherton, J., Papoutsi, C., Lynch, J., Hughes, G., A'Court, C., Hinder, S., Fahy, N., Procter, R., & Shaw, S. (2017). Beyond adoption: A new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *Journal of Medical Internet Research*, *19*(11), e8775. https://doi.org/10.2196/jmir.8775

Guidelines for ethical review in human sciences | Tutkimuseettinen neuvottelukunta. (n.d.). Retrieved November 2, 2020, from https://tenk.fi/en/advice-and-materials/guidelines-ethical-review-human-sciences

Guo, Y., Hong, Y. A., Cai, W., Li, L., Hao, Y., Qiao, J., Xu, Z., Zhang, H., Zeng, C., Liu, C., Li, Y., Zhu, M., Zeng, Y., & Penedo, F. J. (2020). Effect of a WeChat-Based intervention (Run4Love) on depressive symptoms among people living with HIV in China: A randomized controlled trial. *Journal of Medical Internet Research*, *22*(2). https://doi.org/10.2196/16715

Hamilton, M. (1960). A RATING SCALE FOR DEPRESSION. *Journal of Neurology, Neurosurgery & Psychiatry*, *23*(1), 56–62. https://doi.org/10.1136/jnnp.23.1.56

Harvey, A. G., Murray, G., Chandler, R. A., & Soehner, A. (2011). Sleep disturbance as transdiagnostic: consideration of neurobiological mechanisms. *Clinical Psychology Review*, *31*(2), 225–235. https://doi.org/10.1016/J.CPR.2010.04.003

Hayes, S. C., Strosahl, K., & Wilson, K. G. (1999). *Acceptance and Commitment Therapy: An experiential approach to behavior change.* . New York: Guilford Press. https://contextualscience.org/publications/hayes_strosahl_wilson_1999

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect size and Related Estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Heifner, C. (1997). The Male Experience of Depression. *Perspectives in Psychiatric Care*, *33*(2), 10–18. https://doi.org/10.1111/J.1744-6163.1997.TB00536.X

Hernández, N., Yavuz, G., Eşrefoğlu, R., Kepez, T., Özdemir, A., Demiray, B., Alan, H., Ersoy, C., Untersander, S., & Arnrich, B. (2015). Thought and Life Logging: A Pilot Study. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9454*, 26–36. https://doi.org/10.1007/978-3-319-26401-1_3

Hill, C. E. (1996). Dreams and therapy. *Psychotherapy Research*, *6*(1), 1–15. https://doi.org/10.1080/10503309612331331538

Hirschfeld, R. M. A. (2012). The Epidemiology of Depression and the Evolution of Treatment. *The Primary Care Companion for CNS Disorders*, *14*(Suppl 1: Editor Choice), 26328. https://doi.org/10.4088/JCP.11096SU1C.01

Holländare, F., Gustafsson, S. A., Berglind, M., Grape, F., Carlbring, P., Andersson, G., Hadjistavropoulos, H., & Tillfors, M. (2016). Therapist behaviours in internet-based cognitive behaviour therapy (ICBT) for depressive symptoms. *Internet Interventions*, *3*, 1–7. https://doi.org/10.1016/j.invent.2015.11.002

Holmes, E. A., Ghaderi, A., Harmer, C. J., Ramchandani, P. G., Cuijpers, P., Morrison, A. P., Roiser, J. P., Bockting, C. L. H. H., O'Connor, R. C., Shafran, R., Moulds, M. L., & Craske, M. G. (2018). The Lancet Psychiatry Commission on psychological treatments research in tomorrow's science. *The Lancet Psychiatry*, *5*(3), 237–286. https://doi.org/10.1016/S2215-0366(17)30513-8

Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression. John Wiley & Sons, Inc. In *New York* (Vol. 23, Issue 1).

Huibers, M. J. H., Lorenzo-Luaces, L., Cuijpers, P., & Kazantzis, N. (2021). On the Road to Personalized Psychotherapy: A Research Agenda Based on Cognitive Behavior Therapy for Depression. *Frontiers in Psychiatry*, *11*(January), 1–14. https://doi.org/10.3389/fpsyt.2020.607508

Ilanković, A., Damjanović, A., Ilanković, V., Filipović, B., Janković, S., & Ilanković, N. (2014). Polysomnographic Sleep Patterns in Depressive, Schizophrenic and Healthy Subjects. *Psychiatria Danubina*, *26*(1), 20–26.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., & Wang, P. (2010). Research Domain Criteria (RDoC): Toward a new classification framework for research on mental disorders. In *American Journal of Psychiatry* (Vol. 167, Issue 7, pp. 748–751). American Psychiatric Association . https://doi.org/10.1176/appi.ajp.2010.09091379

Ioannidis, J. P. A., Patsopoulos, N. A., & Evangelou, E. (2007). Uncertainty in heterogeneity estimates in meta-analyses. In *British Medical Journal* (Vol. 335, Issue 7626, pp. 914–916). BMJ Publishing Group. https://doi.org/10.1136/bmj.39343.408449.80

Ip, P., Chim, D., Chan, K. L., Li, T. M. H., Ho, F. K. W., Van Voorhees, B. W., Tiwari, A., Tsang, A., Chan, C. W. L., Ho, M., Tso, W., & Wong, W. H. S. (2016). Effectiveness of a culturally attuned Internet-based depression prevention program for Chinese adolescents: A randomized controlled trial. *Depression and Anxiety*, *33*(12), 1123–1131.

https://doi.org/10.1002/da.22554

Jackson, D., Turner, R., Rhodes, K., & Viechtbauer, W. (2014). Methods for calculating confidence and credible intervals for the residual between-study variance in random effects meta-regression models. *BMC Medical Research Methodology*, *14*(1), 103. https://doi.org/10.1186/1471-2288-14-103

Jacobson, N. S., Dobson, K. S., Truax, P. A., Addis, M. E., Koerner, K., Gollan, J. K., Gortner, E., & Prince, S. E. (1996). A component analysis of cognitive - Behavioral treatment for depression. *Journal of Consulting and Clinical Psychology*, *64*(2), 295–304. https://doi.org/10.1037/0022-006X.64.2.295

Jarrett, R. B., Minhajuddin, A., Kangas, J. L., Friedman, E. S., Callan, J. A., & Thase, M. E. (2013). Acute Phase Cognitive Therapy for Recurrent Major Depressive Disorder: Who Drops Out and How Much do Patient Skills Influence Response? *Behaviour Research and Therapy*, *51*(0), 221. https://doi.org/10.1016/J.BRAT.2013.01.006

Jetté, M., Sidney, K., & Blümchen, G. (1990). Metabolic equivalents (METS) in exercise testing, exercise prescription, and evaluation of functional capacity. *Clinical Cardiology*, *13*(8), 555–565. https://doi.org/10.1002/clc.4960130809

Johansson, O., Michel, T., Andersson, G., & Paxling, B. (2015). Experiences of non-adherence to Internet-delivered cognitive behavior therapy: A qualitative study. *Internet Interventions*, *2*(2), 137–142. https://doi.org/10.1016/j.invent.2015.02.006

Johnsen, T. J., & Friborg, O. (2015). The effects of cognitive behavioral therapy as an anti-depressive treatment is falling: A meta-analysis. *Psychological Bulletin*, *141*(4), 747–768. https://doi.org/10.1037/bul0000015

Jonsson, T., Christrup, L. L., Højsted, J., Villesen, H. H., Albjerg, T. H., Ravn-Nielsen, L. V., & Sjøgren, P. (2011). Symptoms and side effects in chronic non-cancer pain: patient report vs. systematic assessment. *Acta Anaesthesiologica Scandinavica*, *55*(1), 69–74. https://doi.org/10.1111/J.1399-6576.2010.02329.X

Karyotaki, E., Ebert, D., Donkin, L., Riper, H., Twisk, J., Burger, S., Rozental, A., Lange, A., Williams, A. D., Zarski, A. C., Geraedts, A., van Straten, A., Kleiboer, A., Meyer, B., Ünlü Ince, B. B., Buntrock, C., Lehr, D., Snoek, F. J., Andrews, G., … Cuijpers, P. (2018). Do guided internet-based interventions result in clinically relevant changes for patients with depression? An individual participant data meta-analysis. *Clinical Psychology Review*, *63*(June), 80–92. https://doi.org/10.1016/j.cpr.2018.06.007

Karyotaki, E., Efthimiou, O., Miguel, C., Bermpohl, F. M. G., Furukawa, T. A., Cuijpers, P., Riper, H., Patel, V., Mira, A., Gemmil, A. W., Yeung, A. S., Lange, A., Williams, A. D., Mackinnon, A., Geraedts, A., Van Straten, A., Meyer, B., Björkelund, C., Knaevelsrud, C., … Forsell, Y. (2021). Internet-Based Cognitive Behavioral Therapy for Depression: A Systematic Review and Individual Patient Data Network Meta-analysis. *JAMA Psychiatry*, *78*(4), 361–371. https://doi.org/10.1001/jamapsychiatry.2020.4364

Karyotaki, E., Kleiboer, A., Smit, F., Turner, D. T., Pastor, A. M., Andersson, G., Berger, T., Botella, C., Breton, J. M., Carlbring, P., Christensen, H., De Graaf, E., Griffiths, K., Donker, T., Farrer, L., Huibers, M. J. H., Lenndin, J., Mackinnon, A., Meyer, B., … Cuijpers, P. (2015). Predictors of treatment dropout in self-guided web-based interventions for depression: An “individual patient data” meta-analysis. *Psychological Medicine*, *45*(13), 2717–2726. https://doi.org/10.1017/S0033291715000665

Karyotaki, E., Riper, H., Twisk, J., Hoogendoorn, A., Kleiboer, A., Mira, A., MacKinnon, A., Meyer, B., Botella, C., Littlewood, E., Andersson, G., Christensen, H., Klein, J. P., Schröder, J., Bretón-López, J., Scheider, J., Griffiths, K., Farrer, L., Huibers, M. J. H. H., … Cuijpers, P. (2017). Efficacy of Self-guided Internet-Based Cognitive Behavioral Therapy in the Treatment of Depressive Symptoms: A Meta-analysis of Individual Participant Data. *JAMA Psychiatry*, *74*(4), 351–359. https://doi.org/10.1001/jamapsychiatry.2017.0044

Karyotaki, E., Smit, Y., Holdt Henningsen, K., Huibers, M. J. H., Robays, J., de Beurs, D., & Cuijpers, P. (2016). Combining pharmacotherapy and psychotherapy or monotherapy for major depression? A meta-analysis on the long-term effects. *Journal of Affective Disorders*, *194*, 144–152. https://doi.org/10.1016/j.jad.2016.01.036

Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. In *Annual Review of Clinical Psychology* (Vol. 3, pp. 1–27). Annual Reviews. https://doi.org/10.1146/annurev.clinpsy.3.022806.091432

Kazdin, A. E. (2009). Understanding how and why psychotherapy leads to change. *Psychotherapy Research: Journal of the Society for Psychotherapy Research*, *19*(4–5), 418–428. https://doi.org/10.1080/10503300802448899

Kazdin, A. E. (2017). Addressing the treatment gap: A key challenge for extending evidence-based psychosocial interventions. *Behaviour Research and Therapy*, *88*, 7–18. https://doi.org/10.1016/J.BRAT.2016.06.004

Kazdin, A. E., & Blase, S. L. (2011). Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspectives on Psychological Science*, *6*(1), 21–37. https://doi.org/10.1177/1745691610393527

Kelders, S. M., Kok, R. N., Ossebaard, H. C., & Gemert-Pijnen, J. E. W. C. van. (2012). Persuasive system design does matter: a systematic review of adherence to web-based interventions. *Journal of Medical Internet Research*, *14*(6), e152. https://doi.org/10.2196/JMIR.2104

Kessler, R. C., & Bromet, E. J. (2013). The epidemiology of depression across cultures. In *Annual Review of Public Health* (Vol. 34, pp. 119–138). Annual Reviews. https://doi.org/10.1146/annurev-publhealth-031912-114409

KFF. (2021). *Mental Health Care Health Professional Shortage Areas (HPSAs)*. https://www.kff.org/other/state-indicator/mental-health-care-health-professional-shortage-areas-hpsas

Kiesler, D. J. (1997). Contemporary interpersonal theory and research: personality, psychopathology, and psychotherapy. *Journal of Psychotherapy Practice and Research*, *6*(4), 339–341.

Kiluk, B. D., Sugarman, D. E., Nich, C., Gibbons, C. J., Martino, S., Rounsaville, B. J., & Carroll, K. M. (2011). A methodological analysis of randomized clinical trials of computer-assisted therapies for psychiatric disorders: Toward improved standards for an emerging field. *American Journal of Psychiatry*, *168*(8), 790–799. https://doi.org/10.1176/appi.ajp.2011.10101443

Kinnunen, H., Rantanen, A., Kentt, T., & Koskim ki, H. (2020). Feasible assessment of recovery and cardiovascular health: Accuracy of nocturnal HR and HRV assessed via ring PPG in comparison to medical grade ECG. *Physiological Measurement*, *41*(4). https://doi.org/10.1088/1361-6579/ab840a

Klein, B., Austin, D., Pier, C., Kiropoulos, L., Shandley, K., Mitchell, J., Gilson, K., & Ciechomski, L. (2009). Internet-based treatment for panic disorder: Does frequency of therapist contact make a difference? *Cognitive Behaviour Therapy*, *38*(2), 100–113. https://doi.org/10.1080/16506070802561132

Klein Hofmeijer-Sevink, M., Batelaan, N. M., Van Megen, H. J. G. M., Penninx, B. W., Cath, D. C., Van Den Hout, M. A., & Van Balkom, A. J. L. M. (2012). Clinical relevance of comorbidity in anxiety disorders: A report from the Netherlands Study of Depression and Anxiety (NESDA). *Journal of Affective Disorders*, *137*(1–3), 106–112. https://doi.org/10.1016/j.jad.2011.12.008

Knowles, S., Lovell, K., Bower, P., Gilbody, S., Littlewood, E., & Lester, H. (2015). Patient experience of computerised therapy for depression in primary care. *BMJ Open*, *5*(11). https://doi.org/10.1136/bmjopen-2015-008581

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

Kok, R. N., Beekman, A. T. F., Cuijpers, P., & van Straten, A. (2017). Adherence to a web-based pre-treatment for phobias in outpatient clinics. *Internet Interventions*, *9*(May), 38–45. https://doi.org/10.1016/j.invent.2017.05.004

Konigbauer, J., Letsch, J., Doebler, P., Ebert, D., & Baumeister, H. (2017). Internet- and mobile-based depression interventions for people with diagnosed depression: A systematic review and meta-analysis. *Journal of Affective Disorders*, *223*, 28–40. https://doi.org/https://dx.doi.org/10.1016/j.jad.2017.07.021

Königbauer, J., Letsch, J., Doebler, P., Ebert, D., & Baumeister, H. (2017). Internet- and mobile-based depression interventions for people with diagnosed depression: A systematic review and meta-analysis. *Journal of Affective Disorders*, *223*(July), 28–40. https://doi.org/10.1016/j.jad.2017.07.021

Kraemer, H., Frank, E., & Kupfer, D. J. (2006). Moderators of Treatment Outcomes: Clinical, Research, and Policy Importance. *JAMA*, *296*(10), 1286–1289. https://doi.org/10.1001/JAMA.296.10.1286

Kraemer, H., Wilson, G. T., Fairburn, C. G., & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, *59*(10), 877–883. https://doi.org/10.1001/ARCHPSYC.59.10.877

Lawler, K., Earley, C., Timulak, L., Enrique, A., & Richards, D. (2021). Dropout from an internet-delivered cognitive behavioral therapy intervention for adults with depression and anxiety: Qualitative study. *JMIR Formative Research*, *5*(11), 1–19. https://doi.org/10.2196/26221

Lee, Y., Brietzke, E., Cao, B., Chen, Y., Linnaranta, O., Mansur, R. B., Cortes, P., Kösters, M., Majeed, A., Tamura, J. K., Lui, L. M. W., Vinberg, M., Keinänen, J., Kisely, S., Naveed, S., Barbui, C., Parker, G., Owolabi, M., Nishi, D., … McIntyre, R. S. (2020). Development and implementation of guidelines for the management of depression: a systematic review. *Bulletin of the World Health Organization*, *98*(10), 683. https://doi.org/10.2471/BLT.20.251405

Lee, Y., Ragguett, R. M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A., Brietzke, E., Lin, K., Pan, Z., Subramaniapillai, M., Chan, T. C. Y., Fus, D., Park, C., Musial, N., Zuckerman, H., Chen, V. C. H., Ho, R., Rong, C., & McIntyre, R. S. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders*, *241*(August), 519–532. https://doi.org/10.1016/j.jad.2018.08.073

Lemmens, L. H. J. M., Müller, V. N. L. S., Arntz, A., & Huibers, M. J. H. (2016). Mechanisms of change in psychotherapy for depression: An empirical update and evaluation of research aimed at identifying psychological mediators. *Clinical Psychology Review*, *50*, 95–107. https://doi.org/10.1016/j.cpr.2016.09.004

Lin, J., Sander, L. B., Paganini, S., Schlicker, S., Ebert, D., Berking, M., Bengel, J., Nobis, S., Lehr, D., Mittag, O., Riper, H., & Baumeister, H. (2017). Effectiveness and cost-effectiveness of a guided internet- and mobile-based depression intervention for individuals with chronic back pain: Protocol of a multi-centre randomised controlled trial. *BMJ Open*, *7*(12), 1–11. https://doi.org/10.1136/bmjopen-2016-015226

Linardon, J., Cuijpers, P., Carlbring, P., Messer, M., & Fuller-Tyszkiewicz, M. (2019). The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. *World Psychiatry*, *18*(3), 325–336. https://doi.org/10.1002/wps.20673

Lopez Pinaya, W. H., Vieira, S., Garcia-Dias, R., & Mechelli, A. (2020). Autoencoders. *Machine Learning: Methods and Applications to Brain Disorders*, 193–208. https://doi.org/10.1016/B978-0-12-815739-8.00011-0

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*(3), 335–343. https://doi.org/10.1016/0005-7967(94)00075-U

Low, C. A., Li, M., Vega, J., Durica, K. C., Ferreira, D., Tam, V., Hogg, M., Zeh, H., Doryab, A., & Dey, A. K. (2021). Digital biomarkers of symptom burden self-reported by perioperative patients undergoing pancreatic surgery: Prospective longitudinal study. *JMIR Cancer*, *7*(2), e27975. https://doi.org/10.2196/27975

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence 2020 2:1*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

MacCallum, R. C., Kim, C., Malarkey, W. B., & Kiecolt-Glaser, J. K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research*, *32*(3), 215–253. https://doi.org/10.1207/s15327906mbr3203_1

MacKenzie, C. S., Reynolds, K., Cairney, J., Streiner, D. L., & Sareen, J. (2012). Disorder-specific mental health service use for mood and anxiety disorders: associations with age, sex, and psychiatric comorbidity. *Depression and Anxiety*, *29*(3), 234–242. https://doi.org/10.1002/DA.20911

Malhi, G. S., & Mann, J. J. (2018). Depression. *The Lancet*, *392*(10161), 2299–2312. https://doi.org/10.1016/S0140-6736(18)31948-2

Mannion, A. F., Junge, A., Grob, D., Dvorak, J., & Fairbank, J. C. T. (2006). Development of a German version of the Oswestry Disability Index. Part 2: sensitivity to change after spinal surgery. *European Spine Journal : Official Publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society*, *15*(1), 66–73. https://doi.org/10.1007/s00586-004-0816-z

Mastoras, R. E., Iakovakis, D., Hadjidimitriou, S., Charisis, V., Kassie, S., Alsaadi, T., Khandoker, A., & Hadjileontiadis, L. J. (2019). Touchscreen typing pattern analysis for remote detection of the depressive tendency. *Scientific Reports 2019 9:1*, *9*(1), 1–12. https://doi.org/10.1038/s41598-019-50002-9

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. *The Journal of Clinical Psychiatry*, *74*(06), 595–602. https://doi.org/10.4088/JCP.12r07757

McQuaid, J. R., Stein, M. B., Laffaye, C., & McCahill, M. E. (1999). Depression in a Primary Care Clinic: the Prevalence and Impact of an Unrecognized Disorder. *Journal of Affective Disorders*, *55*(1), 1–10. https://doi.org/10.1016/S0165-0327(98)00191-8

Mehrotra, A., Hendley, R., & Musolesi, M. (2016). Towards multi-modal anticipatory monitoring of depressive states through the analysis of human-smartphone interaction. *UbiComp 2016 Adjunct - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 1132–1138. https://doi.org/10.1145/2968219.2968299

Meyer, B., Berger, T., Caspar, F., Beevers, C. G., Andersson, G., & Weiss, M. (2009). Effectiveness of a novel integrative online treatment for depression (Deprexis): Randomized controlled trial. *Journal of Medical Internet Research*, *11*(2), 1–18. https://doi.org/10.2196/jmir.1151

Mohr, D. C., Cuijpers, P., & Lehman, K. (2011). Supportive accountability: A model for providing human support to enhance adherence to eHealth interventions. *Journal of Medical Internet Research*, *13*(1). https://doi.org/10.2196/jmir.1602

Mohr, D. C., Ho, J., Duffecy, J., Baron, K. G., Lehman, K. A., Jin, L., & Reifler, D. (2010). Perceived barriers to psychological treatments and their relationship to depression. *Journal of Clinical Psychology*, *66*(4), 394–409. https://doi.org/10.1002/jclp.20659

Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annual Review of Clinical Psychology*, *13*, 23–47. https://doi.org/10.1146/annurev-clinpsy-032816-044949

Möller-Leimkühler, A. M. (2002). Barriers to help-seeking by men: A review of sociocultural and clinical literature with particular reference to depression. In *Journal of Affective Disorders* (Vol. 71, Issues 1–3, pp. 1–9). https://doi.org/10.1016/S0165-0327(01)00379-2

Moshe, I., Terhorst, Y., Cuijpers, P., Cristea, I., Pulkki-Råback, L., & Sander, L. B. (2020). Three Decades of Internet- and Computer-Based Interventions for the Treatment of Depression: Protocol for a Systematic Review and Meta-Analysis. *JMIR Research Protocols*, *9*(3), e14860. https://doi.org/10.2196/14860

Munder, T., & Barth, J. (2018). Cochrane's risk of bias tool in the context of psychotherapy outcome research. *Psychotherapy Research*, *28*(3), 347–355. https://doi.org/10.1080/10503307.2017.1411628

Musca, S. C., Kamiejski, R., Nugier, A., Méot, A., Er-Rafiy, A., & Brauer, M. (2011). Data with Hierarchical Structure: Impact of Intraclass Correlation and Sample Size on Type-I Error. *Frontiers in Psychology*, *2*(APR), 74. https://doi.org/10.3389/fpsyg.2011.00074

Musiat, P., Johnson, C., Atkinson, M., Wilksch, S., & Wade, T. (2022). Impact of guidance on intervention adherence in computerised interventions for mental health problems: a meta-analysis. *Psychological Medicine*, *52*(2), 229–240. https://doi.org/10.1017/S0033291721004621

Musiat, P., & Tarrier, N. (2014). Collateral outcomes in e-mental health: a systematic review of the evidence for added benefits of computerized cognitive behavior therapy interventions for mental health. *Psychological Medicine*, *44*(15), 3137–3150. https://doi.org/10.1017/S0033291714000245

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., & Murphy, S. A. (2018). Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, *52*(6), 446–462. https://doi.org/10.1007/s12160-016-9830-8

National Collaborating Centre for Mental Health (Great Britain). (2021). *The Improving Access to Psychological Therapies Manual*.

Nezlek, J. B. (2001). Multilevel Random Coefficient Analyses of Event- and Interval-Contingent Data in Social and Personality Psychology Research. *Personality and Social Psychology Bulletin*, *27*(7), 771–785. https://doi.org/10.1177/0146167201277001

Nezlek, J. B. (2012). Multilevel modeling for psychologists. In *APA handbook of research methods in psychology, Vol 3: Data analysis and research publication.* (pp. 219–241). American Psychological Association. https://doi.org/10.1037/13621-011

Nezu, A. M. (1986). Efficacy of a Social Problem-Solving Therapy Approach for Unipolar Depression. *Journal of Consulting and Clinical Psychology*, *54*(2), 196–202. https://doi.org/10.1037/0022-006X.54.2.196

NICE. (2017). *NICE guideline for treatment of depression*. https://www.nice.org.uk/guidance/gid-cgwave0725/documents/short-version-of-draft-guideline-2

Nielssen, O., Staples, L. G., Ryan, K., Karin, E., Kayrouz, R., Dear, B. F., Cross, S., & Titov, N. (2022). Suicide after contact with a national digital mental health service. *Internet Interventions*, *28*, 100516. https://doi.org/10.1016/j.invent.2022.100516

Nishiyama, Y., Ferreira, D., Eigen, Y., Sasaki, W., Okoshi, T., Nakazawa, J., Dey, A. K., & Sezaki, K. (2020). IOS Crowd–Sensing Won't Hurt a Bit!: AWARE Framework and

Sustainable Study Guideline for iOS Platform. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12203 LNCS*, 223–243. https://doi.org/10.1007/978-3-030-50344-4_17/FIGURES/11

O'Moore, K. A., Newby, J. M., Andrews, G., Hunter, D. J., Bennell, K., Smith, J., & Williams, A. D. (2018). Internet Cognitive–Behavioral Therapy for Depression in Older Adults With Knee Osteoarthritis: A Randomized Controlled Trial. *Arthritis Care & Research*, *70*(1), 61–70. https://doi.org/10.1002/acr.23257

OECD. (2022). *Suicide rates (indicator)*. https://data.oecd.org/healthstat/suicide-rates.htm

Ormel, J., Cuijpers, P., Jorm, A. F., & Schoevers, R. (2019). Prevention of depression will only succeed when it is structurally embedded and targets big determinants. *World Psychiatry*, *18*(1), 111. https://doi.org/10.1002/WPS.20580

Ormel, J., Kessler, R. C., & Schoevers, R. (2019). Depression: More treatment but no drop in prevalence: How effective is treatment? and can we do better? *Current Opinion in Psychiatry*, *32*(4), 348–354. https://doi.org/10.1097/YCO.0000000000000505

Ormel, J., Vonkorff, M., Ustun, T. B., Pini, S., Korten, A., & Oldehinkel, T. (1994). Common Mental Disorders and Disability Across Cultures: Results From the WHO Collaborative Study on Psychological Problems in General Health Care. *JAMA*, *272*(22), 1741–1748. https://doi.org/10.1001/JAMA.1994.03520220035028

Orrù, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., & Mechelli, A. (2012). Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. *Neuroscience & Biobehavioral Reviews*, *36*(4), 1140–1152. https://doi.org/10.1016/J.NEUBIOREV.2012.01.004

Pagoto, S. L., & Lemon, S. C. (2013). Efficacy vs Effectiveness. *JAMA Internal Medicine*, *173*(13), 1262–1263. https://doi.org/10.1001/JAMAINTERNMED.2013.6521

Pastor, D. A., & Lazowski, R. A. (2018). On the Multilevel Nature of Meta-Analysis: A Tutorial, Comparison of Software Programs, and Discussion of Analytic Choices. *Multivariate Behavioral Research*, *53*(1), 74–89. https://doi.org/10.1080/00273171.2017.1365684

Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology*, *31*(2), 109–118. https://doi.org/10.1037/h0024436

Pearson, J. L., Stanley, B., King, C. A., & Fisher, C. B. (2001). Intervention research with persons at high risk for suicidality: Safety and ethical considerations. *Journal of Clinical Psychiatry*, *62*(SUPPL. 25), 17–26.

Pedregosa, F., Weiss, R., Brucher, M., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://scikit-learn.sourceforge.net.

Pennebaker, J. W. (2004). *Writing to heal : a guided journal for recovering from trauma & emotional upheaval*. New Harbinger Publications.

Perera-Delcourt, R. P., & Sharkey, G. (2019). Patient experience of supported computerized

CBT in an inner-city IAPT service: a qualitative study. *The Cognitive Behaviour Therapist*, *12*. https://doi.org/10.1017/S1754470X18000284

Piccinini, F., Martinelli, G., & Carbonaro, A. (2020). Accuracy of Mobile Applications versus Wearable Devices in Long-Term Step Measurements. *Sensors 2020, Vol. 20, Page 6293*, *20*(21), 6293. https://doi.org/10.3390/S20216293

Pigott, H. E., Leventhal, A. M., Alter, G. S., & Boren, J. J. (2010). Efficacy and Effectiveness of Antidepressants: Current Status of Research. *Psychotherapy and Psychosomatics*, *79*(5), 267–279. https://doi.org/10.1159/000318293

Pollock, K. M. (2001). Exercise in treating depression: Broadening the psychotherapist's role. *Journal of Clinical Psychology*, *57*(11), 1289–1300. https://doi.org/10.1002/JCLP.1097

Poulos, J., & Valle, R. (2018). Missing Data Imputation for Supervised Learning. *Applied Artificial Intelligence*, *32*(2), 186–196. https://doi.org/10.1080/08839514.2018.1448143

Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, *10*(1), 57–71. https://doi.org/10.1002/jrsm.1332

Racine, N., McArthur, B. A., Cooke, J. E., Eirich, R., Zhu, J., & Madigan, S. (2021). Global Prevalence of Depressive and Anxiety Symptoms in Children and Adolescents During COVID-19: A Meta-analysis. *JAMA Pediatrics*, *175*(11), 1142–1150. https://doi.org/10.1001/JAMAPEDIATRICS.2021.2482

Rayner, L., Hotopf, M., Petkova, H., Matcham, F., Simpson, A., & Mccracken, L. M. (2016). Depression in patients with chronic pain attending a specialised pain treatment centre: prevalence and impact on health care costs. *Pain*, *157*(7), 1472. https://doi.org/10.1097/J.PAIN.0000000000000542

Revelle, W. (2020). *Procedures for Psychological, Psychometric, and Personality Research [R package psych version 1.9.12.31]*. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/package=psych

Reynolds, S., Barkham, M., Stiles, W. B., Shapiro, D. A., Hardy, G. E., & Rees, A. (1996). Acceleration of changes in session impact during contrasting time- limited psychotherapies. *Journal of Consulting and Clinical Psychology*, *64*(3), 577–586. https://doi.org/10.1037/0022-006X.64.3.577

Richards, D., Enrique, A., Eilert, N., Franklin, M., Palacios, J., Duffy, D., Earley, C., Chapman, J., Jell, G., Sollesse, S., & Timulak, L. (2020). A pragmatic randomized waitlist-controlled effectiveness and cost-effectiveness trial of digital interventions for depression and anxiety. *Npj Digital Medicine*, *3*(1). https://doi.org/10.1038/s41746-020-0293-8

Rohani, D. A., Faurholt-Jepsen, M., Kessing, L. V., & Bardram, J. E. (2018). Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptoms in patients with affective disorders: Systematic review. *Journal of Medical Internet Research*, *20*(8). https://doi.org/10.2196/mhealth.9691

Rozgonjuk, D., Pruunsild, P., Jürimäe, K., Schwarz, R. J., & Aru, J. (2020). Instagram use frequency is associated with problematic smartphone use, but not with depression and anxiety symptom severity. *Mobile Media and Communication*, *8*(3), 400–418.

https://doi.org/10.1177/2050157920910190

Rubin, D. B. (1996). Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*, *91*(434), 473–489. https://doi.org/10.1080/01621459.1996.10476908

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D. C. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research*, *17*(7), 1–11. https://doi.org/10.2196/jmir.4273

Saletu-Zyhlarz, G., Saletu, B., Anderer, P., Brandstätter, N., Frey, R., Gruber, G., Klösch, G., Mandl, M., Grünberger, J., & Linzmayer, L. (1997). Nonorganic Insomnia in Generalized Anxiety Disorder. *Neuropsychobiology*, *36*(3), 117–129. https://doi.org/10.1159/000119373

Sander, L. B., Gerhardinger, K., Bailey, E., Robinson, J., Lin, J., Cuijpers, P., & Mühlmann, C. (2020). Suicide risk management in research on internet-based interventions for depression: A synthesis of the current state and recommendations for future research. *Journal of Affective Disorders*, *263*, 676–683. https://doi.org/10.1016/J.JAD.2019.11.045

Sander, L. B., Paganini, S., Lin, J., Schlicker, S., Ebert, D., Buntrock, C., & Baumeister, H. (2017). Effectiveness and cost-effectiveness of a guided Internet- and mobile-based intervention for the indicated prevention of major depression in patients with chronic back pain-study protocol of the PROD-BP multicenter pragmatic RCT. *BMC Psychiatry*, *17*(1), 1–13. https://doi.org/10.1186/s12888-017-1193-6

Sander, L. B., Paganini, S., Terhorst, Y., Schlicker, S., Lin, J., Spanhel, K., Buntrock, C., Ebert, D., & Baumeister, H. (2020). Effectiveness of a Guided Web-Based Self-help Intervention to Prevent Depression in Patients with Persistent Back Pain: The PROD-BP Randomized Clinical Trial. *JAMA Psychiatry*, *77*(10), 1001–1011. https://doi.org/10.1001/jamapsychiatry.2020.1021

Sander, L. B., Rausch, L., & Baumeister, H. (2016). Effectiveness of Internet- and mobile-based psychological interventions for the prevention of mental disorders: a systematic review and meta-analysis protocol. *Systematic Reviews*, *5*(1), 30. https://doi.org/10.1186/s13643-016-0209-5

Schmidt, I. D., Forand, N. R., & Strunk, D. R. (2019). Predictors of Dropout in Internet-Based Cognitive Behavioral Therapy for Depression. *Cognitive Therapy and Research*, *43*(3), 620–630. https://doi.org/10.1007/s10608-018-9979-5

Schuch, F. B., Vancampfort, D., Firth, J., Rosenbaum, S., Ward, P. B., Silva, E. S., Hallgren, M., De Leon, A. P., Dunn, A. L., Deslandes, A. C., Fleck, M. P., Carvalho, A. F., & Stubbs, B. (2018). Physical activity and incident depression: A meta-analysis of prospective cohort studies. *American Journal of Psychiatry*, *175*(7), 631–648. https://doi.org/10.1176/appi.ajp.2018.17111194

Seligman, M. E., & Csikszentmihalyi, M. (2000). Positive psychology. An introduction. *The American Psychologist*, *55*(1), 5–14. https://doi.org/10.1037/0003-066X.55.1.5

Selmi, P. M., Klein, M. H., Greist, J. H., Sorrell, S. P., Erdman, H. P., Selmi, M., Sorrell, P., Ph, D., Klein, H., Erdman, P., Greist, H., Ph, D., Selmi, P. M., Klein, M. H., Greist, J. H., Sorrell, S. P., & Erdman, H. P. (1990). Computer-administered cognitive-behavioral therapy for depression. *American Journal of Psychiatry*, *147*(1), 51–56. https://doi.org/10.1176/ajp.147.1.51

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*. https://doi.org/10.1017/S0033291719000151

Shim, M., Mahaffey, B., Bleidistel, M., & Gonzalez, A. (2017). A scoping review of human-support factors in the context of Internet-based psychological interventions (IPIs) for depression and anxiety disorders. *Clinical Psychology Review*, *57*(March), 129–140. https://doi.org/10.1016/j.cpr.2017.09.003

Singal, A. G., Higgins, P. D. R., & Waljee, A. K. (2014). A primer on effectiveness and efficacy trials. *Clinical and Translational Gastroenterology*, *5*(1), e45. https://doi.org/10.1038/ctg.2013.13

Singer, J. D., & Willett, J. B. (2009). Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. In *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195152968.001.0001

Sivertsen, B., Krokstad, S., Øverland, S., & Mykletun, A. (2009). The epidemiology of insomnia: associations with physical and mental health. The HUNT-2 study. *Journal of Psychosomatic Research*, *67*(2), 109–116. https://doi.org/10.1016/J.JPSYCHORES.2009.05.001

Smith, P., Scott, R., Eshkevari, E., Jatta, F., Leigh, E., Harris, V., Robinson, A., Abeles, P., Proudfoot, J., Verduyn, C., & Yule, W. (2015). Computerised CBT for depressed adolescents: Randomised controlled trial. *Behaviour Research and Therapy*, *73*(9kp, 0372477 PG-104–10), 104–110. https://doi.org/10.1016/j.brat.2015.07.009

Spek, V., Cuijpers, P., Nyklícek, I., Riper, H., Keyzer, J., & Pop, V. (2007). Internet-based cognitive behaviour therapy for symptoms of depression and anxiety: a meta-analysis. *Psychological Medicine*, *37*(3), 319–328. https://doi.org/10.1017/S0033291706008944

Spek, V., Nyklíček, I., Cuijpers, P., & Pop, V. (2008). Predictors of outcome of group and internet-based cognitive behavior therapy. *Journal of Affective Disorders*, *105*(1–3), 137–145. https://doi.org/10.1016/j.jad.2007.05.001

Spek, V., Nyklicek, I., Smits, N., Cuijpers, P., Riper, H., Keyzer, J., & Pop, V. (2007). Internet-based cognitive behavioural therapy for subthreshold depression in people over 50 years old: A random controlled clinical trial. *Psychological Medicine*, *37*(12 PG-1797–1806), 1797–1806. https://doi.org/http://dx.doi.org/10.1017/S0033291707000542

Spijker, J., De Graaf, R., Bijl, R. V., Beekman, A. T. F., Ormel, J., & Nolen, W. A. (2002). Duration of major depressive episodes in the general population: Results from the Netherlands Mental Health Survey and Incidence Study (NEMESIS). *The British Journal of Psychiatry*, *181*(3), 208–213. https://doi.org/10.1192/BJP.181.3.208

Statisa. (2020). *Global connected wearable devices 2016-2022 | Statista*. https://www.statista.com/statistics/487291/global-connected-wearable-devices/

Statista. (2022). *Smartphone users worldwide*. https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/

Steffen, A., Nübel, J., Jacobi, F., Bätzing, J., & Holstiege, J. (2020). Mental and somatic comorbidity of depression: A comprehensive cross-sectional analysis of 202 diagnosis groups using German nationwide ambulatory claims data. *BMC Psychiatry*, *20*(1), 1–15. https://doi.org/10.1186/S12888-020-02546-8/FIGURES/4

Stiles, W. B., Honos-Webb, L., & Surko, M. (1998). Responsiveness in Psychotherapy. *Clinical Psychology: Science and Practice*, *5*(4), 439–458. https://doi.org/10.1111/j.1468-2850.1998.tb00166.x

Streiner, D. L. (2002). The 2 "Es" of research: efficacy and effectiveness trials. *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie*, *47*(6), 552–556. https://doi.org/10.1177/070674370204700607

Sucala, M., Schnur, J. B., Constantino, M. J., Miller, S. J., Brackman, E. H., & Montgomery, G. H. (2012). The therapeutic relationship in e-therapy for mental health: a systematic review. *Journal of Medical Internet Research*, *14*(4), e110. https://doi.org/10.2196/jmir.2084

Taylor, D. J., Lichstein, K. L., Durrence, H. H., Reidel, B. W., & Bush, A. J. (2005). Epidemiology of insomnia, depression, and anxiety. *Sleep*, *28*(11), 1457–1464. https://doi.org/10.1093/SLEEP/28.11.1457

Thomas, N., McLeod, B., Jones, N., & Abbott, J. A. (2015). Developing Internet interventions to target the individual impact of stigma in health conditions. *Internet Interventions*, *2*(3), 351–358. https://doi.org/10.1016/J.INVENT.2015.01.003

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/J.2517-6161.1996.TB02080.X

Titov, N. (2011). Internet-delivered psychotherapy for depression in adults. *Current Opinion in Psychiatry*, *24*(1), 18–23. https://doi.org/10.1097/YCO.0B013E32833ED18F

Titov, N., Andrews, G., Davies, M., Mcintyre, K., Robinson, E., & Solley, K. (2010). Internet treatment for depression: A randomized controlled trial comparing clinician vs. technician assistance. *PLoS ONE*, *5*(6), e10939. https://doi.org/10.1371/journal.pone.0010939

Titov, N., Dear, B., Nielssen, O., Staples, L., Hadjistavropoulos, H., Nugent, M., Adlam, K., Nordgreen, T., Bruvik, K. H., Hovland, A., Repål, A., Mathiasen, K., Kraepelien, M., Blom, K., Svanborg, C., Lindefors, N., & Kaldo, V. (2018). ICBT in routine care: A descriptive analysis of successful clinics in five countries. *Internet Interventions*, *13*(July), 108–115. https://doi.org/10.1016/j.invent.2018.07.006

Titov, N., Fogliati, V. J., Staples, L. G., Gandy, M., Johnston, L., Wootton, B., Nielssen, O., & Dear, B. F. (2016). Treating anxiety and depression in older adults: randomised controlled trial comparing guided V. self-guided internet-delivered cognitive–behavioural therapy . *BJPsych Open*, *2*(1), 50–58. https://doi.org/10.1192/bjpo.bp.115.002139

Topooco, N., Riper, H., Araya, R., Berking, M., Brunn, M., Chevreul, K., Cieslak, R., Ebert, D., Etchmendy, E., Herrero, R., Kleiboer, A., Krieger, T., García-Palacios, A., Cerga-Pashoja, A., Smoktunowicz, E., Urech, A., Vis, C., & Andersson, G. (2017). Attitudes towards digital treatment for depression: A European stakeholder survey. *Internet Interventions*, *8*, 1–9. https://doi.org/10.1016/j.invent.2017.01.001

Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Whelan, P., Carvalho, A. F., Keshavan, M., Linardon, J., & Firth, J. (2021). The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry*, *20*(3), 318–335. https://doi.org/10.1002/wps.20883

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), e16. https://doi.org/10.2196/mental.5165

Torous, J., Myrick, K. J., Rauseo-Ricupero, N., & Firth, J. (2020). Digital mental health and COVID-19: Using technology today to accelerate the curve on access and quality tomorrow. *Journal of Medical Internet Research*, *22*(3), 1–6. https://doi.org/10.2196/18848

Trollor, J. N., Anderson, T. M., Sachdev, P. S., Brodaty, H., & Andrews, G. (2007). Prevalence of mental disorders in the elderly: the Australian National Mental Health and Well-Being Survey. *The American Journal of Geriatric Psychiatry : Official Journal of the American Association for Geriatric Psychiatry*, *15*(6), 455–466. https://doi.org/10.1097/JGP.0B013E3180590BA9

UNESCO. (2017). International Standard Classification of Education. In *UNESCO Institute for Statistics*. http://www.uis.unesco.org

Unützer, J., Bruce, M. L., Bauer, M. S., Durham, M., Escobar, J., Ford, D., Hoagwood, K., Horwitz, S., Lawson, W., Lewis, L., McGuire, T., Pincus, H., Scheffler, R., & Smith, W. (2002). The Elderly. *Mental Health Services Research 2002 4:4*, *4*(4), 245–247. https://doi.org/10.1023/A:1020924901595

Van Ballegooijen, W., Cuijpers, P., Van Straten, A., Karyotaki, E., Andersson, G., Smit, J. H., Riper, H., Ballegooijen, W. van, & Straten, A. van. (2014). Adherence to Internet-Based and Face-to-Face Cognitive Behavioural Therapy for Depression: A Meta-Analysis. *PLOS ONE*, *9*(7), e100674. https://doi.org/10.1371/journal.pone.0100674

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Verduijn, J., Verhoeven, J. E., Milaneschi, Y., Schoevers, R. A., van Hemert, A. M., Beekman, A. T. F., & Penninx, B. W. J. H. (2017). Reconsidering the prognosis of major depressive disorder across diagnostic boundaries: full recovery is the exception rather than the rule. *BMC Medicine*, *15*(1). https://doi.org/10.1186/S12916-017-0972-8

Vos, T., Abajobir, A. A., Abbafati, C., Abbas, K. M., Abate, K. H., Abd-Allah, F., Abdulle, A. M., Abebo, T. A., Abera, S. F., Aboyans, V., Abu-Raddad, L. J., Ackerman, I. N., Adamu, A. A., Adetokunboh, O., Afarideh, M., Afshin, A., Agarwal, S. K., Aggarwal, R., Agrawal, A., … Murray, C. J. L. (2017). Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, *390*(10100), 1211–1259. https://doi.org/10.1016/S0140-6736(17)32154-2

Wagner, B., Horn, A. B., & Maercker, A. (2014). Internet-based versus face-to-face cognitive-behavioral intervention for depression: A randomized controlled non-inferiority trial. *Journal of Affective Disorders*, *152–154*(1), 113–121. https://doi.org/10.1016/j.jad.2013.06.032

Wallert, J., Gustafson, E., Held, C., Madison, G., Norlund, F., Von Essen, L., & Olsson, E. M. G. (2018). Predicting adherence to internet-Delivered psychotherapy for symptoms of depression and anxiety after myocardial infarction: Machine learning insights from the U-CARE heart randomized controlled trial. *Journal of Medical Internet Research*, *20*(10). https://doi.org/10.2196/10754

Wampold, B. E. (2001). *The great psychotherapy debate: Models, methods, and findings.* Lawrence Erlbaum Associates Publishers. https://psycnet.apa.org/record/2001-00819-000

Wampold, B. E. (2015). How important are the common factors in psychotherapy? An update. *World Psychiatry*, *14*(3), 270–277. https://doi.org/10.1002/wps.20238

Wang, P. S., Lane, M., Olfson, M., Pincus, H. A., Wells, K. B., & Kessler, R. C. (2005). Twelve-Month Use of Mental Health Services in the United States: Results From the National Comorbidity Survey Replication. *Archives of General Psychiatry*, *62*(6), 629–640. https://doi.org/10.1001/ARCHPSYC.62.6.629

Warmerdam, L., Van Straten, A., Twisk, J., & Cuijpers, P. (2013). Predicting outcome of Internet-based treatment for depressive symptoms. *Psychotherapy Research*, *23*(5), 559–567. https://doi.org/10.1080/10503307.2013.807377

Webb, C. A., Rosso, I. M., & Rauch, S. L. (2017). Internet-Based Cognitive-Behavioral Therapy for Depression: Current Progress and Future Directions. *Harvard Review of Psychiatry*, *25*(3), 114–122. https://doi.org/10.1097/HRP.0000000000000139

Weisel, K. K., Fuhrmann, L. M., Berking, M., Baumeister, H., Cuijpers, P., & Ebert, D. (2019). Standalone smartphone apps for mental health—a systematic review and meta-analysis. *Npj Digital Medicine*, *2*(1), 1–10. https://doi.org/10.1038/s41746-019-0188-8

Weissman, M. M., & Klerman, G. L. (1990). *Interpersonal Psychotherapy for Depression*.

Weisz, J. R., Chorpita, B. F., Palinkas, L. A., Schoenwald, S. K., Miranda, J., Bearman, S. K., Daleiden, E. L., Ugueto, A. M., Ho, A., Martin, J., Gray, J., Alleyne, A., Langer, D. A., Southam-Gerow, M. A., Gibbons, R. D., Glisson, C., Green, E. P., Hoagwood, K. E., Kelleher, K., … Mayberg, S. (2012). Testing Standard and Modular Designs for Psychotherapy Treating Depression, Anxiety, and Conduct Problems in Youth: A Randomized Effectiveness Trial. *Archives of General Psychiatry*, *69*(3), 274–282. https://doi.org/10.1001/ARCHGENPSYCHIATRY.2011.147

Wells, K. B., Golding, J. M., & Burnam, M. A. (1988). Psychiatric disorder in a sample of the general population with and without chronic medical conditions. *American Journal of Psychiatry*, *145*(8), 976–981. https://doi.org/10.1176/ajp.145.8.976

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS One*, *12*(4), e0174944. https://doi.org/10.1371/journal.pone.0174944

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., Burstein, R., Murray, C. J. L., & Vos, T. (2013). Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet (London, England)*, *382*(9904), 1575–1586. https://doi.org/10.1016/S0140-6736(13)61611-6

Williams, C., & Garland, A. (2002). A cognitive–behavioural therapy assessment model for use in everyday clinical practice. *Advances in Psychiatric Treatment*, *8*(3), 172–179. https://doi.org/10.1192/APT.8.3.172

Wing, R. R., Phelan, S., & Tate, D. (2002). The role of adherence in mediating the relationship between depression and health outcomes. *Journal of Psychosomatic Research*, *53*(4), 877–881. https://doi.org/10.1016/S0022-3999(02)00315-X

Wisniewski, H., Henson, P., & Torous, J. (2019). Using a Smartphone App to Identify Clinically Relevant Behavior Trends via Symptom Report, Cognition Scores, and Exercise Levels: A Case Series. *Frontiers in Psychiatry*, *10*, 652. https://doi.org/10.3389/fpsyt.2019.00652

World Health Organization. (2017). *Depression and other common mental disorders: global health estimates.* https://www.who.int/mental_health/management/depression/prevalence_global_health_estimates/en/

Wright, B., Tindall, L., Littlewood, E., Allgar, V., Abeles, P., Trépel, D., & Ali, S. (2017). Computerised cognitive-behavioural therapy for depression in adolescents: Feasibility results and 4-month outcomes of a UK randomised controlled trial. *BMJ Open*, *7*(1), e012834. https://doi.org/10.1136/bmjopen-2016-012834

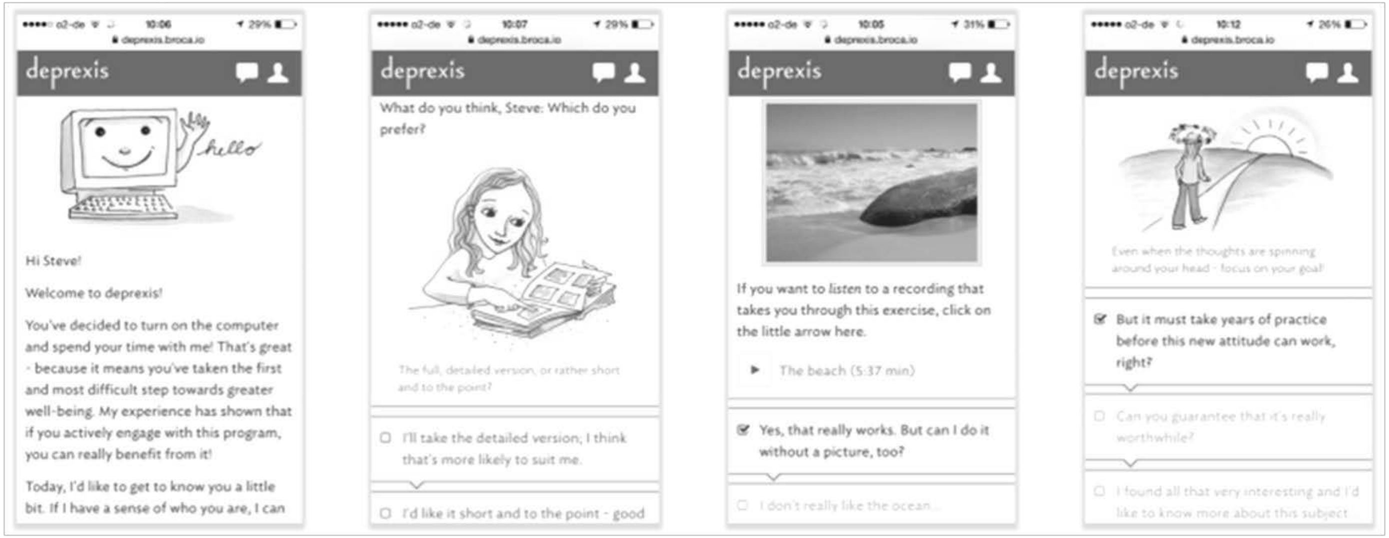
Wright, J. H., Owen, J. J., Richards, D., Eells, T. D., Richardson, T., Brown, G. K., Barrett, M., Rasku, M. A., Polser, G., & Thase, M. E. (2019). Computer-assisted cognitive-behavior therapy for depression: A systematic review and meta-analysis. *Journal of Clinical Psychiatry*, *80*(2). https://doi.org/10.4088/JCP.18r12188

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35. https://doi.org/10.1002/1097-0142

Appendix

Appendix 1. Screenshots from the digital intervention, *deprexis*



Hi Steve!

Welcome to deprexis!

You've decided to turn on the computer and spend your time with me! That's great - because it means you've taken the first and most difficult step towards greater well-being. My experience has shown that if you actively engage with this program, you can really benefit from it!

Today, I'd like to get to know you a little bit. If I have a sense of who you are, I can

What do you think, Steve: Which do you prefer?

The full, detailed version, or rather short and to the point?

□ I'll take the detailed version; I think that's more likely to suit me.

□ I'd like it short and to the point - good

If you want to *listen* to a recording that takes you through this exercise, click on the little arrow here.

▶ The beach (5:37 min)

☑ Yes, that really works. But can I do it without a picture, too?

□ I don't really like the ocean...

Even when the thoughts are spinning around your head - focus on your goal!

☑ But it must take years of practice before this new attitude can work, right?

□ Can you guarantee that it's really worthwhile?

□ I found all that very interesting and I'd like to know more about this subject

Note. An example of a digital intervention for depression delivered via the internet and accessed via a smartphone. In screen one users are given an introduction to the intervention. In screen two, users are given the choice of selecting between a short-form and longer-form version of the intervention content. In screen 3, users are given a guided mindfulness meditation exercise. In screen four, users are presented with an interactive dialogue relating to the core components of CBT.

Source: Meyer, B., Berger, T., Caspar, F., Beevers, C. G., Andersson, G., & Weiss, M. (2009). Effectiveness of a novel integrative online treatment for depression (Deprexis): Randomized controlled trial. *Journal of Medical Internet Research*, *11*(2), 1–18. https://doi.org/10.2196/jmir.1151 Published and reproduced under the terms of Creative Commons Attribution 2.0 License.

Appendix 2. Selected Characteristics of the (k = 83) Studies Included in the Meta-analysis

Study	N	Mean Age (SD)	Control	Outcome Measure	Delivery method	Guidance Type	Hedge's g	Intervention Completion rate (%)	% Completors	Country
Andersson et al. (2005)	117	36.35 (10.71)	WLC	BDI	Internet	GS	0,97	74.0	65.0	SE
Andersson et al. (2013)	69	42.3 (13.5)	gF2F	MADRS-S	Internet	GS	0,39	96.9	87.9	SE
Baumeister et al. (2020)	209	49.9 (9.36)	TAU	HRSD-17	Internet	GS	0,25	-	55.0	DE
Beevers et al. (2017)	376	31.91 (11.2)	WLC	QIDS-SR	Internet	TG	0,82	-	-	US
Beiwinkel et al. (2017)	180	47.74 (10.92)	ATT	PHQ-9	Internet	GS	0,39	-	-	DE
Berger et al. (2011)	51	38.91 (14.16)	WLC	BDI-II	Internet	GS	1,14	85.2	56.0	Multiple
Berger et al. (2011)	51	39.11 (13.71)	WLC	BDI-II	Internet	UG	0,66	68.0	36.0	Multiple
Birney et al. (2016)	300	-	ATT	PHQ-9	Smartphone	TG	0,14	-	-	US
Boele et al. (2017)	89	44.99 (11.99)	WLC	CES-D	Internet	GS	0,63	-	-	NL
Boeschoten et al. (2016)	171	48.9 (10.5)	WLC	BDI-II	Internet	GS	0,08	-	50.6	NL
Buntrock et al. (2015)	406	45.04 (11.89)	ATT	CES-D	Internet	GS	0,66	82.2	68.3	DE
Carlbring et al. (2013)	80	44.4 (13.5)	WLC	MADRS-S	Internet	GS	0,64	72.9	27.5	SE
Choi et al. (2012)	55	39 (11.7)	WLC	CBDI	Internet	GS	0,91	69.5	68.0	ANZ
Christensen et al. (2004)	360	36.07 (9.4)	ATT	CES-D	Internet	TG	0,33	51.0	-	ANZ
Clarke et al. (2002)	299	44.35 (12.2)	TAU	CES-D	Internet	UG	0,25	37.1	-	US
Clarke et al. (2005)	175	47.27 (10.8)	TAU	CES-D	Internet	TG	0,06	84.3	-	US
Clarke et al. (2005)	180	44.73 (10.5)	TAU	CES-D	Internet	TG	-0,20	84.3	-	US

Clarke et al. (2009)	160	22.65 (2.3)	TAU	PHQ-8	Internet	UG	0,16	-	-	US
De Graaf et al. (2009)	203	44.71 (12)	TAU	BDI-II	Internet	UG	0,15	42.5	14.0	NL
Deady et al. (2016)	104	21.74 (2.22)	ATT	PHQ-9	Internet	UG	0,46	37.5	-	ANZ
Ebert et al. (2014)	150	47.1 (8.2)	WLC	CES-D	Internet	GS	0,69	-	60.0	DE
Ebert et al. (2017)	256	50.8 (11.8)	ATT	CES-D	Internet	GS	0,89	-	61.5	DE
Ebert et al. (2018)	204	44.2 (11.73)	WLC	QIDS-C	Internet	GS	0,40	83.3	61.8	DE
Farrer et al. (2011)	73	40.47 (12.13)	TAU	CES-D	Internet	UG	0,78	30.0	15.8	ANZ
Farrer et al. (2011)	80	42.58 (12.2)	TAU	CES-D	Internet	TG	1,08	40.0	17.8	ANZ
Fischer et al. (2015)	90	45.28 (11.99)	WLC	BDI	Internet	UG	0,33	-	-	DE
Flygare et al. (2020)	95	45.3 (12.2)	ATT	MADRS-S	Internet	GS	0,23	73.8	-	SE
Forand et al. (2018)	89	-	WLC	PHQ-9	Internet	GS	1,61	77.8	55.9	US
Forsell et al. (2017)	42	31.01 (4.57)	TAU	MADRS-S	Internet	GS	1,18	53.0	-	SE
Geraedts et al. (2014)	231	43.4 (9.2)	TAU	CES-D	Internet	GS	0,25	-	27.6	NL
Gilbody et al. (2015)	481	39.97 (12.81)	TAU	PHQ-9	Internet	TG	-0,01	-	12.0	UK
Gilbody et al. (2015)	449	-	TAU	PHQ-9	Internet	TG	-0,05	-	14.8	UK
Gladstone et al. (2018)	369	15.4 (1.5)	ATT	CES-D	Internet	TG	-0,28	22.7	-	US
Gladstone et al. (2018)	369	15.4 (1.5)	ATT	CES-D	Internet	TG	-0,06	22.7	-	US
Glozier et al. (2013)	562	57.95 (6.6)	ATT	PHQ-9	Internet	TG	0,16	-	-	ANZ
Guo et al. (2020)	300	28.3 (5.8)	WLC	CES-D	Smartphone	TG	0,63	55.0	-	CN
Hallgren et al. (2015)	629	43 (12)	TAU	MADRS-S	Internet	GS	0,33	60.0	-	SE

Ip et al. (2016)	257	14.63 (0.81)	ATT	CES-D-R	Internet	TG	-0,01	30.0	10.1	CN
Johansson et al. (2012a)	92	45.6 (14)	other	BDI-II	Internet	GS	1,12	-	78.3	SE
Johansson et al. (2012b)	79	44.28 (12.72)	ATT	BDI-II	Internet	GS	0,56	80.7	-	SE
Johansson et al. (2012b)	78	45.22 (11.39)	ATT	BDI-II	Internet	GS	0,83	77.2	-	SE
Johansson et al. (2019a)	54	-	WLC	MADRS-S	Internet	GS	1,27	78.8	54.0	SE
Johansson et al. (2019b)	144	63 (12)	ATT	PHQ-9	Internet	TG	0,44	-	59.7	SE
Kenter et al. (2016)	269	38 (11.4)	ATT	CES-D	Internet	GS	-0,07	-	12.5	NL
Kivi et al. (2014)	79	36.6 (11.3)	TAU	BDI-II	Internet	GS	0,12	72.9	55.6	SE
Lamers et al. (2015)	116	57.09 (9.16)	WLC	CES-D	Internet	GS	0,35	-	-	NL
Lappalainen et al. (2014)	38	44.61 (14.28)	F2F	BDI-II	Internet	GS	-0,15	-	-	FI
Lappalainen et al. (2015)	39	51.9 (12.88)	WLC	BDI-II	Internet	GS	0,61	97.4	94.7	FI
Levin et al. (2011)	191	43.52 (12.93)	TAU	CES-D	Computer	TG	0,44	-	-	US
Lobner et al. (2018)	647	43.89 (13.29)	TAU	PHQ-9	Internet	UG	0,00	-	9.1	DE
Lokman et al. (2017)	329	43.25 (12.94)	WLC	IDS-SR	Internet	UG	0,42	-	-	NL
Meyer et al. (2015)	163	42 (11.39)	TAU	PHQ-9	Internet	UG	0,57	-	-	DE
Meyer et al. (2019)	200	40.3 (13.12)	WLC	PHQ-9	Internet	UG	0,54	-	-	DE
Milgrom et al. (2016)	43	31.6 (4.44)	TAU	BDI-II	Internet	GS	0,81	-	85.7	ANZ
Mira et al. (2017)	80	35.91 (9.94)	WLC	BDI-II	Internet	UG	0,50	73.0	77.8	ES
Mira et al. (2017)	88	35.77 (9.78)	WLC	BDI-II	Internet	TG	0,35	73.0	77.8	ES

Montero-Marin et al. (2016)	198	43.11 (9.49)	TAU	BDI-II	Internet	GS	0,08	-	-	ES
Montero-Marin et al. (2016)	200	42.81 (10.84)	TAU	BDI-II	Internet	UG	0,12	-	-	ES
Moritz et al. (2012)	210	38.57 (13.75)	WLC	BDI	Internet	UG	0,43	63.2	-	DE
Newby et al. (2017)	90	46.7 (12.6)	WLC	PHQ-9	Internet	GS	0,79	-	65.9	ANZ
Nobis et al. (2015)	256	51 (12)	ATT	CES-D	Internet	GS	0,89	-	62.0	DE
Noguchi et al. (2017)	651	43.85 (11.3)	WLC	CES-D	Internet	UG	-0,02	-	-	JP
Nygren et al. (2019)	50	33.86 (8.12)	WLC	BDI-II	Internet	GS	1,23	62.9	36.0	SE
Oehler et al. (2020)	347	-	ATT	IDS-SR	Internet	TG	0,23	91.7	-	DE
O'moore et al. (2018)	69	61.9 (6.92)	TAU	PHQ-9	Internet	TG	1,02	-	84.1	ANZ
Perini et al. (2009)	45	49.29 (12.06)	WLC	PHQ-9	Internet	GS	0,83	-	74.1	ANZ
Pfeiffer et al. (2020)	330	51.6 (14.9)	TAU	QIDS-SR	Internet	GS	0,14	47.5	-	US
Pots et al. (2016)	169	46.9 (11.77)	WLC	CES-D	Internet	GS	0,56	-	73.0	NL
Pugh et al. (2016)	50	-	WLC	EPDS	Internet	GS	1,04	84.6	60.0	CA
Reins et al. (2018)	131	41.6 (10.8)	ATT	HRSD-24	Internet	GS	0,32	-	75.4	DE
Richards et al. (2015)	188	39.86 (10.92)	WLC	BDI-II	Internet	GS	0,65	-	36.0	UK
Roepke et al. (2015)	186	-	WLC	CES-D	Smartphone + Internet	UG	0,31	-	-	US
Rosso et al. (2016)	77	28.99 (7.21)	ATT	HRSD-17	Internet	TG	0,79	-	91.9	US
Ruwaard et al. (2009)	54	42 (9.51)	WLC	BDI-IA	Internet	GS	0,84	-	-	NL

Salamanca-Sanabria et al. (2020)	214	22.15 (4.7)	WLC	PHQ-9	Internet	TG	0,88	-	9.3	CO
Sander et al. (2020)	295	52.8 (7.7)	TAU	PHQ-9	Internet	GS	0,42	67.5	0.5	DE
Schure et al. (2019)	343	42.9 (13.3)	WLC	PHQ-9	Smartphone + Internet	TG	0,50	28.1	-	US
Segal et al. (2020)	460	48.3 (14.9)	TAU	PHQ-9	Internet	TG	0,55	60.0	27.4	CA
Selmi et al. (1990)	24	29.9 (4.41)	WLC	BDI	Computer	TG	1,00	100.0	100.0	US
Selmi et al. (1990)	24	-	F2F	BDI	Computer	TG	0,18	100.0	100.0	US
Selmi et al. (1990)	24	29.9 (4.4)	WLC	BDI	Computer	TG	1,67	100.0	100.0	US
Smith et al. (2015)	112	-	WLC	MFQ-C	Computer	UG	0,82	-	85.5	UK
Smith et al. (2017)	113	39.94 (12.96)	WLC	PHQ-9	Internet	TG	0,87	-	59.3	ANZ
Spek et al. (2007)	202	55 (4.95)	WLC	BDI-II	Internet	UG	0,27	78.1	48.3	NL
Spek et al. (2007)	201	55 (4.95)	gF2F	BDI-II	Internet	UG	-0,06	78.1	48.3	NL
Titov et al. (2010)	86	42.79 (12.91)	WLC	PHQ-9	Internet	GS	1,27	-	69.6	ANZ
Titov et al. (2010)	81	44.99 (12.92)	WLC	PHQ-9	Internet	TG	1,27	-	80.5	ANZ
Titov et al. (2015)	52	-	WLC	PHQ-9	Internet	GS	2,29	-	70.0	ANZ
Ünlü Ince et al. (2013)	96	35.2 (9.3)	WLC	CES-D	Internet	GS	1,51	-	20.4	NL
van Luenen et al. (2018)	188	46.3 (10.63)	ATT	PHQ-9	Internet	TG	0,61	-	-	NL
Vernmark et al. (2010)	58	34.95 (11.86)	WLC	BDI	Internet	GS	0,57	85.7	58.6	SE
Wagner et al. (2014)	62	-	F2F	BDI-II	Internet	GS	-0,01	-	78.1	CH
Warmerdam et al. (2008)	175	-	WLC	CES-D	Internet	GS	0,47	-	37.5	NL

Warmerdam et al. (2008)	175	-	WLC	CES-D	Internet	GS	0,26	-	38.6	NL
Williams et al. (2013)	63	44.76 (12.05)	WLC	BDI-II	Internet	GS	0,97	-	54.3	ANZ
Wright et al. (2017)	91	15.35 (1.3)	ATT	BDI	Computer	TG	0,00	-	62.2	UK

Note. WLC = Waiting list control; gF2F = group Face-to-Face; TAU = Treatment as usual; ATT = attention; GS = guided service; TG = technical guidance; UG = unguided