Generating and interpreting evidence from psychotherapy: An examination of measurement models, missing cases, and classification methods

A thesis submitted for the degree of Doctor of Philosophy

Written by Mr. Eyal Karin (MappStat)

MACQUARIE University
SYDNEY·AUSTRALIA

Department of Psychology

Principal Supervisor: Professor Nickolai Titov, PhD

Adjunct Supervisor: Associate Professor Blake F. Dear, PhD

Adjunct Supervisor: Professor Gillian Z. Heller, PhD

November 2019

Field of research code (170110)

General summary

The primary aim of this thesis is to explore and identify some of the statistical assumptions that underpin the measurement, statistical analysis, and interpretation of the effects of psychotherapy for anxiety and depression. The identification of these statistical assumption are then used to reflect on the suitability of common statistical techniques that underpin quantitative psychotherapy research and treatment evaluation.

A series of five studies are presented, exploring the different statistical assumptions that underpin the measurement of symptom change through treatment (Studies 1 and 2), the handling of missing cases (Studies 3 and 4), and the classification of symptom outcomes into categories that represent the individual impact of treatment (Studies 5).

The clinical datasets employed in these studies are comprised from samples of participants enrolled in randomised controlled trials (n>820) or patients enrolled in routine care (n>6700), who receive internet-delivered cognitive behavioural therapy (iCBT) for anxiety and depression. The iCBT context is used as an exemplar psychotherapy context, with highly protocolised procedures which reduces measurement variance due to therapy type or therapist.

The results of these studies, identify several statistical assumptions that seem to generalise across psychotherapy data; being the proportional reduction of symptom change, the assumption of missing at random that is conditional on treatment adherence, and the occurrence of proportional symptom change that is non-specific to treatment. These results also indicate that the use of conventional methods for reporting treatment efficacy, including Cohen's *d* effect size and the Reliable Change Index (RCI), and for statistically adjusting for data missing from clinical trials, including the missing completely at random assumption (MCAR), may result in error in evaluation and interpretation. Each of the five studies also point to the benefits of selecting alternative statistical methods that better fit the context of psychotherapy data, reduce measurement error, and increase the ability to interpret clinical change with increased validity.

As a body of work, the thesis seeks to point to a set of methods that strike a balance between the competing priorities of researchers to select methods that fit the specific nuances of psychotherapy data, and the selection of methods that enable the comparison and generalisability of psychotherapy outcomes across different contexts (e.g. different symptom scales and treatment types).

The research of this thesis is explored through both statistical and clinical viewpoints but is primarily written and directed for clinical researchers and a clinical audience. Implications for the broader field of mental health research are also discussed.

Statement of Originality

This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

(Signed)

Date: NOVEMBER 1, 2019

Ethics statement

The research of this thesis was conducted under the supervision of the Faculty of Medicine and Health Sciences Low-Risk Ethics Subcommittee (Ref No. 5201600835; 5201100226; 5201100470; 5201100469; 5201300458; 5201300456; 5201200872; 5201400582; 5201100680) and met the requirements set out in the National Statement on Ethical Conduct in Human Research (2007 – Updated May 2015) (the National Statement).

Acknowledgements

Through this PhD project, I was fortunate to be supported by many people, without which this project would not have been possible. I'd like to thank and acknowledge them briefly, although my gratitude extends far beyond these brief statements.

First and foremost, I am grateful for the opportunity to meet and work with my primary supervisors, Prof. Nick Titov, and Assoc. Prof Blake Dear. Through the years of this project, Nick and Blake's mentoring and support was extended in many ways, as expert clinical researchers, leading as clinical managers, and as globally connected researchers. Their clinical vision and cumulative effort over a period of nearly a decade ultimately enabled the large psychotherapy patient records that formed the samples and basis for this methodological project.

On a personal level, I was fortunate to have two supervisors with willingness, dedication and endurance for pursuing the highest possible research standards. This project was a demanding attempt to strike an interdisciplinary balance between the nuances of both clinical and statistical research streams; that often present different language norms, research priorities and research methods. As an interdisciplinary project, the studies and arguments of the thesis did not fit comfortably within the typical clinical or statistical spaces and presented with various ambiguities and challenges. Yet, whilst demanding, the willingness and dedication to pursue new ideas, solutions and research standards, is an example, and by no means an exception, of Nick and Blake's handling of projects from start to end. This project would not have been possible without their dedication, support and contribution, and I consider this thesis theirs just as much as it is mine.

I would also like to thank my adjunct supervisor Prof. Gillian Heller for her dedication, warmth and expertise as a theoretical statistical researcher. Gillian's support and statistical input played an essential role in the design and communication of each of the studies, and her oversight throughout the project was fundamental.

For my supervisors' dedication, mentorship and energy they've poured into this project I am truly indebted and appreciative.

Besides my supervisors, I've also had the fortune to meet and be supported by many other academics along the way. I would like to say thank you to Dr Rony Kayrouz, Prof. Olav Nielssen, Dr Bethany Wootton, and Dr Milena Gandy for being both professional colleagues and good friends. Working with you all throughout this project has been a wellspring of mateship and learning!

I would also like to say thank you to the colleagues and collaborators I've met along the way. At the eCentre and Mindspot Clinics in Sydney – Dr. Luke Johnston, Bareena Johnson, Ashwin Sigh, Dr Matthew Terides, Drs Vincent and Rhiannon Fogliati, Dr Lauren Staples, Dr Joanne Dudeney, Tahlia Ricciardi, and Dr Sarah McDonald; At the Psychology-Wise Lab in Regina – Prof. Heather Hadjistavropoulos, Dr Luke Schneider, Dr Dale Dirkse, Dr Nicole Alberts, Joelle Soucy and Victoria Owens; At the Centre for Psychiatry Research at the Karolinska Institutet in Stockholm – Prof Viktor Kaldo, Dr. Martin Kraepelien, Dr Kerstin Blom, and Dr (nearly) Erik Forsell; Thank you, it's been a great privilege to meet and work with you all!

Finally, and most importantly I would like to thank my wife, Dr Monique Crane for being the force that enabled this project. There were many things that I was proud of, humbled by, and thankful for throughout this project. Having Monique by my side was the most important of them all.

Table of Contents

List of figures

General introduction

Measurement of Symptom Change Following Web-Based Psychotherapy: Statistical Characteristics and Analytical Methods for Measuring and Interpreting Change (Study 1)

Statistical characteristics and analytical methods for measuring and interpreting symptom change in psychotherapy – a replication and elaboration study (Study 2)

"Wish You Were Here": Examining Characteristics, Outcomes, and Statistical Solutions for Missing Cases in Web-Based Psychotherapeutic Trials (Study 3)

Examining Characteristics, Outcomes, and Statistical Solutions for Missing Cases in Web-Based Psychotherapeutic Trials – a replication and extension (Study 4)

Classification of symptom improvement in psychotherapy: A new proposed approach for classifying minimal treatment-related response (Min-TR) (Study 5)

List of tables

General introduction

Measurement of Symptom Change Following Web-Based Psychotherapy: Statistical Characteristics and Analytical Methods for Measuring and Interpreting Change (Study 1)

Statistical characteristics and analytical methods for measuring and interpreting symptom change in psychotherapy – a replication and elaboration study (Study 2)

“Wish You Were Here”: Examining Characteristics, Outcomes, and Statistical Solutions for Missing Cases in Web-Based Psychotherapeutic Trials (Study 3)

Examining Characteristics, Outcomes, and Statistical Solutions for Missing Cases in Web-Based Psychotherapeutic Trials – a replication and extension (Study 4)

Classification of symptom improvement in psychotherapy: A new proposed approach for classifying minimal treatment-related response (Min-TR) (Study 5)

List of Abbreviations

Clinical Disorders & Related Diagnostic Terms:

DSM-IV : Diagnostic and Statistical Manual
GAD : generalised anxiety disorder
MADD : Mixed anxiety and depressive disorder
MDD : Major depressive disorder
MDE : Major depressive episode
MINI : Mini-International Neuropsychiatric Interview
OCD : Obsessive compulsive disorder
Panic/Ag : Panic disorder and/or agoraphobia
PTSD : Posttraumatic stress disorder
SCID : Structured Clinical Interview
SocPhob : Social phobia or social anxiety disorder

Research Guideline Frameworks:

CONSORT : Consolidated Standards of Reporting Trials
JARS: Journal Article Reporting Standards for Research in Psychology
PRISMA : Transparent reporting of Systematic Reviews and Meta-analyses
STRAD : Standards for Reporting of Diagnostic Accuracy Studies
STROBE : The Strengthening the Reporting of Observational Studies in Epidemiology statement

Statistical Modelling Notation:

AIC : Akaike information criterion.
ANCOVA : Analysis of covariance
ANOVA : Analysis of variance
AUC : Area under the curve
BIC : Bayesian information criterion.
BOCF : Baseline observation carried forward
CI : Confidence interval
LOCF : Last observation carried forward
exp(β)*:* Exponentiated model coefficient
GEE : Generalised estimation equation
HLM : Hierarchical Linear Modelling
HMLM : Hierarchical multivariate Linear Modelling
MANOVA - Multivariate analysis of variance
MAR : Missing at random
MCAR : Missing at random
MCID : Minimal clinically important difference
Min-TR : Minimal treatment related change
Mixed : Longitudinal multilevel models
MNAR : Missing not at random
n *:* Sample size
p : P-value, Statistical Type I Error
QIC : Quasilikelihood under the independence model criterion.
Q-Q : Quantile-Quantile
R^2 : Model measurement error squared (variance) explained by a model
RCI : Reliable Change Index
Rep : Replication sub-sample
RRI : Relative risk increment

SD : Standard deviation
SE : Standard error
Δ : Difference quantity – primarily time related
σ^2 : Model measurement error squared (variance)

Symptom Measures Abbreviations:

ADIS-IV: Anxiety Disorders Interview Schedule for DSM-IV
ADIS-R: Anxiety Disorders Interview Schedule-Revised
AKUADS: Aga Khan University Anxiety and Depression Scale
ASI: Anxiety Sensitivity Index, BAI: Beck Anxiety Inventory
BAT: Behavioural Activation Test
BDI-II: Beck Depression Inventory, second edition
BSI : Brief Symptom Inventory
BSQ – Body Sensation Questionnaire
CAQ: Cognitive Anxiety Questionnaire
CESD: Centre of Epidemiological Studies Depression Scale
CGI: Clinical Global Improvement-Patient Rating
CID: Clinical Interview for Depression
CORE-OM: Clinical Outcome's in Routine Evaluation
DASS-21: Depression Anxiety and Stress Scales (DASS) 21-item version
DASS-42: Depression Anxiety and Stress Scales (DASS) 42-item version
EDS: The Edinburgh Depression Scale
EFI: Effects on Life Inventory FI: Fear Inventory
FNE: Fear of Negative Evaluation Scale
FQ: Fear Questionnaire
FQAD: Fear Questionnaire Anxiety–Depression Subscale
FQSP: Fear Questionnaire Social Phobia Subscale
FSS: Fear Survey Schedule
GAD-7: Generalised Anxiety Disorder 7-item scale
GAS: Generalised Anxiety Scale from the Guys/Age Concerned Survey
HADS-A: Hospital Anxiety and Depression Scale — Anxiety Subscale
HADS-D: Hospital Anxiety and Depression Scale — Depression Subscale
HAI: Health Anxiety Inventory
HARS: Hamilton Anxiety Rating Scale
HRSD: Hamilton Rating Scale for Depression
IIEF – International Index of Erectile Function
K-10: Kessler 10-item
KSQ: Kellner's Symptom Questionnaire
LSAS-SR: Liebowitz Social Anxiety Scale self-report version
MADRS-S: Montgomery–Asberg Depression Rating Scale — self rated version
MASQ: Mood and Symptoms Questionnaire
MIA: Mobility Inventory for Agoraphobia
NEO-N: NEO-Five Factor Inventory—Neuroticism Subscale
OCI-R: Obsessive–Compulsive Inventory—Revised Version
PANAS: Positive and negative affect scale
PAS: Panic and Agoraphobia Scale
PCL-C Posttraumatic Stress Disorder Checklist — Civilian
PSC: Psychosomatic Symptom Checklist
PDSS – Panic Disorder Severity Scale
PDSS-SR: Panic Disorder Severity Scale — Self-report
PHQ: Patient Health Questionnaire PRCA:
Personal Report of Communication Apprehension

PRCS: Personal Report of Confidence as a Speaker
PSI: Physical Symptoms inventory
PSWQ: Penn State Worry Questionnaire
QLESQ: Quality of Life Enjoyment and Satisfaction Scale
QOLI: Quality of Life Inventory
SAD: Social Avoidance and Distress Scale
SDS: Sheehan Disability Scale
SIAS: Social Interaction Anxiety Scale
SIAS/SPS6 composite: Social Interaction Anxiety Scale and Social Phobia Scale 6-item composite
SIGH: Structured Interview Guide for the Hamilton Anxiety Rating Scale (A = Anxiety, D = Depression)
SPSQ: Social Phobia Screening Questionnaire
SSPS: Self-Statements During Public Speaking (positive and negative subscales)
SSS: Subjective Symptoms Scale (a measure of interference with daily functioning)
STAI: State-Trait Anxiety Inventory (T = trait, S = state)
TRQ – Tinnitus Reaction Questionnaire
WSAS: Work and Social Adjustment Scale
YBOCS: Yale-Brown Obsessive Compulsive Scale
Zung SRSD: Zung Self-Rating Scale for Depression.

Treatment-Related Terminology

CBT : cognitive behavioral therapy
iCBT : internet-delivered cognitive behavioural therapy
ITT : intention-to-treat analysis
MUM : Macquarie university model
RCT : Randomised controlled trial

Chapter 1

General introduction

The importance of effective treatment for anxiety and depressive disorders

Anxiety and depressive disorders are the most common mental disorders and are a major public health problem worldwide (Kessler, Aguilar-Gaxiola, Alonso, Chatterji, Lee, Ormel, ., ... & Wang, 2009; Whiteford, Degenhardt, Rehm, Baxter, Ferrari, Erskine, ... & Burstein, 2013). For example, in Australia, anxiety and depressive disorders have been estimated to affect more than 1.5 (6.3%) and 1.3 million (5.4%) adults each year, respectively (Ciobanu, Ferrari, Erskine, Santomauro, Charlson, Leung, ... & Baune, 2018), and are much more common than substance use disorders, schizophrenia and bipolar disorder, which affect around 450,000 (1.9%), 200,000 (0.8%) and 100,000 (0.4%), respectively (Ciobanu et al., 2018).

Depressive disorders include a range of syndromes that manifest with pervasively depressed mood and diminished pleasure, as well as changes in several domains that may include appetite and weight, sleep disturbance, loss of energy, indecisiveness and feelings of hopelessness, worthlessness and recurrent thoughts of death (American Psychiatric Association, 2013). The anxiety disorders include several overlapping syndromes that are characterised by excessive fear and worry, autonomic arousal, restlessness, muscle tension, disturbed sleep, impaired concentration, hypervigilance, irritability and fatigue (American Psychiatric Association, 2013).

The anxiety and depressive disorders affect twice as many women as men and usually emerge in early adult life (Slade, Johnston, Oakley Browne, Andrews, & Whiteford, 2009). These disorders frequently overlap and occur together with other mental and physical health disorders (Slade et al., 2009; Teesson, Mitchell, Deady, Memedovic, Slade & Baillie, 2011). Depression and anxiety disorders tend to be recurrent and chronic without effective treatment (Musliner, Munk-Olsen, Laursen, Eaton, Zandi, & Mortensen, 2016; Nierenberg, Petersen & Alpert, 2003), cause significant functional impairment and disability (Ciobanu et al., 2018; Kessler, Heeringa, Lakoma, Petukhova, Rupp, Schoenbaum, ... & Zaslavsky, 2008) and are associated with increased mortality (Cuijpers & Smit, 2002). The significant disability associated with anxiety and depressive disorders is reflected in the finding that they account for around half of the disability-adjusted life years of all the mental health disorders (Ciobanu et al., 2018). The anxiety and depressive disorders, therefore, represent a significant

proportion of the total worldwide burden of all diseases (Whiteford et al., 2013). Looked at another way, among adults in Australia, depressive and anxiety disorders were accompanied by an average of 6.2 and 4.4 days out of role in the previous month, respectively, compared with 1.4 days for those with no mental disorders (Slade et al., 2007). This shows the importance of effective treatment for both the individuals who experience them and for the wider community affected by these conditions.

Treatments for anxiety and depression

The two main approaches to treating anxiety and depression are pharmacotherapy and psychotherapy (Davey & Chanen, 2016; Cuijpers, Sijbrandij, Koole, Andersson, Beekman, & Reynolds, 2014). Anxiolytic and antidepressant medications are among the most frequently prescribed medical treatments in Australia, with more than 10% of the population using these medications at any given time (Davey & Chanen, 2016). The rates of anxiolytic and antidepressant medication use have increased in most high-income countries in the last two decades (Olfson & Marcus, 2009). Psychotherapy, which in this thesis refers to any intervention which aims to ameliorate, manage, or prevent anxiety and depression through modifying psychological processes (Mahoney, 2012; Lambert, 2007; Zeig & Munion, 1990), can be provided on its own or in combination with pharmacotherapy (Cuijpers et al., 2014). Despite the high rates of pharmacotherapy, most patients report a preference for psychotherapy over pharmacotherapy (McHugh, Whitton, Peckham, Welge, & Otto, 2013).

There is now substantial evidence for the efficacy of psychotherapy for anxiety and depression, from the results of literally thousands of randomised controlled trials (Braakmann, 2015; Horvath, 2013; Scull, 2015). There are many models of psychotherapy, including cognitive behavioural therapy, brief psychodynamic psychotherapy, problem-solving therapy, interpersonal psychotherapy, behavioural activation, social skills training and many more, which are based on a range of models of the causes and manifestations of underlying psychological distress and disorder. There is evidence to support the efficacy of many of the models of psychotherapy. Meta-analytic studies have found, although not without some controversy, that the dominant approaches to psychotherapy to be similarly effective in terms of symptom reduction, a result often referred to as the *dodo bird* verdict (Budd & Hughes, 2009;

Cuijpers et al., 1998; Luborsky, Rosenthal, Diguer, Andrusyna, Berman Levitt ... & Krause, 2002). However, it is important to note that, even where psychotherapies are similarly effective in terms of symptom reduction, there are other dimensions on which psychotherapies can be compared, including the cost, how simple they are to deliver, how efficiently they produce treatment effects and their acceptability to patients (Cougle, 2012; Richards, Ekers, McMillan, Taylor, Byford, Warren, ... & O'Mahen, 2016; Cuijpers, Huibers, Ebert, Koole, & Andersson, 2013). Nevertheless, the demonstrated effectiveness of psychotherapy has made it one of the modes of treatment for anxiety and depression in health care systems around the world.

Arguably the most researched and widely adopted model of psychotherapy for mood disorder is cognitive behaviour therapy (CBT) (David, Cristea & Hofmann, 2018). The evidence for the efficacy of CBT for both anxiety and depression is now very large (Cuijpers, Cristea, Karyotaki, Reijnders & Huibers, 2016; Hofmann, Asnaani, Vonk, Sawyer & Fang, 2012). For example, a recent meta-analysis of 144 controlled trials of CBT found large effect sizes for major depressive disorder (Hedges g = 0.75), generalised anxiety disorder (Hedges g = 0.81), panic disorder (Hedges g = 0.81) and social anxiety disorder (Hedges g = 0.88) over waitlist and other control conditions. The evidence base for CBT has led it to be widely considered as the gold-standard psychotherapy against which other psychotherapies are compared and considered (David et al., 2018). Nevertheless, it is also well recognised that the between-groups effect sizes are not as large in trials comparing CBT with treatment-as-usual care and placebo pharmacotherapy (Cuijpers et al., 2016).

Despite the evidence for the efficacy of psychotherapy, community surveys of the prevalence of mental disorder show that many adults with anxiety and depression do not or cannot access effective psychotherapy. For example, in 2007, only 35% of adults identified as having a mental health disorder accessed any form of mental health care, with 24% accessing general practitioners, 13% seeing a psychologist, and 8% seeing a psychiatrist (Burgess, Pirkis, Slade, Johnston, Meadows & Gunn, 2009; Slade et al., 2009). Low rates of treatment are believed to be due to a range of barriers to mental health care, including low perceived need, lack of awareness of the efficacy of treatment, the direct and indirect costs of treatment, the limited availability of services, stigma and strong preferences to self-manage

(Mojtabai, Olfson, Sampson, Jin, Druss, Wang & Kessler, 2011). The high prevalence and burden of anxiety and depression, the low rates of treatment, and the numerous barriers to care, have all driven the calls for innovation in the way psychotherapy for common mental disorders are delivered (Bower & Gilbody, 2005; Clark, Layard, Smithies, Richards, Suckling & Wright, 2009; Kazdin & Blasé, 2011; Whiteford et al., 2013).

Internet-delivered psychotherapy

Delivery of psychotherapy via the internet is a recent approach used to increase access to psychotherapy. Internet-delivered psychotherapy employs the same underlying models and principles as face-to-face psychotherapy but uses technological devices and the internet to deliver therapeutic information and instructions to consumers. A feature of internet-delivered psychotherapy that is less common in services offering face to face care, has been the systematic measurement of symptoms using validated symptom questionnaires both at baseline and to measure progress through treatment and outcome. CBT has been the preferred model for most of the emerging online psychotherapy services, both because of the large evidence base for this model of care, and because it is comparatively easy to understand and readily adapted to online settings that are largely self-guided learning-based programs. Internet-delivered CBT (iCBT) has been the subject of most trials and research, and also adaptation as part of routine care (Andersson & Titov, 2014; Andersson, Titov, Dear, Rozental & Carlbring, 2019). This information is usually provided in the form of online modules, which consumers work through in their own time and without seeing a clinician face-to-face (Titov, Dear, Nielssen, Staples, Hadjistavropoulos, Nugent, ... & Repål, 2018). Internet-delivered psychotherapies can be provided with clinician support, where a clinician engages with consumers via telephone and secure email as an adjunct to the online modules, a model that is often referred to as *clinician-guided* treatment. However, internet-delivered psychotherapies can also involve little or no clinician support, or completely *self-guided* treatment. Whether clinician-guided or self-guided, internet-delivered psychotherapies typically involve only a fraction of the clinician time required for delivery of traditional face-to-face psychotherapies.

There is now a large body of evidence for the efficacy of internet-delivered psychotherapies for anxiety and depression, and their ability to increase access to treatment for consumers who otherwise would not be able to access care (Andersson, Cuijpers, Carlbring, Riper & Hedman, 2014; Andrews, Basu,, Cuijpers, Craske, McEvoy, English & Newby, 2018; Carlbring, Andersson, Cuijpers, Riper & Hedman-Lagerlöf, 2018). For example, a recent meta-analysis (trials = 64) found evidence of moderate to large effect sizes for iCBT for major depressive disorder (Hedges g = 0.67), generalised anxiety disorder (Hedges g = 0.70), panic disorder (Hedges g = 1.31) and social anxiety disorder (Hedges g = 0.92) over control conditions (Andrews et al., 2018). Moreover, meta-analyses indicated that iCBT is as effective as face-to-face CBT (Carlbring et al., 2018) and that results are maintained for several years after treatment (Andrews et al., 2018). Following the success of numerous RCTs, internet-delivered psychotherapy is increasingly being offered as a part of routine care (Titov et al., 2018). For example, the Australian Federal Government has funded the MindSpot Clinic to operate nationally and deliver digital (via telephone and internet) assessment, referral and treatment services to more than 20,000 Australians with anxiety and depression each year (Titov, Dear, Staples, Bennett-Levy, Klein, Rapee, ... & Nielssen, 2017). Similar *digital mental health services* (DMHS) have been established in many other countries including Sweden, Canada, Denmark and Norway (Titov et al., 2018), with similar services planned in other countries. The clinical outcomes of iCBT interventions delivered by these DMHS as part of routine care have replicated the results obtained from the initial controlled clinical trials of iCBT interventions (Staples, Dear, Johnson, Fogliati, Gandy, ,Fogliati, ... & Titov, 2019; Staples, Fogliati, Dear, Nielssen & Titov, 2016).

However, it is important to note that not all attempts at providing internet-delivered psychotherapies as a part of routine care have been successful (Gilbody, Littlewood, Hewitt, Brierley, Tharmanathan, Araya, ... & Kessler, 2015) and there still remains substantial variability between studies and treatments in clinical effectiveness (Cuijpers, Karyotaki, Reijnders, & Ebert, 2019; Ebrahim, Sohani, Montoya, Agarwal, Thorlund, Mills, & Ioannidis, 2014). Despite encouraging outcomes, numerous questions still remain, including which treatments and treatment approaches are the most effective (Boswell, Kraus, Miller, & Lambert, 2015; Cuijpers, van Straten, Bohlmeijer, Hollon &

Andersson, 2010), the characteristics of patients who benefit most and from which intervention (Driessen, Cuijpers, Hollon & Dekker, 2010; Roth & Fonagy, 2013), the models and approaches most effective for offering internet-delivered psychotherapy as part of routine care (Andrews, Bell, Boyce, Gale, Lampe, Marwat, ... & Wilkin, 2018; Andrews & Williams, 2015), and where such services best fit within existing health systems (Delgadillo, McMillan, Leach, Lucock, Gilbody, & Wood, 2014).

The importance of clinical evidence

Clinicians, health system managers, policymakers and even the general public rely on published research to weigh up the evidence regarding the efficacy of psychotherapies in deciding which treatment is worth providing, funding and pursuing (Howard, Moras, Brill, Martinovich & Lutz,1996; Spring, 2007; Roth & Fonagy, 2013; APA Presidential Task Force, 2006). Clinical evidence is generated through outcome audits and research that attempts to measure and interpret the efficacy of treatment under different conditions (Bakker & Wicherts, 2011; Roth & Fonagy, 2013; Wells, 1999). The basic philosophy underlying so-called *evidence-based medicine* is that clinical evidence plays a fundamental role in deciding which treatment to recommend and provide (Greenhalgh, Howick & Maskrey, 2014). This means that new treatments are compared with existing treatments to establish if they are indeed more effective, and to establish the conditions under which they are effective, leading to more informed decisions about the selection and the design of treatment pathways (Atkins, Fink, & Slutsky, 2005; Sackett, 2002).

Clinicians rely on clinical evidence in making decisions about whether psychotherapy is warranted for particular clients, and which particular treatment is most likely to be helpful under which conditions (APA Presidential Task Force, 2006). The father of evidence-based medicine, David Sackett, describes evidence-based medicine as:

> *"... the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence-based medicine means integrating individual clinical expertise with the best available*

> *external clinical evidence from systematic research"* (Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996, *p*.3).

While traditional clinical expertise derived from the often heuristic synthesis of long experience as to what might help an individual patient is still important, evidence-based medicine relies on the application of interventions with proven outcomes alongside clinical expertise. In two influence papers, trying to define empirically supported treatment, Chambless, Hollon, Miller and Robinson set out the role of clinical evidence in psychotherapy (Chambless & Hollon, 1998; Hollon, Miller & Robinson, 2002), arguing that without clinical evidence:

> *"… health care professionals are forced to rely exclusively on their direct experience of the effects of different interventions - an approach that risks erroneous conclusions."* (Hollon, Miller & Robinson, 2002, *p*. 1054)

Without reference to robust clinical evidence, there is an increased likelihood of both poor clinical decisions, unfounded advice and poor treatment outcomes for patients (APA Presidential Task Force, 2006; Chambless & Hollon, 1998; Hollon, Miller & Robinson, 2002). Moreover, that evidence is not just important for clinicians. Modern health system designers and policymakers rely heavily on clinical evidence to improve health care systems and decide which programs to support (e.g., Whiteford, 2019; Jorm, 2018; Pirkis, Burgess, Coombs, 2005; Meurk, Leung, Hall, Head & Whiteford, 2016). Without clinical evidence, health service administrators may not recognise gaps in services or take steps to address those gaps (Jorm & Mahli, 2013; Whiteford et al., 2013)

The challenges in generating clinical evidence in psychotherapy

Generating clinical evidence in the area of psychotherapy relies on the study of concepts and domains (e.g., mental health, self-efficacy, quality of life) that can be difficult to operationalise and measure (Altman & Simera, 2016; Bakker & Wicherts, 2011; Flay, Biglan, Boruch , Castro,

Gottfredson, Kellam, ... & Ji, 2005; Wells, 1999). Gottfredson and colleagues note that clinical evidence in psychotherapy is often:

> "… *specific to the intervention actually tested, the samples (or populations), the point in time and settings from which they were drawn, and the outcomes measured"* (Gottfredson, Cook, Gardner, Gorman-Smith, Howe, Sandler & Zafft, 2015, *p*. 893).

and went on to argue:

> *" … it is essential that conclusions from the research be clear regarding the intervention, population(s), time, and settings, and the outcomes for which efficacy is claimed."* (Gottfredson et al., 2015, *p*.896)

On the face of it, this advice seems straightforward. However, there are a number of factors that need to be taken into consideration in the generation and reporting of clinical evidence. For example, even where the focus of a clinical trial is on a specific condition such as major depressive disorder, the target outcomes of psychotherapy can vary from remission of the clinical diagnoses (McMillan, Gilbody, & Richards, 2010), to a reduction in symptoms (Kroenke, Monahan, & Kean, 2015; Hiller, Schindler, & Lambert, 2012; Sobocki , Ekman, Ågren, Runeson, & Jönsson, 2006), the prevention of relapse (Zhang, Zhang, Zhang, Jin & Zheng, 2018), increased adherence to treatment regimes (Donkin, Christensen, Naismith, Neal, Hickie & Glozier, 2011), the learning and use of particular psychological skills (Hundt, Mignogna, Underhill & Cully, 2013) or the production of some physiological or neurological change (Thomas, Leeson, Larkin, Deng, Pai, Mills & McLennan, 2016; Luna & Foster, 2015). Similarly, again focusing on the example of major depressive disorder, the populations of interest may span different age ranges from children to older adults (Hobbs, Mahoney & Andrews, 2017), whether the treatment is provided in one-on-one or a group-based format (Barkowski, Schwartze, Strauss, Burlingame, Barth & Rosendahl, 2016), to the general population or migrant or minority

populations (Diaz, 2017), or within inpatient or outpatient settings (Byatt, Levin, Ziedonis, Simas & Allison, 2015). Thus, the task of generating clinical evidence is complicated by the huge diversity of conditions, populations, contexts and outcomes relevant to the efficacy of psychotherapy (Kazdin, 1999; Roth & Fonagy, 2013).

Further to the challenges noted above, there is also a diverse range of research designs, methods and measures used to generate clinical evidence. For example, clinical researchers seeking to examine the efficacy of psychotherapy for depression can choose from a number of depression symptom scales, each measuring symptoms of depression in a different way (Choi, Schalet, Cook & Cella, 2014). There are also a large number of different structured clinical interviews, each using different questions and internal logic to identify the presence of clinically important diagnoses (Mitchell, Vaze & Rao, 2009). Moreover, the generation of clinical evidence can occur through both qualitative research methods (e.g., case studies, clinical interviews, single-subject studies) that examine the experience of treatment for specific individuals in some detail, and quantitative methods (e.g., randomised clinical trials) that emphasise the use of statistical estimates to describe the effects of treatment for groups of people (Bauer, Lambert & Nielsen, 2004; Preacher, 2015; Horner, Carr, Halle, McGee, Odom & Wolery, 2005). The diversity of research designs, methods and measures used in the evaluation of psychotherapy further complicates the generation of clinical evidence.

In summary, the diversity inherent in psychotherapy research means that the generation and interpretation of clinical evidence is highly complex, which Clarke (2007) described as:

> *"Every year, millions of journal articles are added to the tens of millions that already exist in the health literature, and tens of millions of web pages are added to the hundreds of millions currently available. Within these, there are many tens of thousands of research studies which might provide the evidence needed to make well-informed decisions about health care. The task of working through all this material is overwhelming enough, without then finding that the studies of relevance to the decision you wish to make all describe their findings in different ways, making*

> *it difficult if not impossible to draw out the relevant information."* (Clarke, 2007, p. 39)

In addition to the challenges in generating and interpreting clinical evidence, researchers have also commented on the effect of the diversity in methods on our ability to weigh the relative effectiveness of different treatments (Lambert & Ogles, 2009; Gottfredson, et al., 2015). The challenge to develop clinical evidence in a way that enables the comparison of different psychotherapies in differing populations and situations, has been described as the challenge to achieve ***external validity*** and ***evidence generalisability*** (Khorsan & Crawford, 2014; Rothwell, 2005; Glasgow, Green, & Ammerman, 2007; von Wolff, Jansen, Hölzel, Westphal, Härter & Kriston, 2014). Without the ability to consolidate and generalise clinical evidence in various settings, evidence-based psychotherapy practice is not possible (Glasgow, et al., 2007). Hence scientific research with the aim of improving and standardising measurement methodology, and creating metrics that can be used to consolidate and compare research findings (Ogles, Lunnen, & Bonesteel, 2001; Yi, Ma, Li, Zhou, Xiao, Zhang, ... & Liu, 2015), is important to establish the evidence base for psychotherapy. This thesis aims to add to that research by proposing methods to deal with several components in the generation of clinical evidence concerning psychotherapy that are not dealt with in a satisfactory way with existing methods.

The movement towards methodological standards

The emphasis on the use of robust research methods, scientific measurement tools and data analytics emerged in the 1950s and has now become widespread in clinical research (Beckstead, 2013, Ross, 1988; Tansella, 2002; Turner, Shamseer, Altman, Weeks, Peters, Kober, ... & Moher, 2012). However, the rapid advancement of scientific methods has created both opportunities and challenges. Advances in measurement methods, including the analysis of the various components of measurement instruments and their changes over time, has meant that data from clinical studies can be more accurately captured, evaluated and understood (Flay, et al., 2005; Gottfredson, et al., 2015). However,

the use of inappropriate methods that do not properly capture features such as the characteristics of the sample, the purpose of treatment, or confounding patterns within the data, can lead to faulty conclusions that threaten the validity of the clinical evidence. An influential figure in the drive to improve the rigour of health research, Doug Altman, has noted the dangers of applying incorrect measurement methods:

> *"What should we think about a doctor who uses the wrong treatment, either wilfully or through ignorance, or who uses the right treatment wrongly (such as by giving the wrong dose of a drug)? Most people would agree that such behavior was unprofessional, arguably unethical, and certainly unacceptable. What, then, should we think about researchers who use the wrong techniques (either wilfully or in ignorance), use the right techniques wrongly, misinterpret their results, report their results selectively, cite the literature selectively, and draw unjustified conclusions? We should be appalled. Yet numerous studies of the medical literature, in both general and specialist journals, have shown that all of the above phenomena are common. This is surely a scandal."* (Altman, 1994, *p*.283)

Altman noted that methodological errors from the use of statistically *wrong techniques*, the use of *right techniques wrongly*, *misinterpretation of results*, *incomplete reporting* or drawing *unjustified conclusions*, were common in published studies, even in top-ranking clinical research journals (Altman, 1994, Chalmers & Altman, 1999; Turner et al., 2012). Altman and others have observed that these errors could be minimised by the integration of certain precautions in the design, measurement, and analysis of clinical research studies aimed at generating clinical evidence (Bell, Fiero, Horton, & Hsu, 2014; Djulbegovic & Guyatt, 2017; Harris, Reeder & Hyun, 2011; Von Elm & Egger, 2004).

In an effort to improve the quality and interpretability of clinical evidence, and reduce the inappropriate use of different measurement methodologies, Altman and colleagues proposed the Consolidated Standards of Reporting Trials (CONSORT) framework and checklist (Begg, Cho, Eastwood, Horton, Moher, Olkin, ... & Stroup, 1996; Moher, Schulz & Altman, 2001; Schulz, Altman,

& Moher, 2010; Glasziou, Altman, Bossuyt, Boutron, Clarke, Julious, ... & Wager, 2014). CONSORT sets out the minimal methodological standards that researchers must both consider when designing research studies and report when publishing results of research (Begg, et al., 1996; Moher, et al., 2001; Schulz, et al., 2010). CONSORT includes the recommendation to:

(1) Use standardized (validated) scales and outcome measures (item 6 in Table 1);
(2) Include statistical analyses with estimates of precision and uncertainty (items 12, 16, 17, 18);
(3) Classify outcomes into interpretable and meaningful categories where possible (item 17, 18);
(4) Appropriately address missing cases and data (Hollis & Campbell, 1999) (item 13);
(5) Report harms and adverse events (items 18-19);
(6) Ensure random sampling and describe the features of any sample in a transparent way (item 4);
(7) Report on any other known sources of measurement bias (items 9, 11, 18).

Through adherence to CONSORT, clinical researchers are required to carefully consider their research methods and to transparently report their measurement methodology and outcomes, which in turn enhances the peer-review process and the replication of findings, resulting in more robust clinical evidence (Plint, Moher, Morrison, Schulz, Altman, Hill, & Gaboury, 2006; Turner, et al., 2012).

Since the establishment and adoption of CONSORT, similar frameworks have been established to guide clinical researchers using other clinical research designs, such as clinical observational studies (STROBE; The Strengthening the Reporting of Observational Studies in Epidemiology statement; Von Elm, Altman, Egger, Pocock, Gøtzsche, Vandenbroucke, & Strobe Initiative, 2007), diagnostics studies (STRAD; Standards for Reporting of Diagnostic Accuracy Studies; Bossuyt, Reitsma, Bruns, Gatsonis, Glasziou, Irwig, ... & Kressel, 2015), nonrandomized trial designs (TREND; transparent Reporting of Evaluations with Nonrandomized Designs; Des Jarlais, Lyles, Crepaz & the TREND Group, 2004), and meta-analytic studies (PRISMA; Transparent reporting of systematic reviews and meta-analyses; Moher, Liberati, Tetzlaff ,& Altman, 2009). These frameworks all highlight the issues relevant to their specific domains. However, as shown in Table 1, several common methodological issues can be

identified across all of the frameworks for research, forming a list of common methodological considerations that can be applied across quantitative research designs. These include the need for:

(1) Standardized outcome scales (STROBE – 7, 11; CONSORT – 6; TREND - 6);
(2) Statistical analyses with precision estimates (STROBE - 12; CONSORT - 12; TREND - 11);
(3) Appropriately addressing missing data (STROBE - 12; CONSORT - 13; TREND - 11);
(4) Reporting of harms or adverse events ((STROBE - 12; CONSORT - 13; TREND - 11); and;
(5) Classify outcomes into interpretable and meaningful categories where possible (STROBE - 12; CONSORT – 17-19; TREND - 11).

Table 1 – methodological checklist items from the STROBE, CONSORT and TREND frameworks

STROBE statement (Von Elm et al., 2007)			**CONSORT statement (Schulz, et al., 2010)**		
Section	**Item**		**Section**	**Item**	
Study design	4	Present key elements of study design early in the paper			a. Description of trial design (such as parallel, factorial) including allocation ratio
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	Trial design	3	b. Important changes to methods after trial commencement (such as eligibility criteria), with reasons
Participants	6	(*a*) *Cohort study*—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up			a. Eligibility criteria for participants
		Case-control study—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls	Participants	4	b. Settings and locations where the data were collected
		Cross-sectional study—Give the eligibility criteria, and the sources and methods of selection of participants	Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered
		(*b*) *Cohort study*—For matched studies, give matching criteria and number of exposed and unexposed	Outcomes	6	a. Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed
		Case-control study—For matched studies, give matching criteria and the number of controls per case			b. Any changes to trial outcomes after the trial commenced, with reasons
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable			a. How sample size was determined
Data sources/ measurement	8	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	Sample size	7	b. When applicable, explanation of any interim analyses and stopping guidelines
Bias	9	Describe any efforts to address potential sources of bias	Randomisation:		

Study size	10	Explain how the study size was arrived at
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding
		(*b*) Describe any methods used to examine subgroups and interactions
		(*c*) Explain how missing data were addressed
		(*d*) *Cohort study*—If applicable, explain how loss to follow-up was addressed *Case-control study*—If applicable, explain how matching of cases and controls was addressed *Cross-sectional study*—If applicable, describe analytical methods taking account of sampling strategy

Sequence generation	8	a. Method used to generate the random allocation sequence b. Type of randomisation; details of any restriction (such as blocking and block size)
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions
Blinding	11	a. If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how b. If relevant, description of the similarity of interventions
Statistical methods	12	a. Statistical methods used to compare groups for primary and secondary outcomes b. Methods for additional analyses, such as subgroup analyses and adjusted analyses

TREND statement (Bossuyt et al., 20105)

Section	Item	
Participants	3	a. Eligibility criteria for participants, including criteria at different levels in recruitment/sampling plan (e.g., cities, clinics, subjects) b. Method of recruitment (e.g., referral, self-selection), including the sampling method if a systematic sampling plan was implemented c. Recruitment setting d. Settings and locations where the data were collected
Interventions	4	a Details of the interventions intended for each study condition and how and when they were actually administered, specifically including: b. Content: what was given? c. Delivery method: how was the content given? d. Unit of delivery: how were subjects grouped during delivery? e. Deliverer: who delivered the intervention? f. Setting: where was the intervention delivered? g. Exposure quantity and duration: how many sessions or episodes or events were intended to be delivered? How long were they intended to last? h. Time span: how long was it intended to take to deliver the intervention to each unit?

		i. Activities to increase compliance or adherence (e.g., incentives)
Objectives	5	Specific objectives and hypotheses
Outcomes	6	a. Clearly defined primary and secondary outcome measures b. Methods used to collect data and any methods used to enhance the quality of measurements c. Information on validated instruments such as psychometric and biometric properties
Sample size	7	How sample size was determined and, when applicable, explanation of any interim analyses and stopping rules
Assignment method	8	a. Unit of assignment (the unit being assigned to study condition, e.g., individual, group, community) b. Method used to assign units to study conditions, including details of any restriction (e.g., blocking, stratification, minimization) c. Inclusion of aspects employed to help minimize potential bias induced due to non-randomization (e.g., matching)
Blinding (masking)	9	Whether or not participants, those administering the interventions, and those assessing the outcomes were blinded to study condition assignment; if so, statement regarding how the blinding was accomplished and how it was assessed
Unit of Analysis	10	a. Description of the smallest unit that is being analyzed to assess intervention effects (e.g., individual, group, or community) b. If the unit of analysis differs from the unit of assignment, the analytical method used to account for this (e.g., adjusting the standard error estimates by the design effect or using multilevel analysis)
Statistical methods	11	a. Statistical methods used to compare study groups for primary methods outcome(s), including complex methods for correlated data b. Statistical methods used for additional analyses, such as subgroup analyses and adjusted analysis c. Methods for imputing missing data, if used d. Statistical software or programs used

CONSORT : Consolidated Standards of Reporting Trials; STROBE : The Strengthening the Reporting of Observational Studies in Epidemiology statement; TREND statement: Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions.

As shown in Table 1, these guidelines and frameworks represent *a technical consensus* about the way to measure, present and interpret evidence about the effects of treatment. The CONSORT checklist has been used to form a consensus about evidence standards in clinical disciplines such as nursing (Smith, Lee, Lee, Choi, Jones, Bausell ,& Broome, 2008), medical trials (Turner, et al.,, 2012) psychiatry (Han, Kwak, Marks, Pae, Wu, Bhatia, ., ... & Patkar, 2009), E-health clinical research (Eysenbach & Consort-EHEALTH Group, 2011), paediatric psychology (Stinson, McGrath, & Yamada, 2003), and sports research (Yoon & Knobloch, 2012).

Since its establishment in the early 2000's the CONSORT has been adopted by the American Psychological Association, and has influenced the development of similar guidelines and frameworks concerned with the generation of clinical evidence in psychotherapeutic trials and evaluations, as described in the Journal Article Reporting Standards for Research in Psychology (JARS; Appelbaum, Cooper, Kline, Mayo-Wilson, Nezu, & Rao, 2018). The JARS effectively replicates the philosophical and methodological principles embodied in CONSORT (Appelbaum et al., 2018) and sets a similar minimal methodological standard for publication in leading psychology journals, including research on the efficacy of psychotherapy. According to the JARS standard (see Table 2) and consistent with CONSORT, clinical evidence in psychotherapy research should include among other things:

(1) The use of validated measures and measurement metrics;
(2) Appropriate reporting and adjustment for missing data and cases; and.
(3) The estimation of outcomes through statistical analysis with measures of uncertainty and inferential statistics.

In summary, recognition of the shortcomings in the methods used in previous research concerning the efficacy of psychotherapy has led to the development of more valid research methods, scientific measurement tools and tools for data analytics. This is reflected in the development and adoption of consensus guidelines for the methodology that must be considered in the generation, reporting and interpretation of clinical evidence. These guidelines are designed to improve the validity

and generalisability of the clinical evidence generated across all fields of health care research, including psychotherapy research.

The limitations of the current guidelines and frameworks

Despite the contribution of the CONSORT and JAR frameworks, it is important to recognise that their existence alone cannot ensure that the clinical evidence generated from research studies is internally valid or externally generalizable (Lang & Altman, 2013; Yoon & Knobloch, 2012). This is because these frameworks typically comprise a list of broad, but nonspecific principles that do not ensure that researchers will choose the methods that are suitable for the features of their data, that is, ensuring internal validity, or produce estimates comparable with other studies or contexts, that is, ensuring external generalisability. Instead, researchers are still required to make informed decisions about the methods and steps they will employ to address the issues raised by these frameworks, and the onus to achieve internal validity and external generalisability remains on researchers (Lang & Altman, 2013). For example, item 11 of the CONSORT checklist requires researchers to report statistical uncertainty through the use of confidence intervals and effect sizes. However, there are a range of statistical techniques that are available for estimating and representing the nature of clinical change associated with treatment, including (1) a categorical binary change in clinical diagnostic status or reliable clinical change; (2) nonlinear estimates of change, such as exponential or multiplicative symptom change estimates; (3) clustered outcomes as latent variables; (4) non-parametric or semi-parametric methods; and (5) standardization of scores or non-parametric ranking scores (Hartmann, van der Kooij ,Zeeck, 2009 ; Haynes, 2012; Preacher, 2015; Verkuilen & Smithson, 2012). The choice of statistical analytics can also include or omit conditional adjustments on key variables, such as demographical covariates, that may or may not be critical for detecting patterns of clinical change (Koutsouleris, Kambeitz-Ilankovic, Ruhrmann, Rosen, Ruef, Dwyer, ... & Schmidt, 2018). Similarly, there are many approaches for handling missing data, including unconditional imputation methods, bootstrapping or last observation carried forward (Woolley, Cardoni & Goethe, 2009), which can increase or decrease the accuracy of the results under different conditions (Bell & Fairclough, 2014;

Little, Jorgensen, Lang & Moore, 2014; Schafer & Graham, 2002). Finally, there are numerous ways of calculating effect sizes, such as partial and semi-partial effect sizes, where each effect represents the clinical outcomes in a different way and are not necessarily comparable (Kelly & Preacher, 2012; Smithson & Shou, 2016). Importantly, while all of these approaches are CONSORT compliant, each has the potential to lead to different estimates of the amount of clinical change, with the potential to affect the conclusions drawn (Smithson & Shou, 2016; Smithson & Verkuilen, 2006). Thus, while consensus frameworks are an important step in improving the quality of reporting of clinical evidence, there is still considerable opportunity for researchers to compromise the clinical evidence they generate through the decisions made in addressing the issues raised in consensus frameworks.

It is important to note that frameworks, such as CONSORT and JARS, are intentionally nonspecific in order to enable researchers to use methods and approaches that are suitable for their data and context (Yoon & Knobloch, 2012; Hopkins, Marshall, Batterham, & Hanin, 2009; Lang & Altman, 2013). All of the dominant frameworks provide researchers with the freedom to select appropriate scales, analytics, and designs for their aims and circumstances rather than mandating specific scales, analytics or designs. The importance of this is reflected in research demonstrating that the quality of clinical evidence can improve when, for example, researchers employ specialised scales that are designed for specific sub-populations (Heisel & Flett, 2016), and bespoke strategies for handling missing cases within their context (Kessler, van Loo, Wardenaar, Bossarte, Brenner, & Nierenberg, 2016). On the other hand, the non-specific nature of the guidelines requires researchers to understand and identify the optimal methods for generating clinical evidence, something researchers are often unable to do (Bell, et al., 2013; Harris, et al., 2011; Sharpe, 2013) and can lead to incorrect clinical conclusions about the effects of treatment (Baldwin, Fellingham, & Baldwin, 2016; Micceri, 1989; Verkuilen & Smithson, 2012; Vickers, 2005). This is also true of the application of inappropriate methods for handling missing cases (Bell & Fairclough, 2014; Li, Stuart & Allison, 2015; Streiner, 2008), the classification of clinical outcomes into unsuitable categories (King, 2011; Ogles, et al., 2001; Ronk, Hooke & Page, 2012), and the use of non-sensitive measurement scales (Angst, 2011; Mokkink,

Terwee, Patrick, Alonso, Stratford, Knol, ... & De Vet, 2010; Wyrwich, Norquist, Lenderking, Acaster, & Industry Advisory Committee of International Society for Quality of Life Research ISOQOL, 2013).

Hence, on the one hand, current frameworks provide a helpful list of considerations when generating and reporting clinical evidence, but on the other, their non-specific and non-directive nature leaves significant room for clinical researchers to generate invalid and unreliable clinical evidence that cannot be compared or replicated (Yoon & Knobloch, 2012). For this reason, the CONSORT and other frameworks need to be viewed as another part of the process used to generate clinical evidence (Lang & Altman, 2013; Yoon & Knobloch, 2012).

As well as addressing each of the items in CONSORT or JARS, researchers must also have some knowledge of, or access to, expert advice that takes into account the specialised literature in the fields of clinical and statistical methodology. Methodological literature offers guidance about the suitability of applying these methodologies in different contexts. An example is the use of longitudinal generalised estimation equation models (GEE models) as a method for statistically analysing trends of clinical change over time (Hubbard, Ahern, Fleischer, Van der Laan, Lippman, Jewell, ... & Satariano, 2010). The GEE method has been introduced to researchers via theoretical papers (Liang & Zeger, 1986), statistical software (Halekoh, Højsgaard, & Yan, 2006) and published examples of application in various journals (Genders, Spronk, Stijnen, Steyerberg, Lesaffre, & Hunink, 2012). However, there are other approaches for analysing longitudinal data, such as panel model designs, growth curve models, and methods that emphasis the partitioning of within-person variance, which are suitable under different conditions and with certain research aims (Little, Deboeck & Wu, 2015; Preacher, 2015). Faced with all of these approaches, clinical researchers must consider the ability of the available statistical methods to represent the trajectory of their patients in the treatment being considered, and must make their own choice about the method that they consider most suitable (Lang & Altman, 2013; Thabane, Mbuagbaw, Zhang, Samaan, Marcucci, Ye, & Debono, 2013), as well as make similar decisions about which outcome measures to use and how missing data should be handled (Lang & Altman, 2013). Together, the available frameworks rely on clinical researchers being sufficiently informed in all the areas of clinical and research methodology to make appropriate decisions.

The onus of choosing a method, and the idiographic-nomothetic axis

The preceding sections summarise several of the key challenges facing researchers and clinicians in their quest to generate robust and reliable clinical evidence. However, an additional challenge faced by clinical researchers involves determining the appropriate balance between internal validity and external generalisability. When studying the effect of psychotherapy, clinical researchers attempt to capture the trajectory of patients with as much accuracy and detail as possible, with the objective of maximum internal validity, while at the same time, measuring and reporting data in a way that can be generalised and compared, ensuring high external validity. The idea that a researcher's approach to measurement needs to balance the demands of internal validity and external generalisability originated in the work of Wilhelm Windelband in the 19th century (Kinzel, 2017; Robinson, 2011), who coined the terms *nomothetic measurement*, that is, the ability to generalise, and idiographic measurement, that is, the ability to specify. Nomothetic measurement aims to capture the generalizable features that are shared between different occurrences of a phenomenon, whereas *idiographic measurement* aims to describe the smallest features within particular phenomena (Robinson, 2011). Knowingly or not, clinical researchers are required to make choices that will define where their clinical evidence will fall on an idiographic-nomothetic axis, and whether they want to capture and emphasise the more specific idiographic features, or the more generalizable nomothetic features.

The ideographic-nomothetic axiom and the challenges associated with it are reflected in clinical research. For example, when researchers select scales for measuring outcomes, they may choose scales specific to a subgroup, such as adolescents with depression, or scales that are designed for general use in whole populations, such as people with depression. Specific measures may be chosen because they capture specific features that are relevant to a subgroup, such as the social challenges of adolescence (Reeve, Thissen, DeWalt, Huang, Liu, Magnus & Haley, 2016), and consequently these scales may be more sensitive to important clinical changes relevant to the subgroup. However, at the same time, the results from the specific scale may limit our ability to compare the clinical outcomes with other studies that use other measures that overlook these features, substantially reducing our ability to compare

outcomes across contexts. Thus, in making methodological decisions, clinical researchers often make important, but sometimes implicit, choices about the internal validity or idiographic, and external generalisability or nomothetic characteristics of the clinical evidence that will be generated.

There are many situations where clinical researchers make decisions along the idiographic-nomothetic axiom, in designing, analysing and reporting clinical evidence that emphasize either internal validity or external generalisability. Measurement metrics, such as Cohen's *d* effect sizes (Ellis, 2010; Lakens, 2013), represent clinical change through standardized unit-free scores in a way that enables the comparison of treatment effects across studies, measures and samples (Baird & Harlow, 2016; Cumming, Fidler, Kalinowski, & Lai, 2012; Lakens, 2013), which emphasises external generalisability and falls on the nomothetic end of the axiom. As a result, however, the standardized metrics can obscure the specificity and interpretability of results (Fried, van Borkulo, Epskamp, Schoevers,Tuerlinckx & Borsboom, 2016; McGrath & Meyer, 2006; Nakagawa, & Schielzeth, 2013; Smithson & Shou, 2016) by overlooking idiographic features of data within studies, such as statistical distributions, functions of change, or the possibility of subgroups with different trajectories of improvement. For example, when researchers employ generalised linear models, they fit more nuanced data features, such as the function of change, scale of scores and the presence of flooring or ceiling effects, in order to gain accuracy and validity through improved measurement accuracy (McLean, Sanders, & Stroup, 1991; Nakagawa & Schielzeth, 2013; Verkuilen & Smithson, 2012). At the same time, however, researchers opting for generalised models may lose the ability to compare the magnitude of change across studies of dissimilar symptom scales, distributions or change functions (Nakagawa & Schielzeth, 2013). In other words, the choice of generalised linear models over standardised metrics can impose a nomothetic-idiographic trade-off, between generalised measurement and interpretation of change, and the specific (ideographic) measurement and interpretation of change in context.

Within psychotherapy, the measurement of clinical evidence can be seen to move between the two ends of the nomothetic- idiographic spectrum (Levine, Sandeen & Murphy, 1992). On the extreme end of the nomothetic-idiographic spectrum are metrics such as clinical diagnosis change and effect sizes that emphasize the comparability of evidence across contexts (Levine et al., 2001; Rumpf, Meyer,

Hapke & John, 2001). Whilst these methods offer generalisability across contexts (Choi et al., 2014), standardized statistics overly simplify and overlook the specific and nuanced features of clinical data such as the heterogeneity of patient symptom remission (Fried, et al., 2016; Keller, 2003; Zimmerman, McGlinchey, Posternak, Friedman, Attiullah, Boerescu & Attiullah, 2006; Zimmerman, Posternak & Chelminski, 2007), or the statistical features of change (Fried, et al., 2016; Hiller et al., 2012; Kraemer, Noda & O'Hara, 2004; Vickers 2005). At the extreme idiographic end of the nomothetic-idiographic spectrum are statistical methods such a data mining, that represent a class of statistical methods for making outcome prediction rules that are specific, complex and fit the granular features of a dataset (Boman, Abdesslem, Forsell, Gillblad, Görnerup, Isacsson, ... & Kaldo, 2019; Chekroud, Zotti, Shehzad, Gueorguieva, Johnson,Trivedi, ... & , Corlett, 2016). These methods are considered to increase the predictive accuracy of individual clinical outcomes within a given context, but are very limited in their ability to create rules and knowledge that generalises across scales and contexts (Castelvecchi, 2016; Chekroud et al., 2016). Through this viewpoint, even the choice of statistical analyses and reporting of results represents a decision on the nomothetic-ideographic axiom. Thus, clinical researchers face numerous such choices when designing studies, generating and interpreting clinical evidence, with each decision affecting either the external or the internal validity of the clinical evidence generated.

The dominance of nomothetic methods in psychotherapy research

Because of the desire to report results that can be readily compared, techniques that emphasise nomothetic over ideographic, or external generalisability over internal validity, have become dominant in psychotherapy research. This dominance is particularly apparent when looking at the statistical analyses employed, the categorisation of clinical outcomes and how missing data is handled. Of all the statistical methods available, effect sizes and in particular Cohen's d have become the most common metric for describing the magnitude of clinical change in psychotherapy trials (Cumming, 2014; Johnston; Alonso-Coello, Friedrich, Mustafa, Tikkinen, Neumann, ... & Dalmau, 2016; Larken, 2013) and most meta-analyses and systematic reviews in the field now describe clinical change for individual

trials and therapies in terms of effect sizes (Sanders & Hunsley, 2018). Moreover, when trying to classify the clinical significance of symptom changes observed in treatment, the dichotomisation of change into specific unit-free categories, for example, reliable improvement versus not improved, has become dominant through the adoption of the methodology termed the Reliable Change Index (RCI) (Jacobson & Traux, 1991; Hiller, et al., 2012). Further to its use in identifying reliable clinical improvement. the RCI has also been recommended as a way to classify people who may have experienced adverse change in their symptoms during treatment (Rozental, Andersson, Boettcher, Ebert, Cuijpers, Knaevelsrud, ... & Carlbring, 2014).

Finally, and notwithstanding the comparatively fewer publications on approaches for handling missing data, a close reading of published clinical trials indicates that the dominant approach to handling missing cases include complete case analysis, last-observation carried-forward, and baseline observation carried forward methods, which do not consider specific features of missing data. A recent review of approaches for handling missing data in studies published in a sample of leading medical journals, found that 62% of studies overlooked missing data altogether and 32% used last-observation carried-forward approaches (Bell et al., 2014; Karyotaki, Riper, Twisk, Hoogendoorn, Kleiboer, Mira, ... & Andersson, 2017). Thus, the vast majority of studies used approaches that did not examine the specific features of missing cases in their datasets, such as systematic dropout (Fernandez, Salem, Swift and Ramatahal, 2015), or choose a method based on a preliminary examination for the reasons for missing cases. The handling of missing cases also reflects the general trend in psychotherapy research towards nomothetic methods that emphasise external validity and generalisability, rather than methods that emphasise internal validity and the idiographic aspects of the research context.

To characterise an exemplary range of methodological decisions and practices clinical researchers use in psychotherapy research, the methods detailed in seventy-four recent psychotherapy studies are reviewed in Table 2. The studies are taken from listed in three meta-analyses that collate evidence about the efficacy of psychotherapy for the treatment of depression and anxiety (Andersson, Cuijpers, Carlbring, Riper & Hedman, 2014; Karyotaki, et al., 2017; Newby, McKinnon, Kuyken, Gilbody & Dalgleish, 2015). One of these meta-analyses (n = 46 studies) explored the efficacy of

transdiagnostic cognitive behaviour therapy for anxiety and depression (Newby, et al., 2015), another (n = 15 studies) compared internet-delivered and face-to-face cognitive behaviour therapy for psychiatric and somatic disorders (Andersson, et al., 2014), and the third explored the efficacy of self-guided cognitive behaviour therapy for depression (Karyotaki, et al., 2017). Together, these meta-analyses studies represent a simple, brief and exemplary listing of psychotherapy studies that are used to evaluate the efficacy of psychotherapy for anxiety and depression within adult populations. The methodological practices within these 74 studies were reviewed to survey: (1) the type of statistical analysis conducted; (2) the symptom scales used; (3) any effort to classify the outcomes of treatment; and (4) the way missing cases were handled; this information is collated in table 2.

Table 2: Disorder, design, sample size, methodological framework and standardized measures used in the list of studies reviewed to survey the clinical evidence.

Study	Year	Disorder	Design	Sample size	Guideline framework	Measures
Barlow et al.	1984	GAD or PAN	RCT	20	N/A	STAI, BDI, CSR, PSC
Kabat-Zinn et al.	1992	GAD, PAN	Open trial	22	N/A	HARS-A, HARS-D, BAI, BDI, MI, FSS
Radley et al.	1997	GAD	Open trial	9	N/A	HAD-A, HAM-A, GAS, STAI, FI, PSI, CAQ, ELI
Barrowclough et al.	2001	GAD, PAN, SocPhob	RCT	55	Not reported	BAI, STAI-T, HARS-A, BDI, GDS
Clarke et al.	2002	DEP	RCT	299	Not reported	CED-S
Patel et al.	2003	Any common mental disorders	RCT	450	Not reported	CISR, BDQ
Proudfoot et al.	2003	GAD and/or DEP	RCT	274	Not reported	BDI-II, BAI, WSAS
Christensen et al.	2004	DEP	RCT	525	Not reported	CED-S
Kenwright et al.	2004	Phobia or PAN	Non-randomised trial	27	Not reported	FQ, WSAS, BDI
Marks et al.	2004	Phobia or PAN	RCT	93	CONSORT	FQ, WSAS
Norton et al.	2004	GAD	RCT	23	Not reported	DASS-42, MASQ
Proudfoot et al.	2004	GAD and/or DEP	RCT	167	Not reported	BDI-II, BAI, WSAS
Anderson et al.	2005	PAN, SocPhob	Open trial	10	Not reported	PRCS, SSPS-pos, SSPS-neg, PRCA
Carlbring et al.	2005	PAN	RCT	49	Not reported	BSQ, ACQ, MI, BAI, BDI, QILI
Clarke et al,	2005	DEP	RCT	255	Not reported	CED-S; SF-12 PCS
Cyranowski et al.	2005	DEP, PAN	Open trial	18	Not reported	HRSD, HARS, BDI, BAI, WLESQ
Gollings et al.	2006	Body dissatisfaction	RCT	40	Not reported	BSQ BIAQ BDI-II STAI, RSE
Craske et al.	2007	PAN	RCT	65	Not reported	ASI, FQ, CSR, BSI, SSS, BATs
Erickson et al.	2007	GAD, PTSD, PAN, OCD, SocPhob	RCT	152	Not reported	GAF, BAI, BDI-II, ASI
Lee et al.	2007	GAD, PAN	RCT	46	Not reported	HAM-A STAI, HAM-D BDI, SCL-90-R
Liu et al.	2007	DEP GAD, PAN, SocPhob, OCD, MADD	RCT	254	Not reported	CISR, HRSD, SF-36,
McEvoy & Nathan	2007	GAD and/or affective disorder	Open trial	143	Not reported	BDI-II, BAI
Paxton et al.	2007	Body dissatisfaction, disordered eating	RCT	79	CONSORT, APA	BSQ BULIT-R BDI-II RSE
Ree and Craigie	2007	GAD and/or DEP	Open trial	26	Not reported	BDI, DASS-42
Spek et al.	2007	DEP	RCT	301	Not reported	BDI-II, EDS, WHO-CIDI
Spek et al.	2007	DEP	RCT	201	Not reported	BDI, EDS, WHO-CIDI
Westra et al.	2007	GAD	Open trial	115	Not reported	BDI-II, BAI
Kaldo et al.	2008	Tinnitus	RCT	51	CONSORT	TRQ
Kiropoulos et al.	2008	PAN	RCT	86	Not reported	PDSS ASP DASS ACQ BVS WHO-QOL
Norton	2008	GAD	Open trial	52	Not reported	STAI-S,
Andersson et al.	2009	Specific phobia (spider)	RCT	30	Not reported	BAT BDI, BAI, FSS-III
Clarke et al.	2009	DEP	RCT	160	CONSORT	PHQ-8
De Graaf et al	2009	DEP	RCT	303	Not reported	BDI-II, WSAS, SF-36, DAS-A, SCL-6

Kim et al.	2009	GAD, PAN	RCT	46	Not reported	HAM-A, BAI, SCL-90-R, HAM-D, BDI
Meyer et al.	2009	DEP	RCT	396	Not reported	BDI, WSAS
Wetherell et al.	2009	GAD	RCT	30	Not reported	HARS, PSWQ, BDI-II, SF-36
Bergström et al.	2010	PAN	RCT	113	CONSORT	PDSS CGI MADRS ASI SDS
Botella et al.	2010	SocPhob	RCT	98	CONSORT	FPSQ
Bressi et al.	2010	GAD and/or DEP	RCT:	60	Not reported	SCL-90-R, CGI, IIP
Ellard et al.	2010	GAD	Open trial	42	Not reported	BDI-II, BAI, PANAS-PA, PANAS-NA, OCI-R, other
Jakupcak et al.	2010	DEP & PTSD	Open trial	7	Not reported	PCL-M, BDI-II, QOLI
Titov et al.	2010	GAD, DEP, PAN, SocPhob	RCT	78	CONSORT	PHQ-9, GAD-7, Social Phobia-12, PDSS-SR, other
Andersson et al.	2011	GAD	Open trial	10	Not reported	CORE-OM, MADRS-S, BAI, QOLI
Andrews et al.	2011	SocPhob	RCT	37	Not reported	SIAS
Berger et al.	2011	DEP	RCT	76	Not reported	BDI-II, BSI, IIP, WHOQOL-BREF, GSI
Carlbring et al.	2011	GAD	RCT:	54	CONSORT	CORE-OM, MADRS-S, BAI, QOLI
Dear et al.	2011	GAD, DEP, PAN, or SocPhob	Open trial	32	Not reported	DASS-21, PHQ-9, PSWQ, PDSS-SR, SP-12, others
Farrer et al	2011	DEP & Psychological distress	RCT	155	CONSORT	CED-S
Hedman,et al.	2011	SocPhob	RCT	126	CONSORT	LSAS
Johnston et al.	2011	GAD, PAN, SocPhob	RCT	139	CONSORT	GAD-7, DASS-21, PSWQ, SIAS-6/SPS-6, others
Nixon & Nearmy	2011	DEP, PTSD	Open trial	20	Not reported	CAPS, PDS, DASS-D, PTCI
Schover et al.	2011	Male sexual dysfunction	RCT	81	Not reported	IIEF
Titov et al.	2011	GAD, PAN, SocPhob	RCT	75	CONSORT	PSWQ, SPSQ, PDSS-SR, GAD-7, PHQ-9, SDS, K-10, DASS-21, NEO
Vollestadet al.	2011	GAD	RCT	76	Not reported	BAI, PSWQ, STAI, BDI-II, SCL-90
Arch et al.	2012	GAD	RCT:	128	Not reported	ADIS CSR, ASI, PSWQ, FQ, QOLI, AAQ,
Brenes et al.	2012	GAD or PAN	RCT:	60	Not reported	PSWQ, STAI-T, ASI, BDI, HARS-A, SF-36
Farchione et al.	2012	GAD	RCT	37	CONSORT	HARS, HRSD, SIGH-A, SIGH-D, BDI-II, BAI, Others
Johansson et al.	2012	DEP & comorbid disorders	RCT	115	CONSORT	BDI-II, MADRS-S, BAI, QOLI
Moritz et al.	2012	DEP	RCT	210	Not reported	BDI, DAS, RSE, SBQ-R, WHOQOL-BREF
Norton	2012	GAD	RCT	87	CONSORT	ADIS CSR, CGI, ADDQ, BAI, PDSS, SPDQ, GAD-Q
Norton & Barrera	2012	GAD, PAN, SocPhob	RCT	46	JARS	ADIS CSR, STAI, PDSS, SPDQ, GADQ-IV, BDI
Schmidt et al.	2012	GAD, PAN, SocPhob	RCT	96	CONSORT	ASI, BDI-II, MI, SDS, SPRAS, CGI
Zou et al.	2012	GAD and/or DEP	Open trial	22	Not reported	GAD-7, PHQ-9, SDS
Arch et al.	2013	PAN, OCD, SocPhob, GAD, PTSD	RCT	105	Not reported	CSR, PSWQ, MASQ-AA, BDI-II
Johansson et al.	2013	GAD, DEP, SocPhob, PAN	RCT	100	CONSORT	GAD-7, PHQ-9
Newby et al.	2013	GAD and/or DEP	RCT	99	CONSORT	PHQ-9, GAD-7, K-10, BDI-II, PSWQ, WHODAS-II
Wuthrich & Rapee	2013	GAD and/or DEP	RCT	60	CONSORT	GDS, CES-D, GAI, PSWQ, SF-12
Berger et al.	2014	GAD, PAN, SocPhob	RCT	88	Not reported	BAI, BDI-II, BSI, SPS, SIAS, MI, PSWQ, BSQ, CAQ

Phillips et al.	2014	GAD and/or DEP, social adjustment	RCT	637	CONSORT	WSAS, CORE-10, GAD7, PHQ9,
Wagner et al.	2014	DEP	RCT	62	CONSORT	BDI, BSI, AHS, ATQ-R
Gilbody et al.	2015	DEP	RCT	691	CONSORT	PHQ-9, SF-36, ED-Q5
Kleiboer et al.	2015	GAD and/or DEP	RCT	537	Not reported	PHQ9, CESD, HADS, BAI
Meyer et al.	2015	DEP	RCT	326	CONSORT (e-Health)	PHQ-9; GAD-7; SF-12
Klein et al.	2016	DEP	RCT	1013	CONSORT (e-Health)	PHQ-9, HDRS-24, QIDS-C16, SF-12

Disorders: GAD = generalised anxiety disorder, MADD = mixed anxiety and depressive disorder, MDD = major depressive disorder, OCD = obsessive compulsive disorder, Panic/Ag = panic disorder and/or agoraphobia, PTSD = posttraumatic stress disorder, SocPhob = social phobia or social anxiety disorder.

Measures: ADIS-IV: Anxiety Disorders Interview Schedule for DSM-IV; ADIS-R: Anxiety Disorders Interview Schedule-Revised; AKUADS: Aga Khan University Anxiety and Depression Scale; ASI: Anxiety Sensitivity Index, BAI: Beck Anxiety Inventory; BAT: Behavioural Activation Test; BDI-II: Beck Depression Inventory, second edition; BSI : Brief Symptom Inventory; BSQ – Body Sensation Questionnaire; CAQ: Cognitive Anxiety Questionnaire; CESD: Centre of Epidemiological Studies Depression Scale; CGI: Clinical Global Improvement-Patient Rating, CID: Clinical Interview for Depression; CORE-OM: Clinical Outcome's in Routine Evaluation; DASS-21: Depression Anxiety and Stress Scales (DASS) 21-item version; DASS-42: Depression Anxiety and Stress Scales (DASS) 42-item version; EDS: The Edinburgh Depression Scale; EFI: Effects on Life Inventory; FI: Fear Inventory; FNE: Fear of Negative Evaluation Scale; FQ: Fear Questionnaire; FQAD: Fear Questionnaire Anxiety–Depression Subscale; FQSP: Fear Questionnaire Social Phobia Subscale; FSS: Fear Survey Schedule; GAD-7: Generalised Anxiety Disorder 7-item scale; GAS: Generalised Anxiety Scale from the Guys/Age Concerned Survey; HADS-A: Hospital Anxiety and Depression Scale — Anxiety Subscale; HADS-D: Hospital Anxiety and Depression Scale — Depression Subscale; HAI: Health Anxiety Inventory; HARS: Hamilton Anxiety Rating Scale; HRSD: Hamilton Rating Scale for Depression; IIEF – International Index of Erectile Function; K-10: Kessler 10-item; LSAS-SR: Liebowitz Social Anxiety Scale self-report version; MADRS-S: Montgomery–Asberg Depression Rating Scale — self rated version; MASQ: Mood and Symptoms Questionnaire; MIA: Mobility Inventory for Agoraphobia; NEO-N: NEO-Five Factor Inventory—Neuroticism Subscale; OCI-R: Obsessive–Compulsive Inventory—Revised Version; PANAS: Positive and negative affect scale; PAS: Panic and Agoraphobia Scale; PCL-C Posttraumatic Stress Disorder Checklist — Civilian; PSC: Psychosomatic Symptom Checklist; PDSS – Panic Disorder Severity Scale; PDSS-SR: Panic Disorder Severity Scale — Self report; PHQ: Patient Health Questionnaire; PRCA: Personal Report of Communication Apprehension; PRCS: Personal Report of Confidence as a Speaker; PSI: Physical Symptoms inventory; PSWQ: Penn State Worry Questionnaire; QLESQ: Quality of Life Enjoyment and Satisfaction Scale; QOLI: Quality of Life Inventory; SAD: Social Avoidance and Distress Scale; SDS: Sheehan Disability Scale; SIAS: Social Interaction Anxiety Scale; SIAS/SPS6 composite: Social Interaction Anxiety Scale and Social Phobia Scale 6-item composite; SIGH: Structured Interview Guide for the Hamilton Anxiety Rating Scale (A = Anxiety, D = Depression); SPSQ: Social Phobia Screening Questionnaire; SQ: Kellner's Symptom Questionnaire; SSPS: Self-Statements During Public Speaking (positive and negative subscales); SSS: Subjective Symptoms Scale (measure of interference with daily functioning) STAI: State-Trait Anxiety Inventory (T = trait, S = state); TRQ – Tinnitus Reaction Questionnaire; WSAS: Work and Social Adjustment Scale; YBOCS: Yale Brown Obsessive Compulsive Scale; Zung SRSD: Zung Self-Rating Scale for Depression.

Designs and guideline frameworks: CONSORT : Consolidated Standards of Reporting Trials; JARS: Journal Article Reporting Standards for research in psychology
PRISMA : Transparent reporting of systematic reviews and meta-analyses; RCT = randomised controlled trial; STRAD : Standards for Reporting of Diagnostic Accuracy Studies; STROBE : The Strengthening the Reporting of Observational Studies in Epidemiology statement;

Table 2 cont.: Measurement methodology, statistical analyses, clinical effect reporting and reporting of categorical outcomes

Study	Year	How was missing data handled	Conducted statistical analysis (type)	How clinical effects were reported (Type of effect sizes)	Reporting of clinical events (Categorical outcomes)
Barlow et al.	1984	N/A	ANOVA	Means	not reported
Kabat-Zinn et al.	1992	Completer's analysis	Repeated measures ANOVA	Means	Proportion remaining clinical
Radley et al.	1997	N/A	Wilcoxon signed ranks test	Means/ Ranks	N/A
Barrowclough et al.	2001	Completer's analysis	ANCOVA	Means, Cohen's D	20% improvement criterion; Proportion remaining clinical
Clarke et al.	2002	Not reported	Mixed linear model; Quadratic effects	Means (effect sizes - not outright specified)	Not reported
Patel et al.	2003	Model-based simulation (MAR - time&condition only)	Mixed linear model	Means, Cohen's D	Proportion remaining clinical
Proudfoot et al.	2003	Completer's analysis	Mixed linear model	Means	Proportion remaining clinical
Christensen et al.	2004	Completer's analysis	ANOVA (Difference scores)	Means, Cohen's D	Proportion remaining clinical
Kenwright et al.	2004	Completer's analysis	Paired T-tests	Means, effect sizes, percentage improvement	Not reported
Marks et al.	2004	LOCF	MANOVA	Means, effect sizes, percentage improvement	Not reported
Norton et al.	2004	not reported	MANOVA	Means, Cohen's D	Proportion remaining clinical
Proudfoot et al.	2004	Completer's analysis	Mixed linear model	Means	Proportion remaining clinical
Anderson et al.	2005	N/A	Repeated measures ANOVA	Means	30% improvement criterion
Carlbring et al.	2005	Not reported	Repeated measures ANOVA	Means, Cohen's D	RCI; Proportion remaining clinical; clincal interverviews SCID
Clarke et al,	2005	Model-based simulation (MAR time&condition only)	Mixed linear model (random slope, intercept)	Means (effect sizes - not outright specified)	Not reported
Cyranowski et al.	2005	LOCF	Mixed linear models	Means, Cohen's D	Proportion remaining clinical; 50% improvement criterion
Gollings et al.	2006	Not reported	Repeated ANOVA	Means, effect sizes (partial $\eta 2$)	Proportion remaining clinical
Craske et al.	2007	BOCF; LOCF	ANOVA	Means, Cohen's D	Proportion remaining clinical
Erickson et al.	2007	Completer's analysis	Repeated measures ANOVA	Means, Cohen's D	Proportion remaining clinical; 50% improvement criterion
Lee et al.	2007	LOCF	Repeated measures ANOVA	Means	Not reported

Liu et al.	2007	Completer's analysis	Repeated measures ANCOVA	Means	Proportion remaining clinical
McEvoy & Nathan	2007	Completer's analysis	MANOVA	Means, Cohen's D	RCI; Proportion remaining clinical
Paxton et al.	2007	Completer's analysis	ANVOA	Means, Cohen's D	RCI; Proportion remaining clinical
Ree and Craigie	2007	Completer's analysis	Paired T-tests	Means, Cohen's D	20% improvement criterion & Proportion remaining clinical
Spek et al.	2007	Model based Multiple imputations (MAR time&condition only)	Observed scores with imputations	Means, Cohen's D	RCI, Proportion remaining clinical
Spek et al.	2007	Model based Multiple imputations (MAR time&condition only)	Observed scores with imputations	Means, Cohen's D	RCI; Proportion remaining clinical
Westra et al.	2007	Not reported	Linear regression	Means, Std, correlation coefficients	1Std/fixed value
Kaldo et al.	2008	LOCF	Repeated ANOVA	Means, Cohen's D	Fixed score criterion
Kiropoulos et al.	2008	BOCF	Repeated measures ANOVAs ranking test	Means, effect sizes (partial $\eta 2$)	Not reported
Norton	2008	Model based MCAR	Mixed linear model	Means, Cohen's D	Not reported
Andersson et al.	2009	Not reported	Post treatment f- test; χ square comparison between RCI	Means, Cohen's D	RCI, Proportion remaining clinical
Clarke et al.	2009	Model based Multiple imputations (MAR time&condition only)	Generalized hierarchical mixed modeling with random slopes	Means, Cohen's D	Proportion remaining clinical
De Graaf et al	2009	Completer's analysis	Repeated measures ANOVA	Means, Cohen's D	RCI; Proportion remaining clinical
Kim et al.	2009	Model based MCAR	Repeated measures ANOVA	Means	Not reported
Meyer et al.	2009	Completer's analysis, LOCF	MANOVA	Means, Cohen's D	RCI, Proportion remaining clinical
Wetherell et al.	2009	N/A	MANOVA	Means, effect sizes Hedge's G	Not reported
Bergström et al.	2010	Model based MCAR	Mixed linear model	Means, Cohen's D	40% improvement criterion; Proportion remaining clinical; clincal interverviews DSM-IV
Botella et al.	2010	Completer's analysis, model based MCAR imputation, LOCF	ANCOVA	Means, effect sizes (partial $\eta 2$)	Proportion remaining clinical; CGI-I rating
Bressi et al.	2010	Model based Multiple imputations (MAR time&condition only)	Paired T-tests	Means, Cohen's D	RCI
Ellard et al.	2010	Completer's analysis	Repeated measures ANOVA	Means, Cohen's D	30% improvement criterion
Jakupcak et al.	2010	N/A	Repeated measures ANOVA	Means	Proportion remaining clinical
Titov et al.	2010	BOCF	ANCOVA	Means, Cohen's D	50% improvement, Proportion remaining clinical
Andersson et al.	2011	N/A	Paired T-tests	Means, Cohen's D	Clinical interview CGI-I rating
Andrews et al.	2011	Not reported	ANCOVA	Means, Cohen's D	Not reported
Berger et al.	2011	LOCF	ANCOVA	Means, Cohen's D	RCI; Proportion remaining clinical
Carlbring et al.	2011	Model based MCAR	Repeated measures ANOVA	Means, Cohen's D	Clinical interview & diagnosis (CGI-I)
Dear et al.	2011	BOCF	Paired T-tests	Means, Cohen's D	Clinical interview & diagnosis; Proportion remaining clinical

Farrer et al	2011	Model based Multiple imputations (MAR time&condition only)	Repeated measures ANOVA	Means, effect sizes (Hedge's G)	Proportion remaining clinical
Hedman,et al.	2011	Completer's analysis	Mixed linear model	Means, Cohen D, inferiority margin	RCI; Proportion remaining clinical
Johnston et al.	2011	BOCF	ANCOVA	Means, Cohen's D	Clinical interview MINI; Proportion remaining clinical
Nixon & Nearmy	2011	LOCF	Paired T-tests	Means, Cohen's D	RCI; Proportion remaining clinical
Schover et al.	2011	Model based Multiple imputations (MAR time&condition only)	Mixed linear model	Means, Cohen's D	Proportion remaining clinical
Titov et al.	2011	BOCF	ANCOVA	Means, Cohen's D	50% improvement, Proportion remaining clinical, Clinical interveiws (MINI)
Vollestadet al.	2011	LOCF, Completer's analysis	ANCOVA	Means, Cohen's D	RCI; Proportion remaining clinical
Arch et al.	2012	Completer's analysis	HLM/HMLM	Means, Cohen's D	Clinical interview & diagnosis (CGI-I)
Brenes et al.	2012	Model based Multiple imputations (MAR time&condition only)	ANCOVA	Means, Cohen's D	RCI; Clinical interview diagnosis (CGI-S)
Farchione et al.	2012	N/A	Linear Regression	Means, effect sizes - Hedges' G	Clinical interview & diagnosis; Proportion remaining clinical
Johansson et al.	2012	Model based simulation (MAR)	Mixed linear models (random intercept/slope)	Means, Cohen's D	Clinical interview CGI-I rating; Proportion remaining clinical
Moritz et al.	2012	Model based Multiple imputations (MAR)	ANCOVAS	Means, Cohen's D	50% improvement
Norton	2012	Model based simulation (MAR); LOCF	Mixed linear model	Means, Cohen's D	Not reported
Norton & Barrera	2012	Model based simulation (MAR)	Mixed linear model	Means, effect sizes (partial η2)	30% improvement criterion
Schmidt et al.	2012	Model based simulation (MAR)	Linear regression	Means	RCI; Proportion remaining clinical
Zou et al.	2012	BOCF	Paired T-tests	Means, Cohen's D	Proportion remaining clinical, 50% improvement, clinical interview (MINI)
Arch et al.	2013	Model based MCAR	HLM	Means, effect sizes (partial η2)	RCI
Johansson et al.	2013	N/A	Mixed linear model	Means, Cohen's D	Clinical interview CGI-I rating; Proportion remaining clinical
Newby et al.	2013	Completer's analysis (study 1); MCAR (Study 2)	Repeated measures ANOVA	Means, effect sizes Hedge's G, Cohen's D	RCI; Clinical interview diagnosis Mini; Proportion remaining clinical
Wuthrich & Rapee	2013	Model based Multiple imputations (MAR)	Mixed linear model	Means, Cohen's D	RCI; Proportion remaining clinical
Berger et al.	2014	Model based simulation (MAR); LOCF	Mix linear model	Means, Cohen's D	Proportion remaining clinical
Phillips et al.	2014	Not reported	Mixed linear model (random intercept)	Means, Cohen's D	Proportion remaining clinical

Wagner et al.	2014	BOCF	MANOVA	Means, Cohen's D	Proportion remaining clinical
Gilbody et al.	2015	Senstivity analysis assuming best/worst cases outcomes	Mixed linear models	Means, effect sizes (Hedge's G)	Proportion remaining clinical
Kleiboer et al.	2015	Model based Multiple imputations (MAR) (no adjustments); Completer's analysis	Linear regression	Means, Cohen's D	Not reported
Meyer et al.	2015	Completer's analysis	Mixed linear model	Means, Cohen's D	Proportion remaining clinical & 50% improvement
Klein et al.	2016	Model based Multiple imputations (MAR); Completer's analysis	Mixed linear model (random intercept)	Means, Cohen's D	Proportion remaining clinical

Statistical abbreviations and measures: ANCOVA : Analysis of covariance; ANOVA : Analysis of variance; BOCF : baseline observation carried forward; CGI - I : Clinical Global Impression-Improvement; DSM-IV : Diagnostic and Statistical Manual; LOCF : last observation carried forward; HLM : Hierarchical Linear Modelling; HMLM : Hierarchical multivariate Linear Modelling; MANOVA - Multivariate analysis of variance; MAR : Missing at random; MCAR : Missing at random; MINI : Mini-International Neuropsychiatric Interview; N/A : Information not available; RCI : Reliable Change Index; SCID : Structured Clinical Interview; Std : Standard deviation

Table 2 cont.: Statistical precautions, missing cases assumption testing, statistical modelling assumption testing

Study	**How was missing data handled?**	**Missing values assumption testing**	**Analytics assumption checking**
Anderson et al.	N/A	N/A	Not reported
Andersson et al.	Not reported	Not reported	Not reported
Andersson et al.	N/A	N/A	Not reported
Andrews et al.	Not reported	Not reported	Not reported
Arch et al.	Model based MCAR	T-tests of subgroups at baseline - no patterns	Not reported
Arch et al.	Completer's analysis	Tested - T-tests of subgroups at baseline - no patterns identified	Tested - identified curvilinear patterns - employed curvilinear terms
Barlow et al.	N/A	N/A	Not reported
Barrowclough et al.	Completer's analysis	Not reported	Tested - Skewness reported
Berger et al.	LOCF	Not reported	Not reported
Berger et al.	Model based simulation (MAR); LOCF	Not reported	Not reported
Bergström et al.	Model based MCAR	Not reported	Not reported
Botella et al.	Completer's analysis, model based MCAR imputation, LOCF	Not reported	Sensitivity of results
Brenes et al.	Model based Multiple imputations (MAR time&condition only)	Not reported	Not reported
Bressi et al.	Model based Multiple imputations (MAR time&condition only)	Tested - Chi square comparison at baseline - no patterns	Not reported
Carlbring et al.	Not reported	Not reported	Not reported
Carlbring et al.	Model based MCAR	Not reported	Not reported

Christensen et al.	Completer's analysis	Not reported	Not reported
Clarke et al,	Model based simulation (MAR time&condition only)	Not reported	Reported Quadratic trends; dosage effects
Clarke et al.	Model based Multiple imputations (MAR time&condition only)	Tested - increased completion with age - no other factors	Reported Quadratic trends; dosage effects
Clarke et al.	Not reported	Tested - Reported increased attrition with increased baseline; no other effects	Reported Quadratic trends; dosage effects
Craske et al.	BOCF; LOCF	Not reported	Not reported
Cyranowski et al.	LOCF	Not reported	Not reported
De Graaf et al	Completer's analysis	N/A	Tested - some deviation from normality
Dear et al.	BOCF	Not reported	Not reported
Ellard et al.	Completer's analysis	Not reported	Not reported
Erickson et al.	Completer's analysis	Tested- identified increased depression at baseline	Not reported
Farchione et al.	N/A	N/A	Not reported
Farrer et al	Model based Multiple imputations (MAR time&condition only)	Tested - adherence and treatment condition effects	Tested - deviations from normality
Gilbody et al.	Sensitivity analysis assuming best/worst cases outcomes	Not reported	Not reported
Gollings et al.	Not reported	t-tests of baseline differences (age effects) (no symptom effects)	Checked - results not reported
Hedman,et al.	Completer's analysis	Not reported	Not reported
Jakupcak et al.	N/A	N/A	Not reported
Johansson et al.	Model based simulation (MAR)	N/A	Tested baseline only - reported normality
Johansson et al.	N/A	N/A	not reported
Johnston et al.	BOCF	Not reported	not reported
Kabat-Zinn et al.	Completer's analysis	not reported	not reported
Kaldo et al.	LOCF	N/A	Not reported
Kenwright et al.	Completer's analysis	not reported	not reported
Kim et al.	Model based MCAR	Not reported	not reported
Kiropoulos et al.	BOCF	Not reported	Checked – positive skewness reported
Kleiboer et al.	Model based Multiple imputations (MAR) (no adjustments); Completer's analysis	Not reported	Not reported
Klein et al.	Model based Multiple imputations (MAR); Completer's analysis	Tested - no variables identified	Not reported
Lee et al.	LOCF	Not reported	not reported
Liu et al.	Completer's analysis	Chi square - comparison at baseline; gender differences	not reported
Marks et al.	LOCF	Not reported	not reported

McEvoy & Nathan	Completer's analysis	Not reported	not reported
Meyer et al.	Completer's analysis, LOCF	Not reported	Not reported
Meyer et al.	Completer's analysis	Not reported	Not reported
Moritz et al.	Model based Multiple imputations (MAR)	Not reported	Not reported
Newby et al.	Completer's analysis (study 1); MCAR (Study 2)	Not reported	not reported
Nixon & Nearmy	LOCF	not reported	Not reported
Norton	Model based MCAR	Not reported	Not reported
Norton	Model based simulation (MAR); LOCF	Not reported	Not reported
Norton & Barrera	Model based simulation (MAR)	Not reported	Not reported
Norton et al.	not reported	Not reported	Not reported
Patel et al.	Model based simulation (MAR - time&condition only)	Not reported	Reported skewness; applied (Bootstrapped)
Paxton et al.	Comnpleters analysis	Not reported	Not reported
Phillips et al.	Not reported	Tested - age decreasing missingness, increased baseline symptoms	Reported Normality of model residuals
Proudfoot et al.	Completer's analysis	Not reported	Not reported
Proudfoot et al.	Completer's analysis	Not reported	Not reported
Radley et al.	N/A	N/A	N/A
Ree and Craigie	Completer's analysis	not reported	Not reported
Schmidt et al.	Model based simulation (MAR)	Not reported	Not reported
Schover et al.	Model based Multiple imputations (MAR time&condition only)	Not reported	Not reported
Spek et al.	Model based Multiple imputations (MAR time&condition only)	Not reported	Not reported
Spek et al.	Model based Multiple imputations (MAR time&condition only)	Not reported	Checked -"close to normal"
Titov et al.	BOCF	Not reported	Not reported
Titov et al.	BOCF	Not reported	Not reported
Vollestadet al.	LOCF, Completer's analysis	Not reported	Not reported
Wagner et al.	BOCF	t-tests of baseline differences (age effects) (no symptom effects)	Not reported
Westra et al.	Not reported	Chi square comparison at baseline - no patterns	Not reported
Wetherell et al.	N/A	N/A	Tested - No dependent variables departed from normality
Wuthrich & Rapee	Model based Multiple imputations (MAR)	Not reported	Not reported
Zou et al.	BOCF	N/A	Not reported

Statistical abbreviations: BOCF : baseline observation carried forward; LOCF : last observation carried forward; MAR : Missing at random; MCAR : Missing completely at random; N/A : Information not available

The summary of the various research practices from Table 2 shows that, 25 of 74 studies (33%) followed the specification of the CONSORT framework, with the number of studies following the framework increasing over time (CONSORT was launched formally in 2001). All 74 studies (100%) employed a standardized scale to assess their primary outcome, and all employed statistical analyses to evaluate treatment-related change. However, among the 74 studies, 122 different standardized outcome scales were employed as primary outcomes, together with 26 different types of analytical methods. Fifty-five of the studies (74%) reported Cohen's *d* effect sizes to convey the magnitude of clinical change, while others used partial eta square and percentage improvement (9/74; 12%). Forty-seven (63%) of the studies reported on the proportion of individuals who remained with clinical levels of symptoms following treatment, with 18/74 (24%) using RCI and 13/74 (18%) using percentage improvement to identify proportions of people making clinical improvements as a result of treatment. Similarly, missing data was handled in several ways, as 31% (23/74) employed imputation methods that assume data is missing entirely at random, 36% (27/74) employing last-observation or baseline-observation carried forward solutions (LOCF), and 33% (24/74) did not take any steps to address missing data in their analyses. Thus, the majority of studies failed to use appropriate methods for handling missing cases; that is, the selection of methods based on the careful evaluation of patterns with missing data (Bell et al., 2014; Little, 1995; Little et al., 2014).

Finally, it can also be noted that the studies surveyed in Table 2 tended to employ Cohen's d effect sizes, to categorise clinical events using RCI methodology and to use complete-case and LOCF methodologies for missing data. Taken together, while there is some variation in the research methods employed, there is a strong tendency towards the nomothetic approach and the routine adoption of some methods across studies for generating comparable metrics for clinical evidence.

Strengths and weaknesses with the current nomothetic research methods and clinical metrics

The use of nomothetic research methods and clinical metrics can, of course, offer clear advantages for psychotherapy research by allowing comparisons of treatment efficacy between studies

of different populations (Ellis, 2010; Laken, 2013; Kelley & Preacher, 2012). The Cohen's *d* effect size in particular allows a ready comparison of the efficacy of a range of therapies as diverse as cognitive behaviour therapy, brief psychodynamic psychotherapy, problem solving therapy, interpersonal psychotherapy and behavioural activation using random-effects meta-analysis, whereas it would be almost impossible to compare all of these therapies in a single study. However, it is possible that the ***current*** nomothetic approach to reporting data may be at the cost of internal validity of studies of psychotherapy, and hence the value of the clinical evidence generated by those studies.

Several recent studies have raised concerns about the current nomothetic methods of Cohen's *d* effect sizes, RCI and LOCF, including, Bower and colleagues (Bower, Kontopantelis, Sutton, Kendrick, Richards, Gilbody, , ... & Meyer, 2013) echoing earlier studies (Driessen, Cuijpers, Hollon & Dekker, 2010; Kroenke & Spitzer, 2002; Kroenke, Spitzer, Williams, Monahan, & Löwe, 2007) showing that Cohen's *d* effects are highly dependent on the severity of baseline symptoms, and samples with more severe baseline symptoms achieve larger Cohen's *d* effect sizes. Hence, the value of effect sizes for comparing different treatments is limited where samples differ in their composition and severity. Moreover, the statistical assumptions required for the Cohen's *d* effect sizes, including the need for symptom scores to be normally distributed and for change in symptom scores to be linear (Laken, 2013; Ng & Cribbie, 2017). In this way, such metrics may not be suitable where the symptom scales used are designed to produce bounded outcome scores, that is scores which are bounded at maximum and minimum values (Baldwin et al., 2016; Verkuilen & Smithson, 2012). Further, given that some symptom measures are designed in such a way that there is a limit in how high a person can score, called a *ceiling effect* as well as a limit in how low a person can score, called a *flooring effect*, symptom scales in the psychotherapy context may systematically produce bounded outcome scores. If there is a mismatch between the features of the data and the assumptions required, then the selection of metrics such as Cohen's *d* for the measurement and interpretation of evidence may result in misleading evidence around the relative efficacy of different treatments.

Similar to effect sizes, emerging methodological research also suggests that the commonly used methods to categorise clinical change and manage missing data may also be suboptimal. Hiller and

colleagues (2012) suggest that the use of linear cut-offs inherent in the RCI methodology to identify clinically meaningful change, such as five points of change on a depression scale, may incorrectly over-classify those individuals with high baseline symptoms as making clinical changes, whilst also incorrectly under-classifying individuals with less severe symptoms on baseline who have made clinical changes. Although little empirical data is available to corroborate this argument, the RCI approach for identifying clinical change is only statistically appropriate where the underlying function of change is linear and the distribution of scores is normal (Jacobson, Roberts, Berns, & McGlinchey, 1999; Jaboson & Traux, 1991). If symptom change is linear, then all patients undertaking therapy would be expected to change by the same amount (e.g., 5 points) whether their baseline symptoms were mild, moderate or severe. However, if symptom change is not linear or symptoms are not normally distributed, then the use of cutoff scores are likely to result in classification errors and erroneous conclusions.

Furthermore, with findings from recent meta-analytic studies suggesting that missing cases drop out of treatment in a systematic way, and that dropping out of treatment is associated with lower treatment adherence and more severe baseline symptoms (Fernandez, et al.,, 2015; Karyotaki, et al., … , & Cuijpers, 2015), some doubt is cast about the suitability of the predominant MCAR missing cases assumption and the LOCF approach for missing cases handling. Surprisingly, however, there is currently little empirical evidence regarding the suitability of dominant approaches for handling missing data in psychotherapy research. However, within the wider literature concerning missing cases it is widely understood that the strategies commonly used in the psychotherapy literature, in particular LOCF and MCAR , often overlook important features of missing data and introduce measurement bias and estimation error (Little 1995; Little et al., 2012 ;Little, 1995; Little, D'agostino, Cohen, Dickersin, Emerson, Farrar , ... , & Neaton, 2012; Sullivan, White, Salter, Ryan & Lee, 2018). This raises further grounds to question the suitability of dominant nomothetic methods clinical researchers are using to generate clinical evidence concerning psychotherapy, particularly some of the clinical metrics used, such as the use of Cohen's *d*, RCI, and the LOCF and MCAR approaches to handling missing data.

It should be recognised that doubts about the dominant methods used in psychotherapy research are not new. Several decades ago, Wilder (1965) and Micciri (1989) and Norman (Norman, Sloan &

Wyrwich, 2003) identified that the description of symptom change as a linear amount using Cohen's *d* effect sizes, could be replaced with methods that consider change as proportional and relative to baseline, using generalised linear methods. Similarly, Rubin (1976) and Little (1995) have demonstrated that simple missing data approaches, such as MCAR and LOCF, pose risks to the estimation of outcomes and the validity of conclusions in clinical trials. However, this methodological research is often published in specialist journals focused purely on statistical and methodological issues that are often too technical and overlooked by a clinician or practitioner audience (Harlow, 2017; Sharpe, 2013).

Together, the assumption of methodology researchers is that clinical researchers are aware of their work, understand the issues raised, and will carefully consider it in the context of their own fields when conducting research and generating clinical evidence is not supported in practice. In contrast, as reflected within several of the methodological and statistical guidelines, much of the methodological literature assumes that clinical researchers will choose between appropriate methods by identifying the features of their clinical data, such as statistical distributions and functions of change and choose methods that best fits the features of their data (Lang & Altman, 2013; Sharpe, 2013). This discrepancy between the predominant choices clinical researchers make in their research, and the recommendations of methodology guidelines make, can be seen to imply that methodologists take a much more idiographic, data-centric and context-specific view to the generation of clinical evidence, which is the exact opposite of the tendency to use more nomothetic approaches in psychotherapy research.

The need for valid and generalisable research methods and metrics

Given the complexity of selecting appropriate outcome measures, statistical models and metrics noted in previous sections, it is problematic to expect that clinical researchers could simply follow the current approach taken in the statistical guidelines, and make appropriate decisions in their efforts to generate valid and generalisable clinical evidence. This is evident from the studies summarised in Table 2, where nomothetic methods and metrics (emphasising generalisability) have been employed without exploring their underlying assumptions and ensuring their validity. Further, as shown in Table 2, only

9/74 (12%) of studies took the appropriate steps to screen for patterns in their missing data and only 15/74 (20%) attempted to check the statistical suitability of their analytic approach. This is concerning given much of the available psychotherapy evidence is based on several widely used nomothetic methods and metrics, such as Cohen's *d*, RCI, LOCF, which may or may not be appropriate. This problematic expectation is also reflected in numerous examples of clinical reviews, where clinical researchers do not check, or at least report to check, the suitability of their data for the statistical methods and clinical metrics they employ (Bell & Fairclough, 2014; Nieminen & Kaur, 2019; Nieminen, Virtanen & Vähänikkilä, 2017). This situation raises the need for accessible methodological research that bridges the gap between the statistical and methodological literature, and provides guidance to clinical researchers about the most suitable approaches for generating clinical evidence in the psychotherapy field.

The risks of using inappropriate research methods and metrics include generating clinical evidence that leads to incorrect conclusions about the efficacy of psychotherapy, the relative efficacy of different psychotherapies, and about who does and does not benefit from psychotherapy. These risks have been demonstrated in a recent meta-analysis concerning psychotherapy for depression, which reported substantially reduced effects sizes when more advanced effect size metrics, which account for sample size (i.e., Hedge's *g*) are applied to the estimation of clinical effects rather than the commonly used Cohen's *d* (Cuijpers, Karyotaki, Reijnders & Ebert, 2019; Cuijpers, et al., 2010). More broadly, a recent initiative to replicate 100 historically significant studies in psychology produced substantially different results in over half of the 100 studies (Open Science Collaboration, 2015). These findings and others have led to some researchers to claim a large proportion of evidence in psychology may be unreliable because of poor methodical and statistical decisions (Carpenter, 2012; Francis, 2012; Faulkner, Fidler & Cumming, 2008), and that the resultant research evidence may be potentially invalid (Ioannidis, 2005; 2012; Yarkoni & Westfall, 2017). While these claims are often disputed as overblown and overly critical (Pashler & Harris, 2012; Stroebe & Strack, 2014), questions about the internal validity of clinical evidence undermines the credibility of the evidence base for psychotherapy and psychology in general (Carpenter, 2012; Pashler & Wagenmakers, 2012; Sanders & Hunsley, 2018).

The Present Thesis

The limited research concerning the appropriateness of dominant measurement methods and metrics (Cohen's d, RCI, MCAR and LOCF) used in psychotherapy research reflects a critical ***research gap***. There is limited available research about: (1) the *appropriateness* of different measurement methods and metrics, and the validity of different types of practices for evaluating the efficacy of psychotherapy; (2) the impact of different *measurement methodologies and metrics* on of the validity and generalisability of psychotherapy evidence; or (3) the potential of alternative methods and metrics for achieving greater validity and generalisability. These areas can be addressed by targeted methodological research which explores whether the current measurement methods, metrics and practices in psychotherapy optimally capture the effects of treatment (i.e., measure with validity), and whether these methods optimally allow the comparison of psychotherapy effects across contexts (i.e., enable comparability and generalisability).

To start to address this critical research gap, the thesis employs a novel approach. It first focusses on exploring core statistical features of anxiety and depressive symptoms (as measured with standardised symptom scales) and how symptoms change over time in the context of psychotherapy. Then, these results are used to compare and critique the relative validity of different measurement methods and metrics that could be used as a shared metric for evaluating psychotherapy outcomes. The overarching aim of doing this was to start to optimise the validity and generalisability of measurement methods and metrics used within psychotherapy research by identifying methods and metrics that would best reflect the features of psychotherapy data.

This aim of maximising validity and generalisability has not yet been systematically pursued in the psychotherapy literature. Efforts to maximise both validity and generalisability are not common in either the broader statistical (Baldwin et al., 2016) or clinical literature (Lambert & Ogles, 2009). For example, selecting methods that do not fit the features of the data are well recognised within the statistical literature as representing a threat to the internal validity of evidence (Cohen, 2017; Cumming, 2013, Ng & Cribbie, 2013). For this reason, an approach that emphasized validity would prioritise the

selection of methods that fit the context, with little regard to the methods and metrics commonly used in the broader literature. Within the clinical literature, however, there is much less consideration for the idiosyncratic features of the data (Lambert & Ogles, 2009; Thompson, 2002; Sharpe, 2013). Instead much more emphasis is placed on the use of widely-used methods and metrics, which enable the comparison of outcomes across contexts (Clarke, 2007). Thus, clinical researchers often adopt the methodological approaches used by other researchers' in their field without necessarily considering the validity of those methods and metrics in their context.

To achieve the aims of this thesis, five studies were conducted. These studies were designed to address the main issues identified in the review of the literature, and in the review of studies reported in Table 2. Studies 1 and 2 explore the different statistical features of clinical change in symptoms of anxiety and depression as a result of psychotherapy. Studies 3 and 4 explore different approaches for the handling of missing cases in psychotherapy research for anxiety and depression. Study 5 explores methods for the identification and classification of clinical change in symptoms of anxiety and depression as a result of treatment. By engaging with the three areas of symptom measurement, missing cases and outcome classification, the current thesis aimed to provide data that would inform and guide the decisions of clinical researchers along several of the steps in the process of measuring clinical evidence for psychotherapy. It is hoped that the studies in this thesis will help improve the ability of clinical researchers to generate valid and reliable clinical evidence, and potentially new techniques that more suitably describe and model clinical change.

The thesis uses a large clinical dataset (n > 820) obtained from a series of randomised controlled trials conducted by a specialist research clinic that develops and evaluates new internet-delivered psychological treatments for anxiety and depression. An additional second dataset comprises a large sample treated as part of routine care (n = 6701) provided by a national digital mental health service, which delivers internet-delivered psychological treatments for large numbers of adults with anxiety and depression every year. The use of these datasets, being data from internet-delivered psychotherapy, was seen as advantageous given that this form of therapy is associated with high levels of control, standardisation and fidelity, reducing the measurement variance associated with traditional, therapist-

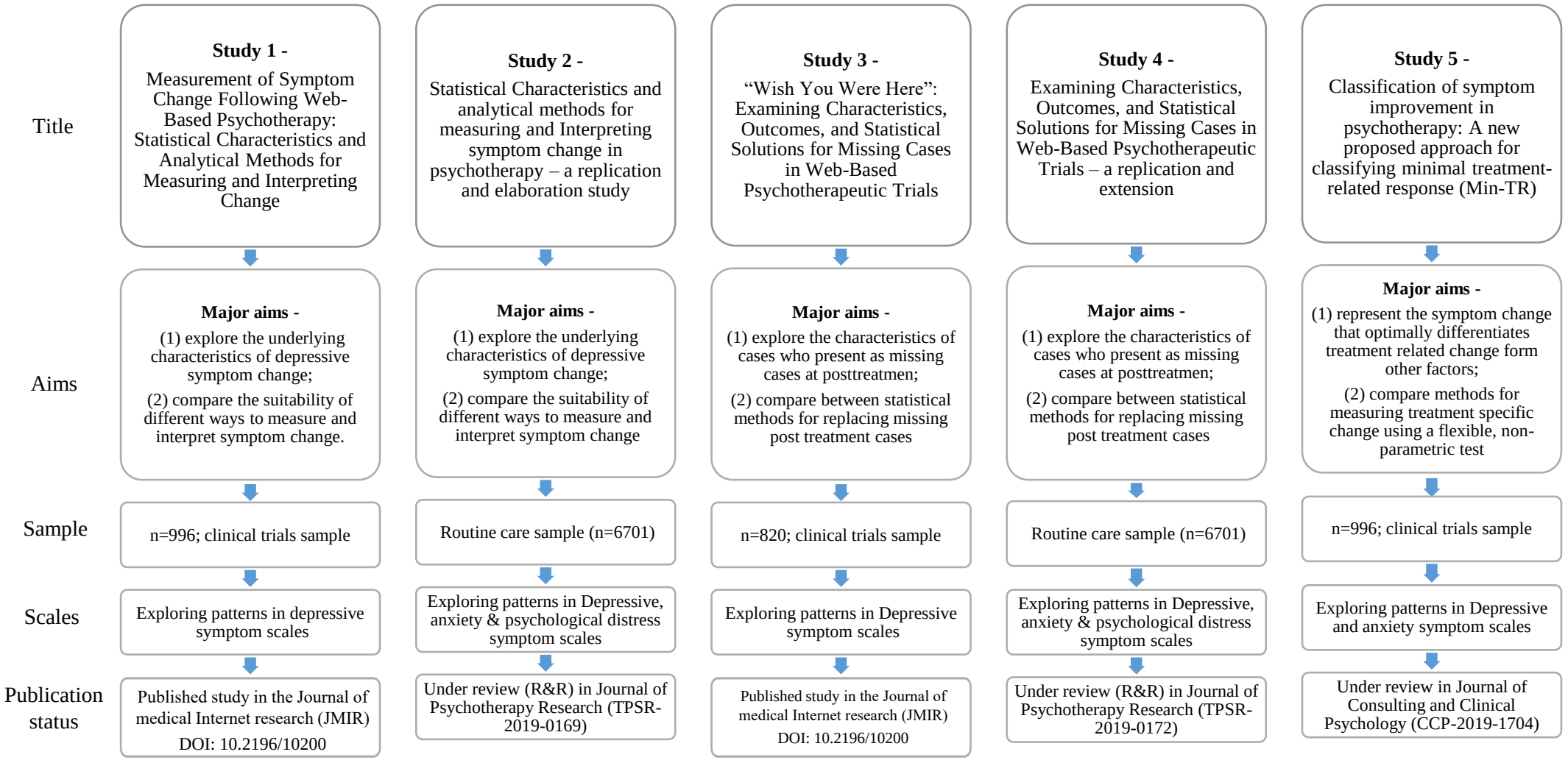
dependent, psychotherapies (Murphy & Hutton, 2018). For this reason, the use of data from internet-delivered psychotherapy provided a valuable opportunity to investigate statistical measurement principles and the impact of measurement methodology on treatment with less impact from other factors that may be present in traditional face-to-face psychotherapy.

Flow and sequence of studies

The major research components of this thesis will be structured in a way that follows the process of measurement researchers engage with when they operationalise statistical evidence about symptom change in psychotherapy; that is, the core methodological steps that are mandated under frameworks such as STORBE, CONSORT, TREND and JARs. Study 1 is a published pilot study, investigating the topic of measurement models for symptom change, with Study 2 aiming to replicate and elaborate the findings in a routine care context. Study 3 is a published pilot study, investigating the statistical suitability of approximating and replacing the outcomes of missing cases, with Study 4 aiming to replicate and elaborate the findings in a routine care context. Study 5 is a pilot study, investigating the classification of treatment outcomes categories of response and non-response. The research of this thesis aims to explore these topics through statistical and clinical viewpoints, but the research studies are primarily designed for clinical researchers and a clinical audience. Each topic is explored and discussed separately but integrated as a body of work in the General Discussion. The flow of studies is further detailed in the chart below (Figure 1) including the title and topic of each paper (tier 1- Top) each paper's aims, (tier 2) sample used (tier 3), symptom scales explored (tier 4), and current status of publication (tier 5)

.

Figure 1: Flow of thesis studies; including titles (tier 1- Top) aims, (tier 2) samples (tier 3), symptom scales explored (tier 4), and current status of each paper (tier 5)



References

Altman, D. (1994). The scandal of poor medical research. BMJ, 308(6924), 283-284.
Altman, D. G., & Simera, I. (2016). A history of the evolution of guidelines for reporting medical research: the long road to the EQUATOR Network. Journal of the Royal Society of Medicine, 109(2), 67-77.
American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders. BMC Med, 17, 133-137.
Andersson, G., & Titov, N. (2014). Advantages and limitations of Internet-based interventions for common mental disorders. World Psychiatry, 13(1), 4-11.
Andersson, G., Carlbring, P., Berger, T., Almlöv, J., & Cuijpers, P. (2009). What makes internet therapy work?. Cognitive behaviour therapy, 38(S1), 55-60.
Andersson, G., Cuijpers, P., Carlbring, P., Riper, H., & Hedman, E. (2014). Guided Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: a systematic review and meta-analysis. World Psychiatry, 13(3), 288-295.
Andersson, G., Cuijpers, P., Carlbring, P., Riper, H., & Hedman, E. (2014). Guided Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: a systematic review and meta-analysis. World Psychiatry, 13(3), 288-295.
Andersson, G., Titov, N., Dear, B. F., Rozental, A., & Carlbring, P. (2019). Internet-delivered psychological treatments: from innovation to implementation. World Psychiatry, 18(1), 20-28.
Andrews, G., & Williams, A. D. (2015). Up-scaling clinician assisted internet cognitive behavioural therapy (iCBT) for depression: A model for dissemination into primary care. Clinical psychology review, 41, 40-48.
Andrews, G., Basu, A., Cuijpers, P., Craske, M. G., McEvoy, P., 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, 70-78.
Andrews, G., Bell, C., Boyce, P., Gale, C., Lampe, L., Marwat, O., ... & Wilkins, G. (2018). Royal Australian and New Zealand College of Psychiatrists clinical practice guidelines for the treatment of panic disorder, social anxiety disorder and generalised anxiety disorder. Australian & New Zealand Journal of Psychiatry, 52(12), 1109-1172.
Angst, F. (2011). The new COSMIN guidelines confront traditional concepts of responsiveness. BMC medical research methodology, 11(1), 152.
APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. The American Psychologist, 61(4), 271.
Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. American Psychologist, 73(1), 3.
Atkins, D., Fink, K., & Slutsky, J. (2005). Better information for better health care: the Evidence-based Practice Center program and the Agency for Healthcare Research and Quality. Annals of Internal Medicine, 142(12_Part_2), 1035-1041.
Baird, G. L., & Harlow, L. L. (2016). Does One Size Fit All? A Case for Context-Driven Null Hypothesis Statistical Testing. Journal of Modern Applied Statistical Methods, 15(1), 7.
Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. Behavior research methods, 43(3), 666-678.
Baldwin, S. A., Fellingham, G. W., & Baldwin, A. S. (2016). Statistical models for multilevel skewed physical activity data in health research and behavioral medicine. Health Psychology, 35(6), 552.
Barkowski, S., Schwartze, D., Strauss, B., Burlingame, G. M., Barth, J., & Rosendahl, J. (2016). Efficacy of group psychotherapy for social anxiety disorder: A meta-analysis of randomized-controlled trials. Journal of anxiety disorders, 39, 44-64.
Bauer, S., Lambert, M. J., & Nielsen, S. L. (2004). Clinical significance methods: A comparison of statistical techniques. Journal of Personality Assessment, 82(1), 60-70.
Beckstead, J. W. (2013). On measurements and their quality: Paper 2: Random measurement error and the power of statistical tests. International journal of nursing studies, 50(10), 1416-1422.
Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., ... & Stroup, D. F. (1996). Improving the quality of reporting of randomized controlled trials: the CONSORT statement. Jama, 276(8), 637-639.
Bell, M. L., & Fairclough, D. L. (2014). Practical and statistical issues in missing data for longitudinal patient-reported outcomes. Statistical methods in medical research, 23(5), 440-459.
Bell, M. L., Fiero, M., Horton, N. J., & Hsu, C. H. (2014). Handling missing data in RCTs; a review of the top medical journals. BMC medical research methodology, 14(1), 118.
Bell, M. L., Olivier, J., & King, M. T. (2013). Scientific rigour in psycho-oncology trials: why and how to avoid common statistical errors. Psycho-Oncology, 22(3), 499-505.

Boman, M., Abdesslem, F. B., Forsell, E., Gillblad, D., Görnerup, O., Isacsson, N., ... & Kaldo, V. (2019). Learning machines in Internet-delivered psychological treatment. Progress in Artificial Intelligence, 1-11.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., ... & Kressel, H. Y. (2015). STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Radiology, 277(3), 826-832.

Boswell, J. F., Kraus, D. R., Miller, S. D., & Lambert, M. J. (2015). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. Psychotherapy research, 25(1), 6-19.

Bower, P., & Gilbody, S. (2005). Stepped care in psychological therapies: access, effectiveness and efficiency: narrative literature review. The British Journal of Psychiatry, 186(1), 11-17.

Bower, P., Kontopantelis, E., Sutton, A., Kendrick, T., Richards, D. A., Gilbody, S., ... & Meyer, B. (2013). Influence of initial severity of depression on effectiveness of low intensity interventions: meta-analysis of individual patient data. Bmj, 346, f540.

Braakmann, D. (2015). Historical paths in psychotherapy research. In Psychotherapy Research (pp. 39-65). Springer, Vienna.

Budd, R., & Hughes, I. (2009). The Dodo Bird Verdict—controversial, inevitable and important: a commentary on 30 years of meta-analyses. Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice, 16(6), 510-522.

Burgess, P. M., Pirkis, J. E., Slade, T. N., Johnston, A. K., Meadows, G. N., & Gunn, J. M. (2009). Service use for mental health problems: findings from the 2007 National Survey of Mental Health and Wellbeing. Australian and New Zealand Journal of Psychiatry, 43(7), 615-623.

Byatt, N., Levin, L. L., Ziedonis, D., Simas, T. A. M., & Allison, J. (2015). Enhancing participation in depression care in outpatient perinatal care settings: a systematic review. Obstetrics and gynecology, 126(5), 1048.

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.

Carpenter, S. (2012). Psychology's bold initiative. Science, 335, 1558–1560.

Castelvecchi, D. (2016). Can we open the black box of AI?. Nature News, 538(7623), 20.

Chalmers, I., & Altman, D. G. (1999). How can medical journals help prevent poor medical research? Some opportunities presented by electronic publishing. The Lancet, 353(9151), 490-493.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. Journal of consulting and clinical psychology, 66(1), 7.

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., ... & Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: a machine learning approach. The Lancet Psychiatry, 3(3), 243-250.

Choi, S. W., Schalet, B., Cook, K. F., & Cella, D. (2014). Establishing a common metric for depressive symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS Depression. Psychological assessment, 26(2), 513.

Ciobanu, L. G., Ferrari, A. J., Erskine, H. E., Santomauro, D. F., Charlson, F. J., Leung, J., ... & Baune, B. T. (2018). The prevalence and burden of mental and substance use disorders in Australia: Findings from the Global Burden of Disease Study 2015. Australian & New Zealand Journal of Psychiatry, 52(5), 483-490.

Clark, D. M., Layard, R., Smithies, R., Richards, D. A., Suckling, R., & Wright, B. (2009). Improving access to psychological therapy: Initial evaluation of two UK demonstration sites. Behaviour research and therapy, 47(11), 910-920.

Clarke, M. (2007). Standardising outcomes for clinical trials and systematic reviews. Trials, 8(1), 39.

Cuijpers P, Cristea IA, Karyotaki E, Reijnders M, Huibers MJ. How effective are cognitive behavior therapies for major depression and anxiety disorders? a meta-analytic update of the evidence. World Psychiatry. 2016;15(3):245-258.

Cuijpers, P. (1998). Minimising interventions in the treatment and prevention of depression. Taking the consequences of the'Dodo Bird Verdict'. Journal of Mental Health, 7(4), 355-365.

Cuijpers, P., & Smit, F. (2002). Excess mortality in depression: a meta-analysis of community studies. Journal of affective disorders, 72(3), 227-236.

Cuijpers, P., Karyotaki, E., Reijnders, M., & Ebert, D. D. (2019). Is psychotherapy effective? Pretending everything is fine will not help the field forward. Epidemiology and Psychiatric Sciences, 1-2.

Cuijpers, P., Sijbrandij, M., Koole, S. L., Andersson, G., Beekman, A. T., & Reynolds III, C. F. (2014). Adding psychotherapy to antidepressant medication in depression and anxiety disorders: a meta-analysis. Focus, 12(3), 347-358.

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(02), 211-223.

Cumming, G. (2014). The new statistics: Why and how. Psychological science, 25(1), 7-29.

Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the American psychological association publication manual: Effect sizes, confidence intervals, and meta-analysis. Australian Journal of Psychology, 64(3), 138-146.

Davey, C. G., & Chanen, A. M. (2016). The unfulfilled promise of the antidepressant medications. Medical Journal of Australia, 204(9), 348-350.

David, D., Cristea, I., & Hofmann, S. G. (2018). Why cognitive behavioral therapy is the current gold standard of psychotherapy. Frontiers in psychiatry, 9, 4.

Delgadillo, J., McMillan, D., Leach, C., Lucock, M., Gilbody, S., & Wood, N. (2014). Benchmarking routine psychological services: a discussion of challenges and methods. Behavioural and cognitive psychotherapy, 42(01), 16-30.

Des Jarlais, D. C., Lyles, C., Crepaz, N., & Trend Group. (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. American journal of public health, 94(3), 361-366.

Diaz, R. (2017). Immigration and Depression in Canada: Is there really a Healthy Immigrant Effect? What is the Pattern of Depression by Time since Immigration? (Doctoral dissertation, University of Calgary).

Djulbegovic, B., & Guyatt, G. H. (2017). Progress in evidence-based medicine: a quarter century on. The Lancet, 390(10092), 415-423.

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.

Driessen, E., Cuijpers, P., Hollon, S. D., & Dekker, J. J. (2010). Does pretreatment severity moderate the efficacy of psychological treatment of adult outpatient depression? A meta-analysis.

Ebrahim, S., Sohani, Z. N., Montoya, L., Agarwal, A., Thorlund, K., Mills, E. J., & Ioannidis, J. P. (2014). Reanalyses of randomized clinical trial data. Jama, 312(10), 1024-1032.

Ellis, P. D. (2010). The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results. Cambridge University Press.

Eysenbach, G., & Consort-EHEALTH Group. (2011). CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. Journal of medical Internet research, 13(4), e126.

Faulkner, C., Fidler, F., Cumming, G. (2008). The value of RCT evidence depends on the quality of statistical analysis. Behaviour Research and Therapy, 46, 270–281.

Faulkner, C., Fidler, F., Cumming, G. (2008). The value of RCT evidence depends on the quality of statistical analysis. Behaviour Research and Therapy, 46, 270–281. doi:10.1016/j.brat.2007.12.001

Fernandez, E., Salem, D., Swift, J. K., & Ramtahal, N. (2015). Meta-analysis of dropout from cognitive behavioral therapy: Magnitude, timing, and moderators. Journal of Consulting and Clinical Psychology, 83(6), 1108.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., ... & Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. Prevention science, 6(3), 151-175.

Francis, G. (2012). Replication initiative: Beware misinterpretation. Science, 336(6083), 802-802.

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

Genders, T. S., Spronk, S., Stijnen, T., Steyerberg, E. W., Lesaffre, E., & Hunink, M. M. (2012). Methods for calculating sensitivity and specificity of clustered data: a tutorial. Radiology, 265(3), 910-916.

Gilbody, S., Littlewood, E., Hewitt, C., Brierley, G., Tharmanathan, P., Araya, R., ... & Kessler, D. (2015). Computerised cognitive behaviour therapy (cCBT) as treatment for depression in primary care (REEACT trial): large scale pragmatic randomised controlled trial. Bmj, 351, h5627.

Glasgow, R. E., Green, L. W., & Ammerman, A. (2007). A focus on external validity. Evaluation & the Health Professions, 30(2), 115-117.

Glasziou, P., Altman, D. G., Bossuyt, P., Boutron, I., Clarke, M., Julious, S., ... & Wager, E. (2014). Reducing waste from incomplete or unusable reports of biomedical research. The Lancet, 383(9913), 267-276.

Gottfredson, D. C., Cook, T. D., Gardner, F. E., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. Prevention Science, 16(7), 893-926.

Greenhalgh, T., Howick, J., & Maskrey, N. (2014). Evidence based medicine: a movement in crisis?. Bmj, 348, g3725.

Halekoh, U., Højsgaard, S., & Yan, J. (2006). The R package geepack for generalized estimating equations. Journal of Statistical Software, 15(2), 1-11.

Han, C., Kwak, K. P., Marks, D. M., Pae, C. U., Wu, L. T., Bhatia, K. S., ... & Patkar, A. A. (2009). The impact of the CONSORT statement on reporting of randomized clinical trials in psychiatry. Contemporary clinical trials, 30(2), 116-122.

Harlow, L. L. (2017, March). The making of Psychological Methods. In European Congress of Methodology, VII, Jul, 2016, Palma de Mallorca, Spain; Portions of an earlier version of this paper were presented in Harlow, LL,“20

Years of Psychological Methods" at the aforementioned congress. (Vol. 22, No. 1, p. 1). American Psychological Association.
Harris, A., Reeder, R., & Hyun, J. (2011). Survey of editors and reviewers of high-impact psychology journals: statistical and research design problems in submitted manuscripts. The Journal of psychology, 145(3), 195-209.
Hartmann, A., van der Kooij, A. J., and Zeeck, A. (2009). Exploring nonlinear relations: models of clinical decision making by regression with optimal scaling. Psychother. Res. 19, 482–492.
Haynes, R. B. (2012). Clinical epidemiology: how to do clinical practice research. Lippincott williams & wilkins.
Heisel, M. J., & Flett, G. L. (2016). Investigating the psychometric properties of the Geriatric Suicide Ideation Scale (GSIS) among community-residing older adults. Aging & mental health, 20(2), 208-221.
Hiller, W., Schindler, A. C., & Lambert, M. J. (2012). Defining response and remission in psychotherapy research: A comparison of the RCI and the method of percent improvement. Psychotherapy Research, 22(1), 1-11.
Hobbs, M. J., Mahoney, A. E., & Andrews, G. (2017). Integrating iCBT for generalized anxiety disorder into routine clinical care: Treatment effects across the adult lifespan. Journal of anxiety disorders, 51, 47-54.
Hofmann SG, Asnaani A, Vonk IJJ, Sawyer AT, Fang A. The efficacy of cognitive behavioral therapy: a review of meta-analyses. Cognit Ther Res. 2012;36(5):427-440.
Hollon, D., Miller, I. J., & Robinson, E. (2002). Criteria for evaluating treatment guidelines. American Psychologist, 57(12), 1052-1059.
Hopkins, W., Marshall, S., Batterham, A., & Hanin, J. (2009). Progressive statistics for studies in sports medicine and exercise science. Medicine+ Science in Sports+ Exercise, 41(1), 3.
Horner, Robert H., Edward G. Carr, James Halle, Gail McGee, Samuel Odom, and Mark Wolery. "The use of single-subject research to identify evidence-based practice in special education." Exceptional children 71, no. 2 (2005): 165-179.
Horvath, A. O. (2013). You can't step into the same river twice, but you can stub your toes on the same rock: Psychotherapy outcome from a 50-year perspective.
Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness, and patient progress. American psychologist, 51(10), 1059.
Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Lippman, S. A., Jewell, N., ... & Satariano, W. A. (2010). To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. Epidemiology, 21(4), 467-474.
Hundt, N. E., Mignogna, J., Underhill, C., & Cully, J. A. (2013). The relationship between use of CBT skills and depression treatment outcome: A theoretical and methodological review of the literature. Behavior therapy, 44(1), 12-26.
Ioannidis, J. P. (2005). Why most published research findings are false. PLos med, 2(8), e124.
Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. Perspectives on Psychological Science, 7, 645–654
Jacobson, N. S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. Journal of consulting and clinical psychology, 59(1), 12.
Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: description, application, and alternatives. Journal of consulting and clinical psychology, 67(3), 300.
Johnston, B. C., Alonso-Coello, P., Friedrich, J. O., Mustafa, R. A., Tikkinen, K. A., Neumann, I., ... & Dalmau, G. M. (2016). Do clinicians understand the size of treatment effects? A randomized survey across 8 countries. Canadian Medical Association Journal, 188(1), 25-32.
Jorm, AF (2018) Australia's 'Better Access' Scheme: Has it had an impact on population mental health? Australian and New Zealand Journal of Psychiatry 52: 1057–1062.
Karyotaki, E., Riper, H., Twisk, J., Hoogendoorn, A., Kleiboer, A., Mira, A., ... & Andersson, G. (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.
Kazdin, A. E. (1999). The meanings and measurement of clinical significance. Journal of Consulting and Clinical Psychology, 67(3), 332-339.
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.
Keller, M. B. (2003). Past, present, and future directions for defining optimal treatment outcome in depression: remission and beyond. Jama, 289(23), 3152-3160.
Kelley, K., & Preacher, K. J. (2012). On effect size. Psychological Methods, 17(2), 137-152.
Kessler, 2007; Clinical evidence—key definitions and concepts. May 2007. SG5/N1R8.
Kessler, R. C., Aguilar-Gaxiola, S., Alonso, J., Chatterji, S., Lee, S., Ormel, J., ... & Wang, P. S. (2009). The global burden of mental disorders: an update from the WHO World Mental Health (WMH) surveys. Epidemiologia e psichiatria sociale, 18(01), 23-33.

Kessler, R. C., Heeringa, S., Lakoma, M. D., Petukhova, M., Rupp, A. E., Schoenbaum, M., ... & Zaslavsky, A. M. (2008). Individual and societal effects of mental disorders on earnings in the United States: results from the national comorbidity survey replication. American Journal of Psychiatry, 165(6), 703-711.

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Cai, T., ... & Nierenberg, A. A. (2016). Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. Molecular psychiatry, 21(10), 1366.

Khorsan, R., & Crawford, C. (2014). External validity and model validity: a conceptual approach for systematic review methodology. Evidence-Based Complementary and Alternative Medicine, 2014.

King, M. T. (2011). A point of minimal important difference (MID): a critique of terminology and methods. Expert review of pharmacoeconomics & outcomes research, 11(2), 171-184.

Kinzel, K. (2017). Wilhelm Windelband and the problem of relativism. British Journal for the History of Philosophy, 25(1), 84-107.

Koutsouleris, N., Kambeitz-Ilankovic, L., Ruhrmann, S., Rosen, M., Ruef, A., Dwyer, D. B., ... & Schmidt, A. (2018). Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression: a multimodal, multisite machine learning analysis. JAMA psychiatry, 75(11), 1156-1172.

Kraemer, H. C., Noda, A., & O'Hara, R. (2004). Categorical versus dimensional approaches to diagnosis: methodological challenges. Journal of Psychiatric Research, 38(1), 17-25.

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: a new depression diagnostic and severity measure. Psychiatry Ann, 32(9), 1-7.

Kroenke, K., Monahan, P. O., & Kean, J. (2015). Pragmatic characteristics of patient-reported outcome measures are important for use in clinical practice. Journal of clinical epidemiology, 68(9), 1085-1092.

Kroenke, K., Spitzer, R. L., Williams J. B. W, Monahan, P. O., & Lowe B. (2007). Anxiety Disorders in Primary Care: Prevalence, Impairment, Comorbidity, and Detection, Annuals of Internal Medicine, Vol. 146, No. 5, pp. 317-325.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Frontiers in psychology, 4, 863.

Lambert, M. (2007). Presidential address: What we have learned from a decade of research aimed at improving psychotherapy outcome in routine care. Psychotherapy research, 17(1), 1-14.

Lambert, M. J. (2015). Outcome Research: methods for improving outcome in routine care. In Psychotherapy Research (pp. 593-610). Springer, Vienna.

Lambert, M. J., & Ogles, B. M. (2009). Using clinical significance in psychotherapy outcome research: The need for a common procedure and validity data. Psychotherapy Research, 19(4-5), 493-501.

Lang, T. A., & Altman, D. G. (2013). Basic statistical reporting for articles published in biomedical journals: the "Statistical Analyses and Methods in the Published Literature" or the SAMPL Guidelines". Handbook, European Association of Science Editors, 23-26

Levine, F. M., Sandeen, E., & Murphy, C. M. (1992). The therapist's dilemma: Using nomothetic information to answer idiographic questions. Psychotherapy: Theory, Research, Practice, Training, 29(3), 410.

Li, P., Stuart, E. A., & Allison, D. B. (2015). Multiple imputation: a flexible tool for handling missing data. Jama, 314(18), 1966-1967.

Liang, K. Y., Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika 73:13–22.

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. Journal of the American Statistical Association, 90(431), 1112-1121.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., ... & Neaton, J. D. (2012). The prevention and treatment of missing data in clinical trials. New England Journal of Medicine, 367(14), 1355-1360.

Little, T. D., Deboeck, P., & Wu, W. (2015). Longitudinal data analysis. Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource, 1-17.

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2014). On the joys of missing data. Journal of pediatric psychology, 39(2), 151-162.

Luborsky, L., Rosenthal, R., Diguer, L., Andrusyna, T. P., Berman, J. S., Levitt, J. T., ... & Krause, E. D. (2002). The dodo bird verdict is alive and well—mostly. Clinical Psychology: Science and Practice, 9(1), 2-12.

Luna, R. A., & Foster, J. A. (2015). Gut brain axis: diet microbiota interactions and implications for modulation of anxiety and depression. Current opinion in biotechnology, 32, 35-41.

Mahoney, M. (Ed.). (2012). Psychotherapy process: Current issues and future directions. Springer Science & Business Media.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: the case of r and d. Psychological methods, 11(4), 386.

McHugh, R. K., Whitton, S. W., Peckham, A. D., Welge, J. A., & Otto, M. W. (2013). Patient preference for psychological vs. pharmacological treatment of psychiatric disorders: a meta-analytic review. The Journal of clinical psychiatry, 74(6), 595.

McLean, R. A., Sanders, W. L., & Stroup, W. W. (1991). A unified approach to mixed linear models. The American Statistician, 45(1), 54-64.

McMillan, D., Gilbody, S., & Richards, D. (2010). Defining successful treatment outcome in depression using the PHQ-9: a comparison of methods. Journal of affective disorders, 127(1), 122-129.

Meurk, C., Leung, J., Hall, W., Head, B. W., & Whiteford, H. (2016). Establishing and governing e-mental health care in Australia: a systematic review of challenges and a call for policy-focussed research. Journal of Medical Internet Research, 18(1), e10.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological bulletin, 105(1), 156.

Mitchell, A. J., Vaze, A., & Rao, S. (2009). Clinical diagnosis of depression in primary care: a meta-analysis. The Lancet, 374(9690), 609-619.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Annals of internal medicine, 151(4), 264-269.

Moher, D., Schulz, K. F., Altman, D. G., & Consort Group. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials.

Mojtabai, R., Olfson, M., Sampson, N. A., Jin, R., Druss, B., Wang, P. S., ... & Kessler, R. C. (2011). Barriers to mental health treatment: results from the National Comorbidity Survey Replication. Psychological medicine, 41(8), 1751-1761.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., ... & De Vet, H. C. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research, 19(4), 539-549.

Murphy, R., & Hutton, P. (2018). Practitioner Review: Therapist variability, patient-reported therapeutic alliance, and clinical outcomes in adolescents undergoing mental health treatment–a systematic review and meta-analysis. Journal of Child Psychology and Psychiatry, 59(1), 5-19.

Musliner, K. L., Munk-Olsen, T., Laursen, T. M., Eaton, W. W., Zandi, P. P., & Mortensen, P. B. (2016). Heterogeneity in 10-year course trajectories of moderate to severe major depressive disorder: a danish national register-based study. JAMA psychiatry, 73(4), 346-353.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in Ecology and Evolution, 4(2), 133-142.

Newby, J. M., McKinnon, A., Kuyken, W., Gilbody, S., & Dalgleish, T. (2015). Systematic review and meta-analysis of transdiagnostic psychological treatments for anxiety and depressive disorders in adulthood. Clinical psychology review, 40, 91-110.

Ng, V. K., & Cribbie, R. A. (2017). Using the Gamma Generalized Linear Model for modeling continuous, skewed and heteroscedastic outcomes in psychology. Current Psychology, 36(2), 225-235.

Nierenberg, A. A., Petersen, T. J., & Alpert, J. E. (2003). Prevention of relapse and recurrence in depression: the role of long-term pharmacotherapy and psychotherapy. Journal of Clinical Psychiatry, 64(15), 13-17.

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. Medical Care, 41, 582–592.

Ogles, B. M., Lunnen, K. M., & Bonesteel, K.(2001). Clinical significance: History, application, and current practice. Clinical Psychology Review, 21, 421–446.

Olfson, M., & Marcus, S. C. (2009). National patterns in antidepressant medication treatment. Archives of general psychiatry, 66(8), 848-856.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. Perspectives on Psychological Science, 7(6), 531-536.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?. Perspectives on Psychological Science, 7(6), 528-530.

Pirkis, J, Burgess, P, Coombs, T. (2005) Routine measurement of outcomes in Australia's public sector mental health services. Australia and New Zealand Health Policy 2: 8.

Plint, A. C., Moher, D., Morrison, A., Schulz, K., Altman, D. G., Hill, C., & Gaboury, I. (2006). Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. Medical Journal of Australia, 185(5), 263.

Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. Annual Review of Psychology, 66, 825-852.

Reeve, B. B., Thissen, D., DeWalt, D. A., Huang, I. C., Liu, Y., Magnus, B., ... & Haley, S. (2016). Linkage between the PROMIS® pediatric and adult emotional distress measures. Quality of Life Research, 25(4), 823-833.

Robinson, O. C. (2011). The idiographic/nomothetic dichotomy: Tracing historical origins of contemporary confusions. History & Philosophy of Psychology, 13(2), 32-39.
Ronk, F. R., Hooke, G. R., & Page, A. C.(2012). How consistent are clinical significance classifications when calculation methods and outcome measures differ? Clinical Psychology, Science and Practice, 19, 167–179.
Rosser, R.M. "A History of the Development of Health Indices. " in Teeling-Smith, G. (ed.). Measuring the Social Benejits of Medicine, Office of Health Economics, London, 1988.
Roth, A., & Fonagy, P. (2013). What works for whom?: a critical review of psychotherapy research. Guilford Publications.
Rothwell, P. M. (2005). External validity of randomised controlled trials:"to whom do the results of this trial apply?". The Lancet, 365(9453), 82-93.
Rozental, A., Andersson, G., Boettcher, J., Ebert, D. D., Cuijpers, P., Knaevelsrud, C., ... & Carlbring, P. (2014). Consensus statement on defining and measuring negative effects of Internet interventions. Internet Interventions, 1(1), 12-19.
Rubin, D. B. (1976), "Inference and Missing Data," Biometrika, 63, 581-592.
Sackett, D. L. (2002). Clinical epidemiology: what, who, and whither. Journal of Clinical Epidemiology, 55(12), 1161-1166.
Sackett, D. L., Rosenberg, W. M., Gray, J. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't.
Sanders, S. G., & Hunsley, J. (2018). The new Caucus-race: Methodological considerations for meta-analyses of psychotherapy outcome. Canadian Psychology/psychologie canadienne, 59(4), 387.
Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. Psychological methods, 7(2), 147. http://dx.doi.org/10.1037/1082-989x.7.2.147
Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMC medicine, 8(1), 18.
Scull, A. (2015). Madness in civilization: A cultural history of insanity, from the Bible to Freud, from the madhouse to modern medicine. Princeton University Press.
Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. Psychological Methods, 18(4), 572-582.
Slade, T., Johnston, A., Oakley Browne, M. A., Andrews, G., & Whiteford, H. (2009). 2007 National Survey of Mental Health and Wellbeing: methods and key findings. Australian and New Zealand Journal of Psychiatry, 43(7), 594-605.
Smith, B. A., Lee, H. J., Lee, J. H., Choi, M., Jones, D. E., Bausell, R. B., & Broome, M. E. (2008). Quality of reporting randomized controlled trials (RCTs) in the nursing literature: application of the consolidated standards of reporting trials (CONSORT). Nursing outlook, 56(1), 31-37.
Smithson, M. & Shou, Y. (2016). Moderator Effects Differ on Alternative Effect-Size Measures. Behavioral Research Methods, doi: 10.3758/s13428-016-0735-z
Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychological methods, 11(1), 54.
Sobocki , P., Ekman, M., Ågren, H., Runeson, B., & Jönsson, B. (2006). The mission is remission: health economic consequences of achieving full remission with antidepressant treatment for depression. International journal of clinical practice, 60(7), 791-798.
Spring, B. (2007). Evidence-based practice in clinical psychology: what it is, why it matters; what you need to know. Journal of clinical psychology, 63(7), 611-631.
Staples, L. G., Dear, B. F., Johnson, B., Fogliati, V., Gandy, M., Fogliati, R., ... & Titov, N. (2019). Internet-delivered treatment for young adults with anxiety and depression: evaluation in routine clinical care and comparison with research trial outcomes. Journal of affective disorders.
Staples, L. G., Fogliati, V. J., Dear, B. F., Nielssen, O., & Titov, N. (2016). Internet-delivered treatment for older adults with anxiety and depression: implementation of the Wellbeing Plus Course in routine clinical care and comparison with research trial outcomes. BJPsych open, 2(5), 307-313.
Stinson, J. N., McGrath, P. J., & Yamada, J. T. (2003). Clinical trials in the Journal of Pediatric Psychology: Applying the CONSORT statement. Journal of Pediatric Psychology, 28(3), 159-167.
Streiner, D. L. (2008). Missing data and the trouble with LOCF. Evidence-based mental health, 11(1), 3.
Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. Perspectives on Psychological Science, 9(1), 59-71.
Sullivan, T. R., White, I. R., Salter, A. B., Ryan, P., & Lee, K. J. (2018). Should multiple imputation be the method of choice for handling missing data in randomized trials?. Statistical methods in medical research, 27(9), 2610-2626.
Tansella, M. (2002). The scientific evaluation of mental health treatments: an historical perspective. Evidence-based mental health, 5(1), 4-5.

Teesson, M., Mitchell, P. B., Deady, M., Memedovic, S., Slade, T., & Baillie, A. (2011). Affective and anxiety disorders and their relationship with chronic physical conditions in Australia: findings of the 2007 National Survey of Mental Health and Wellbeing. Australian & New Zealand Journal of Psychiatry, 45(11), 939-946.

Thabane, Lehana, Lawrence Mbuagbaw, Shiyuan Zhang, Zainab Samaan, Maura Marcucci, Chenglin Ye, Marroon Thabane et al. "A tutorial on sensitivity analyses in clinical trials: the what, why, when and how." BMC medical research methodology 13, no. 1 (2013): 92.

Thomas, M. L. (2019). Advances in applications of item response theory to clinical assessment. Psychological assessment.

Thomas, S. J., Leeson, P. R., Larkin, T., Deng, C., Pai, B. N., Mills, J., & McLennan, P. (2016). Modelling complex relationships between physiological and psychosocial factors in depression. International Journal of Psychophysiology, 100(108), 107.

Titov, N., Dear, B. F., Staples, L. G., Bennett-Levy, J., Klein, B., Rapee, R. M., ... & Nielssen, O. B. (2017). The first 30 months of the MindSpot Clinic: Evaluation of a national e-mental health service against project objectives. Australian & New Zealand Journal of Psychiatry, 51(12), 1227-1239.

Titov, N., Dear, B., Nielssen, O., Staples, L., Hadjistavropoulos, H., Nugent, M., ... & Repål, A. (2018). ICBT in routine care: a descriptive analysis of successful clinics in five countries. Internet interventions, 13, 108-115.

Turner, L., Shamseer, L., Altman, D. G., Weeks, L., Peters, J., Kober, T., ... & Moher, D. (2012). Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. The Cochrane Library.

Verkuilen, J. and Smithson, M. (2012). Mixed and mixture regression models for continuous bounded responses using the beta distribution. Journal of Educational and Behavioral Statistics, 37, 82-113.

Vickers, A. J. (2005). Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. BMC medical research methodology, 5(1), 35. doi:10.1186/1471-2288-5-35

Von Elm, E., & Egger, M. (2004). The scandal of poor epidemiological research.

Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gotzsche, P. C., & Vandenbroucke, J. P. (2007). Policy and practice-The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. Bulletin of the World Health Organization, 85(11), 867-872.

von Wolff, A., Jansen, M., Hölzel, L. P., Westphal, A., Härter, M., & Kriston, L. (2014). Generalizability of findings from efficacy trials for chronic depression: An analysis of eligibility criteria. Psychiatric services, 65(7), 897-904.

Wells, K. B. (1999). Treatment research at the crossroads: the scientific interface of clinical trials and effectiveness research. American Journal of Psychiatry, 156(1), 5-10.

Whiteford H, Harris M and Diminic S (2013) Mental health service system improvement: Translating evidence into policy. Australian and New Zealand Journal of Psychiatry 47: 703–706.

Whiteford, H. (2019). We have increased access to mental health treatment, but it needs to be effective treatment. Australian & New Zealand Journal of Psychiatry, 53(3), 257-258.

Whiteford, H. A., Degenhardt, L., Rehm, J., Baxter, A. J., Ferrari, A. J., Erskine, H. E., ... & Burstein, R. (2013). Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. The Lancet, 382(9904), 1575-1586.

Wilder, J. (1965). Pitfalls in the methodology of the Law of Initial Value. American Journal of Psychotherapy, 19, 577–584.

Woolley, S. B., Cardoni, A. A., & Goethe, J. W. (2009). Last-Observation-Carried-Forward Imputation Method in Clinical Efficacy Trials : Review of 352 Antidepressant Studies. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy, 29(12), 1408-1416

Wyrwich, K. W., Norquist, J. M., Lenderking, W. R., Acaster, S., & Industry Advisory Committee of International Society for Quality of Life Research (ISOQOL. (2013). Methods for interpreting change over time in patient-reported outcome measures. Quality of Life Research, 22(3), 475-483.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. Perspectives on Psychological Science, 12(6), 1100-1122.

Yi, D., Ma, D., Li, G., Zhou, L., Xiao, Q., Zhang, Y., ... & Liu, L. (2015). Statistical use in clinical studies: is there evidence of a methodological shift?. PloS one, 10(10), e0140159.

Yoon, U., & Knobloch, K. (2012). Quality of reporting in sports injury prevention abstracts according to the CONSORT and STROBE criteria: an analysis of the World Congress of Sports Injury Prevention in 2005 and 2008. British journal of sports medicine, 46(3), 202-206.

Zeig, J. K., & Munion, W. (1990). What is psychotherapy?: Contemporary perspectives. Jossey-Bass.

Zhang, Z., Zhang, L., Zhang, G., Jin, J., & Zheng, Z. (2018). The effect of CBT and its modifications for relapse prevention in major depressive disorder: a systematic review and meta-analysis. BMC psychiatry, 18(1), 50.

Zimmerman, M., Posternak, M. A., & Chelminski, I. (2007). Heterogeneity among depressed outpatients considered to be in remission. Comprehensive psychiatry, 48(2), 113-117.

Zimmerman, M., McGlinchey, J. B., Posternak, M. A., Friedman, M., Boerescu, D., & Attiullah, N. (2006). Discordance between self-reported symptom severity and psychosocial functioning ratings in depressed outpatients: implications for how remission from depression should be defined. Psychiatry Research, 141(2), 185-191.

Appendix A – references used in the review of psychotherapy measurement methodology

Anderson, P. L., Zimand, E., Hodges, L. F., & Rothbaurn, B. O. (2005). Cognitive behavioral therapy for public-speaking anxiety using virtual reality for exposure. Depression and Anxiety, 22(3), 156-158.

Andersson G, Waara J, Jonsson U, et al. Internet-based self-help vs. one-session exposure in the treatment of spider phobia: a randomized controlled trial. Cogn Behav Ther. 2009;38:114–20.

Andersson, G., Estling, F., Jakobsson, E., Cuijpers, P., & Carlbring, P. (2011). Can the patient decide which modules to endorse? An open trial of tailored internet treatment of anxiety disorders. Cognitive Behaviour Therapy, 40(1), 57-64.

Andrews G, Davies M, Titov N. Effectiveness randomized controlled trial of face to face versus Internet cognitive behaviour therapy for social phobia. Aust N Z J Psychiatry. 2011;45:337–40.

Arch, J. J., Ayers, C. R., Baker, A., Almklov, E., Dean, D. J., & Craske, M. G. (2013). Randomized clinical trial of adapted mindfulness-based stress reduction versus group cognitive behavioral therapy for heterogeneous anxiety disorders. Behaviour Research and Therapy, 51(4-5), 185-196.

Arch, J. J., Eifert, G. H., Davies, C., Vilardaga, J. C. P., Rose, R. D., & Craske, M. G. (2012). Randomized clinical trial of cognitive behavioral therapy (CBT) versus acceptance and commitment therapy (ACT) for mixed anxiety disorders. Journal of Consulting and Clinical Psychology, 80(5), 750-765.

Barlow, D. H., Cohen, A. S., Waddell, M. T., Vermilyea, B. B., Klosko, J. S., Blanchard, E. B., et al. (1984). Panic and generalized anxiety disorders: Nature and treatment. Behavior Therapy, 15(5), 431-449.

Barrowclough, C., King, P., Colville, J., Russell, E., Burns, A., & Tarrier, N. (2001). A randomized trial of the effectiveness of cognitive-behavioral therapy and supportive counseling for anxiety symptoms in older adults. Journal of Consulting and Clinical Psychology, 69(5), 756-762.

Berger T, Hämmerli K, Gubser N, Andersson G, Caspar F. Internet-based treatment of depression: a randomized controlled trial comparing guided with unguided self-help. Cogn Behav Ther. 2011;40 (4):251-266.

Berger, T., Boettcher, J., & Caspar, F. (2014). Internet-based guided self-help for several anxiety disorders: A randomized controlled trial comparing a tailored with a standardized disorder-specific approach. Psychotherapy (Chic), 51(2), 207-219.

Bergström J, Andersson G, Ljótsson B, et al. Internet-versus group-administered cognitive behaviour therapy for panic disorder in a psychiatric setting: a randomised trial. BMC Psychiatry. 2010;10:54.

Botella C, Gallego MJ, Garcia-Palacios A, et al. An Internet-based self-help treatment for fear of public speaking: a controlled trial. Cyberpsychol Behav Soc Netw. 2010;13:407–21.

Brenes, G. A., Miller, M. E., Williamson, J. D., Mccall, W. V., Knudson, M., & Stanley, M. A. (2012). A randomized controlled trial of telephone-delivered cognitive-behavioral therapy for late-life anxiety disorders. The American Journal of Geriatric Psychiatry, 20(8), 707-716.

Bressi, C., Porcellana, M., Marinaccio, P. M., Nocito, E. P., & Magri, L. (2010). Short-term psychodynamic psychotherapy versus treatment as usual for depressive and anxiety disorders: A randomized clinical trial of efficacy. Journal of Nervous and Mental Disease, 198(9), 647-652.

Carlbring P, Nilsson-Ihrfelt E, Waara J, et al. Treatment of panic disorder: live therapy vs. self-help via Internet. Behav Res Ther. 2005;43:1321–33.

Carlbring, P., Maurin, L., Torngren, C., Linna, E., Eriksson, T., Sparthan, E., et al. (2011). Individually-tailored, internet-based treatment for anxiety disorders: A randomized controlled trial. Behaviour Research and Therapy, 49(1), 18-24.

Christensen H, Griffiths KM, Jorm AF. Delivering interventions for depression by using the internet: randomised controlled trial. BMJ. 2004;328(7434): 265.

Clarke G, Eubanks D, Reid E, et al. Overcoming Depression on the Internet (ODIN) (2): a randomized trial of a self-help depression skills program with reminders.J Med Internet Res. 2005; 7(2):e16.

Clarke G, Kelleher C, Hornbrook M, Debar L, Dickerson J, Gullion C. Randomized effectiveness trial of an Internet, pure self-help, cognitive behavioral intervention for depressive symptoms in young adults. Cogn Behav Ther. 2009;38(4):222-234.

Clarke G, Reid E, Eubanks D, et al. Overcoming depression on the Internet (ODIN): a randomized controlled trial of an Internet depression skills intervention program.J Med Internet Res. 2002;4 (3):E14.

Craske, M. G., Farchione, T. J., Allen, L. B., Barrios, V., Stoyanova, M., & Rose, R. (2007). Cognitive behavioral therapy for panic disorder and comorbidity: More of the same or less of more? Behaviour Resarch and Therapy, 45(6), 1095-1109.

Cyranowski, J. M., Frank, E., Shear, M., Swartz, H., Fagiolini, A., Scott, J., et al. (2005). Interpersonal psychotherapy for depression with panic spectrum symptoms: A pilot study. Depression and Anxiety, 21(3), 140-142.

de Graaf LE, Gerhards SA, Arntz A, et al. Clinical effectiveness of online computerised cognitive-behavioural therapy without support for depression in primary care: randomised trial. Br J Psychiatry. 2009;195(1):73-80.

Dear, B. F., Titov, N., Schwencke, G., Andrews, G., Johnston, L., Craske, M. G., et al. (2011). An open trial of a brief transdiagnostic internet treatment for anxiety and depression. Behaviour Research and Therapy, 49(12), 830-837.

Ellard, K. K., Fairholme, C. P., Boisseau, C. L., Farchione, T. J., & Barlow, D. H. (2010). Unified protocol for the transdiagnostic treatment of emotional disorders: Protocol development and initial outcome data. Cognitive and Behavioral Practice, 17(1), 88-101.

Erickson, D. H., Janeck, A. S., & Tallman, K. (2007). A cognitive-behavioral group for patients with various anxiety disorders. Psychiatric Services, 58(9), 1205-1211.

Farchione, T. J., Fairholme, C. P., Ellard, K. K., Boisseau, C. L., Thompson-Hollands, J., Carl, J. R., et al. (2012). Unified protocol for transdiagnostic treatment of emotional disorders: A randomized controlled trial. Behavior Therapy, 43(3), 666-678.

Farrer L, Christensen H, Griffiths KM, Mackinnon A. Internet-based CBT for depression with and without telephone tracking in a national helpline: randomised controlled trial. PLoS One. 2011;6(11):e28099.

Gilbody S, Littlewood E, Hewitt C, et al; REEACT Team. Computerised cognitive behaviour therapy (cCBT) as treatment for depression in primary care (REEACT trial): large scale pragmatic randomised controlled trial. BMJ. 2015;351:h5627.

Gollings EK, Paxton SJ. Comparison of internet and face-to-face delivery of a group body image and disordered eating intervention for women: a pilot study. Eat Disord. 2006;14:1–15.

Hedman E, Andersson G, Ljótsson B, et al. Internet-based cognitive behavior therapy vs. cognitive behavioral group therapy for social anxiety disorder: a randomized controlled non-inferiority trial. PLoS One. 2011;6:e18001.

Jakupcak, M., Wagner, A., Paulson, A., Varra, A., & Mcfall, M. (2010). Behavioral activation as a primary care-based treatment for ptsd and depression among returning veterans. J Trauma Stress, 23(4), 491-495.

Johansson, R., Björklund, M., Hornborg, C., Karlsson, S., Hesser, H., Ljótsson, B., et al. (2013). Affect-focused psychodynamic psychotherapy for depression and anxiety through the internet: A randomized controlled trial. PeerJ, 1, e102. doi: 10.7717/peerj.102.

Johansson, R., Sjoberg, E., Sjogren, M., Johnsson, E., Carlbring, P., Andersson, T., et al. (2012). Tailored vs. Standardized internet-based cognitive behavior therapy for depression and comorbid symptoms: A randomized controlled trial. PLoS ONE, 7(5), e36905.

Johnston, L., Titov, N., Andrews, G., Spence, J., & Dear, B. F. (2011). A rct of a transdiagnostic internet-delivered treatment for three anxiety disorders: Examination of support roles and disorder-specific outcomes. PLoS ONE, 6(11).

Kabat-Zinn, J., Massion, A. O., Kristeller, J., Peterson, L. G., Fletcher, K. E., Pbert, L., et al. (1992). Effectiveness of a meditation-based stress reduction program in the treatment of anxiety disorders. American Journal of Psychiatry, 149(7), 936-943.

Kaldo V, Levin S, Widarsson J, et al. Internet versus group cognitive-behavioral treatment of distress associated with tinnitus. A randomised controlled trial. Behav Ther. 2008;39:348–59.

Kenwright, M., Marks, I. M., Gega, L., & Mataix-Cols, D. (2004). Computer-aided self-help for phobia/panic via internet at home: A pilot study. The British Journal of Psychiatry, 184(5), 448-449.

Kim, Y. W., Lee, S.-H., Choi, T. K., Suh, S. Y., Kim, B., Kim, C. M., et al. (2009). Effectiveness of mindfulness-based cognitive therapy as an adjuvant to pharmacotherapy in patients with panic disorder or generalized anxiety disorder. Depression and Anxiety, 26(7), 601-606.

Kiropoulos LA, Klein B, Austin DW, et al. Is internet-based CBT for panic disorder and agoraphobia as effective as face-to-face CBT? J Anxiety Disord. 2008;22:1273–84.

Kleiboer A, Donker T, Seekles W, van Straten A, Riper H, Cuijpers P. A randomized controlled trial on the role of support in Internet-based problem solving therapy for depression and anxiety. Behav Res Ther. 2015;72(6):63-71.

Klein JP, Berger T, Schröder J, et al. Effects of a psychological internet intervention in the treatment of mild to moderate depressive symptoms: results of the EVIDENT Study, a randomized controlled trial. Psychother Psychosom. 2016;85(4):218-228.

Lee, S. H., Ahn, S. C., Lee, Y. J., Choi, T. K., Yook, K. H., & Suh, S. Y. (2007). Effectiveness of a meditation-based stress management program as an adjunct to pharmacotherapy in patients with anxiety disorder. Journal of Psychosomatic Research, 62(2), 189-195.

Liu, S.-I., Huang, H.-C., Yeh, Z.-T., Hwang, L.-C., Tjung, J.-J., Huang, C.-R., et al. (2007). Controlled trial of problem-solving therapy and consultation-liaison for common mental disorders in general medical settings in taiwan. General Hospital Psychiatry, 29(5), 402-408.

Marks, I. M., Kenwright, M., Mcdonough, M., Whittaker, M., & Mataix-Cols, D. (2004). Saving clinicians' time by delegating routine aspects of therapy to a computer: A randomized controlled trial in phobia/panic disorder. Psychological Medicine, 34(1), 9-17.

McEvoy, P. M., & Nathan, P. (2007). Effectiveness of cognitive behavior therapy for diagnostically heterogeneous groups: A benchmarking study. Journal of Consulting and Clinical Psychology, 75(2), 344-350.

Meyer B, Berger T, Caspar F, Beevers CG, Andersson G, Weiss M. Effectiveness of a novel integrative online treatment for depression (Deprexis): randomized controlled trial.J Med Internet Res. 2009;11(2):e15.

Meyer B, Bierbrodt J, Schröder J, et al. Effects of an internet intervention (Deprexis) on severe depression symptoms: randomized controlled trial. Internet Interventions. 2015;2(1):48-59.

Moritz S, Schilling L, Hauschildt M, Schröder J, Treszl A. A randomized controlled trial of internet-based therapy in depression. Behav Res Ther. 2012;50(7-8):513-521.

Newby, J. M., Mackenzie, A., Williams, A. D., Mcintyre, K., Watts, S., Wong, N., et al. (2013). Internet cognitive behavioural therapy for mixed anxiety and depression: A randomized controlled trial and evidence of effectiveness in primary care. Psychological Medicine, 1-14.

Nieminen, P., & Kaur, J. (2019). Reporting of data analysis methods in psychiatric journals: Trends from 1996 to 2018. International journal of methods in psychiatric research, e1784-e1784.

Nieminen, P., Virtanen, J. I., & Vähänikkilä, H. (2017). An instrument to assess the statistical intensity of medical research papers. PloS one, 12(10), e0186882.

Nixon, R. D., & Nearmy, D. M. (2011). Treatment of comorbid posttraumatic stress disorder and major depressive disorder: A pilot study. Journal of Traumatic Stress, 24(4), 451-455.

Norton, P. J. (2008). An open trial of a transdiagnostic cognitive-behavioral group therapy for anxiety disorder. Behavior Therapy, 39(3), 242-250.

Norton, P. J. (2012). A randomized clinical trial of transdiagnostic cognitve-behavioral treatments for anxiety disorder by comparison to relaxation training. Behavior Therapy, 43(3), 506–517.

Norton, P. J., & Barrera, T. L. (2012). Transdiagnostic versus diagnosis-specific cbt for anxiety disorders: A preliminary randomized controlled noninferiority trial. Depression and Anxiety, 29(10), 874-882.

Norton, P. J., Hayes, S. A., & Hope, D. A. (2004). Effects of a transdiagnostic group treatment for anxiety on secondary depression. Depression and Anxiety, 20(4), 198-202.

Patel, V., Chisholm, D., Rabe-Hesketh, S., Dias-Saxena, F., Andrew, G., & Mann, A. (2003). Efficacy and cost-effectiveness of drug and psychological treatments for common mental disorders in general health care in goa, india: A randomised, controlled trial. Lancet, 361(9351), 33-39.

Paxton SJ, McLean SA, Gollings EK, et al. Comparison of face-to-face and internet interventions for body image and eating problems in adult women: an RCT. Int J Eat Disord. 2007;40:692–704.

Phillips R, Schneider J, Molosankwe I, et al. Randomized controlled trial of computerized cognitive behavioural therapy for depressive symptoms: effectiveness and costs of a workplace intervention. Psychol Med. 2014;44(4):741-752.

Proudfoot, J., Goldberg, D., Mann, A., Everitt, B., Marks, I., & Gray, J. (2003). Computerized, interactive, multimedia cognitive-behavioural program for anxiety and depression in general practice. Psychological Medicine, 33(2), 217-227.

Proudfoot, J., Ryden, C., Everitt, B., Shapiro, D. A., Goldberg, D., Mann, A., et al. (2004). Clinical efficacy of computerised cognitive-behavioural therapy for anxiety and depression in primary care: Randomised controlled trial. The British Journal of Psychiatry, 185(1), 46-54.

Radley, M., Redston, C., Bates, F., & Pontefract, M. (1997). Effectiveness of group anxiety management with elderly clients of a community psychogeriatric team. International Journal of Geriatric Psychiatry, 12(1), 79-84.

Ree, M. J., & Craigie, M. A. (2007). Outcomes following mindfulness-based cognitive therapy in a heterogeneous sample of adult outpatients. Behaviour Change, 24(02), 70-86.

Sanders, S. G., & Hunsley, J. (2018). The new Caucus-race: Methodological considerations for meta-analyses of psychotherapy outcome. Canadian Psychology/psychologie canadienne, 59(4), 387.

Schmidt, N. B., Buckner, J. D., Pusser, A., Woolaway-Bickel, K., Preston, J. L., & Norr, A. (2012). Randomized controlled trial of false safety behavior elimination therapy: A unified cognitive behavioral treatment for anxiety psychopathology. Behavior Therapy, 43(3), 518-532.

Schover LR, Canada AL, Yuan Y, et al. A randomized trial of internet-based versus traditional sexual counseling for couples after localized prostate cancer treatment. Cancer. 2012;118:500–9

Spek V, Nyklícek I, Smits N, et al. Internet-based cognitive behavioural therapy for subthreshold depression in people over 50 years old: a randomized controlled clinical trial. Psychol Med. 2007;37(12):1797-1806.

Spek V, Nyklicek I, Smits N, et al. Internet-based cognitive behavioural therapy for subthreshold depression in people over 50 years old: a randomized controlled clinical trial. Psychol Med. 2007;37:1797–806.

Titov, N., Andrews, G., Johnston, L., Robinson, E., & Spence, J. (2010). Transdiagnostic internet treatment for anxiety disorders: A randomized controlled trial. Behaviour Research and Therapy, 48(9), 890-899.

Titov, N., Dear, B. F., Schwencke, G., Andrews, G., Johnston, L., Craske, M. G., et al. (2011). Transdiagnostic internet treatment for anxiety and depression: A randomised controlled trial. Behaviour Research and Therapy, 49(8), 441-452.

Vollestad, J., Sivertsen, B., & Nielsen, G. H. (2011). Mindfulness-based stress reduction for patients with anxiety disorders: Evaluation in a randomized controlled trial. Behaviour Research and Therapy, 49(4), 281-288.

Wagner B, Horn AB, Maercker A. Internet-based versus face-to-face cognitive-behavioral intervention for depression: a randomized controlled non-inferiority trial. J Affect Disord. 2014;152-154:113–21.

Westra, H. A., Dozois, D. J. A., & Marcus, M. (2007). Expectancy, homework compliance, and initial change in cognitive-behavioral therapy for anxiety. Journal of Consulting and Clinical Psychology, 75(3), 363-373.

Wetherell, J. L., Ayers, C. R., Sorrell, J. T., Thorp, S. R., Nuevo, R., Belding, W., et al. (2009). Modular psychotherapy for anxiety in older primary care patients. The American Journal of Geriatric Psychiatry, 17(6), 483-492.

Wuthrich, V. M., & Rapee, R. M. (2013). Randomised controlled trial of group cognitive behavioural therapy for comorbid anxiety and depression in older adults. Behaviour Research and Therapy, 51(12), 779-786.

Zou, J. B., Dear, B. F., Titov, N., Lorian, C. N., Johnston, L., Spence, J., et al. (2012). Brief internet-delivered cognitive behavioral therapy for anxiety in older adults: A feasibility trial. Journal of Anxiety Disorders, 26(6), 650-655.

Chapter 2

Measurement of Symptom Change Following Web-Based Psychotherapy: Statistical Characteristics and Analytical Methods for Measuring and Interpreting Change (Study 1)

This chapter concerns a first and fundamental step in the process of measuring and interpreting psychotherapy evidence - the choice of analytical methods for the measurement of symptom change in treatment. The chapter includes a study that explored the statistical characteristics of depressive symptom change, and used these features to compare comment on the suitability of different ways to measure and interpret symptom change. The major aim of the study was to explore the features of depressive symptom change, and the suitability of different measurement models and metrics that measure and interpret the phenomenon of clinical symptom change. The exploration and results sought to contribute to the limited available research about the choice of measurement analytics in the psychotherapy treatment evaluation context. The study was published in the Journal of Medical Internet Research (JMIR) Mental Health.

Karin, E., Dear, B. F., Heller, G. Z., Gandy, M., & Titov, N. (2018). Measurement of symptom change following web-based psychotherapy: Statistical characteristics and analytical methods for measuring and interpreting change. *JMIR Mental Health*, *5*(3), e10200.

Author contribution:

Mr Eyal Karin designed, analysed, and wrote the study.
Associate Professor Blake F. Dear and Dr Milena Gandy provided the dataset, assisted with the refinement of the manuscript, and helped frame the methodological content for a clinical audience. Professor Gillian Heller oversaw the choice of statistical methodology and assisted with the drafting of the manuscript. Professor Nick Titov oversaw the conception of the project and the drafting of the manuscript.

Original Paper

Measurement of Symptom Change Following Web-Based Psychotherapy: Statistical Characteristics and Analytical Methods for Measuring and Interpreting Change

Eyal Karin[1], MApStat; Blake F Dear[1,2], PhD; Gillian Z Heller[3], PhD; Milena Gandy[1], PhD; Nickolai Titov[1,2], PhD

[1]eCentreClinic, Department of Psychology, Macquarie University, Sydney, Australia

[2]Mindspot Clinic, Macquarie University, Sydney, Australia

[3]Department of Statistics, Faculty of Science and Engineering, Macquarie University, Sydney, Australia

Corresponding Author:
Eyal Karin, MApStat
eCentreClinic
Department of Psychology
Macquarie University
Building C3A
First Walk Macquarie University NSW
Sydney, 2109
Australia
Phone: 61 298508657
Email: eyal.karin@mq.edu.au

Abstract

Background: Accurate measurement of treatment-related change is a key part of psychotherapy research and the investigation of treatment efficacy. For this reason, the ability to measure change with accurate and valid methods is critical for psychotherapy.

Objective: The aims of this study were to (1) explore the underlying characteristics of depressive symptom change, measured with the nine-item Patient Health Questionnaire (PHQ-9), following psychotherapy, and (2) compare the suitability of different ways to measure and interpret symptom change. A treatment sample of Web-based psychotherapy participants (n=1098) and a waitlist sample (n=96) were used to (1) explore the statistical characteristics of depressive symptom change, and (2) compare the suitability of two common types of change functions: linear and proportional change.

Methods: These objectives were explored using hypotheses that tested (1) the relationship between baseline symptoms and the rate of change, (2) the shape of symptom score distribution following treatment, and (3) measurement error associated with linear and proportional measurement models.

Results: Findings demonstrated that (1) individuals with severe depressive baseline symptoms had greater reductions in symptom scores than individuals with mild baseline symptoms (11.4 vs 3.7); however, as a percentage measurement, change remained similar across individuals with mild, moderate, or severe baseline symptoms (50%-55%); (2) positive skewness was observed in PHQ-9 score distributions following treatment; and (3) models that measured symptom change as a proportional function resulted in greater model fit and reduced measurement error (<30%).

Conclusions: This study suggests that symptom scales, sharing an implicit feature of score bounding, are associated with a proportional function of change. Selecting statistics that overlook this proportional change (eg, Cohen *d*) is problematic and leads to (1) artificially increased estimates of change with higher baseline symptoms, (2) increased measurement error, and (3) confounded estimates of treatment efficacy and clinical change. Implications, limitations, and idiosyncrasies from these results are discussed.

(JMIR Ment Health 2018;5(3):e10200) doi:10.2196/10200

KEYWORDS

clinical measurement; treatment evaluation; symptom change; symptom scales; psychotherapeutic change

Introduction

Accurate measurement of treatment-related change is a key part of psychotherapy research [1-3] and the investigation of treatment efficacy [4-6]. For example, measurable change in symptoms of anxiety and depression is often used as the primary means to research and test the safety of emerging treatments [7]. Reporting symptom change in anxiety and depression has been shown to describe the clinical trajectory of participants in treatment [8], illustrate the cost-effectiveness of treatment [9], and compare treatments [10]. For this reason, the ability to measure change with accurate and valid methods is critical for psychotherapy [6,11].

Several statistical and clinical methods are employed to increase the validity and accuracy of change measurement in psychotherapy. The most common methodology in psychotherapy research is the combined use of standardized scales, such as standardized symptom scales of anxiety [12] or depression [1,13], and the use of statistical analyses, such as Cohen *d* effect sizes, that measure and interpret the rate of change in treatment [4-6]. Many types of standardized scales are available for measuring and interpreting change in treatment (eg, clinical interviews, measurement of behavior or quality of life [14]), and that change can be statistically estimated through various statistical methods [15]. However, from the wide range of possible methods for measuring treatment outcomes [16], the use of standardized scales, primarily symptom scales, in combination with effect sizes, primarily Cohen *d*, are the most influential. For example, symptom scales and effect sizes are used to evaluate treatment-related change and treatment efficacy within psychotherapy trials [17-19], epidemiological studies [20,21], meta-analytic studies of various treatments [22], and are even mandated within clinical guidelines for reporting in clinical trials, such as Consolidated Standards of Reporting Trials (CONSORT) [19], Transparent Reporting of Evaluations with Nonrandomized Designs (TREND) [23], Strengthening the Reporting of Observational studies in Epidemiology (STROBE) [24], and others [11].

Notwithstanding the common use of both symptom scales and effect sizes for measuring psychotherapeutic-related change, little research is currently available to verify or refute the use of different statistical methods for measuring and interpreting symptom change [25,26]. For example, the use of effect sizes, such as Cohen *d*, is based on statistical assumptions that change is linear. In technical terms, by employing effect sizes, researchers assume that the symptom change that follows treatment is average, constant, and representative of the average change experienced by any participating individual [18,27]. Put another way, if an average individual with moderate depressive symptoms prior to treatment, such as a score between 10 and 15 on the nine-item Patient Health Questionnaire (PHQ-9), would improve by 5 points on a symptom scale, an individual with severe baseline symptoms (eg, PHQ-9 score of 20-27) would be expected to demonstrate the same rate of improvement (eg, 5 points). Similarly, under the linear assumption, a group of participants with different baseline symptoms (eg, mild, moderate, or severe baseline symptoms) undertaking the same therapy would be expected to have similar effect sizes between groups (eg, 1.0). However, in contrast to the common use of statistics that assume change is linear, there are two lines of research to suggest that real-world symptom change may occur as a proportional function from baseline. First, psychological treatment studies often describe an increased rate of clinical change within samples of increased baseline symptom severity [20,28]. Second, common symptom scales, such as the PHQ-9 [29], the Generalized Anxiety Disorder seven-item scale (GAD-7) [30], and prominent others (eg, Kessler Psychological Distress scale) [31], often demonstrate an implicit design feature of score bounding at minimal symptoms. This bounding within symptom scales should theoretically imply that, under effective treatment, all individuals would reduce their symptoms down to the same endpoint of minimal levels [1,9] and that the rate of change would systematically depend on an individual's symptoms at baseline [32,33].

From a statistical point of view, identifying the characteristics of symptom change, and employing a suitable statistical analysis that captures the underlying function of change, can fundamentally impact both the measurement and interpretation of clinical outcomes [15,34,35]. For example, under circumstances in which change is proportional in nature, the selection of a proportional statistical analysis can greatly increase the accuracy and validity of estimating longitudinal clinical change [34,35]; the detection of moderators of symptom change [36]; the classification of subgroups, such as remitters or nonresponders [37]; as well as the ability to research other objectives [38]. For this reason, the function of symptom change must be researched and more clearly understood. Such research could verify, refute, and draw out the implication for using well-established statistical methods (eg, effect sizes, linear statistics) and emerging alternatives (eg, percentage improvement, generalized linear statistics) for measuring and interpreting change in treatment. In addition, researching the function and characteristics of symptom change has the potential to inform researchers and the broader community about the type of change individuals in treatment are likely to experience.

This Study

This study aims to (1) explore the fundamental statistical characteristics of treatment-related depressive symptom change and (2) compare the implications from measuring and interpreting clinical change through effect sizes, such as Cohen *d*, against emerging alternatives, such as percentage improvement (proportional, generalized longitudinal linear statistics) [25,26].

This study employed a large sample of individuals (N=1098) who underwent Web-based psychotherapy (Internet-delivered cognitive behavioral therapy [ICBT]) [39] for symptoms of depression (PHQ-9 [29]). Although Web-based psychotherapy represents a distinct type of psychotherapy, the use of Web-based treatments, which standardizes treatment materials and participant engagement through automatization, can be seen as an opportunity for researching symptom change with high internal validity and minimum methodological interference.

The statistical characteristics of symptom change were explored with three steps. Initially, the relationship between baseline symptoms and the rate of change was explored. In line with

previous clinical studies that suggest that more severely symptomatic participants demonstrate increased effect sizes [20,32], it was hypothesized that individuals with increased symptoms at baseline would also demonstrate increased rates of symptom change (hypothesis 1). Second, the shape of symptom score distribution before and following treatment were explored. In line with the suggestion that symptoms scores are bounded at minimal symptoms [29,30], the distributions of pretreatment and posttreatment depression symptom levels were hypothesized to show evidence of positive skewness and kurtosis at both pretreatment and posttreatment (hypothesis 2). Third, the measurement error associated with linear and proportional measurement models was compared. In line with the characterization of symptom change as proportional, it was hypothesized that those statistical methods that measure symptom change as a proportional function would be associated with reduced measurement error and indicate greater statistical fit to real symptom data in treatment (hypothesis 3). Finally, an additional effort was taken to explore the patterns of depressive symptom change within a control group (n=96). This addition was designed to explore the pattern of symptom change that is not specific to treatment.

Methods

The Sample

This study combined clinical data from three published randomized controlled trials, all of which evaluated ICBT for symptoms of depression and anxiety [39,40]. These interventions were almost identical in structure and therapeutic content. All trials were delivered using the same evidence-based online treatment approach [7] and were conducted within the same research clinic, the eCentreClinic [41]. A precautionary test, aiming to compare the symptom reduction rates between the individual trials, demonstrated similarities across all three interventions. Specifically, a generalized estimated equation (GEE) model [35], testing the longitudinal symptom change of each trial, resulted in slight differences in the estimates of symptom change across trials (PHQ-9 range 5.23-6.29 points); differences were not statistically significant (group × time: Wald $\chi^2_{2,2368}$=5.0, P=.08).

Together, these trials represent a large random intake of self-selecting adults into treatment over a period of 2 years with a total of 1262 adult participants, of whom 1098 (87.01%) were successfully assessed at both pretreatment and posttreatment time points. Additional information about recruitment, advertising, treatment materials, and additional treatment procedures can be found within additional eCentreClinic publications [7,41].

To be included in these trials, participants were selected on the basis of (1) demonstrating at least mild symptoms of depression or anxiety (a minimum score ≥5 on either the PHQ-9 or the GAD-7), (2) older than 18 years and younger than 65 years, (3) being an Australian resident, and (4) having Internet access for the period of the trial. In addition, applicants who reported a score of 3 (considered severe) on item 9 of the PHQ-9 measuring suicidal risk, were referred to another service.

Additional demographic and symptom characteristics are shown in Table 1 for both the treatment and waitlist control conditions.

Table 1. Sample demographics (N=1194).

Demographics	Collated treatment sample (n=1098)	Control sample (n=96)
Gender (male), n (%)	330 (30.1)	51 (53.1)
Age (years), mean (SD)	52.8 (14.2)	56.3 (13.0)
Using medication during the course, n (%)	351 (31.9)	51 (53.1)
Married, n (%)	713 (64.9)	45 (46.9)
Employed, n (%)	636 (57.9)	49 (51.0)
Education, n (%)		
High school	176 (16.0)	39 (40.6)
Vocational education	307 (27.9)	24 (25.0)
Degree	615 (56.0)	37 (38.5)
PHQ-9[a], mean (SD)		
Before treatment	11.73 (4.83)	10.95 (4.73)
following treatment)	5.60 (4.58)	11.00 (5.04)
GAD-7[b], mean (SD)		
Before treatment	10.91 (4.53)	9.5 (4.53)
Following treatment	5.47 (4.35)	8.83 (4.67)

[a]PHQ-9: nine-item Patient Health Questionnaire..

[b]GAD-7: seven-item Generalized Anxiety Disorder scale.

Symptom Measure

The PHQ-9 was employed as the primary outcome variable, measuring the presence and severity of depressive symptoms [29]. The PHQ-9 is widely used in clinical trials [7,16], comprising nine items, with high internal consistency and high sensitivity to the presence and change of clinical depression diagnoses [29]. Scores on the PHQ-9 correspond to the cumulative experience of common depressive symptoms over the preceding 2-week period. Cumulative scores range from 0 to 27 and scores are clinically interpreted as falling within five categories: (1) no depression symptoms (total score: 0-4), (2) mild depression symptoms (total score: 5-9), (3) moderate depression symptoms (total score: 10-14), (4) moderately severe depression symptoms (total score: 15-19), and (5) very severe depression symptoms (total scores: 20-27). Symptom scores were modified with a small constant added (0.001) to ensure that plausible values of zero symptoms at posttreatment were represented in the model when statistically modeling proportional functions, such as logarithmic link functions.

Analytical Plan

The function of symptom change was explored with three separate steps, corresponding to the three hypotheses.

The first hypothesis that individuals with increased symptoms at baseline would also demonstrate increased rates of symptom change was tested by examining the relationship between baseline symptoms and the rate of symptom change. Symptom change was examined within the five subgroups of individuals of different baseline PHQ-9 score bands (eg, minimal to no symptoms to very severe depression symptoms). Within each subgroup, the rate of change was approximated with GEE models, multilevel models [34], and raw means. These methods represent common longitudinal statistical methods in clinical trials [42]. The estimation of change through all three GEE, mixed models, and raw scores was designed to clarify that the underlying function of symptom change could be identified when using various statistical models.

Under a linear pattern of symptom change, participants of any baseline symptoms would be expected to show a similar rate of improvement overall. That is, an average symptom change score that would be observed across individuals, irrespective of the severity of their symptoms at baseline [18]. In contrast, under a proportional pattern of symptom change, participants presenting with increased baseline symptom severity would likely show larger symptom change compared to those individuals with mild or moderate baseline symptoms [15].

To test the second hypothesis that distributions of pretreatment and posttreatment depression symptom levels would show evidence of positive skewness and kurtosis, the distributions of depression symptoms scores at both pretreatment and posttreatment were evaluated for evidence of skewness. In this step, if the dataset would present with statistically normal distribution of symptom scores at both time points, the symptom change over time would be considered as linear. In contrast, if symptoms changed as a proportional function from baseline, positive skewness should be observed, particularly at posttreatment, where individuals from various baseline symptoms would shift and concentrate around the symptom score band of minimal symptoms. Graphical and numerical explorations of pre-post score distributions were included.

To test the third hypothesis that statistical methods measuring symptom change as a proportional function would be associated with reduced measurement error and indicate greater statistical fit to real symptom data in treatment, the relative measurement accuracy of models that represent either linear or proportional symptom change were compared. Specifically, this step compared model fit statistics and the remaining unexplained (residual) variance associated with each function of change. Both mixed models and GEE models were run initially as models that assume change was linear, represented through models that specified a normal scale of the dependent variable and an identify link function. Following this, alternative statistical models were compared, which specified a gamma scale and a log link function; representing models that assumed change was proportional. Generally, the gamma scale is considered a suitable method for data showing signs of skewness and multiplicative change function [15]; however, the selection of the gamma scale does not imply that alternative multiplicative statistical methods (eg, negative binomial scale, Poisson scale, or zero inflated models) would be less effective.

Formulas emphasizing the difference in statistical notation between the multiplicative model (Equations 1.1-1.2) and the linear model (Equations 1.3-1.5) are presented in Figure 1. With more formal statistical notation, the multiplicative effect within the log link model is created when the intercept, β_0, or baseline symptoms, is multiplied by the treatment effect, β_{tj}, the estimate of exponential change following treatment (Equations 1.6-1.8 in Figure 1).

The suitability of either model type was evaluated through model fit statistics, generated using SAS 9.4 software. Specifically, the quasilikelihood under the independence model criterion (QIC) statistic [43] for GEE models, and Akaike information criterion (AIC) and Bayesian information criterion (BIC) for mixed effects models [44], compared between linear (additive) and generalized linear (proportional) models. Within all AIC, BIC, and QIC model fit estimates, relatively lower scores imply overall reduced variance, and overall increase measurement accuracy.

Figure 1. Equations 1.1-1.8.

Multiplicative model

(1.1) $Y_{ij} \sim Gamma\ (\mu_{ij}, \alpha)$

(1.2) $log\ (\mu_{ij}) = \beta 0 + \beta_{tj} + \epsilon_{ij}$

Linear additive model

(1.3) $Y_{ij} = \beta_0 + \beta_{tj} + \epsilon_{ij}$

(1.4) $\epsilon_{ij} \sim N\ (0, 1)$

(1.5) $i = 1,...,1098;\ j = 0,1$

$t_j = \{ 0\ (\text{time} = \text{pre-treatment});\ 1\ (\text{time} = \text{post treatment})$

β_0 is the random intercept at pre-treatment;

and β_{tj} is the treatment effect of change over time

Linear additive model

(1.6) $\hat{\mu}_{baseline} = e^{\hat{\beta} 0}$

(1.7) $\hat{\mu}_{postreatment} = e^{\hat{\beta} 0} * e^{\hat{\beta} t1}$

(1.8) $\hat{\mu}_{postreatment} = \hat{\mu}_{baseline} * e^{\hat{\beta} t1}$

In addition to model fit statistics, the measurement error associated with the assumption that symptom change was either a fixed average score, or a percentage improvement score, was compared. In this step, measurement error was created for each participant by comparing the predicted posttreatment score under each change assumption (eg, PHQ-9 change of 5 points or 50% from baseline) against a known participant outcome score at posttreatment. The difference between the expected symptom outcome and actual treatment outcome effectively represents measurement error under the two change assumptions, akin to residual scores and measurement error variance. The pattern of residuals created under either assumption of symptom change was explored in two ways. First, the total quantity of error variance under each function was compared. Second, measurement residuals were graphically explored under each function of symptom change by comparing the increase or decrease of residuals for individuals with different baseline symptom score.

Results

In the first step (operationalizing the first hypothesis that individuals with increased symptoms at baseline would also demonstrate increased rates of symptom change), the relationship between baseline symptom severity and the quantity of symptom change was explored graphically. Figure 2, illustrating PHQ-9 change as a linear function, and Figure 3, illustrating PHQ-9 change as a proportional change from baseline, both demonstrate the symptom change on the y-axis within each of the PHQ-9 baseline symptom bands (x-axis). In addition, the symptom change observed within the waitlist condition is included as a dotted trend line, illustrating the trend of nonspecific change in symptoms within each bands of symptom severity at baseline.

Figure 2 illustrates an increased rate of symptom change that was associated closely with increased baseline symptoms. In Figure 2, individuals with severe baseline symptoms were observed to reduce by as much as threefold compared to individuals with mild baseline symptoms (11.4 vs 3.7, respectively). In addition, participants with severe symptoms in the control group demonstrated a sizable reduction in symptoms even when treatment was not applied. This nonspecific symptom-related change was pronounced to the extent that individuals with severe baseline symptoms in the control group demonstrated higher symptom reduction than individuals with moderate symptoms in treatment (7 points vs 6 points, respectively). That is, as a linear effect, the nonspecific symptom change within the control condition was larger than the treatment-related symptom change of individuals with moderate symptoms.

Figure 3 illustrates the proportional percentage change of symptoms within each of the mild, moderate, moderately severe, and severe subgroups. The figure illustrates that as a proportional change, an average treatment-related change of 50% to 55% was observed across all subgroups of individuals who started with at least mild symptoms at baseline. Of note, the rate of proportional improvement in treatment (50%-55%) was greater than the nonspecific change experienced by individuals with severe baseline symptoms in the waitlist conditions (35%). That is, the measurement of change as a percentage change resulted in a clearer differentiation of treatment-specific and nonspecific change.

Figure 2. Measurement of mean treatment-related PHQ-9 symptom change per initial pretreatment symptom severity band; whiskers represent 95% CI s. Symptom change observed under control conditions indicated by a solid trend line.

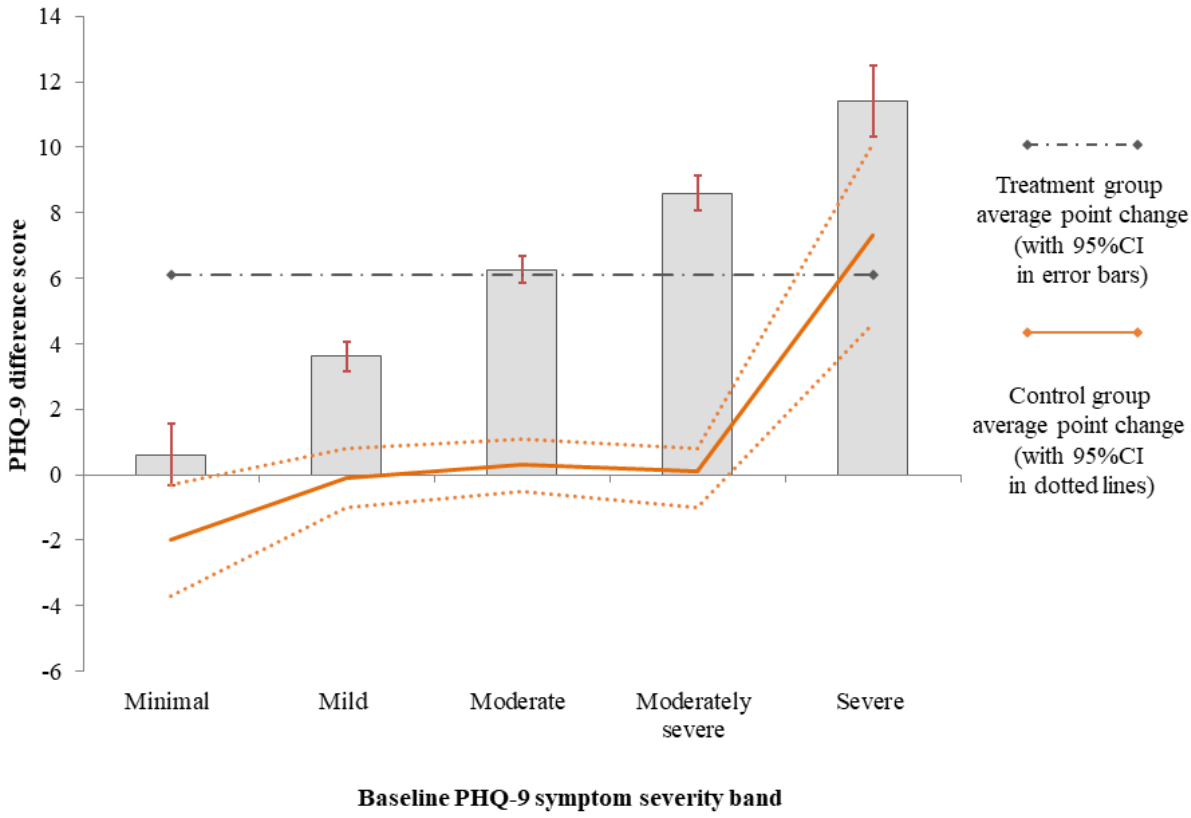


Figure 3. Measurement of mean treatment-related PHQ-9 symptom change as a proportional pattern of remission (52%); per initial pretreatment symptom severity; whiskers represent 95% CIs. Symptom change observed under control conditions indicated by a solid trend line.

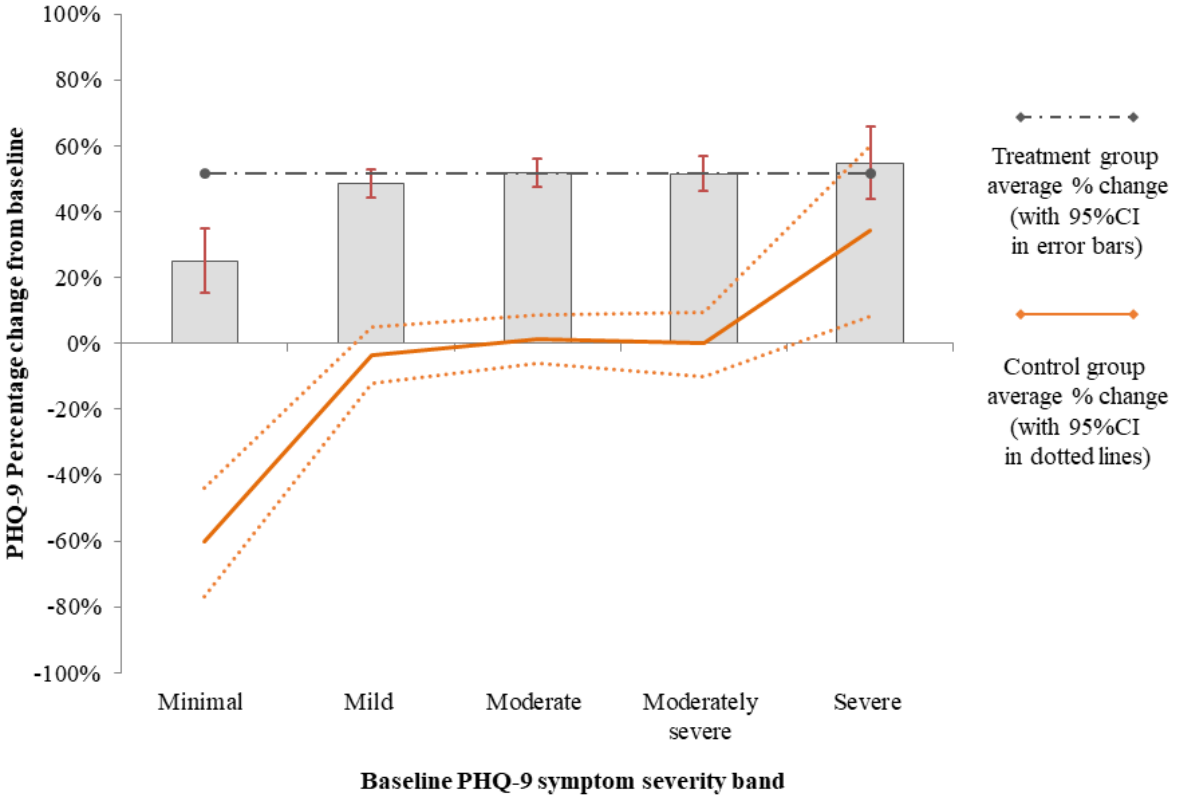


Table 2 includes the numerical descriptions of change for both the treatment and control conditions. Table 2 also includes effect sizes that were calculated within the treatment group as a whole and the effect size demonstrated by individuals in the mild, moderate, moderately severe, and severe bands of baseline symptoms. Individuals with mild depressive symptoms showed smaller effects (1.59) compared to individuals with more severe symptoms (3.9).

In a second step, the second hypothesis that distributions of pretreatment and posttreatment depression symptom levels would show evidence of positive skewness and kurtosis was operationalized with an exploration of the distribution of pretreatment and posttreatment symptom scores. Figure 4 illustrates the distribution of PHQ-9 symptom scores, both before and following treatment. These histograms illustrate a slight positive skewness of scores at pretreatment, with fewer individuals presenting within the severely symptomatic band as compared to the mild and moderate bands. In contrast, at posttreatment, increasing positive skewness was observed, where most individuals who reduced their symptoms became concentrated within the mild to minimal symptom ranges. The numerical estimates of the skewness are collated in Table 3.

Taken together, both numerically and graphically, the distributions of symptom scores demonstrated significant positive skewness that increased at posttreatment.

In a third step, the third hypothesis that statistical methods measuring symptom change as a proportional function would be associated with reduced measurement error and indicate greater statistical fit to real symptom data in treatment was operationalized, seeking to explore the model fit of the linear and the multiplicative statistical models of symptom change. Table 4 collates the goodness-of-fit statistics from models that specified either a proportional or linear function of change.

In Table 4, models that specified a proportional function of symptom change demonstrated a several-fold improvement in the model fit statistics within both the GEE and mixed models, including reduced QIC statistics, reduced AIC, and reduced BIC estimates. Table 4 also collated the measurement error associated with the prediction that change occurred as a linear change of six points, or as a percentage improvement (52% reduction from baseline). A notable reduction in the total estimate of PHQ-9 error variance was evident when a proportional function of change was assumed (σ^2=16.716 vs σ^2=24.122). This result demonstrated that by characterizing change as a proportional function, the measurement error and remaining unknown individual variation reduced by more than 30%.

Table 2. Rates of change of nine-item Patient Health Questionnaire (PHQ-9) scores associated with linear and proportional change functions; estimates per initial baseline symptom subgroups.

PHQ-9 and change functions	Initial symptom severity					Total
	Minimal (n=72)	Mild (n=345)	Moderate (n=381)	Moderately severe (n=244)	Severe (n=56)	Overall sample (treatment) scores
Observed PHQ-9, mean (SD)						
Pretreatment	2.83 (1.25)	7.32 (1.33)	12.07 (1.40)	16.67 (1.41)	20.86 (0.84)	11.41 (4.79)
Posttreatment	2.22 (2.61)	3.71 (3.3)	5.81 (3.92)	8.07 (5.41)	9.45 (4.99)	5.59 (4.57)
GEE[a] (95% CI)[b]						
Additive change estimate	0.61 (–0.30 to 1.18)	3.66 (3.30 to 4.02)	6.22 (5.82 to 6.62)	8.66 (7.98 to 9.34)	11.43 (10.14 to 12.73)	6.00 (5.71 to 6.28)
Percent proportional change estimate	21% (–1 to 39)	50% (45 to 54)	52% (48 to 55)	52% (48 to 56)	55% (48 to 61)	52 (50 to 54)
Effect size, Cohen *d* (95% CI)	0.32 (0.01 to 0.63)	1.59 (1.43 to 1.74)	2.34 (2.19 to 2.49)	2.54 (2.33 to 2.74)	3.90 (3.45 to 4.36)	1.27 (1.21 to 1.34)
Control group						
Change[c] (95% CI)[b]	–2 (–27 to –1.24)	–0.1 (–0.76 to 0.53)	0.29 (–0.68 to 1.28)	0.48 (–1.01 to 1.15)	7.37 (5.14 to 9.51)	0.68 (–0.37 to 0.16)
Percent proportional change estimate, GEE (95% CI)[b]	–61 (–78 to –44)	–4 (–12 to 5)	1 (–6 to 9)	0 (–10 to 10)	34 (8 to 60)	0% (–1 to 1)

[a]GEE: generalized estimated equation.

[b]Confidence intervals based on modeled marginal means.

[c]Control group change is nonspecific effect.

Figure 4. Dispersion of symptom scores (nine-item Patient Health Questionnaire, PHQ-9) at pretreatment (in light bars) and posttreatment scores (in dark bars). The dotted trend lines are indicative of the shape of each distribution.

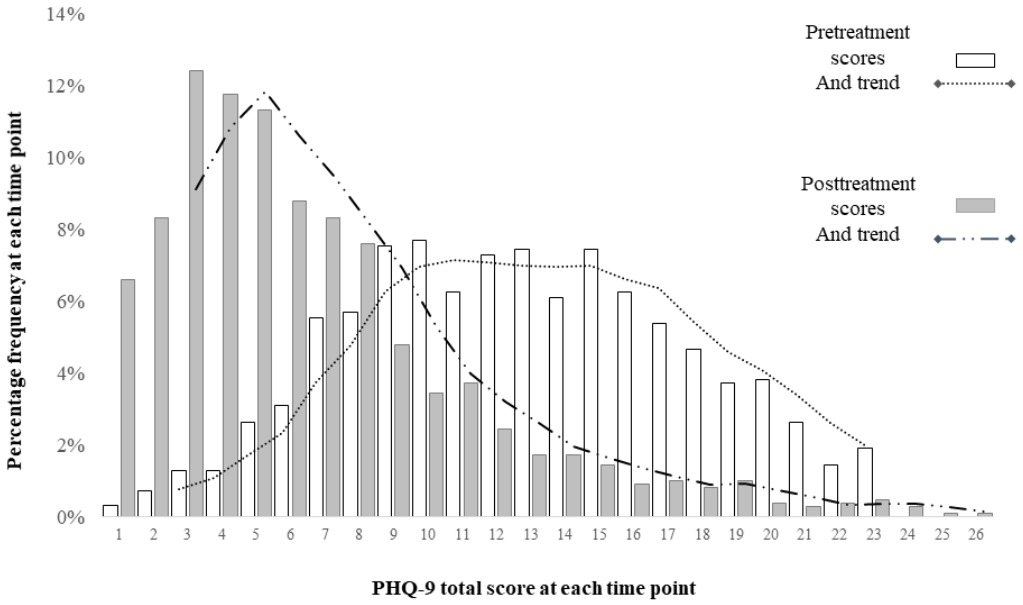


Table 3. Symptom score distributions statistics

Sample and time point	Skewness (SE)	Baseline symptoms, mean (SD)	Effect size, Cohen *d* (95% CI)
Treatment sample (n=1098)			1.27 (1.21 to 1.34)
Pretreatment	0.271 (0.071)[a]	11.73 (4.83)	
Posttreatment	1.359 (0.076)[a]	5.60 (4.58)	
Control sample depression (n=96)			–0.04 (–0.24 to 0.16)
Pretreatment	0.178 (0.109)	10.91 (4.53)	
Posttreatment	0.228 (0.109)	11.00 (5.04)	

[a]Statistical significance beyond .05 alpha on a Shapiro-Wilk test for distribution normality; significance is indicative that normal distribution is not supported within the observed sample.

Table 4. Model fit statistics and dispersion of model residuals for the treatment sample (n=1098). Model fit criterion was derived from SAS software, version 9.3.

Method of change specified	QIC[a,b] (GEE[c] model)	AIC[d,b] (Mixed)	BIC[e,b] (Mixed)	Total variance (PHQ-9 σ^2)
Linear (normal scale)	52457.6	14059.8	14071.3	16.716
Proportional (gamma scale)	2020.5	4041.8	4053.3	24.122

[a]QIC: quasilikelihood under the independence model criterion.

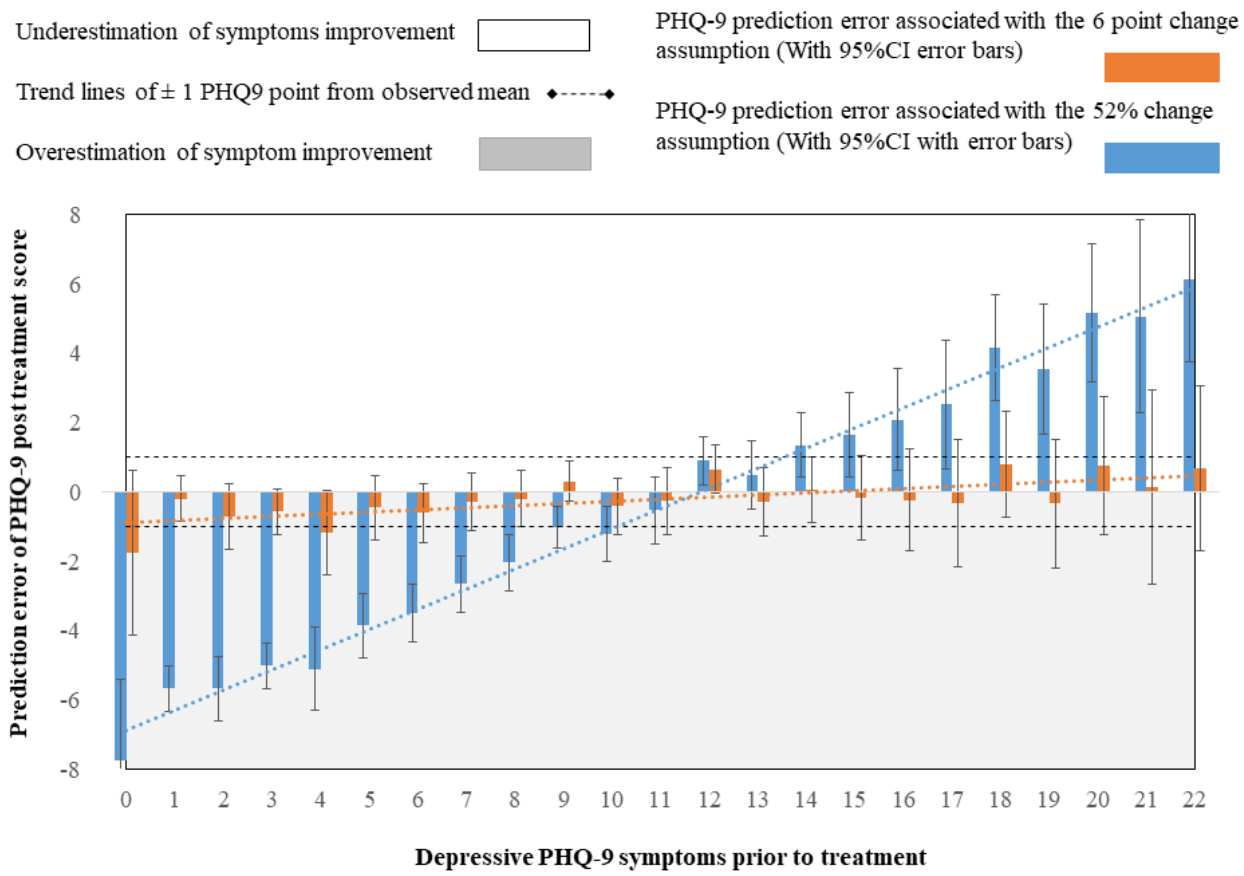
[b]Confidence intervals based on the multiplicative longitudinal GEE model specified in the analytical plan.

[c]GEE: generalized estimated equation.

[d]AIC: Akaike information criterion.

[e]BIC: Bayesian information criterion.

Figure 5. PHQ-9 estimation error (residual) following fixed (linear) and relative (proportional) change assumption.



The measurement error associated with either assumption that change was linear (6 points) or proportional (52%) were graphically explored. Figure 5 illustrates the residual error (y-axis) across individuals who started treatment with different baseline symptoms (x-axis). In the figure, individuals with mild and severe baseline symptoms can be observed to substantially underestimate or overestimate the rate of symptom change when linear change (6 points) was predicted. In contrast, when change was predicted to be proportional (52%), baseline symptoms no longer associated with the rate measurement error. Further, under the proportional assumption, the predicted symptom outcome could be accurately predicted within a single point across individuals with different baselines (marked with dots horizontal lines). In contrast, under the linear assumption, the prediction of symptom outcome become systematically erroneous with baseline severity (a range of up to 16 points between mild and severe).

Discussion

This study aimed to investigate the statistical characteristic of symptom change in treatment and compare different ways to measure and interpret symptom change. Using a Web-based psychotherapy sample (n=1098), as well as a waitlist control condition (n=96), the statistical characterization of depressive symptom change (PHQ-9) was explored in three steps, corresponding to three proposed hypotheses.

Testing of the first hypothesis demonstrated support for the characterization of symptom change as a proportional function through a clear association between symptom severity at baseline and the rate of change. In contrast, as a proportional estimate of change, individuals in treatment demonstrated a consistent rate of proportional symptom change within all subgroups with mild, moderate, moderately severe, and severe baseline symptom (50%-55%). Critically, the dependency between symptom change and baseline symptom severity was also observed in the waitlist condition, with mild and severe participants changing proportionally in their symptoms even when treatment was not applied. Testing of the second and third hypotheses also illustrated support for the characterization of symptom change as proportional function, with symptom score distributions presenting with positive skewness, particularly following treatment (H2). Similarly, increased model fit, and reduced measurement error was observed when the treatment sample was statistically modeled with an underlying proportional function of change (H3).

The analyses within this study are novel in that they characterize the function of depressive symptom change and compare different statistical methods for measuring and interpreting symptom change within treatment as well as nontreatment conditions. The findings suggest that common psychotherapy symptom scales (eg, PHQ-9) are impacted by a feature of natural bounding at minimal symptoms, which is the suspected culprit for the resulting (1) nonnormal distributions at posttreatment,

(2) the dependency between baseline symptoms and rate of change, and (3) the improved model fit for techniques that assume longitudinal change is proportional to baseline.

These findings raise two potentially critical implications for the ability to measure and interpret psychotherapy change in combination with symptom scales. First, the inappropriate use of linear statistics, such as Cohen *d*, when change is proportional would lead to artificially higher estimates of clinical efficacy, both in treatment and in control conditions. For example, in this study, individuals with severe baseline symptoms demonstrated effect sizes that increased by nearly threefold (3.9) when compared to individuals with mild symptoms (1.6), even when the same treatment was applied. This is problematic because linear estimates of change such as Cohen *d* are strongly associated with baseline severity and not with quality or the effectiveness of treatment. This finding is broadly consistent with the data within previous psychotherapy studies showing increased effect sizes with samples of increased symptoms, even when similar treatments are applied [20,29,32].

Second, these findings support a well-established statistical idea posing that the selection of a statistical analysis must match the characteristics of the dataset in order to arrive at valid and accurate statistical measurement, interpretation, and conclusions [4,45]. In this context of depressive symptom scales, the use of proportional statistical analyses resulted in (1) improved statistical modeling of treatment effects, (2) an improved ability to determine what a treatment effect is (50%-55%) and what a nontreatment effect is (35%), as well as for (3) establishing a clinical effect that is robust across individuals with various baseline symptoms (50%-55%). The measurement and interpretation of change as proportional improvement from baseline can also be concretely and easily interpreted as an estimate of change (eg, percentage improvement). Further, in the context of treatment, percentage improvement and percentage change estimates seem to reflect the ideal of treatment (reducing symptoms to minimal) [1,9]. For these reasons, measuring and interpreting change as a fundamentally proportional function can hold critical implications for clinical research that is reliant on accurate and interpretable measurement. For example, researchers seeking to identify clinical moderators, compare between treatments, estimate cost-effectiveness, or classify individual effects are likely to be positively impacted with a suitable choice of analytics that capture the underlying statistical function of change [36,37].

Although the measurement and interpretation of symptom change as a proportional change show promise to increase the accuracy and interpretability of clinical change, several statistical and clinical limitations should be considered about the results of this study. Primarily, the results of this study should be considered as (1) preliminary, (2) specific to a symptom scale of depressive symptoms (PHQ-9), and (3) specific to one kind of treatment model (the Macquarie University online model). Specifically, albeit the strengths of this study as an exploration of change within a large and standardized sample, it is unclear to what extent the 50% to 55% symptom change is specific to this treatment model and to the PHQ-9 scale.

To address these limitations, statistical replication is needed across different symptom scales and treatment models. Specifically, the characterization of symptom change must be observed within other psychotherapy treatment models before more generalizable comments can be made about symptom change and measurement. Future similar studies seeking to characterize and compare symptom change and measurement models could determine to what extent the proportional change pattern generalizes as a measurement principle, across different treatment models and across different symptom scales. In addition, future studies seeking to research this pattern of change could also attempt to compile a meta-analytical characterization of proportional and linear change across different scales and treatment models.

Further, it is important to consider that measurement and interpretation of symptom change as a proportional function is at odds with the widely accepted use of linear statistics in psychotherapy. From one point of view, linear statistics, such as Cohen *d*, are successful as an established measurement standard that can be used to compare change estimates between trials and across clinical instruments [2]. This use of effect sizes has resulted in both enormous amounts of aggregated evidence about the effects of psychotherapy [22] and, for this reason, it is understandable clinical researchers would continue to use this standard for measuring and interpreting symptom change. However, should symptom change occur as a proportional function, the measurement and interpretation of treatment-related change would substantially improve by matching appropriate statistical analysis to the characteristics of the function of symptom change [15,45,46]. A possible solution to this dilemma would be to report both the effect size and percentage estimates of change side by side. In this way, the change that occurs in treatment can be more accurately reported, evaluated, and compared between trials.

Finally, this study does not weigh whether the change rate of 50% to 55% could be evaluated as the same treatment-related effect across individuals with severe or mild baseline symptoms. For example, a symptom reduction demonstrated by individuals with severe baseline symptoms could be interpreted as a more substantive clinical effect than an equivalent symptom reduction achieved with individuals with mild or moderate symptoms [47]. To address these limitations, additional research into the experience of individuals in treatment could determine whether individuals with different baseline symptoms consider the proportional remission pattern an equally satisfactory treatment outcome. For example, Zimmerman and colleagues [48] consider the measurement of patient functionality, positive mental health, and optimism alongside the reduction in depressive symptoms. These additional measures could verify and elaborate on the experience of individuals in treatment and nontreatment conditions, within various symptom bands, shedding more light on the universality or segmentation of the 50% to 55% improvement effect.

In summary, this study aimed to explore the underlying pattern of symptom change and compare different methods for measuring and interpreting depressive symptom change that follows treatment (Web-based psychotherapy). This study has combined evidence of increased rate of change with increased

baseline symptoms (hypothesis 1), score distributions that become increasingly skewed following treatment (hypothesis 2), and increased measurement accuracy achieved by statistical methods that assume change is proportional (hypothesis 3) to suggest that the fundamental function of symptom change is proportional. The promise of matching these characteristics of proportional symptom change to a suitable statistical analysis is important for all (1) statistical modeling and the prediction of treatment effects, (2) an improved ability to differentiate treatment and nonspecific symptom change, as well as for (3) determining an estimate of treatment-related change that will not sway with increased baseline symptoms. Replication of these preliminary findings are essential within additional depressive symptom scales, other types of psychological conditions, and across different treatment modalities.

Acknowledgments

The authors would like to acknowledge Monique Crane, Pery Karin, and the team of reviewers for their helpful and meticulous feedback.

Conflicts of Interest

None declared.

References

1. Kroenke K, Monahan PO, Kean J. Pragmatic characteristics of patient-reported outcome measures are important for use in clinical practice. J Clin Epidemiol 2015 Sep;68(9):1085-1092 [FREE Full text] [doi: 10.1016/j.jclinepi.2015.03.023] [Medline: 25962972]
2. Spring B. Evidence-based practice in clinical psychology: what it is, why it matters; what you need to know. J Clin Psychol 2007 Jul;63(7):611-631. [doi: 10.1002/jclp.20373] [Medline: 17551934]
3. Wise E. Methods for analyzing psychotherapy outcomes: a review of clinical significance, reliable change, and recommendations for future directions. J Pers Assess 2004 Feb;82(1):50-59. [doi: 10.1207/s15327752jpa8201_10] [Medline: 14979834]
4. Flay BR, Biglan A, Boruch RF, Castro FG, Gottfredson D, Kellam S, et al. Standards of evidence: criteria for efficacy, effectiveness and dissemination. Prev Sci 2005 May 16;6(3):151-175. [doi: 10.1007/s11121-005-5553-y] [Medline: 28116558]
5. Gottfredson DC, Cook TD, Gardner FE, Gorman-Smith D, Howe GW, Sandler IN, et al. Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: next generation. Prev Sci 2015 Apr 7;16(7):893-926. [doi: 10.1007/s11121-015-0555-x] [Medline: 25846268]
6. Laurenceau J, Hayes AM, Feldman GC. Some methodological and statistical issues in the study of change processes in psychotherapy. Clin Psychol Rev 2007 Jul;27(6):682-695 [FREE Full text] [doi: 10.1016/j.cpr.2007.01.007] [Medline: 17328996]
7. Titov N, Dear BF, Staples LG, Bennett-Levy J, Klein B, Rapee RM, et al. MindSpot Clinic: an accessible, efficient, and effective online treatment service for anxiety and depression. Psychiatr Serv 2015 Oct;66(10):1043-1050. [doi: 10.1176/appi.ps.201400477] [Medline: 26130001]
8. Gunn J, Elliott P, Densley K, Middleton A, Ambresin G, Dowrick C, et al. A trajectory-based approach to understand the factors associated with persistent depressive symptoms in primary care. J Affect Disord 2013 Jun;148(2-3):338-346. [doi: 10.1016/j.jad.2012.12.021] [Medline: 23375580]
9. Sobocki P, Ekman M, Agren H, Runeson B, Jönsson B. The mission is remission: health economic consequences of achieving full remission with antidepressant treatment for depression. Int J Clin Pract 2006 Jul;60(7):791-798. [doi: 10.1111/j.1742-1241.2006.00997.x] [Medline: 16846399]
10. Gyani A, Shafran R, Layard R, Clark DM. Enhancing recovery rates: lessons from year one of IAPT. Behav Res Ther 2013 Sep;51(9):597-606 [FREE Full text] [doi: 10.1016/j.brat.2013.06.004] [Medline: 23872702]
11. Altman DG, Simera I. A history of the evolution of guidelines for reporting medical research: the long road to the EQUATOR Network. J R Soc Med 2016 Feb;109(2):67-77. [doi: 10.1177/0141076815625599] [Medline: 26880653]
12. Choi S, Schalet B, Cook KF, Cella D. Establishing a common metric for depressive symptoms: linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. Psychol Assess 2014 Jun;26(2):513-527 [FREE Full text] [doi: 10.1037/a0035768] [Medline: 24548149]
13. Schalet BD, Cook KF, Choi SW, Cella D. Establishing a common metric for self-reported anxiety: linking the MASQ, PANAS, and GAD-7 to PROMIS Anxiety. J Anxiety Disord 2014 Jan;28(1):88-96 [FREE Full text] [doi: 10.1016/j.janxdis.2013.11.006] [Medline: 24508596]
14. Snyder C, Aaronson NK, Choucair AK, Elliott TE, Greenhalgh J, Halyard MY, et al. Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations. Qual Life Res 2012 Oct;21(8):1305-1314 [FREE Full text] [doi: 10.1007/s11136-011-0054-x] [Medline: 22048932]

15. Baldwin SA, Fellingham GW, Baldwin AS. Statistical models for multilevel skewed physical activity data in health research and behavioral medicine. Health Psychology 2016;35(6):552-562. [doi: 10.1037/hea0000292] [Medline: 26881287]
16. Clarke M. Standardising outcomes for clinical trials and systematic reviews. Trials 2007 Nov 26;8(1):39 [FREE Full text] [doi: 10.1186/1745-6215-8-39] [Medline: 18039365]
17. Horn SD, Gassaway J. Practice-based evidence study design for comparative effectiveness research. Med Care 2007 Oct;45(10 Supl 2):S50-S57. [doi: 10.1097/MLR.0b013e318070c07b] [Medline: 17909384]
18. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 2013 Nov 26;4:863 [FREE Full text] [doi: 10.3389/fpsyg.2013.00863] [Medline: 24324449]
19. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMC Med 2010 Mar 24;8:18 [FREE Full text] [doi: 10.1186/1741-7015-8-18] [Medline: 20334633]
20. Bower P, Kontopantelis E, Sutton A, Kendrick T, Richards DA, Gilbody S, et al. Influence of initial severity of depression on effectiveness of low intensity interventions: meta-analysis of individual patient data. BMJ 2013 Feb 26;346(feb26 2):f540-f540. [doi: 10.1136/bmj.f540] [Medline: 23444423]
21. Clark DM. Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: the IAPT experience. Int Rev Psychiatry 2011 Aug;23(4):318-327 [FREE Full text] [doi: 10.3109/09540261.2011.606803] [Medline: 22026487]
22. Newby J, McKinnon A, Kuyken W, Gilbody S, Dalgleish T. Systematic review and meta-analysis of transdiagnostic psychological treatments for anxiety and depressive disorders in adulthood. Clin Psychol Rev 2015 Aug;40:91-110 [FREE Full text] [doi: 10.1016/j.cpr.2015.06.002] [Medline: 26094079]
23. Des Jarlais DC, Lyles C, Crepaz N. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND Statement. Am J Public Health 2004 Mar;94(3):361-366. [doi: 10.2105/AJPH.94.3.361]
24. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. Int J Surg 2014 Dec;12(12):1495-1499 [FREE Full text] [doi: 10.1016/j.ijsu.2014.07.013] [Medline: 25046131]
25. Hiller W, Schindler AC, Lambert MJ. Defining response and remission in psychotherapy research: a comparison of the RCI and the method of percent improvement. Psychother Res 2012 Jan;22(1):1-11. [doi: 10.1080/10503307.2011.616237] [Medline: 21943215]
26. McMillan D, Gilbody S, Richards D. Defining successful treatment outcome in depression using the PHQ-9: a comparison of methods. J Affect Disord 2010 Dec;127(1-3):122-129 [FREE Full text] [doi: 10.1016/j.jad.2010.04.030] [Medline: 20569992]
27. Ellis PD. The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results. Cambridge, UK: Cambridge University Press; 2010.
28. Boettcher J, Hasselrot J, Sund E, Andersson G, Carlbring P. Combining attention training with internet-based cognitive-behavioural self-help for social anxiety: a randomised controlled trial. Cogn Behav Therapy 2013 Jul 30;43(1):34-48. [doi: 10.1080/16506073.2013.809141] [Medline: 23898817]
29. Kroenke K, Spitzer RL, Williams JB. The PHQ-9. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x]
30. Kroenke K, Spitzer RL, Williams JB, Monahan PO, Löwe B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Ann Intern Med 2007 Mar 06;146(5):317. [doi: 10.7326/0003-4819-146-5-200703060-00004]
31. Kessler R, Andrews G, Colpe L, Hiripi E, Mroczek D, Normand S, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002;32(6):959-976. [doi: 10.1017/S0033291702006074]
32. Driessen E, Cuijpers P, Hollon SD, Dekker JJ. Does pretreatment severity moderate the efficacy of psychological treatment of adult outpatient depression? A meta-analysis. J Consult Clin Psych 2010;78(5):668-680. [doi: 10.1037/a0020570] [Medline: 20873902]
33. Thase M, Simons AD, Cahalane J, McGeary J, Harden T. Severity of depression and response to cognitive behavior therapy. Am J Psychiatry 1991 Jun;148(6):784-789 [FREE Full text] [doi: 10.1176/ajp.148.6.784] [Medline: 2035722]
34. Fitzmaurice GM. In: Laird NM, Ware JH, editors. Applied Longitudinal Analysis. Philadelphia, PA: John Wiley & Sons; 2012.
35. Liang K, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986 Apr;73(1):13. [doi: 10.2307/2336267]
36. Castellani B, Rajaram R, Gunn J, Griffiths F. Cases, clusters, densities: Modeling the nonlinear dynamics of complex health trajectories. Complexity 2015 Sep 25;21(S1):160-180. [doi: 10.1002/cplx.21728] [Medline: 25855820]
37. Panagiotakopoulos TC, Lyras DP, Livaditis M, Sgarbas KN, Anastassopoulos GC, Lymberopoulos DK. A contextual data mining approach toward assisting the treatment of anxiety disorders. IEEE Trans Inf Technol Biomed 2010 May;14(3):567-581. [doi: 10.1109/TITB.2009.2038905] [Medline: 20071265]
38. Pocock S, Clayton TC, Stone GW. Challenging issues in clinical trial design: part 4 of a 4-part series on statistics for clinical trials. J Am Coll Cardiol 2015 Dec 29;66(25):2886-2898 [FREE Full text] [doi: 10.1016/j.jacc.2015.10.051] [Medline: 26718676]

39. Titov N, Dear BF, Staples LG, Terides MD, Karin E, Sheehan J, et al. Disorder-specific versus transdiagnostic and clinician-guided versus self-guided treatment for major depressive disorder and comorbid anxiety disorders: A randomized controlled trial. J Anxiety Disord 2015 Oct;35:88-102 [FREE Full text] [doi: 10.1016/j.janxdis.2015.08.002] [Medline: 26422822]
40. Dear B, Staples LG, Terides MD, Karin E, Zou J, Johnston L, et al. Transdiagnostic versus disorder-specific and clinician-guided versus self-guided internet-delivered treatment for generalized anxiety disorder and comorbid disorders: A randomized controlled trial. J Anxiety Disord 2015 Dec;36:63-77 [FREE Full text] [doi: 10.1016/j.janxdis.2015.09.003] [Medline: 26460536]
41. eCentreClinic. URL: https://www.ecentreclinic.org/ [accessed 2018-06-07] [WebCite Cache ID 700kbKH6d]
42. Hubbard AE, Ahern J, Fleischer NL, Van der Laan M, Lippman SA, Jewell N, et al. To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. Epidemiology 2010 Jul;21(4):467-474. [doi: 10.1097/EDE.0b013e3181caeb90] [Medline: 20220526]
43. Pan W. Akaike's information criterion in generalized estimating equations. Biometrics 2001;57(1):120-125 [FREE Full text] [doi: 10.1111/j.0006-341X.2001.00120.x]
44. Akaike H. Information theory and an extension of the maximum likelihood principle. 1973 Presented at: 2nd International Symposium on Information Theory; Sep 2-8, 1971; Tsahkadsor, Armenia, USSR.
45. Field AP, Wilcox RR. Robust statistical methods: a primer for clinical psychology and experimental psychopathology researchers. Behav Res Ther 2017 Nov;98:19-38. [doi: 10.1016/j.brat.2017.05.013] [Medline: 28577757]
46. Verkuilen J, Smithson M. Mixed and mixture regression models for continuous bounded responses using the beta distribution. J Educ Behav Stat 2016 Aug 26;37(1):82-113. [doi: 10.3102/1076998610396895]
47. Judd LL, Schettler PJ, Rush AJ, Coryell WH, Fiedorowicz JG, Solomon DA. A new empirical definition of major depressive episode recovery and its positive impact on future course of illness. J Clin Psychiatry 2016 Aug;77(8):1065-1073. [doi: 10.4088/JCP.15m09918] [Medline: 26580150]
48. Zimmerman M, McGlinchey JB, Posternak MA, Friedman M, Attiullah N, Boerescu D. How should remission from depression be defined? The depressed patient's perspective. Am J Psychiatry 2006 Jan;163(1):148-150. [doi: 10.1176/appi.ajp.163.1.148] [Medline: 16390903]

Abbreviations

AIC: Akaike information criterion
BIC: Bayesian information criterion
GEE: generalized estimated equation
ICBT: Internet-delivered cognitive behavioral therapy
QIC: quasilikelihood under the independence model criterion

Edited by J Lipschitz; submitted 23.02.18; peer-reviewed by M Subotic-Kerry, BT Tulbure; comments to author 13.04.18; revised version received 03.05.18; accepted 07.05.18; published 12.07.18

Please cite as:
Karin E, Dear BF, Heller GZ, Gandy M, Titov N
Measurement of Symptom Change Following Web-Based Psychotherapy: Statistical Characteristics and Analytical Methods for Measuring and Interpreting Change
JMIR Ment Health 2018;5(3):e10200
URL: http://mental.jmir.org/2018/3/e10200/
doi:10.2196/10200
PMID:30001999

Chapter 3

Statistical characteristics and analytical methods for measuring and interpreting symptom change in psychotherapy – a replication and elaboration study (Study 2)

This chapter concerns a first and fundamental step in the process of measuring and interpreting psychotherapy evidence, that is, the choice of analytical methods for the measurement of symptom change in treatment. The chapter describes a replication study aiming to identify the statistical features that affect the suitability and impact of different ways to measure and interpret symptom change. The study is set in a context of psychotherapy routine care and includes the examination of statistical features of symptom change across additional symptom scales. In this way, the replication study seeks not only to replicate Study 1 but to comment on the generalisability of the findings in additional psychotherapy contexts. Together, studies one and two aim to identify the features of symptom change, and the opportunities for achieving measurement that is more suited to the features of symptom change as a phenomena (gauging internal validity), as well as the application of these features across clinical contexts (gauging external validity).

Publication Status

This chapter has been submitted for publication to the *Journal of Psychotherapy Research* and is currently under review (TPSR-2019-0169).

Author contribution:

Mr Eyal Karin has designed, analysed, and wrote the study. Dr Monique Crane, Associate Professor Blake F. Dear, Professor Olav Nielssen and Dr Rony Kayrouz provided the dataset, assisted with the refinement of the manuscript, and helped frame the methodological content for a clinical audience. Professor Nick Titov oversaw the conception of the project and the drafting of the manuscript.

STUDY 2 - MEASURING AND INTERPRETING SYMPTOM CHANGE

Abstract

Background: Accurate measurement of symptom change is critical for psychotherapy research and treatment evaluation. Despite the reliance on change measurement, little research is available to explore the features of symptom change and compare the suitability of different statistical measurement models.

Objective: To explore the function of symptom change that occurs following psychotherapy, and compare the suitability of two conventional models for symptom change measurement; linear and proportional symptom change.

Methods: A treatment sample of web-based psychotherapy participants (n=6701), were used to (1) explore the statistical characteristics of depressive, anxiety and psychological distress symptom change, and (2) evaluate the fit of the linear and proportional measurement models.

Results: Findings demonstrated a strong relationship between pre-treatment scores and change magnitude; however, as a percentage change, change remained consistent across individuals of mild, moderate, and severe pre-treatment symptoms. Additional statistical features such as distributional skewness and improved outcome prediction support the measurement and symptom change through a proportional function, and the interpretation of change through percentage metrics.

Conclusions: This study suggests additional evidence about the features of symptom change, and point to new opportunities for increasing the accuracy, interpretability generalisability of clinical evidence across contexts. Implications and limitations from these results are further discussed.

Introduction

The efficacy of psychotherapy is frequently determined using standardised symptom scales which patients complete several times across the course of treatment (Laurenceau, Hayes, & Feldman, 2007; Kroenke, Monahan & Kean, 2015). Using these scales, various estimates of change are then applied to evaluate whether a particular treatment is effective, and to compare the efficacy of different types of psychotherapy (Boswell, Kraus, Miller, & Lambert, 2015; Kazdin, 1999; 2014; Lakens 2013; Rozental, Andersson, Boettcher, Ebert, Cuijpers, Knaevelsrud, & Carlbring, 2014).

Although the accurate measurement and description of symptom change is a fundamental aspect of psychotherapy research, and evidence-based clinical practice (Choi, Schalet, Cook & Cella, 2014; Frost, Reeve, Liepa, Stauffer, Hays, & Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group, 2007; Fried, Borkulo, Epskamp, Schoevers, Tuerlinckx & Borsboom, 2016), measuring and interpreting the change in symptoms through treatment is not always straightforward. In order to accurately represent symptom change, researchers are required to reduce multiple longitudinal patient scores into statistical estimates, that approximate and represent the change experienced as a result of treatment (Erekson, Horner & Lambert, 2016; Fried, van Borkulo, Epskamp, ,Schoevers, Tuerlinckx & Borsboom, 2016; Gunn, Elliott, Densley, Middleton, Ambresin, Dowrick, ., ... & Griffiths, 2013). Ideally, the selection of statistical methods would match the statistical features of the data (e.g., distributions, functions of change, ceiling and floor effects) (Hayes, Laurenceau, Feldman, Strauss, & Cardaciotto, 2007; Lin, Huang, Simon, & Liu 2016) and the purpose of treatment (e.g., remission of symptoms, prevent relapse) (Sobocki, et al., 2006; Thompson, 2002). In this way, the measurement of change will represent the effect of treatment with accuracy and validity (Lang & Altman, 2013; Loken & Gelman, 2017; Beckstead, 2014; Maul, Irribarra & Wilson, 2016; Podsakoff, MacKenzie & Podsakoff, 2012). However, the choice researchers

make to operationalise measurement can vary (Lang & Altman, 2013; Nieminen & Kaur, 2019) and result in varying degrees of measurement error and threats to the validity of evidence (Baldwin, Fellingham & Baldwin, 2016; Ebrahim, Sohani, Montoya, Agarwal, Thorlund, Mills & Ioannidis, 2014; Ng & Cribbie, 2017; Silberzahn, Uhlmann, Martin, Anselmi, Aust, Awtrey,., ... & Carlsson, 2018).

To date, few studies have explored the statistical characteristics of symptom change or tested the relative suitability of different statistical approaches to measure and interpret symptom change. Thus, there are knowledge gaps about: (1) the kind of statistical features that are likely to generalise across the scales and psychotherapy contexts, (2) the validity of frequently used statistical approaches to measure and interpret symptom change, or (3) how to quantify and describe the effect of suitable or unsuitable statistical choice.

The aim of this paper is to test, compare and evaluate the suitability of different statistical approaches for measuring symptom change and efficacy of treatment. The method involves exploring features of symptom change, such as the statistical function of symptom change (denoted as link function in statistical modelling), the distribution of scores on common symptom scales (denoted as scale in statistical modelling), and the reduction in the measurement error associated with different statistical models (denoted as model fit in statistical modelling). These features are often overlooked in clinical research (Miettunen, Nieminen, & Isohanni, 2002; Nieminen & Kaur, 2019), yet the ability to detect these features and their generalisability in the psychotherapy context can verify or refute the suitability of certain measurement approaches for treatment evaluation, and research concerning change over time (Blackwell, Honaker & King, 2017; Field & Wilcox, 2017; Lang & Altman, 2013).

Current methods for analysing symptom change in the context of psychotherapy

In psychotherapy treatment evaluation two broad statistical approaches are typically chosen by researchers to measure and interpret symptom change: (1) linear approaches and (2) proportional measurement approaches (Hiller, Schindler, & Lambert, 2012; McMillan, Gilbody, & Richards, 2010; Ng & Cribbie, 2017). These two approaches reflect distinct statistical modelling methods for longitudinal data (e.g., general or generalised statistical modelling (Fitzmaurice, Laird & Ware, 2012; Liang & Zeger, 1986), and in the context of psychotherapy, these functions represent two distinct ways to measure and evaluate symptom change and clinical efficacy (Field & Wilcox, 2017; Hiller et al., 2012).

Of the two approaches, linear approaches appear to be the most commonly used, and include methods such as linear regression, analysis of variance, *t*-tests (Miettunen, Nieminen, & Isohanni, 2002; Nieminen & Kaur, 2019), metrics such as Cohen's *d* effect sizes (Ellis, 2010; Lakens, 2013; Pek & Flora, 2018), and the reliable change index (Jacobson & Truax, 1991; Lambert & Ogles, 2009; de Beurs, Barendregt, de Heer, van Duijn, Goeree, Kloos, & Merks, 2016). From a statistical perspective, the linear approach assumes that symptom change should be measured and interpreted as a fixed change score; for example, participants improved by 5 points on a symptom scale. The markers of linear change are verified with statistical features such as the occurrence of an average symptom change quantity irrespective of the pre-treatment symptom level, and the presentation of normally distributed symptom scores, prior and following treatment (Field & Wilcox, 2017).

The second approach, the proportional approach, describes the magnitude of treatment efficacy as measured and interpreted as a percentage reduction from pre-treatment. Under the proportional approach, change is considered respective to pre-treatment symptom scores, and consequently, the efficacy of treatment is measured and interpreted as a proportional pattern

where the magnitude of change is relative to the individual symptom score prior to treatment. The markers of proportional change are verified with statistical features such as (1) non-normal distributions, (2) change magnitude that is dependent on pre-treatment scores, and (3) the presence of ceiling and floor effects (Baldwin, Fellingham, & Baldwin, 2016; Ng & Cribbie, 2017; Verkuilen & Smithson, 2012). This approach has been previously effectively employed to measure and interpret psychotherapy (Cuijpers, Karyotaki, Weitz, Andersson, Hollon, & van Straten, 2014; McMillan et al., 2010; Hiller et al., 2012) as well as several related clinical research streams including pain perception (Price, McGrath, Rafii, & Buckingham, 1983; Farrar, Portenoy, Berlin, Kinman & Strom, 2000), suicidal ideation (Bruce, Ten Have, Reynolds, Katz, Schulberg, Mulsant, ... & Alexopoulos, 2004), psychopharmacology trials (Rush, Kraemer, Sackeim, Fava, Trivedi, Frank & Schatzberg, 2006) and health psychology (Baldwin, et al, 2016).

These two approaches reflect two distinct and competing ways to statistically model, measure, and interpret the longitudinal change through therapy (Hiller et al., 2012; Ng & Cribbie, 2017). In the broader statistical literature, the distinction between the two approaches is referred to as the difference between general or generalized statistical modelling (Fitzmaurice, Laird & Ware, 2012; Liang & Zeger, 1986). To date, however, whilst these two approaches are common and well established, little research has explored their relative suitability in the context of psychotherapy and the measurement of treatment efficacy.

A preliminary study conducted in the context of web-based psychotherapy (Karin et al., 2018), on which this study is based, tested the suitability of different measurement approaches by comparing the statistical characteristics of symptom change in a large psychotherapy sample (n=996), and the effect different statistical choices have on the validity and accuracy of results. Through this exploration, statistical features such as (1) symptom scores distribution pre and post-treatment, (2) the function of symptom change over time, or (3) the measurement error

that results from different approaches, were used to support or challenge the suitability of different measurement approaches. This research identified implicit and potentially generalizable statistical features about symptom change. Specifically, strong evidence of skewness in symptom scores at pre-treatment and post-treatment, and evidence of better model fit, together implied that the proportional approach was more valid, accurate and suitable for evaluating the efficacy of psychotherapy for depression. In contrast, the selection of the linear alternative for that context resulted in: (1) substantial increase in the measurement of error (an overall increase of 38%), (2) effects such as regression to the mean, and (3) varied conclusions about the efficacy of treatment even when the same treatment was applied. These findings highlighted how symptom change might be dependent on initial symptom severity, which violates one of the major assumptions of the linear approach (Lang & Altman, 2013; Ng & Cribbie, 2017).

The present study

The present study aims to replicate and extend the earlier work (Karin et al., 2018) by comparing the suitability of the linear and proportional approaches using a large naturalistic sample of patients from a national online mental health service providing psychological treatment for anxiety and depression. This study will also test whether the characteristics of change symptom (e.g. function of symptom change), previously observed in depressive symptom scales (Patient health questionnaire 9 item; PHQ-9; Kroenke, Spitzer & Williams, 2001), in the context of web-based clinical trials (Karin et al., 2018), also apply to symptoms of psychological distress as measured by the Kessler-10 Item Scale (K-10; Kessler, Andrews, Colpe, Hiripi, Mroczek, Normand, & Zaslavsky, 2002), and anxiety as measured by the Generalized Anxiety Disorder - 7 Item Scale

(GAD-7; Spitzer, Kroenke, Williams, & Löw, 2006). Further, the current study also investigates the replicability of these features within a cross-fold replication analysis.

Based on the findings of previous research (Karin et al., 2018) it was hypothesised that; (H1) participants with different levels of initial symptom severity would exhibit proportional change in their symptoms consistent with the proportional, rather than linear approach (H2). Second, consistent with the clinical rationale that effective therapy can reduce symptoms from a diverse range of pre-treatment severity towards the same endpoint of minimal symptoms (Karin et al., (Study 1); Sobocki, Ekman, Ågren, Runeson, & Jönsson, 2006), the statistical distributions of symptoms are hypothesized to show floor effects, and consequent positive distributional skewness. Third, it was hypothesized that analytical methods that account for the proportional remission would increase the accuracy for predicting symptom treatment outcomes (H3) in line with studies such as Baldwin and colleagues (Baldwin, et al., 2016).

Method

The sample

The present study employed clinical participant data from a national digital mental health service, the MindSpot Clinic (mindspot.org.au). The psychotherapeutic intervention used to generate the data was delivered over eight weeks and comprises five lessons that covered: (1) the cognitive-behavioral model and symptom identification; (2) thought monitoring and challenging; (3) de-arousal strategies and behavioral activation; (4) graduated exposure; and (5) relapse prevention. More detail about the assessment procedures, treatment courses, and the methods of maintaining patient safety are described elsewhere (Nielssen et al., 2015; Titov et al., 2016).

The use of online psychotherapeutic data is a unique opportunity for investigating the trajectory of change over time through treatment. This is because the presentation of treatment materials, prescribed patient tasks, introductory and reminder emails, web-based engagement, and outcome measurement surveys, are highly standardized through protocols and automated (Nielssen et al., 2015; Titov et al., 2016). By reducing the variability associated with treatment delivery, through standardized protocols and online materials, the resulting variance in treatment outcomes is attributable more to the individual differences of cases, rather than the delivery of treatment.

The total employed sample consistent of 6701 participants who initiated treatment, with 64% of the sample also available at post-treatment (n=4271). The symptom change estimates were captured by measuring symptoms before the beginning of treatment (considered pre-treatment), and after the eight-week course (considered post-treatment).

This sample was then randomly allocated into five subgroups, each including over 1340 participants at pre-treatment, and over 840 complete measurements at post-treatment. These random subsamples were used to cross-validate several of the results, such as the characteristics of symptom change, in order to establish such characteristics as a genuine phenomenon.

In combination, this sample includes a sample of treatment-seeking individuals registering for treatment within a time window of 36 months (Jan 2014- Dec 2016). Sample demographic information is presented in Table 1, including test statistics obtained to check for successful randomisation within each stratum of the sample.

Table 1: Randomisation of cross-validation samples and participant descriptives

		Available sample at post treatment	Randomisation test statistic
Total sample (n=6701)		n= 4271 (64%)	
Replication sample 1 (n=1341)		n=842 (64%)	
Replication sample 2 (n=1340)		n=846 (64%)	$\chi2$ =3.768, p = 0.438
Replication sample 3 (n=1340)		n=843 (64%)	
Replication sample 4 (n=1340)		n=846 (64%)	
Replication sample 5 (n=1340)		n=848 (64%)	
Demographical variables	Mean (SD)	Count (% of total)	Randomisation test statistic
Average age (SD)	37.57 (10.9)		$\chi2$ =3.768, p = 0.438
Completed 1/5 modules		513 (8%)	
Completed 2/5 modules		715 (11%)	
Completed 3/5 modules		718 (11%)	$\chi2$ =7.533, p = 0.962
Completed 4/5 modules		653 (10%)	
Completed course (5/5)		4102 (61%)	
In a relationship		4458 (67%)	$\chi2$ =0.546, p = 0.969
Employment (employed)		4908 (73%)	$\chi2$ =0.755, p = 0.944
Education (Tertiary)		3239 (49%)	$\chi2$ =3.952, p = 0.413
Gender (Female)		4866 (73%)	$\chi2$ =6.803, p = 0.147
Comorbidity (GAD-7 $\leq$ 8 and PHQ-9 $\leq$ 10)		3437 (51%)	$\chi2$ =2.976, p = 0.562

PHQ-9 - Patient health questionnaire, nine-item scale; GAD-7 – Generalised anxiety disorder scale, seven-item scale.

Symptom measures

The primary outcome measures for the sample were comprised of standardised symptom scales scores. These included the PHQ-9, GAD-7, and the K-10. The psychometric properties of the scales, derived from the current sample, are presented below in Table 2.

Table 2: Assessed Psychometric properties of symptom measures

Scale	Primary pathology measured	Cited origin	Range	Interpretation of symptom severity bands	Cut-off indicative of the clinical range	Internal consistency (Cronbach's α)**	Intraclass correlation coefficient*
Patient Health Questionnaire (9 items)	Depression	Kroenke et al., 2001	0-27	0-4 Minimal 5-9 Mild 10-14 Moderate 15-19 Moderately severe 20-27 Severe	10	0.848	0.716
Generalized Anxiety Disorder scale (7-items)	Anxiety	Spitzer et al., 2006	0-21	0-4 Minimal 5-9 Mild 10-14 Moderate 15-21 Severe	8	0.849	0.744
The Kessler Psychological Distress Scale (10 items)	Psychological distress	Kessler et al,. 2002	0-40 (denoted 10-50)***	0 - 5 Low 6-11 Moderate 12-19 High 20-40 Very high	20 (denoted 10) ***	0.83	0.71

*Estimates identified by comparing patient scores during assessment intake and then again at the point of pre-treatment scores 4-8 weeks later; Estimate is based on a two-way random, single score analysis of items over time; **Proposed cut-offs by the authors of the original papers; *** Values were reduced to by a constant of 10 to reduce the redundant floor value of 10 and model the effective range of the data.

Finally, all three dependent variables were added a small constant (0.001), to ensure that plausible values of zero symptoms were represented in the model when using a proportional function with a logarithmic link function.

Analytical plan

All statistical analyses were conducted using SPSS software version 22 (IBM, 2013) and R version. Consistent with the previous trial (Karin et al., 2018), the features of symptom change were explored through three separate steps, each corresponding to the three hypotheses.

In the first step, the H1 was tested to explore the magnitude of symptom change in groups of individuals that presented with different pre-treatment symptom severity bands. Under the assumption of proportional symptom change, a positive relationship was predicted between pre-treatment severity and the rate of symptom change, such that the rate of change would be greater for individuals with higher pre-treatment severity. For example, within the PHQ-9, the amount of symptom reduction experienced by individuals with moderate pre-treatment depression symptoms (with a score band of 10 to 14), would increase as an absolute amount in each category relatively to the individuals with levels of severe depression symptoms (within the score band of 20 to 27). However, under the proportional approach, the rate of symptom improvement was hypothesised to present as a constant proportional, percentage change across individuals with different pre-treatment symptom severities.

The magnitude of clinical improvement was statistically approximated through longitudinal generalised estimated equation models (GEE; Liang & Zeger, 1986), and examined within each severity band of pre-treatment severity. GEE models are suitable for measuring the average longitudinal rate of symptom change without additional interpretation constraints that can occur under alternative models such as mixed models (e.g. random intercept/slope), however both methods are highly consistent in their ability to estimate and

test the change over time that occurs between two time points (Hubbard et al., 2010; Karin et al., 2018). Estimates of proportional improvement were generated through longitudinal GEE models, specifying a *log* link function and a gamma scale. Under this generalised longitudinal analytical approach, symptom improvement was measured as the relative change from pre-treatment scores (e.g. 50% of pre-treatment scores) estimated through the log-linear change coefficient (exp β). This was done for each of the five-replication sub-samples. In contrast, the rate of symptom improvement as a fixed score (e.g., 5 points on the PHQ-9) was estimated through linear longitudinal GEE models specifying *identity* link function and normal scale. Within these linear models, the estimate of symptom change is represented as an average fixed score (e.g., 5 points), which is meant to reflect the amount of symptom change for the entire sample. These linear longitudinal models were also conducted separately for each band of pre-treatment severity score and within each of the five replication sub-samples.

In the second step, to test H2 that distributions of pre-treatment and post-treatment depression, anxiety, and psychological distress symptoms show positive skew, the distributions of all scale scores at both pre-treatment and post-treatment were tested for the magnitude and significance of score distribution skewness. In this step, if the dataset would present normal distribution of symptom scores at both time points, the symptom change over time would be considered as linear. In contrast, if symptoms changed as a proportional function from pre-treatment, positive skewness should be observed, particularly at post-treatment, where individuals from various initial pre-treatment symptoms scores are likely to concentrate towards the same minimal symptoms score band, that is, a *floor effect* (Baldwin et al., 2016). Graphical and numerical explorations of pre-post score distributions are included.

In the third step, H3 was tested by comparing the measurement error associated with modelling symptom change using either the linear or the proportional approaches. The model measurement error from each approach was estimated by comparing the two (linear and

proportional) predicted post-treatment score, against the known outcome for each patient at post-treatment. The difference between the expected symptom outcome and actual treatment outcome for each approach represents the measurement error for each approach, akin to residual score and measurement error variance.

The pattern of residuals from each approach was then explored in three ways. First, the quantity of residuals was compared as an overall measurement error quantity, akin to the estimation of total model variance. Second, the measurement residuals were plotted graphically against the pre-treatment symptoms. Any association between the rate of measurement error and pre-treatment score is considered indicative of systematic measurement bias associated with the selection of either the linear or the proportional approach. Third, the dispersions of the respective model residuals were explored through quantile-quantile plots, indicating support for the presence of normal (linear approach) or gamma (proportional approach) distribution.

Results

H1: exploring the magnitude of symptom change

H1 was tested by investigating the symptom change of individuals with different pre-treatment symptom severities. Results indicated that change in symptoms over time was dependent on pre-treatment symptoms; supporting the use of the proportional approach. Table 3.1 collates the estimates of change in depressive symptoms (PHQ-9) by initial pre-treatment symptom severity. The estimate of change as a fixed score was illustrated with a column marked absolute symptom score, and a linear measure of effect sizes (Cohen's *d*), while proportional change estimates are marked under a percentage improvement estimate. The estimate of change was also reported for each band of symptom severity and within each of the replication samples.

Within-group effect sizes were calculated according to the formula: $\frac{X_1 - X_2}{SD_{pooled}}$ (Lakems, 2013) where X1 is the pre-treatment score and X2 is the post-treatment score of the group; SDpooled was calculated as: $\sqrt{\frac{(N_1 - 1) \times {SD_1}^2 + (N_2 - 1) \times {SD_2}^2}{N_1 + N_2 - 2}}$, where N1 is the sample size at pre-treatment, N2 is the sample size at post-treatment, SD1 is the standard deviation at pre-treatment, and SD2 is the standard deviation of the post-treatment. Percentage change and difference scores were taken from the GEE model estimated marginal means.

Table 3.1 - PHQ-9 depressive symptom remission estimated as either fixed average or proportional percentage improvement

Baseline Severity score band at pre-treatment		Difference score *[95%CI]*	Percentage improvement *[95%CI]*	effect size *(Cohen's d) [95%CI]*
Severe	Replication 1 (n=171)	10.04 [8.69,11.4]	45% [39%,51%]	2.28 [1.46, 2.58]
	Replication 2 (n=172)	10.27 [8.85,11.69]	46% [39%,52%]	2.26 [1.43, 2.57]
	Replication 3 (n=172)	10.71 [9.3,12.12]	48% [42%,54%]	2.33 [1.49, 2.64]
	Replication 4 (n=172)	10.71 [9.46,11.96]	49% [43%,54%]	2.61 [1.69, 2.94]
	Replication 5 (n=170)	11.89 [10.56,13.21]	54% [47%,59%]	2.77 [1.77, 3.1]
Moderately severe / High	Replication 1 (n=303)	8.54 [7.75,9.32]	51% [46%,55%]	2.47 [1.59, 2.7]
	Replication 2 (n=302)	8.3 [7.59,9.01]	49% [45%,53%]	2.61 [1.7, 2.85]
	Replication 3 (n=302)	8.29 [7.52,9.06]	49% [44%,53%]	2.42 [1.56, 2.66]
	Replication 4 (n=302)	7.77 [7.01,8.53]	47% [42%,51%]	2.3 [1.49, 2.53]
	Replication 5 (n=304)	8.6 [7.83,9.38]	51% [46%,56%]	2.5 [1.6, 2.74]
Moderate	Replication 1 (n=405)	5.44 [4.93,5.96]	46% [41%,50%]	1.91 [1.28, 2.09]
	Replication 2 (n=407)	5.7 [5.18,6.22]	47% [42%,51%]	1.96 [1.32, 2.14]
	Replication 3 (n=405)	6.23 [5.75,6.72]	52% [48%,56%]	2.3 [1.57, 2.5]
	Replication 4 (n=406)	6.17 [5.7,6.65]	51% [47%,55%]	2.32 [1.6, 2.52]
	Replication 5 (n=404)	5.94 [5.42,6.46]	50% [45%,54%]	2.08 [1.4, 2.27]
Mild	Replication 1 (n=338)	3.17 [2.76,3.58]	44% [38%,50%]	1.37 [1, 1.55]
	Replication 2 (n=334)	3.56 [3.14,3.97]	49% [43%,54%]	1.53 [1.1, 1.72]
	Replication 3 (n=337)	3.62 [3.22,4.01]	50% [44%,56%]	1.59 [1.16, 1.78]
	Replication 4 (n=334)	3.2 [2.76,3.64]	44% [38%,50%]	1.31 [0.93, 1.49]
	Replication 5 (n=335)	2.83 [2.35,3.3]	40% [33%,46%]	1.09 [0.76, 1.26]
Minimal	Replication 1 (n=124)	0.67 [0.18,1.16]	26% [4%,43%]	0.37 [0.29, 0.64]
	Replication 2 (n=125)	0.62 [0.16,1.08]	23% [4%,38%]	0.36 [0.28, 0.64]
	Replication 3 (n=124)	0.74 [0.25,1.23]	28% [7%,44%]	0.39 [0.31, 0.66]
	Replication 4 (n=126)	0.85 [0.39,1.31]	30% [12%,44%]	0.5 [0.38, 0.77]
	Replication 5 (n=127)	0.73 [0.31,1.16]	29% [10%,43%]	0.45 [0.36, 0.72]
Overall sample (all severity bands)	Replication 1 (n=1341)	5.67 [5.34,6.01]	47% [45%,50%]	1.05 [1.11, 1.14]
	Replication 2 (n=1340)	5.84 [5.5,6.17]	48% [45%,51%]	1.08 [1.15, 1.17]
	Replication 3 (n=1340)	6.09 [5.76,6.42]	50% [48%,53%]	1.12 [1.2, 1.21]
	Replication 4 (n=1340)	5.85 [5.52,6.17]	48% [46%,51%]	1.1 [1.19, 1.19]
	Replication 5 (n=1340)	5.98 [5.65,6.31]	50% [47%,52%]	1.11 [1.21, 1.2]

PHQ-9 - Patient Health Questionnaire -9 Item

The data presented in Table 3.1 and Table 3.2 indicate that pre-treatment symptom severity was highly related to the rate of symptom change when presented as an average fixed score. Specifically, the average amount of symptom change increases as pre-treatment symptom severity increases with, for example, approximately 4 times as much change among patients with severe symptoms (PHQ-9 range: 10.04 to 11.89; GAD-7 range: 8.16 to 8.34), compared with mild symptoms (PHQ-9 range: 2.83 to 3.62; GAD-7 range: 2.81 to 3.23). This pattern was also reflected in the use of Cohen's *d* effect size. In contrast, the estimate of change using the proportional approach, expressed as a percentage of change, was relatively consistent for all participants irrespective of initial pre-treatment severity. For example, people with severe symptoms exhibited similar amounts of change (PHQ-9 range: 45% to 54%; GAD-7 range: 47% to 54%) to those with mild symptoms (PHQ-9 range: 40% to 50%; GAD-7 range: 41% to 46%). Within the minimal pre-treatment symptom band, lower and slightly dissimilar estimates of percentage change were noted (23%-30%) although these were also minimal as shifts of absolute scores (0.67-0.85 of a point).

The data presented in Table 3.3, detailing the change in the K-10, measuring general psychological distress, was consistent with that found in Tables 3.1 and 3.2 for anxiety and depression. Specifically, the amount of change increased with pre-treatment severity when expressed as an averaged fixed score or effect size. However, when the amount of symptom change was measured as a percentage metric (i.e., proportional approach), a much more uniform estimate of change was observed, irrespective of initial symptom severity.

Table 3.2 – GAD-7 anxiety symptom remission estimated as either fixed average or proportional percentage improvement

Baseline Severity score band at pre-treatment		Difference score *[95%CI]*	Percentage improvement *[95%CI]*	effect size *(Cohen's d) [95%CI]*
Severe	Replication 1 (n=362)	8.16 [7.41,8.9]	47% [43%,51%]	2.19 [1.46, 2.4]
	Replication 2 (*n*=366)	8.9 [8.22,9.58]	51% [47%,55%]	2.44 [1.68, 2.65]
	Replication 3 (*n*=397)	9.02 [8.32,9.71]	52% [48%,56%]	2.47 [1.66, 2.68]
	Replication 4 (*n*=380)	9.34 [8.7,9.97]	54% [50%,57%]	2.68 [1.85, 2.9]
	Replication 5 (*n*=375)	9.1 [8.41,9.8]	52% [48%,56%]	2.48 [1.68, 2.69]
Moderate	Replication 1 (*n*=438)	5.83 [5.37,6.29]	49% [45%,52%]	2.11 [1.45, 2.29]
	Replication 2 (*n*=432)	5.64 [5.14,6.15]	47% [43%,51%]	2.02 [1.35, 2.21]

	Replication 3 (*n*=417)	6.18 [5.7,6.66]	52% [48%,56%]	2.25 [1.53, 2.44]
	Replication 4 (*n*=424)	5.88 [5.37,6.38]	49% [44%,53%]	2.09 [1.4, 2.27]
	Replication 5 (*n*=402)	6.19 [5.69,6.7]	51% [47%,55%]	2.27 [1.54, 2.47]
Mild	Replication 1 (*n*=393)	2.88 [2.48,3.29]	41% [35%,47%]	1.25 [0.88, 1.42]
	Replication 2 (*n*=407)	3.23 [2.87,3.6]	46% [41%,51%]	1.46 [1.06, 1.63]
	Replication 3 (*n*=396)	2.95 [2.58,3.33]	43% [37%,48%]	1.33 [0.95, 1.5]
	Replication 4 (*n*=397)	3.23 [2.85,3.6]	46% [40%,51%]	1.45 [1.05, 1.63]
	Replication 5 (*n*=429)	2.81 [2.43,3.18]	41% [35%,46%]	1.2 [0.86, 1.36]
Minimal	Replication 1 (*n*=148)	0.52 [0,1.03]	18% [-2%,34%]	0.26 [0.19, 0.51]
	Replication 2 (*n*=135)	0.87 [0.42,1.33]	30% [12%,44%]	0.51 [0.39, 0.78]
	Replication 3 (*n*=130)	0.88 [0.49,1.27]	34% [17%,47%]	0.55 [0.45, 0.82]
	Replication 4 (*n*=138)	0.41 [-0.11,0.93]	14% [-6%,31%]	0.21 [0.15, 0.47]
	Replication 5 (*n*=134)	0.37 [-0.2,0.94]	13% [-10%,31%]	0.18 [0.13, 0.45]
Combined sample	Replication 1 (*n*=1341)	4.98 [4.68,5.29]	46% [43%,49%]	1.03 [1.08, 1.12]
	Replication 2 (*n*=1340)	5.31 [5.02,5.6]	49% [46%,51%]	1.11 [1.18, 1.2]
	Replication 3 (*n*=1340)	5.52 [5.22,5.81]	50% [47%,53%]	1.15 [1.24, 1.24]
	Replication 4 (*n*=1340)	5.5 [5.21,5.79]	50% [47%,52%]	1.14 [1.24, 1.24]
	Replication 5 (*n*=1340)	5.33 [5.04,5.62]	49% [46%,51%]	1.09 [1.19, 1.19]

GAD-7 – Generalized Anxiety Disorder-7-Item Scale

Table 3.3 – K-10 psychological distress symptom remission estimated as either fixed average or relative improvement

Baseline Severity score band at pre-treatment		Difference score *[95%CI]*	Percentage improvement *[95%CI]*	effect size (*Cohen's d*) *[95%CI]*
Severe	Replication 1 (*n*=579)	8.16 [7.41,8.9]	36% [32%,40%]	1.51 [1.11, 1.66]
	Replication 2 (*n*=572)	8.9 [8.22,9.58]	39% [35%,42%]	1.5 [1.08, 1.65]
	Replication 3 (*n*=594)	9.02 [8.32,9.71]	37% [34%,41%]	1.6 [1.16, 1.75]
	Replication 4 (*n*=598)	9.34 [8.7,9.97]	37% [34%,40%]	1.59 [1.17, 1.74]
	Replication 5 (*n*=574)	9.1 [8.41,9.8]	40% [37%,44%]	1.64 [1.21, 1.79]
Moderately severe/High	Replication 1 (*n*=487)	5.83 [5.37,6.29]	33% [29%,37%]	1.24 [0.85, 1.39]
	Replication 2 (*n*=504)	5.64 [5.14,6.15]	39% [34%,43%]	1.48 [1.02, 1.63]
	Replication 3 (*n*=473)	6.18 [5.7,6.66]	39% [35%,43%]	1.51 [1.06, 1.67]
	Replication 4 (*n*=471)	5.88 [5.37,6.38]	39% [34%,43%]	1.47 [1.02, 1.63]
	Replication 5 (*n*=492)	6.19 [5.69,6.7]	37% [33%,41%]	1.43 [0.98, 1.59]
Moderate	Replication 1 (*n*=209)	2.88 [2.48,3.29]	32% [21%,41%]	0.75 [0.51, 0.97]
	Replication 2 (*n*=215)	3.23 [2.87,3.6]	33% [25%,41%]	0.99 [0.7, 1.21]
	Replication 3 (*n*=220)	2.95 [2.58,3.33]	40% [32%,47%]	1.22 [0.86, 1.44]
	Replication 4 (*n*=211)	3.23 [2.85,3.6]	33% [24%,41%]	0.95 [0.66, 1.17]
	Replication 5 (*n*=217)	2.81 [2.43,3.18]	25% [17%,33%]	0.69 [0.48, 0.89]
Minimal	Replication 1 (*n*=62)	0.52 [0,1.03]	26% [3%,43%]	0.5 [0.36, 0.88]
	Replication 2 (*n*=48)	0.87 [0.42,1.33]	1% [-43%,31%]	-0.04 [-0.03, 0.39]
	Replication 3 (*n*=53)	0.88 [0.49,1.27]	-10% [-64%,27%]	-0.13 [-0.09, 0.3]
	Replication 4 (*n*=59)	0.41 [-0.11,0.93]	37% [22%,50%]	0.83 [0.77, 1.23]
	Replication 5 (*n*=56)	0.37 [-0.2,0.94]	5% [-36%,33%]	0.08 [0.06, 0.48]
Combined sample	Replication 1 (*n*=1341)	6.03 [5.48,6.56]	35% [31%,38%]	0.87 [0.84, 0.96]
	Replication 2 (*n*=1340)	6.7 [6.14,7.24]	38% [35%,41%]	0.94 [0.9, 1.03]
	Replication 3 (*n*=1340)	6.69 [6.14,7.21]	38% [35%,41%]	0.96 [0.94, 1.05]
	Replication 4 (*n*=1340)	6.59 [6.05,7.1]	37% [34%,40%]	0.93 [0.91, 1.02]
	Replication 5 (*n*=1340)	6.54 [6.01,7.05]	38% [35%,40%]	0.94 [0.94, 1.03]

K-10 - The Kessler 10-Item Scale

H2: The distributions of symptoms are hypothesized to show floor effects and consequent positive skewness

Confirming H2, histograms (Figures 1, 2 and 3) show patterns of symptom score dispersion with clear positive skewness at post-treatment, as well as bounding of symptoms at the minimal symptom range. In contrast, pre-treatment scores are dispersed between different bands of pre-treatment severity, with the majority of scores dispersing around moderate symptoms and no clear evidence of multi-modality or skewness. Figures 1-3 illustrate the distribution of depression symptom scores, anxiety symptom scores, and levels of psychological distress for pre-treatment and post-treatment (PHQ-9, GAD-7, and K-10, respectively). The red lines are the pre-treatment score dispersions for the five replications and the blue lines reflect the distribution of the post-treatment symptom scores for each of the five replications. As can be observed, across the five replications, the pre-treatment distribution demonstrates a normal distribution of scores for all three symptom outcomes measures. In contrast to pre-treatment scores however, the distribution of post-treatment scores demonstrate positive skewness as symptoms show average reductions in severity, and the proportion of people continuing to experience higher levels symptoms reduces.

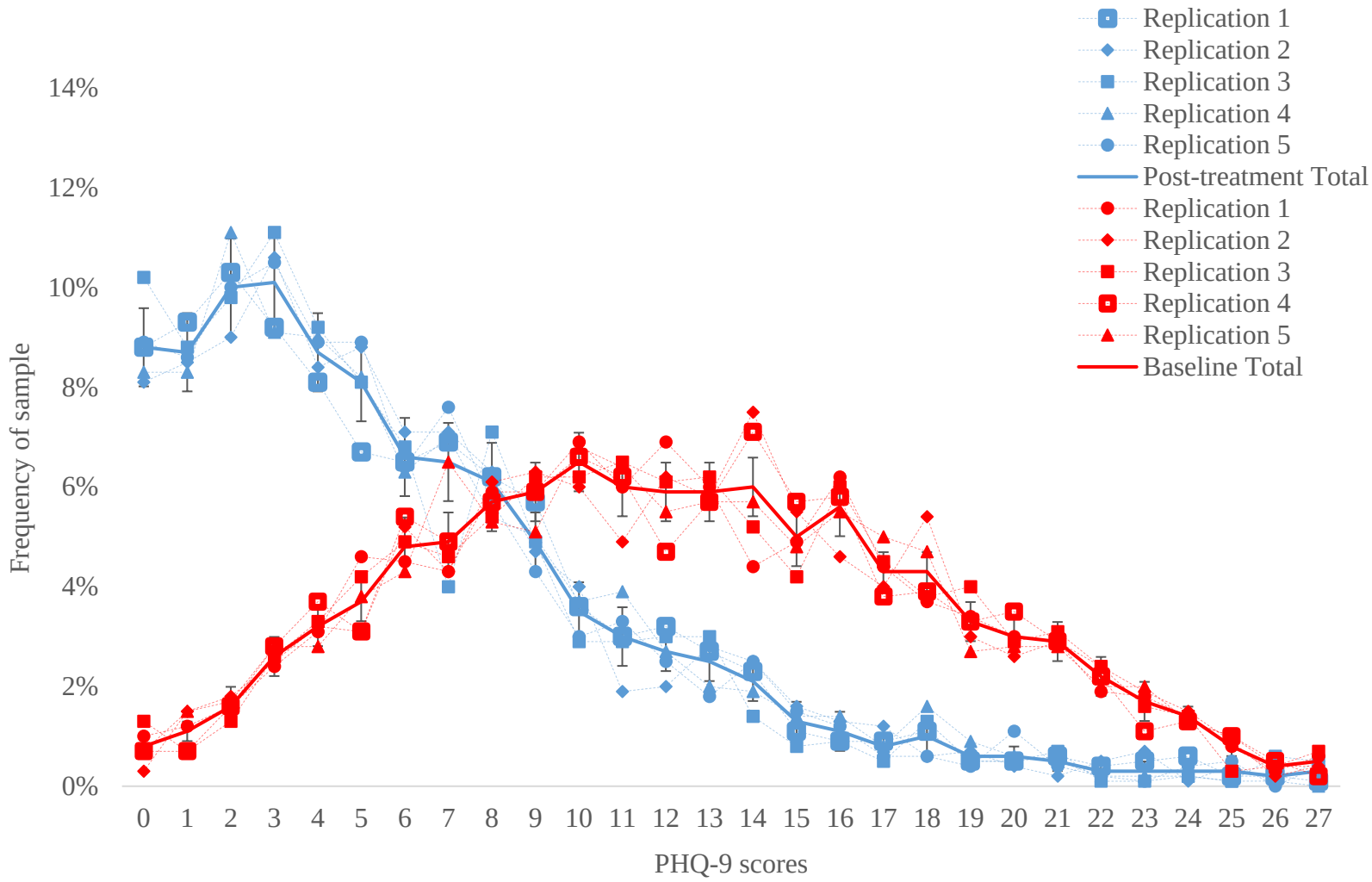


Figure 1: *Depressive symptom (PHQ-9) score distribution; prior and following treatment*

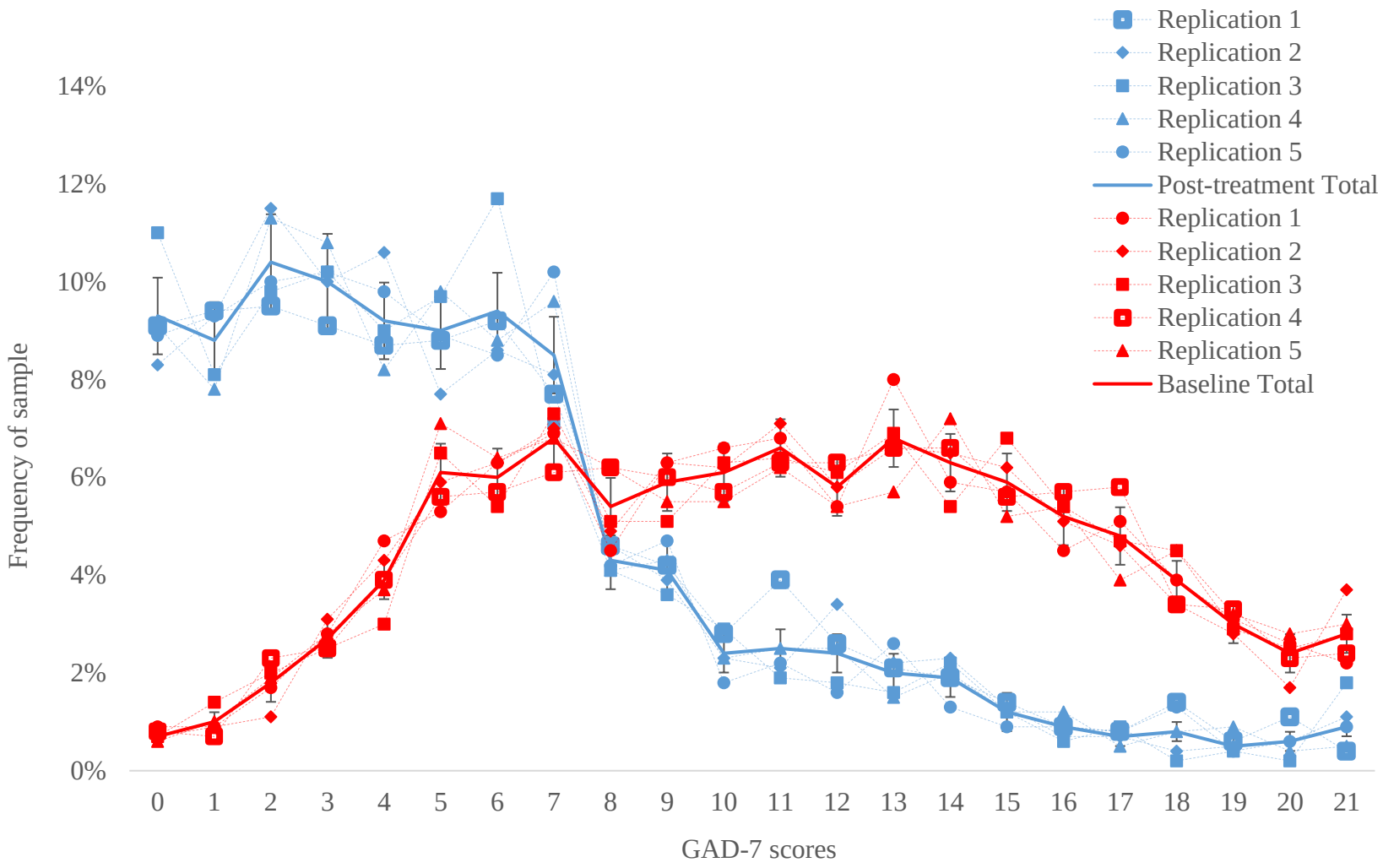


Figure 2: *Generalised anxiety symptom (GAD-7) score distribution; prior and following treatment*

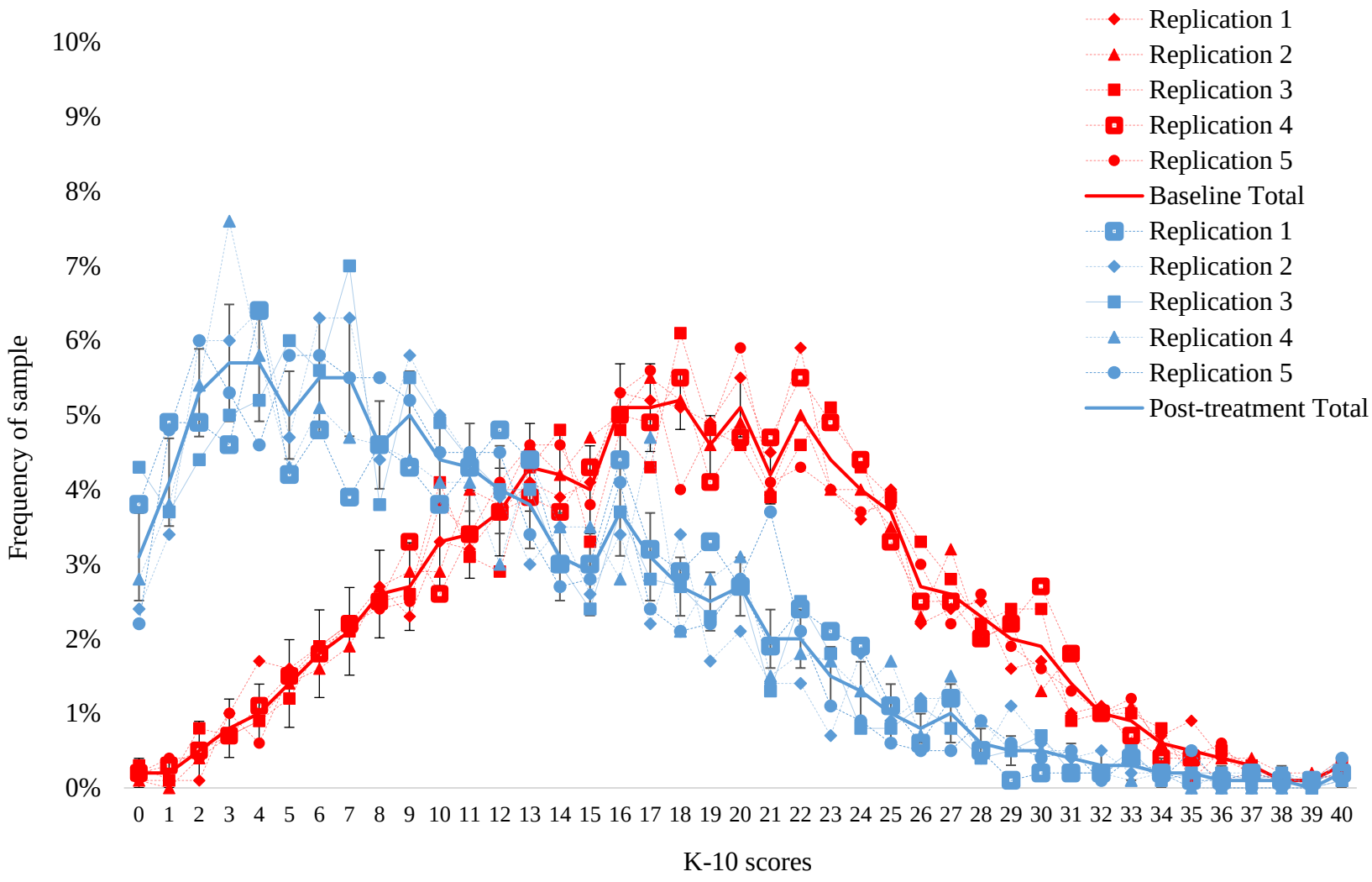


Figure 3: *Psychological distress (K-10) score distribution; prior and following treatment*

Means and SDs and estimates of skewness statistics and standard error (SE) are shown in Table 4, illustrating statistically significant positive skewness at post-treatment, for each of the scales and within each of the cross-fold replication. Taken together, both graphical and numerical descriptions of the symptom score distributions, which indicate significant skewness within the distributions, support the selection of the proportional approach for measuring and predicting symptom change.

Table 4: Overall characteristics of symptom presentations and symptom change

	Sample	Pre-treatment average (SD)	Pre-treatment skewness statistics (SE)	Post-treatment average (SD)	Post-treatment skewness statistics (SE)
PHQ-9	Replication 1	12.23 (5.88)	0.19 (0.067)**	6.25 (5.26)	1.159 (0.084)**
	Replication 2	12.39 (5.83)	0.156 (0.067)**	6.28 (5.28)	1.299 (0.084)**
	Replication 3	12.3 (5.88)	0.159 (0.067)**	5.96 (5.24)	1.369 (0.084)**
	Replication 4	12.3 (5.78)	0.173 (0.067)**	6.21 (5.11)	1.084 (0.084)**
	Replication 5	12.25 (5.89)	0.176 (0.067)**	6.05 (5.16)	1.305 (0.084)**
GAD-7	Replication 1	10.98 (5.07)	0.026 (0.067)	5.86 (4.75)	1.007 (0.084)**
	Replication 2	11.04 (5.05)	0.083 (0.067)	5.62 (4.6)	1.14 (0.084)**
	Replication 3	11.15 (5.15)	-0.007 (0.067)	5.43 (4.52)	1.232 (0.084)**
	Replication 4	11.12 (5.07)	-0.01 (0.067)	5.57 (4.47)	1.091 (0.084)**
	Replication 5	11.03 (5.19)	0.085 (0.067)	5.62 (4.59)	1.187 (0.084)**
K-10	Replication 1	18.14 (7.5)	0.134 (0.067)	11.41 (8.02)	0.682 (0.084)**
	Replication 2	18.32 (7.5)	0.196 (0.067)**	11.02 (8.14)	1.007 (0.084)**
	Replication 3	18.41 (7.61)	0.068 (0.067)	10.98 (7.89)	0.855 (0.084)**
	Replication 4	18.29 (7.55)	0.042 (0.067)	11.11 (7.75)	0.62 (0.084)**
	Replication 5	18.14 (7.62)	0.167 (0.067)**	10.88 (7.72)	0.848 (0.084)**

** Statistical significance at $\alpha<0.01$; SD – Standard deviation; SE – Standard error

The distributions of post-treatment symptom scores were also explored through quantile-quantile plots (Loy, Follett & Hoffmann, 2016). These plots evaluate the dispersion density of observed scores against a theoretical distribution, and in this way, are evaluated as following a normal distribution, or an alternate gamma distribution. Figure 4, (depressive scores; PHQ-9), Figure 5, (psychological distress scores; K-10), and Figure 6 (GAD-7 scores; anxiety) illustrate the deviation of post-treatment scores from either the normal or Gamma distribution, where minimal deviation implies improved fit. Together, these figures demonstrate an overall closer fit of post-treatment scores to the gamma scale, than the normal scale.

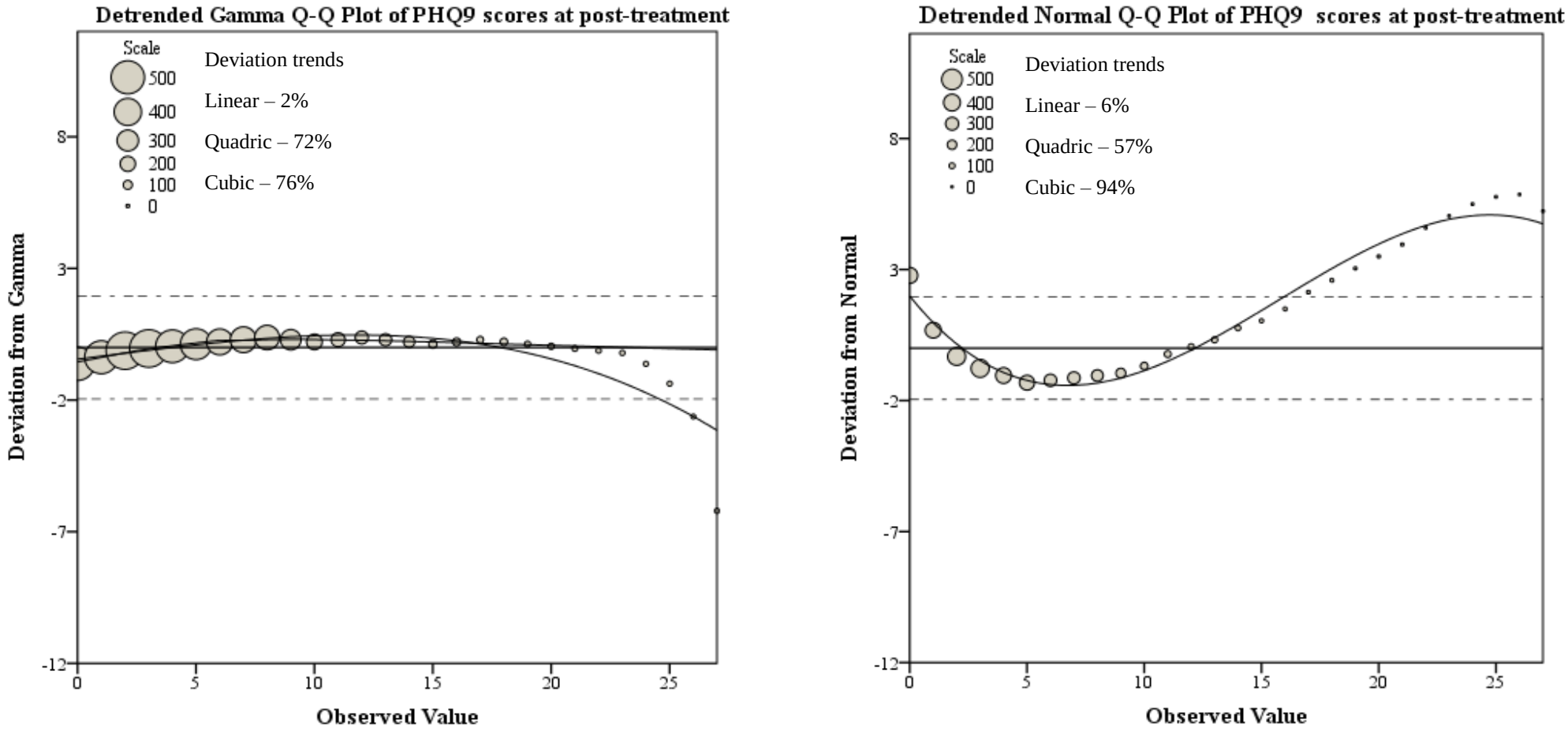


Figure 4: *De-trended Quantile-Quantile plots of post treatment PHQ-9 depressive symptoms; the fit of the observed scores and the Gamma distribution is presented on the left; the fit of the observed scores with a normal distribution is fitted on the right; Q-Q : Quantile-Quantile*

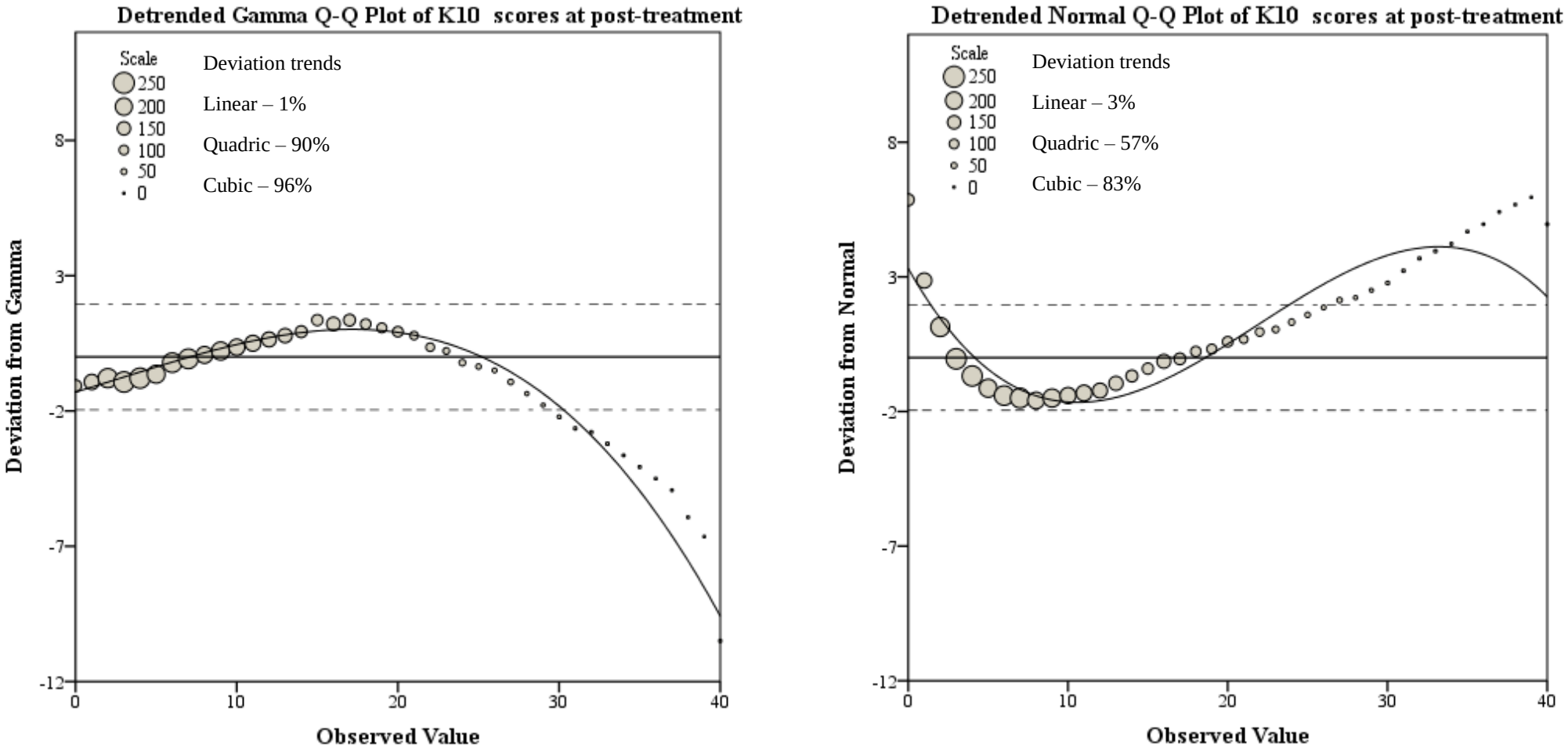


Figure 5*: De-trended Quantile-Quantile plots of post treatment K-10 psychological symptoms; the fit of the observed scores and the Gamma distribution is presented on the left; the fit of the observed scores with a normal distribution is fitted on the right; Q-Q : Quantile-Quantile*

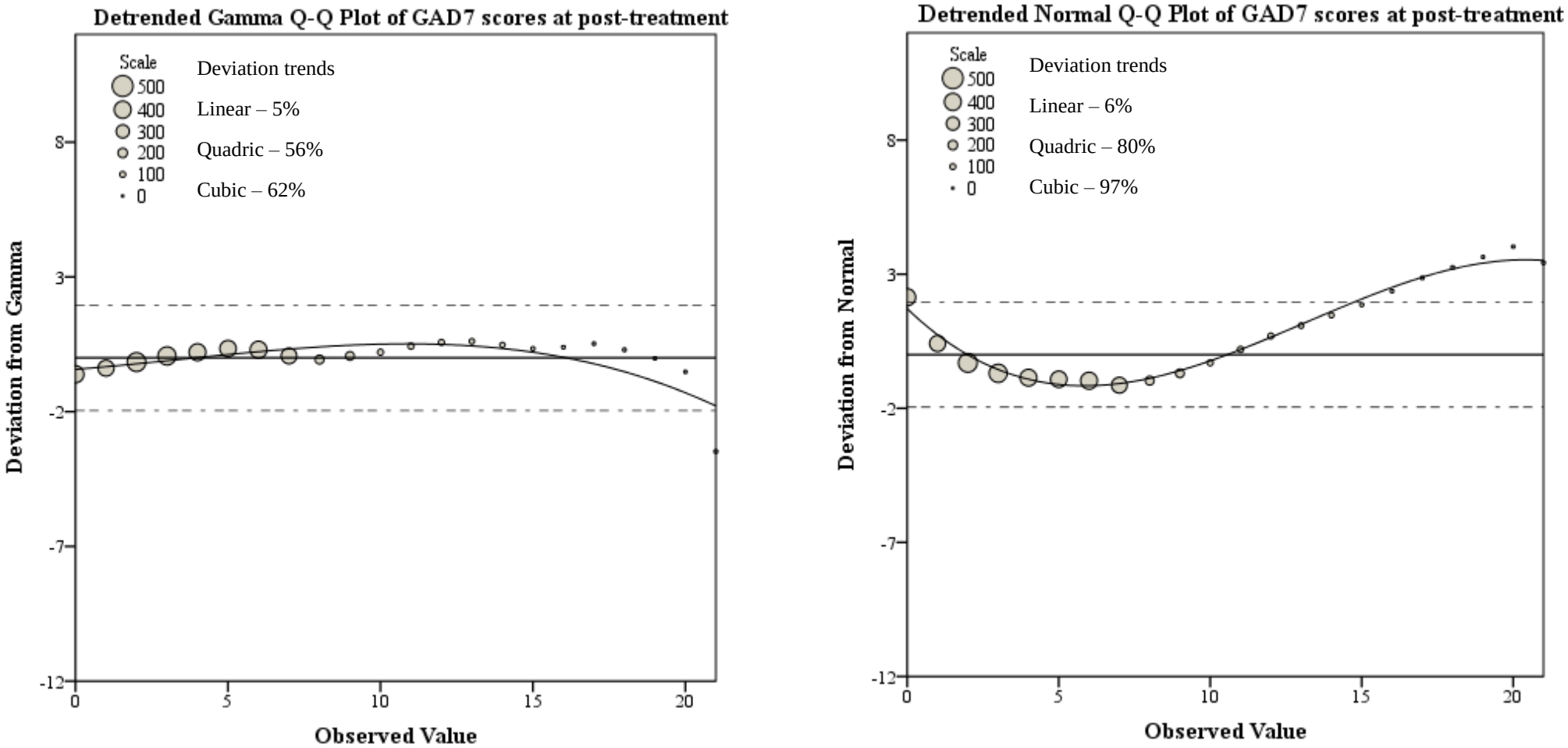


Figure 6: *De-trended Quantile-Quantile plots of post treatment GAD-7 anxiety symptoms; the fit of the observed scores and the Gamma distribution is presented on the left; the fit of the observed scores with a normal distribution is fitted on the right; Q-Q : Quantile-Quantile*

Figures 4,5 and 6 illustrate that the observed dispersion of symptom scores at post-treatment all deviate from the normal distribution. The deviation of scores from the normal distribution is particularly evident with a higher than expected dispersion density around low symptom scores (zero) and high symptoms. In these Figures, deviations that exceed the dotted line, denoting 1.96 standard deviations, are considered to represent significant deviation. In contrast to the normal distribution, the dispersion of post-treatment scores followed the theoretical gamma distribution more closely, with fewer deviations observed around very high symptoms.

H3: Analytical methods that account for the proportional remission would increase the accuracy for predicting symptom treatment outcomes

In support of H3, analytical methods that fit the characteristics of the data resulted in a greater ability to predict the outcomes of patients following treatment and consequently reduced measurement error. Measurement error was explored using the two approaches for modelling symptom change. Table 5 reports the estimation error created when symptom change was estimated as an averaged fixed score (under the linear approach), or as a percentage effect (under the proportional approach). The estimates of error in Table 5 show a sizable reduction in the total variance estimate is evident for depression, anxiety, and psychological distress scores when the proportional approach was taken. Specifically, the total symptom change variance decreased under the proportional remission assumption by 32% for depression symptoms (σ^2 of 19.42 vs. 28.48), 34% for anxiety symptoms (σ^2 of 16.27 vs. 24.34) and 17% for psychological distress (σ^2 of 41.62 vs. 50.03).

Table 5: Measurement error under the assumptions that change is either proportional or linear; within subgroups of differing pre-treatment severity

Measurement approach	Measure	**Total variance** of post-treatment scores (σ^2)	Prediction error for the **Minimal / Low*** symptom group	Prediction error for **Mild*** symptom group	Prediction error for the **Moderate*** symptom group	Prediction error **Moderately severe** / High symptom group	Prediction error for the **Severe*** symptom group
Proportional measurement approach (48%)	PHQ-9	19.42	0.58 [-0.79,-0.37]	0.2 [-0.4,-0.01]	-0.15 [-0.08,0.38]	-0.23 [-0.12,0.57]	-0.09 [-0.51,0.7]
Linear measurement approach (6 points)	PHQ-9	28.48	-5.31 [-5.5,-5.1]	-2.77 [-2.97,-2.6]	-0.12 [-0.35,0.12]	2.27 [1.92,2.62]	4.72 [4.12,5.32]
Proportional measurement approach (49%)	K-10	41.62	0.88 [0.37,1.39]	--	0.48 [0.15,0.81]	0.11 [-0.18,0.4]	0.06 [-0.32,0.45]
Linear measurement approach (6.77 points)	K-10	50.03	4.82 [4.29,5.35]	--	2.45 [2.11,2.78]	-0.5 [-0.79,-0.21]	-4.11 [-4.5,-3.7]
Proportional measurement approach (38%)	GAD-7	16.27	0.74 [0.3,1.19]	-0.41 [0.07,0.76]	0.04 [-0.49,0.41]	--	0.36 [-0.98,0.25]
Linear measurement approach (5.33points)	GAD-7	24.34	-4.68 [-4.91,-4.44]	-2.36 [-2.5,-2.2]	0.59 [0.36,0.81]	--	3.53 [3.22,3.84]

*Pre-treatment scores; Multiplicative GEE models for quantifying change over time specified Yij ~ log (μij) = β0 + β1tj + error; with *mi * mi* working correlation matrix for each Yij, Var(Yij) = var(Scaler paramter_ij) y; where the scale parameter and v(·) are based on a gamma variance function Gamma (μij,α);~ N (0, σ2). Linear GEE models for quantifying change over time specified Yij = β0 + β1tj + error; with mi mi working correlation matrix for each Yij, Var(Yij) = v(Scaler parameter_ij) y; where the scaler parameter and v(·) is a normal distribution; ~ N (0, σ*2*). For all models i= 1,...,6071 (respective to the subsample used); Timej= Pre-treatment; Mid treatment; post treatment.

Additional syntax for each of the models is denoted in Appendix B, *p.228*

Table 5 also reports the predictive accuracy for each of the modelling approaches, as a total sample and within each pre-treatment severity band. As a total sample, the prediction accuracy under a proportional approach was reduced to an average of a single residual point (>0.88) between the predicted outcome and actual outcome. This predictive accuracy was observed for the sample as a whole and within each pre-treatment severity band. In contrast, under the average fixed score approach, more substantial and systematic prediction inaccuracies are observed (<5.31). Specifically, increased predictive error was observed with individuals from the severe and milder pre-treatment symptom bands.

This comparison of linear and multiple measurements is also illustrated in Figures 7, 8, and 9. In these figures, using the linear approach, the severity of symptoms pre-treatment is associated with the degree of prediction error. That is, under the linear approach, the association between pre-treatment severity and the prediction error accounted for 30-40% of the total measurement error. The majority of the error associated with the linear prediction occurred for individuals with mild or severe pre-treatment symptoms. The implication of this finding is that the use of a change function that does not suit the characteristics of the data results in prediction error that is at least 30%. In contrast, the measurement error under a proportional change approach reduced the measurement error associated with pre-treatment severity entirely (R^2 <1%; for each PHQ-9, GAD-7, K-10 scale). Under the proportional change approach, error was overall lower and evenly distributed across all levels of pre-treatment depressive symptoms. In other words, the proportional model was able to more accurately predict the average outcomes of groups from any pre-treatment severity score.

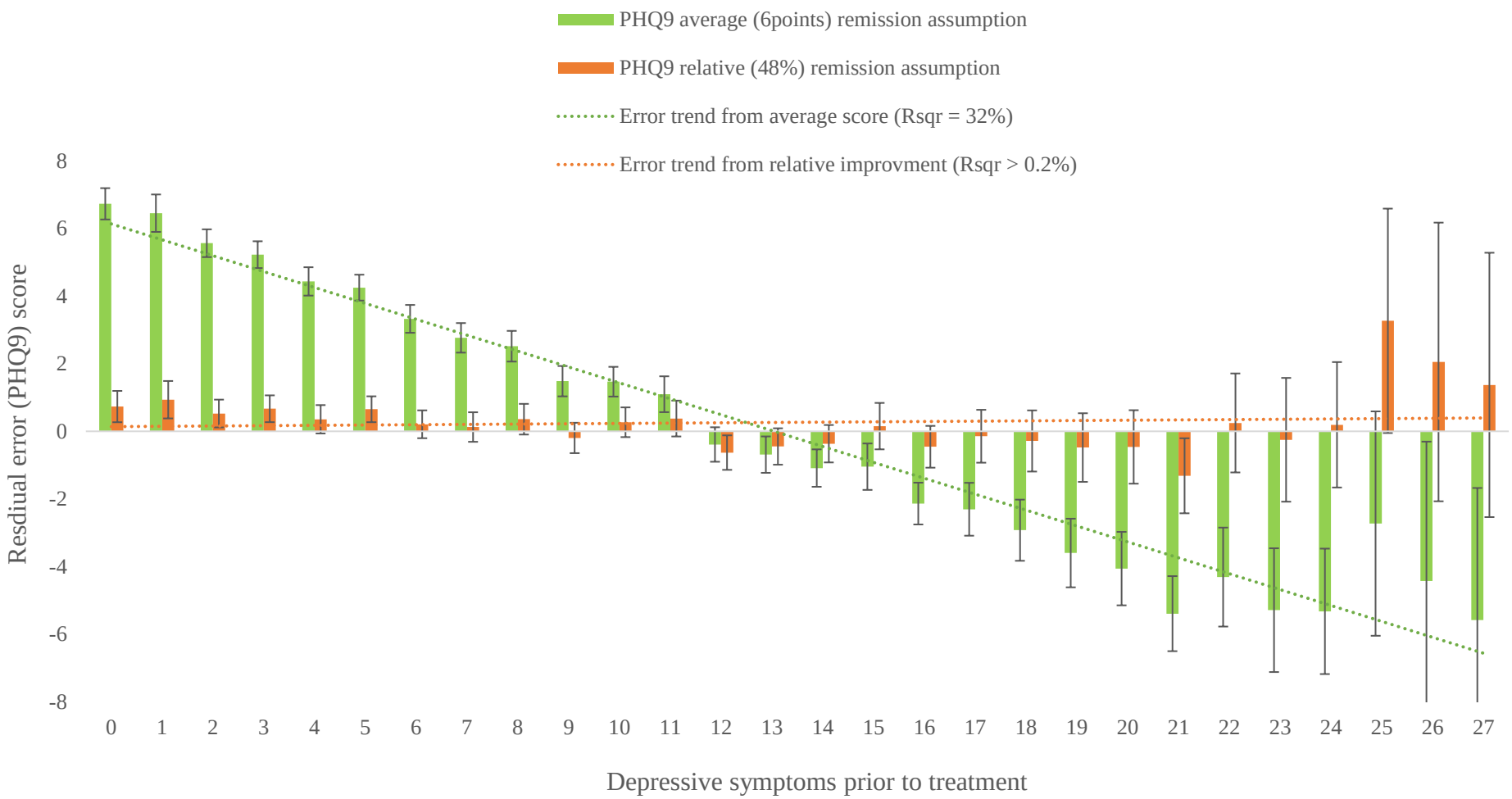


Figure 7: *PHQ-9 estimation error (residual) of post-treatment scores following fixed and relative remission assumptions*

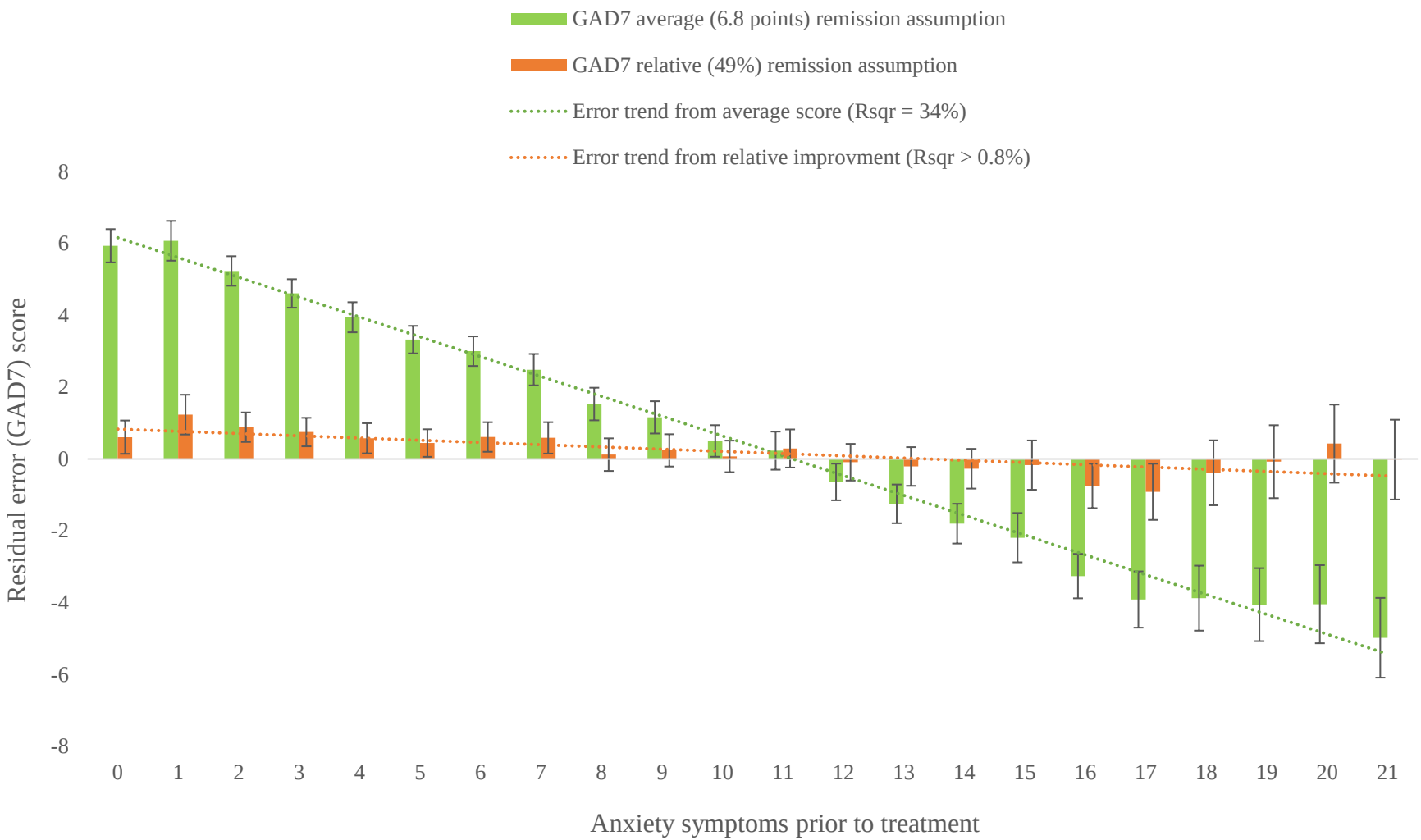


Figure 8: *GAD-7 estimation error (residual) of post-treatment scores that follow the assumption of relative and fixed change*

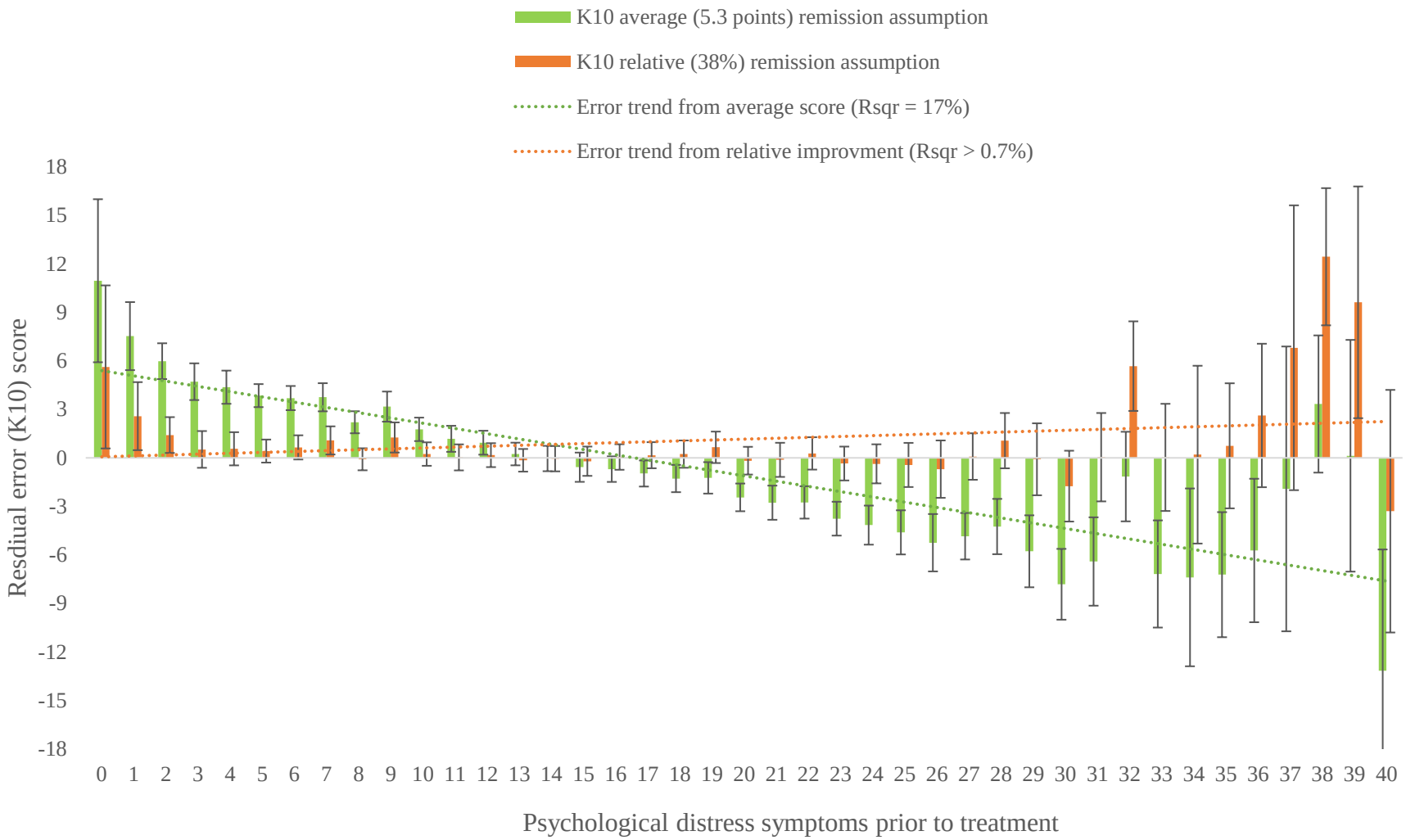


Figure 9: *K-10 estimation error (residual) of post-treatment scores that follow the assumption of relative and fixed change*

Discussion

Summary of findings

The ability to choose an appropriate statistical method is fundamental to the accurate and valid measurement and interpretation of the impact of treatment (Lang & Altamn, 2013; Ebrahim et al., 2013). This study explored the statistical patterns of symptom change through treatment using two common statistical methods that approximate change, that is, linear and proportional change.

Linear methods dominate psychotherapeutic research (Nieminen & Kaur, 2019; Pek & Flora, 2018), and therefore the ability to measure, interpret and statistically model symptom change. However, when the features of clinical data were examined in detail during this study, symptom change appeared proportional, and no linear.

Specifically, evidence for the proportionality of symptom change was identified within three symptom scales (PHQ-9, GAD-7, K-10) and replicated within five large and randomised samples. That is, the results of the study illustrated a high association between pre-treatment symptoms and the quantity of symptom change (H1), positively skewed distributions (H2) and improved measurement accuracy using a proportional model (H3). In combination, these results further support the occurrence of an underlying proportional function of symptom change. In this way, the evidence of the study supported the three hypotheses proposed about the proportionality of symptom improvement from pre-treatment presentations. Furthermore, the study was also able to demonstrate that the use of the proportional model resulted in significantly reduced measurement error compared to the linear model.

This pattern of proportionally changing symptoms is consistent with previous studies that show increased effect sizes in sub-samples with more severe symptoms, even when similar treatments are applied (Bower, et al., 2013; Driessen, et al., 2010; Kroenke, et al., 2001; Paykel, et al., 1995). Further, in the context of treatment, percentage improvement, and percentage

change estimates reflect the ideal reduction of symptoms to a minimal range (Kroenke et al., 2015; Sobocki et al., 2006), and a metric of treatment change that is directly interpretable.

Practical implications of findings for research

The identification of a proportional function of symptom change raises two major implications for the psychotherapy literature and other scientific fields. First, the study has demonstrated that the use of a linear change metric, which is arguably the most common approach (Lakens, 2013), can result in systematically biased estimates of the effect of treatment. For example, this study identified that high (1.2) or low (0.3) Cohen's *d* effect size metrics were entirely dependent on the severity of the sample pre-treatment rather than on the quality or potential of treatment. This variance in outcome was observed with highly differing effect sizes under the same treatment, replicating the finding by Karin and colleagues (2018) which demonstrated that large effect size was associated with elevated pre-treatment scores even under control conditions, indicating the effect of regression to the mean. Together, the relationship between large effect sizes for elevated pre-treatment symptoms resulting from measuring symptom change using linear functions may reflect an artefact of measurement methodology. Similarly, the measurement error resulting from the application of a linear metric (e.g. Cohen's *d*) may also limit the ability to compare the effect of treatment between trials with differing pre-treatment symptoms. For example, in the context of meta-analyses exploring the overall efficacy of treatment, the comparison of effect sizes may be highly influenced by the application of a linear metric and the symptoms at pre-treatment, rather than the therapeutic efficacy of different treatments. This association between pre-treatment symptom severity and the linear measure of effects were narrowly explored by Bower and colleagues (2013) and Cuijpers and colleagues (2010), which did not, however, consider any alternative method to the linear metric. In contrast, as demonstrated in this study, the application of a percentage

change metric may be more suitable as a metric of treatment evaluation given its ability to result in more consistent and statistically accurate description of symptom change across the range of mild, moderate or severe symptom bands.

Second, the measurement and interpretation of symptom change as a proportional function of change shows promise for more accurate and sensitive statistical modelling. For example, the use of a proportional model for the measurement of GAD-7 symptom change resulted in a model that was more accurate in its outcome predictions by nearly 40%, compared to a linear model. This result indicates that the proportional model was able to statistically capture the trajectories of patients with substantially less error; in contrast, a linear model inflated the error without the ability to account for the increased rate of change with increased severity. In this way, the statistical accuracy of proportional measurement may be critical for clinical research seeking to evaluate treatments that enhance the rate of clinical change. For example, researchers seeking to compare the magnitude of symptom change in different treatments, or identify clinical moderators or estimate cost-effectiveness, would likely identify predictors, moderators or mediators that associated with the magnitude of clinical efficacy, rather than merely baseline symptoms. Considered another way, clinical analysts who employ unsuitable linear models are likely to find that their research results correlate predictors of change with pre-treatment symptom severity, rather than correlates that predict the quality of treatment. For this reason, the ability to fully understand the statistical properties of symptom change, and consequently choose an optimal statistical method, is likely to be important for the ability to conduct research about both the clinical effects of psychotherapy (Sanders & Hunsley, 2018), processes of change in psychotherapy (Laurenceau et al., 2007), the characterisations of subgroups of patients (Gunn, et al., 2013; Lin et al., 2016), and the identification of patients at risk (Rozental et al. 2014).

Limitations and implications for future research

Although the results of the study were replicated across three symptom scales, and using a cross-validation analysis, several methodological and clinical limitations should be considered about the design of this study, and the ability to generalise the conclusion that symptom change follows proportional features. First, in this study, the proposed mechanism that shapes the 50% rate of change was argued to result from scale bounding (floor effect) and the relative change that patients demonstrate when they improve their underlying causes of symptoms. However, the occurrence of scale bounding cannot fully account for the consistency of the rate of change. For example, if patients were to fully resolve their symptoms to a minimum, the rate of change should result in a proportional pattern that is incremental and higher than 50%. For this reason, the results of this study cannot fully explain why a specific proportional pattern of 50% would occur across different types of symptoms scales, and different types of subgroups of different pre-treatment symptoms. Rather, the 50% change may be due to both the occurrence of scale bounding and the mixture of subgroups including those with large symptom improvements (e.g. 70%) and other subgroups who improve by lesser rates (e.g. 30%). This collapse of multiple trends into a single estimate (50%) may reflect the so-called Simpson paradox (Kievit, Frankenhuis, Waldorp, & Borsboom, 2013) that conceal important and distinct types of outcomes with the use of a single average metric (e.g., deterioration, minimal response, remission). Although this 50% estimate would adequately describe the overall change in the context of randomised control trials, additional research about subgroups with differing magnitude of change could further enhance the ability to detect and describe more nuanced but important patterns of change.

Second, it is important to note that the current study employed data from patients undergoing a particular type of psychological treatment, which was delivered via the internet and telephone. It is possible that the specific proportional pattern of symptom change observed

in this study and the previous study (Karin et al., 2018) specific to web-based psychotherapy, and that other types of psychotherapies may be associated with other types of symptom change patterns. For example, web-based psychotherapy may be associated with the improvement in particular symptom depression scale items (loss of interest in activity, hopelessness vs. fatigue, appetite), or specific patterns of early fixed change in symptoms (Delgadillo, McMillan, Lucock, Leach, Ali & Gilbody, 2014), where other treatments may be associated with alternative patterns. For this reason, caution and further study are needed before the results could be generalised to the measurement of treatment outcomes with other scales and other types of treatment. Similarly, the current study only examined symptom change in a narrow context of anxiety, depression, and general psychological distress. Moreover, the study only presented one measure for each type of outcome. For this reason, the generalisability of the statistical characteristics of symptom change, and the superior performance of the proportional measurement approach, are still uncertain and pending on replication in other symptoms scales, clinical outcomes (e.g., panic symptoms, social anxiety symptoms) and across different clinical contexts (e.g. different treatments, subgroups).

Third, the current study weighed the validity of the linear and proportional measurement approaches through the comparison of ***statistical*** accuracy. Although statistical prediction and accurate measurement are critical for measurement validity, the clinical utility of the proportional approach may be separate from any significant improvement to the ability to measure change with reduced error (Peeters, 2016; Ronk, Hooke, & Page, 2016; Thompson, 2002). That is, the percentage change approach may, or may not, be associated with qualitative markers of clinical improvement, such as a change in patient or clinician perceptions. In order to determine that the proportional measurement of symptom change can translate into an improved and clinically valid way to measure and interpret clinical change additional research is needed to link the two measurements with a measure that qualitatively evaluate the efficacy

of treatment (e.g. satisfaction with treatment, clinical diagnoses). This research would reinforce the features of proportionally changing symptoms and their improved ability to capture evidence of treatment efficacy.

Conclusions

In summary, this study explored the statistical characteristics of symptom change during psychotherapy and compared two approaches for measuring and interpreting symptom change. The findings of the current study highlight the importance of selecting the correct approach for modelling and estimating symptom change. Although the research is preliminary, this type of methodological research holds potential for more accurate and valid measurement of change, and consequently more accurate and valid metrics for the evaluation of treatment, the comparison of treatment effects, the research of patient trajectories, or other research topics that rely on the measurement of symptoms as a primary outcome.

References

Baldwin, S. A., Fellingham, G. W., & Baldwin, A. S. (2016). Statistical models for multilevel skewed physical activity data in health research and behavioral medicine. Health Psychology, 35(6), 552.

Beckstead, J. W. (2013). On measurements and their quality: Paper 1: Reliability–history, issues and procedures. International journal of nursing studies, 50(7), 968-973.

de Beurs, E., Barendregt, M., de Heer, A., van Duijn, E., Goeree, B., Kloos, M., ... & Merks, A. (2016). Comparing methods to denote treatment outcome in clinical research and benchmarking mental health care. *Clinical psychology & psychotherapy*, *23*(4), 308-318.

Blackwell, M., Honaker, J., & King, G. (2017). A unified approach to measurement error and missing data: overview and applications. Sociological Methods & Research, 46(3), 303-341.

Boswell, J. F., Kraus, D. R., Miller, S. D., & Lambert, M. J. (2015). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. Psychotherapy research, 25(1), 6-19.

Bower, P., Kontopantelis, E., Sutton, A., Kendrick, T., Richards, D. A., Gilbody, S., ... & Meyer, B. (2013). Influence of initial severity of depression on effectiveness of low intensity interventions: meta-analysis of individual patient data. Bmj, 346, f540.

Bruce, M. L., Ten Have, T. R., Reynolds III, C. F., Katz, I. I., Schulberg, H. C., Mulsant, B. H., ... & Alexopoulos, G. S. (2004). Reducing suicidal ideation and depressive symptoms in depressed older primary care patients: a randomized controlled trial. Jama, 291(9), 1081-1091.

Choi, S. W., Schalet, B., Cook, K. F., & Cella, D. (2014). Establishing a common metric for depressive symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS Depression. Psychological assessment, 26(2), 513.

Cuijpers, P., Karyotaki, E., Weitz, E., Andersson, G., Hollon, S. D., & van Straten, A. (2014). The effects of psychotherapies for major depression in adults on remission, recovery and improvement: a meta-analysis. *Journal of affective disorders*, *159*, 118-126.

Dear, B. F., Gandy, M., Karin, E., Staples, L. G., Johnston, L., Fogliati, V. J., ... & Sharpe, L. (2015). The Pain Course: a randomised controlled trial examining an internet-delivered pain management program when provided with different levels of clinician support. Pain, 156(10), 1920.

Delgadillo, J., McMillan, D., Lucock, M., Leach, C., Ali, S., & Gilbody, S. (2014). Early changes, attrition, and dose–response in low intensity psychological interventions. British Journal of Clinical Psychology, 53(1), 114-130.

Driessen, E., Cuijpers, P., Hollon, S. D., & Dekker, J. J. (2010). Does pre-treatment severity moderate the efficacy of psychological treatment of adult outpatient depression? A meta-analysis. Journal of Consulting and Clinical Psychology, Vol 78(5), Oct 2010, 668-680. http://dx.doi.org/10.1037/a0020570

Ebrahim, S., Sohani, Z. N., Montoya, L., Agarwal, A., Thorlund, K., Mills, E. J., & Ioannidis, J. P. (2014). Reanalyses of randomized clinical trial data. *Jama*, *312*(10), 1024-1032.

Erekson, D. M., Horner, J., & Lambert, M. J. (2016). Different lens or different picture? Comparing methods of defining dramatic change in psychotherapy. Psychotherapy Research, 1-11.

Feingold, A. (2009). Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. Psychological methods, 14(1), 43.

Fidler, F. (2002). The fifth edition of the APA Publication Manual: Why its statistics recommendations are so controversial. Educational and Psychological Measurement, 62(5), 749-770.

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour research and therapy*, *98*, 19-38.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). Applied longitudinal analysis (Vol. 998). John Wiley & Sons.

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

Frost, M. H., Reeve, B. B., Liepa, A. M., Stauffer, J. W., Hays, R. D., & Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. (2007). What is sufficient evidence for the reliability and validity of patient-reported outcome measures?. Value in Health, 10, S94-S105.

Gunn, J., Elliott, P., Densley, K., Middleton, A., Ambresin, G., Dowrick, C., ... & Griffiths, F. (2013). A trajectory-based approach to understand the factors associated with persistent depressive symptoms in primary care. Journal of affective disorders, 148(2), 338-346.

Hageman, W. J., & Arrindell, W. A. (1999). Establishing clinically significant change: Increment of precision and the distinction between individual and group level of analysis. Behaviour Research and Therapy, 37(12), 1169-93.

Hayes, A. M., Laurenceau, J. P., Feldman, G., Strauss, J. L., & Cardaciotto, L. (2007). Change is not always linear: The study of nonlinear and discontinuous patterns of change in psychotherapy. Clinical psychology review, 27(6), 715-723.

Hiller, W., Schindler, A. C., & Lambert, M. J. (2012). Defining response and remission in psychotherapy research: A comparison of the RCI and the method of percent improvement. Psychotherapy Research, 22(1), 1-11.

Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Lippman, S. A., Jewell, N., ... & Satariano, W. A. (2010). To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. Epidemiology, 21(4), 467-474

Karin, E., Dear, B. F., Heller, G. Z., Gandy, M., & Titov, N. (2018). Measurement of symptom change following web-based psychotherapy: Statistical characteristics and analytical methods for measuring and interpreting change. JMIR mental health, 5(3), e10200.

Kazdin, A. E. (1999). The meanings and measurement of clinical significance. Journal of consulting and clinical psychology, 67(3), 332-339.

Kazdin, A. E. (2015). Evidence-based psychotherapies II: changes in models of treatment and treatment delivery. South African Journal of Psychology, 45(1), 3-21.

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S. L., ... & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychological medicine, 32(6), 959-976.

Kievit, R., Frankenhuis, W. E., Waldorp, L., & Borsboom, D. (2013). Simpson's paradox in psychological science: a practical guide. Frontiers in psychology, 4, 513.

King, M. T. (2011). A point of minimal important difference (MID): a critique of terminology and methods. Expert review of pharmacoeconomics & outcomes research, 11(2), 171-184.

Kroenke, K., Monahan, P. O., & Kean, J. (2015). Pragmatic characteristics of patient-reported outcome measures are important for use in clinical practice. Journal of clinical epidemiology, 68(9), 1085-1092.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The Phq-9. Journal of general internal medicine, 16(9), 606-613. http://www.jstor.org/stable/3768417

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Frontiers in psychology, 4, 863.

Lang, T. A., & Altman, D. G. (2013). Basic statistical reporting for articles published in biomedical journals: the "Statistical Analyses and Methods in the Published Literature" or the SAMPL Guidelines". Handbook, European Association of Science Editors, 23-26.

Laurenceau, J. P., Hayes, A. M., & Feldman, G. C. (2007). Some methodological and statistical issues in the study of change processes in psychotherapy. Clinical psychology review, 27(6), 682-695.

Lambert, M. J., & Ogles, B. M. (2009). Using clinical significance in psychotherapy outcome research: The need for a common procedure and validity data. *Psychotherapy Research*, *19*(4-5), 493-501.

Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73:13–22.

Lin, Y., Huang, S., Simon, G. E., & Liu, S. (2016). Analysis of depression trajectory patterns using collaborative learning. Mathematical biosciences, 282, 191-203.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. Science, 355(6325), 584-585.

Maul, A., Irribarra, D. T., & Wilson, M. (2016). On the philosophical foundations of psychological measurement. Measurement, 79, 311-320.

Loy, A., Follett, L., & Hofmann, H. (2016). Variations of Q–Q Plots: The power of our eyes!. *The American Statistician*, *70*(2), 202-214.

McMillan, D., Gilbody, S., & Richards, D. (2010). Defining successful treatment outcome in depression using the PHQ-9: a comparison of methods. Journal of affective disorders, 127(1), 122-129.

Ng, V. K., & Cribbie, R. A. (2017). Using the Gamma Generalized Linear Model for modeling continuous, skewed and heteroscedastic outcomes in psychology. Current Psychology, 36(2), 225-235.

Nielssen, O., Dear, B. F., Staples, L. G., Dear, R., Ryan, K., Purtell, C., & Titov, N. (2015). Procedures for risk management and a review of crisis referrals from the MindSpot Clinic, a national service for the remote assessment and treatment of anxiety and depression. BMC psychiatry, 15(1), 304.

Nieminen, P., & Kaur, J. (2019). Reporting of data analysis methods in psychiatric journals: Trends from 1996 to 2018. *International journal of methods in psychiatric research*, e1784-e1784.

Peeters, M. J. (2016). Practical significance: moving beyond statistical significance. *Currents in Pharmacy Teaching and Learning*, *8*(1), 83-89.

Paykel, E. S., Ramana, R., Cooper, Z., Hayhurst, H., Kerr, J., & Barocka, A. (1995). Residual symptoms after partial remission: an important outcome in depression. Psychological medicine, 25(6), 1171-1180.

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, *23*(2), 208.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. Journal of applied psychology, 88(5), 879.

Price, D. D., McGrath, P. A., Rafii, A., & Buckingham, B. (1983). The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. Pain, 17(1), 45-56.

Ronk, F. R., Hooke, G. R., & Page, A. C. (2016). Validity of clinically significant change classifications yielded by Jacobson-Truax and Hageman-Arrindell methods. *BMC psychiatry*, *16*(1), 187.

Rozental, A., Andersson, G., Boettcher, J., Ebert, D. D., Cuijpers, P., Knaevelsrud, C., ... & Carlbring, P. (2014). Consensus statement on defining and measuring negative effects of Internet interventions. Internet Interventions, 1(1), 12-19.

Rush, A. J., Kraemer, H. C., Sackeim, H. A., Fava, M., Trivedi, M. H., Frank, E., ... & Regier, D. A. (2006). Report by the ACNP Task Force on response and remission in major depressive disorder. Neuropsychopharmacology, 31(9), 1841.

Sanders, S. G., & Hunsley, J. (2018). The new Caucus-race: Methodological considerations for meta-analyses of psychotherapy outcome. Canadian Psychology/psychologie canadienne, 59(4), 387.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... & Carlsson, R. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337-356.

Sobocki, P., Ekman, M., Ågren, H., Runeson, B., & Jönsson, B. (2006). The mission is remission: health economic consequences of achieving full remission with antidepressant treatment for depression. International journal of clinical practice, 60(7), 791-798.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. Archives of internal medicine, 166(10), 1092-1097.

Thompson, B. (2002). "Statistical,""practical," and "clinical": How many kinds of significance do counselors need to consider?. Journal of Counseling & Development, 80(1), 64-71.

Titov, N., Dear, B. F., Staples, L. G., Bennett-Levy, J., Klein, B., Rapee, R. M., ... & Nielssen, O. B. (2016). The first 30 months of the MindSpot Clinic: evaluation of a national e-mental health service against project objectives. Australian & New Zealand Journal of Psychiatry, 0004867416671598.

Verkuilen, J. and Smithson, M. (2012). Mixed and mixture regression models for continuous bounded responses using the beta distribution. Journal of Educational and Behavioral Statistics, 37, 82-113.

Young, C. (2018). Model uncertainty and the crisis in science. *Socus*, *4*, 2378023117737206.

Chapter 4

"Wish You Were Here": Examining Characteristics, Outcomes, and Statistical Solutions for Missing Cases in Web-Based Psychotherapeutic Trials (Study 3)

This chapter concerns a second and incremental step in the process of measuring and interpreting psychotherapy evidence, that is, the approximation of outcomes for cases who lapse out of contact with the research team and become missing cases. The chapter includes a study that explored the characteristics of missing cases, explore evidence of their likely trajectories through treatment, and used these feature to comment on the suitability of different statistical approaches for handling missing cases. The study sought to contribute to the limited literature available about the features of missing cases, the assumptions that can be made about their outcomes, and consequently the suitability of different statistical methods for replacing their outcomes without increasing measurement bias.

The major aim of the study was to identify those features of missing cases that dominantly define the trajectory and outcome of missing web-based cases through treatment as a phenomena. The study then explores the venerability of these features and their impact on internal and external validity in a subsequent replication study.

Publication Reference:

Karin, E., Dear, B. F., Heller, G. Z., Crane, M. F., & Titov, N. (2018). "Wish You Were Here": Examining Characteristics, Outcomes, and Statistical Solutions for Missing Cases in Web-Based Psychotherapeutic Trials. *JMIR mental health, 5*(2), e22.

Author contribution:

Mr Eyal Karin has designed, analysed, and wrote the study. Professor Gillian Heller oversaw the choice of statistical methodology and assisted with the drafting of the manuscript. Associate Professor Blake F. Dear and Dr Monique Crane provided the dataset, assisted with the refinement of the manuscript, and helped frame the methodological content for a clinical audience. Professor Nick Titov oversaw the conception of the project and the drafting of the manuscript.

Original Paper

“Wish You Were Here”: Examining Characteristics, Outcomes, and Statistical Solutions for Missing Cases in Web-Based Psychotherapeutic Trials

Eyal Karin[1], MAppStats; Blake F Dear[1,2], PhD; Gillian Z Heller[3], PhD; Monique F Crane[4], PhD; Nickolai Titov[1,2], PhD

[1]eCentreClinic, Department of Psychology, Macquarie University, Sydney, Australia
[2]Mindspot Clinic, Department of Psychology, Macquarie University, Sydney, Australia
[3]Department of Statistics, Macquarie University, Sydney, Australia
[4]Department of Psychology, Macquarie University, Sydney, Australia

Corresponding Author:
Eyal Karin, MAppStats
eCentreClinic
Department of Psychology
Macquarie University
4 First Walk
Sydney, 2109
Australia
Phone: 61 298508657
Email: eyal.karin@mq.edu.au

Abstract

Background: Missing cases following treatment are common in Web-based psychotherapy trials. Without the ability to directly measure and evaluate the outcomes for missing cases, the ability to measure and evaluate the effects of treatment is challenging. Although common, little is known about the characteristics of Web-based psychotherapy participants who present as missing cases, their likely clinical outcomes, or the suitability of different statistical assumptions that can characterize missing cases.

Objective: Using a large sample of individuals who underwent Web-based psychotherapy for depressive symptoms (n=820), the aim of this study was to explore the characteristics of cases who present as missing cases at posttreatment (n=138), their likely treatment outcomes, and compare between statistical methods for replacing their missing data.

Methods: First, common participant and treatment features were tested through binary logistic regression models, evaluating the ability to predict missing cases. Second, the same variables were screened for their ability to increase or impede the rate symptom change that was observed following treatment. Third, using recontacted cases at 3-month follow-up to proximally represent missing cases outcomes following treatment, various simulated replacement scores were compared and evaluated against observed clinical follow-up scores.

Results: Missing cases were dominantly predicted by lower treatment adherence and increased symptoms at pretreatment. Statistical methods that ignored these characteristics can overlook an important clinical phenomenon and consequently produce inaccurate replacement outcomes, with symptoms estimates that can swing from −32% to 70% from the observed outcomes of recontacted cases. In contrast, longitudinal statistical methods that adjusted their estimates for missing cases outcomes by treatment adherence rates and baseline symptoms scores resulted in minimal measurement bias (<8%).

Conclusions: Certain variables can characterize and predict missing cases likelihood and jointly predict lesser clinical improvement. Under such circumstances, individuals with potentially worst off treatment outcomes can become concealed, and failure to adjust for this can lead to substantial clinical measurement bias. Together, this preliminary research suggests that missing cases in Web-based psychotherapeutic interventions may not occur as random events and can be systematically predicted. Critically, at the same time, missing cases may experience outcomes that are distinct and important for a complete understanding of the treatment effect.

(JMIR Ment Health 2018;5(2):e22) doi:10.2196/mental.8363

KEYWORDS

psychotherapy; treatment adherence and compliance; missing data; treatment efficacy; statistical bias

Introduction

Background

Missing cases are often encountered in Web-based psychotherapeutic trials, with the likely frequency of participants to become absent from posttreatment surveys ranging from 1 in every 5, to 1 in every 3 patients [1,2]. Missing cases present a significant challenge to the accuracy of results by reducing the sample size and the statistical power available to estimate the effects of treatment [3]. Furthermore, missing cases can produce measurement bias by systematically concealing important clinical information such as the experience of negative outcomes in treatment.

Although multiple definitions of missing cases are possible (eg, unit, item) [4], this paper will consider missing cases as those individuals who conceal their treatment outcomes as absent cases at the point of posttreatment surveys. Without any information about the outcomes of missing cases, the challenge that these cases pose is that the clinical effect itself cannot be completely understood [3].

The problems associated with missing data are well recognized in the clinical literature, and reflecting this, requirements to account for missing cases are embedded in leading guidelines such as the Consolidated Standards of Reporting Trials statement [5] and other methodological guidelines [6-9]. Such guidelines require clinical researchers to make estimates about the treatment outcomes for missing cases and incorporate these estimates in the measurement and evaluation of treatment effects [7]. The statistical methods employed to account for missing cases' outcomes typically attempt to mimic the remaining observed cases and simulate replacement treatment outcomes [6]. Examples of such statistical methods include model-based imputations and multiple imputations [9,10]. These statistical methods aim to resolve both issues of reduced sample size and potential measurement bias associated with overlooking missing cases outcomes [6,9,11].

When attempting to approximate and replace missing cases outcomes, statistical and methodological guidelines first advise that research explore for evidence about the characteristics and likely outcomes of missing cases. This is a first and pivotal step in the process of handling missing cases, which can lead to a more educated guess about the kind of clinical outcomes missing cases that would have likely occurred [6,9,12,13]. In more statistical terms, researchers are required to make an informed assumption about the unknown outcomes for missing cases and effectively decide whether missing cases are a distinct subgroup with distinct and important outcomes or a random and ignorable extension of the whole sample [6,13]. It is also important to note that any characterization of missing cases and the replacement of their outcomes is made under one of three possible assumptions [8]. First, the assumption that missing cases and their outcomes are comparable with the characteristics and outcomes of the overall reaming sample is named the missing completely at random (MCAR) assumption. Similarly, the assumption when missing cases show some distinct characteristics but are assumed to be comparable in outcomes with a similar subgroup of remaining cases (stratified subgroup) is named the missing at random (MAR) assumption. Alternatively, if missing cases are assumed to have characteristics and outcomes that are not comparable to any subset of the reaming cases, the assumption of missing not at random (MNAR) is made.

Notwithstanding the range of statistical solutions [14], guidelines [15], and theoretical discussions [4] about missing cases in psychotherapy or Web-based psychotherapy, questions remain about the characteristics and solutions that could be applied to missing cases following treatment.

The first question regards the characteristics of missing cases and the ability to identify any systematic predictors of missingness. Currently, no concerted empirical studies are available to identify and assess those participant characteristics that are likely to increase the likelihood of becoming missing at posttreatment. As separate from the dropout and treatment adherence literature [2,16-18], factors that predict whether a case will become missing have not been explored within large-scale psychotherapeutic studies; although it is conceivable that these overlap [19].

A second related question concerns the ability to identify variables that describe why missing cases occurred and at the same time give reason to suspect that the outcomes for missing cases are distinct from the overall sample [2,6,7]. For example, if missing cases were characterized by lower treatment adherence, the treatment outcomes of missing cases should also be impacted by lower treatment dosage. This hypothetical example illustrates a scenario where individuals with poorer fit to treatment remove themselves from treatment, conceal their outcomes as missing cases, and leave the evaluation of treatment results to be determined by a margin of people to whom the treatment appeals. In these circumstances, recognizing the role of predictors, such as treatment adherence, is critical for the ability to detect both the increased risk of cases to become missing, as well as for the ability to approximate accurate replacement outcomes for such cases [6,9,10,15].

A third consequent unanswered question concerns the relative accuracy of replacing missing psychotherapy cases under different statistical missing cases strategies and assumptions. Without studies that investigate missing cases and their likely outcomes in the context of psychotherapy, Web-based psychotherapy, or other similar clinical fields, uncertainty remains about the ability to replace and handle missing cases [9]. To explore the suitability of different missing cases solutions, comprehensive clinical research is required that can compare simulated outcomes for missing cases against a proximal outcome of missing cases. Currently, no solutions are available within the Web-based psychotherapy literature to suggest a benchmark for proximally measuring the outcomes for missing cases. As a consequence, no evidence is currently available to support or refute the suitability of any type of

statistical strategy or quantify the implications missing cases have for the estimation of treatment effects.

This Study

The primary aim of this study was to empirically explore evidence from a large naturalistic Web-based psychotherapy sample and provide evidence toward three interrelated questions about missing cases. Specifically, this study sought to (1) identify the characteristics and dominant predictors of missing cases, (2) identify predictors that may have joint influence on likelihood of missing cases and clinical outcomes, and (3) identify a suitable clinical measurement benchmark that can then be used to test the accuracy and suitability of different statistical replacements strategies.

Three hypotheses were made about the characteristics of missing cases and the ability to approximate their outcomes. Consistent with previous theoretical discussions of missing cases in psychotherapy [2,19] and clinical trials [8,20], it was hypothesized that missing cases do not occur as a random event (H1), and participant and treatment features such as treatment adherence would predict the likelihood of participants to present as missing cases following treatment. Second, consistent with the dropout and adherence literature [2,19], it was hypothesized that cases that became missing during posttreatment would be characterized with lower treatment adherence (H2). Third, consistent with statistical guidelines [9,15], it was hypothesized that the replacement of clinical outcomes for missing cases would be made with minimal measurement bias, on the condition of adjusting for key predictors (H3).

Methods

The Sample

This study employed clinical data from three large randomized controlled trials (RCTs; n = 820) investigating the efficacy of Web-based cognitive behavioral therapy (CBT) interventions for reducing symptoms of anxiety and depression [21-23]. These trials employed a similar recruitment methodology and treatment procedures under the Macquarie University Web-based Model (MUM) [24], involving the weekly delivery of Web-based materials organized into psychotherapeutic lessons, together with notifications, emails, and survey reminders over a period of 8 weeks. Telephone contact by a trained clinician was attempted in combination with reminder emails in efforts to engage participants and increase survey participation following treatment. This contact protocol was uniformly applied before treatment, at the end of treatment, and at the point of 3-month follow-up to facilitate participant engagement and adherence.

To be included in these trials, participants were selected on the basis of (1) Demonstrating at least minimal symptoms of anxiety or depression, as determined by the presence of at least mild symptoms of depression or anxiety (a minimum score ≥5 on either the Patient Health Questionnaire 9-item, PHQ-9 [25]; or the Generalized Anxiety Disorder Scale 7-item, GAD-7 [26]); (2) Being over the age of 18 years; (3) Being an Australian resident; and (4) Having Internet access for the period of the trial. In addition, applicants who reported a score of 3 (considered severe) on item 9 of the PHQ-9 measuring suicide-risk were referred to another service.

In combination, these trials represent a random intake of adults seeking treatment for symptoms of depression and anxiety over a period of 2 years within the eCentreClinic [27]. The demographic and symptom characteristics of the participating sample are shown in Table 1.

It is important to note that Web-based psychotherapy data can present a unique opportunity for investigating missing cases and their trajectories in treatment. The standardization of treatment engagement and materials can be considered to reduce the outcome measurement variance associated with treatment delivery. With reduced treatment related variance, the individual's response to treatment remains the main source of statistical variation. In more statistical terms, this sample represents a unique opportunity to measure missing cases influences and outcomes with increased internal validity and within a large sample, enabling a robust statistical testing of the first and second hypotheses. In addition, this sample collates a unique subsample of individuals who are missing at posttreatment but are successfully recontacted during a clinical follow-up, enabling a niche subsample that can be used to test the third hypothesis.

Measures

The primary outcome measure for this study was the PHQ-9, a quantitative measure of depressive symptoms [25]. The PHQ-9 is widely used in psychotherapy and Web-based psychotherapy, is sensitive to the presence and severity of depressive symptoms, and is illustrative of high internal consistency [24,28]. Total scores range from 0 to 27, and the scale comprises 9 items, each offering four responses ranging from 0 to 3. Total scores are clinically interpreted: no depression (total score: 0-4), mild depression (total score: 5-9), moderate depression (total score: 10-14), moderately severe depression (total score: 15-19), and very severe depression (total scores: 20-27). PHQ-9 baseline symptom of the sample are presented in Table 1.

The PHQ-9 scale was administered to measure symptoms at pretreatment (baseline), posttreatment, and again 3 months after the completing of treatment. The original trials comprising the dataset all demonstrated significant and similar average symptom reductions from baseline to posttreatment (46%-53%), which were maintained at 3-month follow-up (50%-53%).

Comorbidity, demographic measures, and treatment adherance were also included as independent variables, aiming to predict missing cases and their clinical trajectories through treatment.

Comorbidity

Participants were defined as having comorbidity if they demonstrated scores of anxiety and depression above a predetermined clinical threshold (GAD-7≥8 and PHQ-9≥10 at baseline; GAD-7 [25]; PHQ-9 [29]).

Table 1. Demographic and clinical sample characteristics. GAD-7: Generalized Anxiety Disorder Scale 7-item; N/A: not applicable; PHQ-9: Patient Health Questionnaire 9-item.

Characteristics	Total sample collated, value (n=820)	Completers[a], value (n=682)	Missing cases[b], value (n=138)	Recontacted cases[c], value (n=55)
Gender, n (%)				
Female	606 (73.9)	465 (75.2)	95 (68.8)	39 (71)
Male	214 (26.1)	153 (24.8)	43 (31.2)	16 (29)
Age, mean (SD)	43.2 (11.1)	44.1 (11.4)	40.4 (11.1)	38.1 (11.4)
Treatment adherence, n (%)				
Completed (1 of 5)	65 (7.9)	9 (1.5)	49 (35.5)	14 (25)
Completed (2 of 5)	53 (6.5)	26 (4.2)	24 (17.4)	6 (11)
Completed (3 of 5)	76 (9.3)	39 (6.3)	23 (16.7)	13 (24)
Completed (4 of 5)	145 (17.7)	101 (16.3)	22 (15.9)	10 (18)
Completed all modules	481 (58.7)	443 (71.7)	20 (14.5)	12 (22)
Relationship status, n (%)				
Otherwise	306 (37.3)	215 (34.8)	62 (44.9)	23 (42)
In a relationship	514 (62.7)	403 (65.2)	76 (55.1)	32 (58)
Education, n (%)				
Non-tertiary	356 (43.4)	254 (41.1)	71 (51.4)	26 (47)
Tertiary	464 (56.6)	364 (58.9)	67 (48.6)	29 (53)
GAD-7 baseline, mean (SD)	11.3 (4.6)	11.0 (4.6)	12.0 (4.6)	11.9 (4.8)
PHQ-9 baseline, mean (SD)	12.3 (4.7)	11.9 (4.8)	13.9 (4.4)	13.7 (4.5)
Comorbidity, n (%)				
None	345 (42.1)	277 (44.8)	44 (31.9)	17 (31)
Comorbid	475 (57.9)	341 (55.2)	94 (68.1)	38 (69)
Missing at posttreatment, n (%)	138 (16.8)	N/A[d]	N/A	55 (40)
Missing at follow-up, n (%)	147 (17.9)	N/A	N/A	N/A

[a]Individuals that completed all surveys.

[b]Individuals with any missing posttreatment data.

[c]Individuals recontacted at 3-month follow-up (n=55).

[d]N/A: not applicable.

Demographic Measures

Age in years at the start of treatment, relationship status, pretreatment symptom scores, pretreatment anxiety scores, and education background were considered. The categories created to measure levels of education, relationship status, treatment adherence, and gender are presented in Table 1.

Treatment Adherence

Under the MUM Internet CBT (iCBT) model, treatment material was organized through five Web-based lessons over a period of 8 weeks. Each lesson comprised introductory CBT explanations, homework assignments, cases stories, and other materials [24]. Participants were required to complete each of the five Web-based lessons in sequence to gain access to the subsequent lesson. Adherence to treatment was therefore measured in this study as the incremental indication that an individual has logged on to the assigned secured website and accessed the Web-based material as these were made available over time. In this way, treatment adherence was measured as the minimal but continued progression of participants through the intended course design.

Recontacted Follow-Up Cases as a Proximal Outcome for Missing Cases at Posttreatment

A key subsample of interest in this study were those participants who presented as missing cases at posttreatment but recontacted at follow-up. In total, 83.2% of participants (682/820) completed the self-report symptom questionnaires at posttreatment. Out of those 138 participants who did not complete the posttreatment survey, 60.1% (83/138) also did not complete questionnaires at the 3-month follow-up. However, 40.0% (55/138) of participants who were missing at posttreatment were successfully surveyed through a 3-month clinical follow-up effort. These recontacted individuals were considered as cases who were partly missing at posttreatment, who would have been completely missing

within study designs that followed a pre-post only protocol. Recontacted cases could be used as a proximal measurement of missing posttreatment outcomes, on the condition that recontacted cases show similarities to cases who were missing at both post and follow-up; as individuals belonging to a broader category of individuals with missing cases.

Analytical Plan

Statistical analysis was conducted with three steps. The first step aimed to characterize missing cases by testing for significant predictors of missing cases (H1, H2). Initially, all possible predictors of missing cases were tested through separate logistic regression models. Within those logistic regression models, missing posttreatment cases versus nonmissing were the binary dependent variable. Following a series of univariate models, a stepwise model building analysis was attempted with the intention to identify a multivariate but parsimonious model of missing cases predictors. This was done by considering all possible predictors in a saturated binary logistic model, including treatment adherence, baseline depression score, baseline anxiety score, and demographic variables of gender, age, employment status, education status, and relationship status. Following, a stepwise variable selection strategy was taken, as outlined by Harrell [30], where predictors that increased the odds of becoming a missing case were retained in a final model. These remaining predictors were interpreted as dominant predictors that statistically characterize the features of missing cases. Each possible predictor of missing cases was assessed for statistical significance at an adjusted *P* value of .01 or less. In addition, the pseud- *R* squared, associated with each missing cases predictor was reported, aiming to convey the known, or model related, proportion of missing cases probability variance; with larger pseud- *R* squared indicating greater outcome predicative success, with a maximum of 1 [31]. In parallel to the prediction of missing cases, longitudinal models of symptom remission were conducted. These models intended to identify those participant characteristics that jointly predict missing cases and increased or decreased rate of symptom improvement following treatment. Longitudinal predictors of symptom change were examined with generalized estimating equation (GEE) models [32], as a series of separate univariate models. In combination, this step intended to test the ability of any one variable to predict missing cases likelihood, as well the outcomes those individuals were likely to experience at posttreatment.

In a second step, the 55 participants who were missing at posttreatment, but successfully recontacted at the 3-month follow-up, were also tested for their ability to represent missing cases who remained missing at both posttreatment and follow-up. The intention of this step was to suggest evidence that recontacted cases could be used as a proxy for missing posttreatment cases as a broader group. This was achieved by (1) Comparing the baseline symptom scores of cases with complete information ("completers"), missing cases at both time points ("completely missing cases"), and cases who are missing at post but are recontacted at 3-month follow-up ("recontacted cases"); (2) The characteristics of recontacted cases and completely missing cases were compared in a binary logistic regression seeking to test for differences between those recontacted cases and cases who were missing at both time points; and (3) To determine whether scores at 3-month follow-up could approximate posttreatment scores more broadly, a comparison between posttreatment and follow-up scores was conducted. In other words, testing whether missing cases who were recontacted at 3-month follow-up were likely to have similar treatment outcomes at posttreatment. Overall symptom change between post treatment and follow-up was tested with a longitudinal GEE model, testing for any additional symptom change between posttreatment and follow-up symptom outcomes.

In a third step, the third hypothesis was operationalized. This step compared simulated replacement scores, approximated by various adjusted models, against known outcome scores from recontacted cases. The aim of the third step was to quantify and test the relative accuracy of predicted replacement scores against known, proximal recontacted cases outcomes. Simulated follow-up scores were generated using longitudinal GEE and mixed models [33] as common longitudinal methods in clinical trials [34]. All models included a gamma scale, unstructured pattern of within subjects' correlation over time, and log link function to account for positive skewness and proportional remitting symptoms from baseline [21-24].

Various simulated scores were evaluated as either overestimating, underestimating, or being equivalent to recontacted cases scores in accordance to the degree they predicted the observed outcomes of recontacted cases. Specifically, if the mean CI of the simulated symptom replacement scores included the mean symptom outcome of the recontacted cases, statistical equivalence was interpreted [35]. If the CI interval of the mean replacement scores would exclude the mean of the recontacted cases, the simulation models were considered to overestimate or underestimate the outcomes of missing cases.

Statistical analysis was conducted using Statistical Package for the Social Sciences (SPSS) [36] version 22 (IBM Corp).

Results

Step 1 (H1, H2)—Joint Predictors of Missing Cases and Clinical Outcomes

Results from the first step, testing for predictors of missing values at posttreatment through univariate and multiple logistic regression models, are presented in Table 2.

These results demonstrate that as separate univariate models, and as a multivariate model, the stepwise variable selection identified baseline depressive symptoms (Wald χ^2_1=152.4, P<.001) and treatment adherence (Wald χ^2_4=10.1, P<.01) were the dominant predictors of missing cases probability. Together, these variables predicted 40.3% of the probability variance (Nagelkerke pseudo R squared=0.403), with treatment adherence accounting for the majority of that variance as a single dominant predictor (39%).

The impact of increased baseline severity demonstrated that for every one additional unit on the PHQ-9 at baseline, the odds of a participant to become a missing posttreatment case increased relatively by 8.4% (1.5% as a relative risk). The predictor of

treatment adherence demonstrated a strong but nonlinear predictor of missing cases probability. Specifically, participants who completed the entire program had only a 4% probability of becoming missing at posttreatment. In contrast, participants who completed only one lesson were over 70 times more likely to have missing posttreatment values relative to participants who attempted all five lessons (odds ratio=0.014).

An interaction between depressive baseline severity and treatment adherence was also explored and was found to be nonsignificant (Wald χ^2_4=3.0, P=.56). The nonsignificant interaction implies that baseline severity and treatment adherence were separate in their influences on missing cases.

Variables that influenced (moderated) the rate of symptom improvement were also tested. These analyses aimed to identify those participant characteristics that predicted the likelihood of an individual to become missing at posttreatment and at the same time, predict an individual's clinical outcome. Each of the nine variables were examined for their ability to predict increased symptom reduction following treatment through the statistical testing of a time by covariate interaction term. These interaction coefficients are presented in Table 3.

Table 2. Logistical regression model testing for predictor of missing cases of posttreatment. GAD-7: Generalized Anxiety Disorder Scale 7-item; PHQ-9: Patient Health Questionnaire 9-item.

Predictors of missing values	Univariate models				Multivariate models (P=.05)[a]		
	P	Odds ratio	Percentage of missing cases[b] (95% CI)	Variance explained, %	P	Percentage of missing cases[b] (95% CI)	Variance explained, %
Demographic					—	—	
Age (% per year)	<.001	0.97	−3 (−2 to −5]	3			
Gender					—	—	
Female	.14		16 (13-19)				
Male		0.74	20 (15 to 20)	<1			
Relationship status					—	—	
In a relationship	.04		15 (12 to 18)				
Otherwise		1.46	20 (16 to 25)	1			
Education level					—	—	
Tertiary education	.047		14 (12 to 18)				
Otherwise		1.48	20 (16 to 24)	1			
Initial severity			0 (0 to 0)				
Baseline anxiety symptoms (% per GAD-7 point)	.03	1.05	5 (0.5 to 9)	1	—	—	
Baseline depression symptoms (% per PHQ-9 point)	<.001	1.09	9 (5 to 14)[c]	4	.002	8 (3 to 14)[c]	40
Comorbidity at baseline: (PHQ-9≥10 and GAD-7≥8)	.01		20 (16 to 24)		—	—	
None		0.59	13 (10 to 17)	2			
Treatment adherence							
Completed all modules	<.001		4 (3 to 6)[c]	39		4 (3 to 6)[c]	40
Completed (4 of 5)		4.12	15 (10 to 22)		<.001	14 (9 to 21)	
Completed (3 of 5)		10	30 (21 to 41)		<.001	27 (18 to 38)	
Completed (2 of 5)		19.08	45 (33 to 59)		<.001	42 (29 to 56)	
Completed (1 of 5)		70.59	75 (64 to 84)		<.001	75 (63 to 84)	

[a]All models are based on a logistic regression model, including a log link function. Overall model accuracy for classification of missing values outcomes was 87.4%, with a specificity of 96.6% and sensitivity of 42%.

[b]Percentage of relative risk of an individual to become becoming missing at posttreatment.

[c]Relative odds of an individual to become a missing case with every additional unit increase.

Table 3. Association of predictor variables with clinical symptom change from baseline. GAD-7: Generalized Anxiety Disorder Scale 7-item; GEE: generalized estimating equation; PHQ-9: Patient Health Questionnaire 9-item.

Predictor of rate of clinical change	GEE[a] univariate models			Mixed univariate models		
	Moderation of symptom change (Time×IV) at posttreatment			Moderation of symptom change (Time×IV) at posttreatment		
	P	Wald chi-square (degrees of freedom)	Percentage change[b,c] (95% CI)	*P*	*F* statistic (degrees of freedom)	Percentage change[b,c] (95% CI)
Demographic						
Age (years, % per year)	.03	7.1 (1)	−1 (0 to −2)	.007	1.8 (1,1071)	<1 (<1 to <1)
Gender						
Female (versus male)	.43	1.7 (1)		.27	1.3 (1,1071)	
Relationship status						
In a relationship (versus otherwise)	.21	3.2 (1)		.22	1.5 (1,1071)	
Education level						
Tertiary (versus otherwise)	.17	3.5 (1)		.15	1.9 (1,1071)	
Initial severity						
Baseline anxiety symptoms (% per GAD-7 point)	>.99	0.1 (1)		.98	<0.1 (1,1071)	
Baseline depression symptoms (% per PHQ-9 point)	<.001	22.3 (1)	2 (1 to 3)	<.001	11.6 (1,1071)	2 (1 to 3)
Comorbidity at baseline: (PHQ-9≥10 and GAD-7≥8)	.16	3.6 (1)		.10	2.3 (1,1071)	
None						
Treatment adherence	<.001	39.0 (4)		<.001	3.6 (4,1071)	
Completed all modules			49 (45 to 52)			40 (16 to 56)
Completed (4 of 5)			40 (32 to 47)			29 (0 to 50)
Completed (3 of 5)			46 (36 to 55)			26 (−7 to 49)
Completed (2 of 5)			42 (27 to 53)			35 (2 to 57)
Completed (1 of 5)			21 (−8 to 43)			19 (−11 to 41)

[a]All models are based on a GEE model of change over time, interacting with a covariate.

[b]Percentage indication of a change from baseline.

[c]Marginal means reported for predictors with statistical significance (P<.05).

From Table 3, treatment adherence, baseline symptom levels, and age significantly moderated rate of symptom improvement following therapy. Greater rates of symptom improvement were observed with higher levels of treatment adherence and higher baseline depression scores.

Taken together, the predictors of treatment adherence and baseline PHQ-9 symptoms demonstrated a joint association with both the rate of clinical improvement and the likelihood of missing data at posttreatment. The ability of treatment adherence and PHQ-9 baseline symptoms to influence both clinical outcomes and missing cases probability is graphically illustrated in Figure 1 (missing cases likelihood and symptom change trends associated with program adherence) and Figure 2 (missing cases likelihood and symptom outcome trends associated with baseline severity).

Step 2—Testing Recontacted Cases as a Proxy of the Broader Group of Missing Cases

This step intended to establish evidence that recontacted cases at 3-month follow-up could be used as a proxy for the unknown outcomes of posttreatment missing cases. Initially, the baseline symptoms scores of the 3 missing cases subgroups were compared with a simple analysis of variance. A pairwise comparison of the PHQ-9 baseline symptom scores among the 3 groups indicated that participants who completed the surveys at both time points demonstrated overall lower PHQ-9 symptoms at baseline (PHQ-9 of 12.0; 95% CI 11.6-12.3) compared with recontacted cases (PHQ-9 of 13.7; 95% CI 12.5-15.0; P<0.001) and cases who were missing at posttreatment and 3-month follow-up (PHQ-9 of 13.6; 95% CI 12.6-14.1; P<0.001). However, participants who were recontacted at follow-up demonstrated equivalent symptom scores (P=0.54) to those participants who were completely missing. This finding indicated that missing cases and recontacted cases shared

similarities as a group of individuals who present with missing cases.

A second analysis was conducted attempting to identify differences between those individuals who were missing cases and recontacted (55/138) and those individuals who were missing at posttreatment and follow-up (83/138). A logistic regression that specified recontacts and completely missing cases as its binary outcome was conducted. All possible predictors of missing cases were considered and assessed for statistical significance at a *P* value of .05 or less to account for the size of the subgroup (n=138). The resulting logistic regression models did not identify any one predictor that could explain the probability of missing or recontacted status.

A third longitudinal GEE analysis was conducted to corroborate that posttreatment and follow-up symptom scores were similar enough on average to be used interchangeably. Consistent with previous findings [23], a 45% reduction in symptoms was observed from baseline (PHQ-9 of 12.3 [95% CI 12.0-12.7]) to posttreatment (PHQ-9 of 6.4; 95% CI 6.0-6.8; Wald χ^2_2=572.1; *P*< 0.001), with only a smaller (>7%) but significant additional improvement (PHQ-9 of 5.9; 95% CI 5.6-6.3; Wald χ^2_2=6.4; *P*< 0.001) detected between posttreatment and follow-up time points.

Figure 1. Treatment adherence (Completion out of five modules) and the likelihood of missing cases or symptom improvement from pretreatment levels (%); dotted lines illustrate 95% CI of the estimate. PHQ-9: Patient Health Questionnaire 9-item.

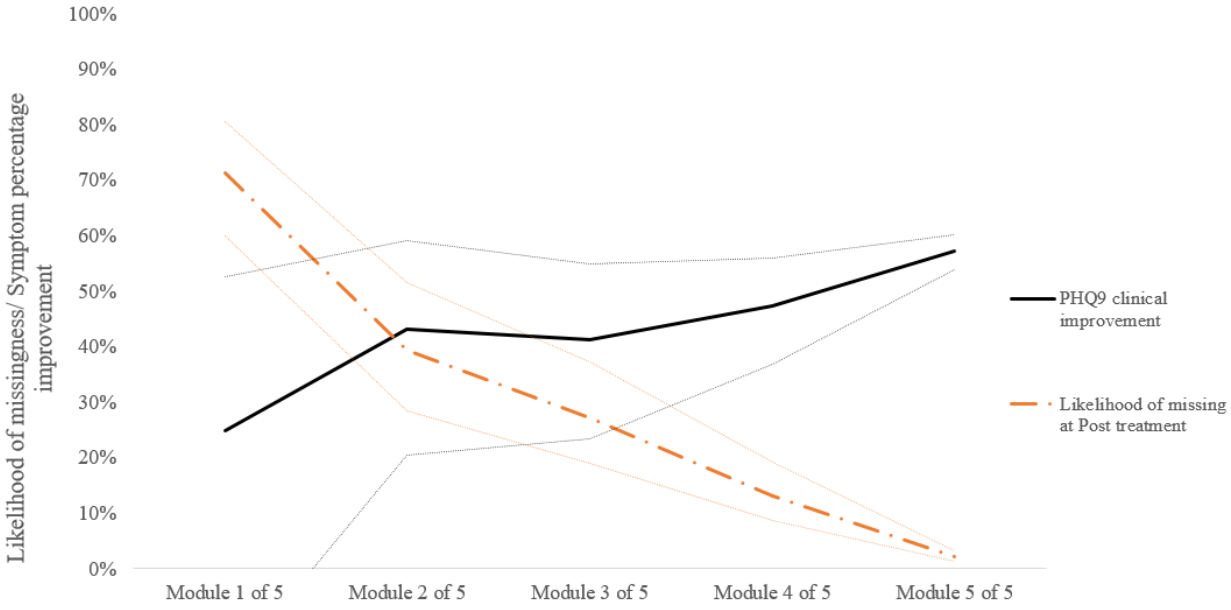
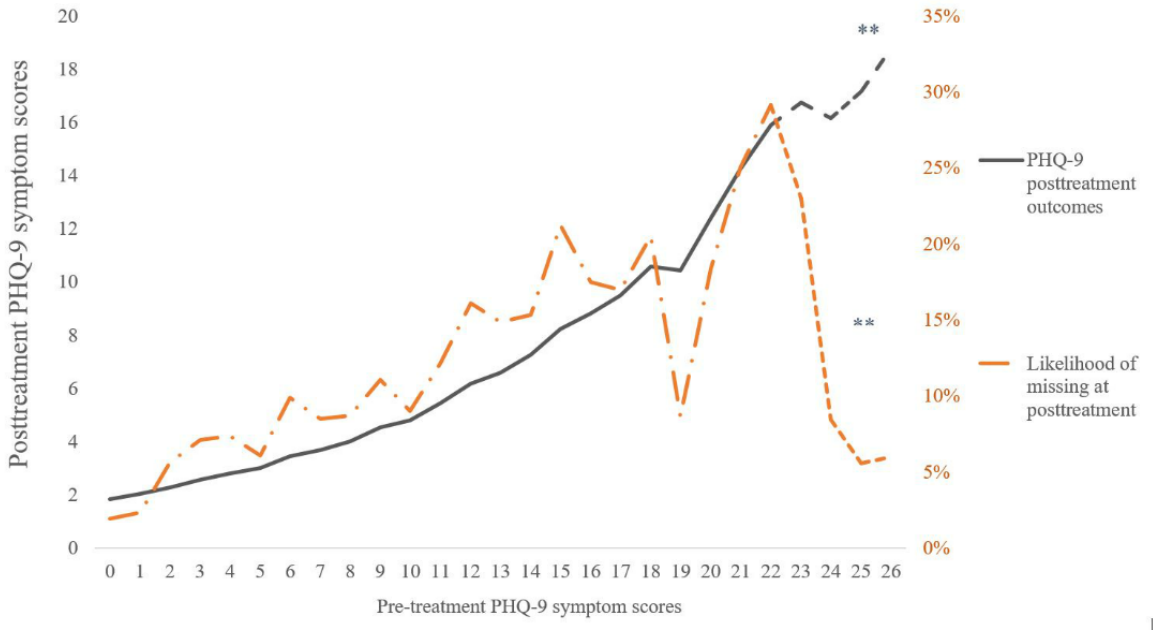


Figure 2. Pretreatment Patient Health Questionnaire 9-item (PHQ-9) symptoms influencing likelihood of missing cases or symptom outcomes. The **-dotted line implies a sample size of <10 participants from the sample of 820.



Together, these 3 results illustrated that the recontacted follow-up cases of this study present as a close, albeit imperfect, proxy for the outcomes of the broader group of individuals with missing posttreatment cases.

Step 3 (H3)—Using Recontacted Cases to Test the Accuracy of Simulated Replacement Score Under the Missing at Random, Missing Completely at Random, and Missing Not at Random Assumptions

In this step, the suitability of simulated replacement scores was explored by comparing the various predicted replacement scores against the known follow-up symptom outcome scores from recontacted individuals (Mean=8.11, 95% CI 6.53-10.07).

Table 4 presents the simulated mean PHQ-9 scores and CIs for replacement scores generated under different unadjusted and adjusted statistical models, as well as through the last observation carried forward (LOCF) and baseline observation carried forward (BOCF) methodology.

Table 4 illustrates those models that overlooked missing cases characteristics and did not adjust the approximation of missing cases; *underestimated* the symptom outcome scores of recontacted cases by as much as 30%. Similarly, replacement methods such as LOCF and BOCF both produced *significantly higher* estimates of symptom outcomes following treatment (24% and 69%, respectively).

Table 5 presents the mean and CIs generated through models that conditionally adjusted their estimation of missing cases outcomes. The approximated scores generated from each model are presented in Table 5 in descending order of accuracy; relative to the actual scores observed for recontacted cases. These results demonstrated that from the range adjusted models, models that included either treatment adherence or baseline severity in the prediction of outcomes could be interpreted as statistically equivalent to actual scores observed at 3-month follow-up. Specifically, both the GEE model and mixed model that adjust their estimates for treatment adherence and baseline severity resulted in the minimal approximation error (8%) relatively to the observed mean from actual outcomes.

Together, given some of the adjusted models were able to capture close approximations of the observed recontacted cases outcomes, the assumption of that missing cases cannot be conditionally compared with the remaining cases was refuted (MNAR).

Table 4. Depression (Patient Health Questionnaire 9-item, PHQ-9) simulate (approximated) replacement scores—unadjusted (missing completely at random, MCAR) models, last observation carried forward (LOCF), and baseline observation carried forward (BOCF). GEE: generalized estimating equation; N/A: not applicable.

Source of PHQ-9 estimates	Mean (95% CI)	Relative percentage accuracy from recontacted cases (95% CI)	Conclusion drawn about accuracy[a]
Recontacted cases	8.11 (6.53-10.07)	N/A	
BOCF	13.75 (12.57-15.03)	69 (55-85)	Significant overestimation
LOCF	9.96 (8.65-11.48)	24 (7-42)	Significant overestimation
MCAR (GEE)	5.93 (5.58-6.3)	−27 (−22 to −31)	Significant underestimation
MCAR (mixed)	5.96 (5.62-6.34)	−26 (−14 to −37)	Significant underestimation

[a]Relative accuracy from observed recontacted cases following a clinical follow-up.

Table 5. Depression (Patient Health Questionnaire 9-item; PHQ-9) simulate (approximated) replacement scores from various adjusted models. GAD-7: Generalized Anxiety Disorder Scale 7-item; GEE: generalized estimating equation; MAR: missing at random; N/A: not applicable.

Source of PHQ-9 estimates	Mean score (95% CI)	Relative percentage accuracy from recontacted cases (95% CI)	Conclusion drawn about accuracy[a]
Observed symptom score from recontacted cases	8.11 (6.53-10.07)	N/A	N/A
MAR PHQ-9 baseline and treatment adherence (GEE)	7.47 (6.84-8.15)	−8 (−16 to 0)	Statistical equivalence
MAR PHQ-9 baseline and treatment adherence (mixed)	7.5 (6.89-8.16)	−8 (−19 to 6)	
MAR GAD-7 baseline and treatment adherence (GEE)	7.15 (6.7-7.63)	−12 (−24 to 3)	Statistical equivalence
MAR GAD-7 baseline and treatment adherence (mixed)	7.28 (6.81-7.78)	−10 (−23 to 4)	
MAR treatment adherence (GEE)	6.91 (6.6-7.25)	−15 (−19 to −11)	Statistical equivalence
MAR treatment adherence (mixed)	7.06 (6.71-7.43)	−13 (−26 to 3)	
MAR PHQ-9 baseline (GEE)	6.57 (6.07-7.12)	−19 (−25 to −12)	Statistical equivalence
MAR PHQ-9 baseline (mixed)	6.54 (6.05-7.07)	−19 (−30 to −7)	
MAR comorbidity and education, and age (GEE)	6.31 (5.99-6.65)	−22 (−26 to −18)	Significant underestimation
MAR comorbidity and education, and age (mixed)	6.33 (6.01-6.66)	−22 (−34 to −8)	
MAR comorbidity (GEE)	6.23 (5.9-6.57)	−23 (−27 to −19)	Significant underestimation
MAR comorbidity (mixed)	6.24 (5.93-6.58)	−23 (−35 to −9)	
MAR GAD-7 baseline (GEE)	6.09 (5.82-6.37)	−25 (−28 to −21)	Significant underestimation
MAR GAD-7 baseline (mixed)	6.12 (5.85-6.4)	−25 (−36 to −10)	
MAR age (GEE)	6.03 (5.97-6.08)	−26 (−26 to −25)	Significant underestimation
MAR age (mixed)	6.07 (6.01-6.13)	−25 (−39 to −8)	
MAR Marital Status (GEE)	6 (5.83-6.17)	−26 (−28 to −24)	Significant underestimation
MAR marital Status (mixed)	6.03 (5.87-6.2)	−26 (−38 to −10)	
MAR education (GEE)	5.96 (5.88-6.04)	−27 (−28 to −26)	Significant underestimation
MAR education (mixed)	5.99 (5.91-6.07)	−26 (−40 to −9)	
MAR gender (GEE)	5.94 (5.89-6)	−27 (−27 to −26)	Significant underestimation
MAR gender (mixed)	5.98 (5.93-6.03)	−26 (−40 to −9)	

[a]Relative accuracy from observed recontacted cases following a clinical follow-up.

Discussion

Principal Findings

The primary aim of this study was to examine the characteristics, likely clinical outcomes, and statistical solutions that could be applied when missing cases in Web-based psychotherapy are encountered. This was done by first exploring the characteristics of missing cases within a large, naturalistic Web-based treatment sample; specifically identifying those participant characteristics that could predict the likelihood an individual would become missing following treatment, and at the same time, predict the outcomes such individual was likely to experience. In addition, this study attempted to test the suitability and accuracy of different statistical solutions for replacing missing cases (eg, adjusted and unadjusted model approximations, LOCF, and BOCF replacement strategies) through the comparison of statistically approximated outcomes against known outcomes from cases who were missing and were successfully recontacted (recontacted cases). The results were organized with three interrelated steps.

In a fundamental first step, the features of treatment adherence rates and baseline symptom severity were identified as predictors that can significantly increase the likelihood of participants to become missing at posttreatment. Together, treatment adherence and baseline symptoms explained 41% of the probability variance of missing cases status and were identified as the dominant predictor from a range of alternatives predictors initially included in the model. In this way, the first hypothesis, stating that missing cases were not occurring at random, was supported. This result demonstrated support for the first hypothesis, stating that missing cases were not occurring at random.

Critically, the variables of treatment adherence and baseline symptoms also shaped the clinical outcomes missing cases were likely to experience. Specifically, poorer treatment adherence was also associated with increased symptoms and distinct symptom outcomes. Similarly, higher pretreatment symptoms were associated with higher symptoms following treatment. This finding supported the second hypothesis and is consistent with research about the role of dosage, adherence, and treatment

outcomes [37-39]. At the same time, the association of increased symptoms and missing cases is also in line with previous research, suggesting that severely depressed participants are more likely to drop out [1,2,38]; and in parallel, an association between baseline severity and increased residual symptoms at posttreatment [40]. Recognizing treatment adherence and baseline severity as variables that predict both who will become missing and their likely clinical outcomes is key for understanding the likely clinical trajectory of missing cases.

In more statistical terms, this study demonstrated that missing cases cannot be assumed to be a random portion of the overall sample (MCAR), and overlooking the specific pattern of treatment adherence and baseline severity can result in overestimation of treatment efficacy and underestimate remaining symptom. The additional comparison of proximal recontacted cases with replacement methods such as LOCF and BOCF also demonstrated significant measurement error with overestimation that is as high as 70%; consistent with previous research [41,42], indicating these methods lead to overly conservative underestimates of treatment benefits.

Finally, testing of the third hypothesis demonstrated that missing cases could be predicted with minimal error; however, only by accounting for the specific variables that influence both missing cases likelihood and clinical outcomes. Specifically, among all the available model-based approximation methods, models that adjust their estimate of clinical outcomes by treatment adherence and baseline symptom severity demonstrated acceptable statistical accuracy. Using either GEE or mixed methodologies, models that adjusted for both treatment adherence and baseline severity of symptoms resulted in prediction that were only 8% lower than actual values of recontacted cases and were considered statistically equivocal. This result can also be interpreted as a verification of the suitability of replacing missing cases through adjusted replacement strategies under conditional MAR assumption; that is, given that the approximation of missing cases outcomes resulted in minimum differences from the observed outcomes of recontacted cases, the suitability of the statistical approximation is supported. In addition, these results could be interpreted as refuting of the MNAR assumption, given that missing cases were accurately captured under conditionally adjusted models (adjusted for treatment adherence and baseline symptoms).

To our knowledge, this is the first study to use naturalistic measurement to verify whether missing psychotherapy cases conceal poorer clinical outcomes, as well as explore both the bias and underpinning causes. These findings are, however, consistent with current thinking about the potential causes of, and outcomes for, missing cases [1,2,9,20], as well as a long standing statistical requirement to take steps to identify and resolve missing cases bias [6,9,10].

The importance of recognizing key predictors of missing cases, as well as their clinical outcomes can be considerable. Missing cases in psychotherapy research are common [1,2] and can pose a fundamental challenge for measurement and interpretation of clinical effects [43]. On the basis of the present findings, researchers seeking to produce accurate and more complete estimates of treatment outcomes should consider whether missing cases in their own datasets show an association with variables such as treatment adherence and baseline treatments. If these trends are present, missing cases and their outcomes may not be random, and further steps would be needed to truly estimate the effects treatment. Although the implications missing cases pose for other aspects of clinical measurement is beyond the scope of this paper, the pattern of results demonstrated in this paper may certainly impact additional clinical measurement practices. For example, research aiming to identify clinical moderators, quantify patient risk, evaluate treatment efficacy, or make treatment comparison may certainly be impacted by missing cases patterns, such as those identified in this study, or additional patterns that could be identified through similar other research.

Limitations and Future Directions

Although this study relied on a large clinical sample with high internal reliability, the results and conclusions drawn must be considered with several limitations. First, and foremost, the demonstration of missing cases characteristics, their approximated outcomes, and the suitability of replacing missing cases is preliminary and specific to a treatment model (iCBT) [30]. As shown by previous research [1], the proportion of missing values and clinical outcomes vary widely between trials. This variability may suggest that different clinical samples could also show both different predictors of missing cases and different outcome trajectories experienced by missing cases. However, broadly speaking, given that treatment engagement and initial depressive symptom rate commonly associated with both treatment adherence [2] and outcomes [41], these variable may reflect a critical starting point for the examination of missing cases in other Web-based psychotherapy trials, if not psychotherapy in general.

A second limitation relates to the use of recontacted cases to verify the suitability of statistical methods to replace missing cases. This sample of recontacted cases relied on a modest sample of 55 recontacted cases. Despite efforts to empirically compare recontacted cases with completely missing cases, recontacted cases can only be assumed to represent the larger group of missing cases. Albeit the uncertainty associated with recontacted cases, it is important to note that recontacted cases embody naturally occurring proximal outcomes that cannot be researched with artificial statistical studies. Given that no alternative is currently available to verify the outcomes for missing cases, recontacted cases may prove a novel future measurement proxy for missing cases as a broader group.

To address both limitations, replication of these missing patterns and research methodology in other similar treatment samples is key. It is important to note that investigating missing cases in naturalistic, clinical settings, as well as collating a sizeable group of recontacted cases is not straightforward given their rarity (eg, 55/820). However, increasingly large and standardized psychotherapy databases are becoming available [1], and these large databases may enable to similarly research methodology and exploration of predictors, outcomes and proximal measurements for missing cases.

In addition, it is important to acknowledge that this study does not pertain to exhaust the theoretical causes, or the identification

of predictors that may underpin missing cases and their outcomes. Other alternative important participant variables could indeed play a role in underpinning why cases become missing and how their outcomes should be approximated. For example, the presence of a major depression diagnosis [39], credibility, or motivation [38] may lead to different rates of treatment adherence and at the same time, better capture the trajectory of missing cases in treatment. For this reason, similar future studies may consider a more direct measurement of participant engagement that may underpin their trajectory in treatment. For example, measurements of motivation, enthusiasm, clinical barriers, treatment credibility, or other clinical consideration may offer a more interruptible means to profile missing cases and their likely clinical outcomes.

Furthermore, it is important to consider that the ability to use adjusted approximation models that factor both treatment adherence and baseline symptoms may not be realistic in small samples. For example, a psychotherapy sample of 30 or less, may be underpowered, or show insufficient variance for the use of complex adjusted statistical models. For this reason, more parsimonious and more robust solutions for replacing missing cases in smaller samples should be explored. For example, methods that are less statistically demanding, such as the application of LOCF for cases who do not complete treatment, could be coupled with unadjusted (MCAR), approximation of outcome for those cases that adhere to treatment in full. This type of hybrid solution may result in a less statistically demanding strategy, which hyphenates the LOCF overly conservative approximation of outcomes, with the MCAR assumption, which is overly liberal as a method that underestimates symptom outcomes. Such solutions are beyond the scope of this paper; however, the application of corrective missing cases methods for small samples may be key for psychotherapy trials such as pilots and small RCTs.

Finally, it is important to note that results of this study imply that within Web-based psychotherapeutic interventions such as CBT-based interventions, the role of adherence and baseline symptoms could likely be important and implicit. Recognizing such patterns can lead to clearer understanding of missing cases, the assumptions that can be made about missing cases, and a more accurate consideration of their outcomes. Although these results should be considered as possible fundamental pattern in the application of any statistical replacement strategy, it important to note that this study does not advocate the use of any one statistical approach over another as means for handling missing cases. Rather, this study intended to explore the implicit characteristics that influence Web-based psychotherapy cases and suggest those measurement considerations that would likely improve the application of missing cases strategies.

In summary, this research aimed to create a more concrete awareness of missing cases and ways to handle missing cases in Web-based psychotherapeutic trials. Using concrete and transparent statistical modeling, this research demonstrated that missing cases can occur systematically and with clinical outcomes that are dissimilar to the outcomes of those individuals who are surveyed following treatment. This study also offered (1) a research design framework that can concretely quantify the outcome bias associated with naturalistically occurring missing cases, (2) highlight important predictors that explain both missing cases and their outcomes, and (3) suggest a naturalistic benchmark (recontacted cases) that could be conditionally used for quantifying the outcomes for missing cases and verifying the suitability of various statistical solutions that approximate missing cases. Together, all three aspects of characteristics, bias in outcomes, and methods to resolve the bias in outcomes should be considered preliminary and pendent on future replication.

Conflicts of Interest

None declared.

References

1. Karyotaki E, Kleiboer A, Smit F, Turner DT, Pastor AM, Andersson G, et al. Predictors of treatment dropout in self-guided web-based interventions for depression: an 'individual patient data' meta-analysis. Psychol Med 2015 Oct;45(13):2717-2726. [doi: 10.1017/S0033291715000665] [Medline: 25881626]
2. Fernandez E, Salem D, Swift JK, Ramtahal N. Meta-analysis of dropout from cognitive behavioral therapy: magnitude, timing, and moderators. J Consult Clin Psychol 2015 Dec;83(6):1108-1122. [doi: 10.1037/ccp0000044] [Medline: 26302248]
3. Nich C, Carroll KM. Intention-to-treat meets missing data: implications of alternate strategies for analyzing clinical trials data. Drug Alcohol Depend 2002 Oct 1;68(2):121-130 [FREE Full text] [Medline: 12234641]
4. Cavanagh K. Turn on, tune in and (don't) drop out: engagement, adherence, attrition, and alliance with internet-based interventions. In: Bennett-Levy J, Richards D, Farrand P, Christensen H, Griffiths K, Kavanagh D, et al, editors. Oxford Guide to Low Intensity CBT Interventions. Oxford, London: Oxford University Press; 2010.
5. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. The Lancet 2001 Apr;357(9263):1191-1194. [doi: 10.1016/S0140-6736(00)04337-3]
6. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med 2012 Oct 4;367(14):1355-1360 [FREE Full text] [doi: 10.1056/NEJMsr1203730] [Medline: 23034025]
7. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ 1999;319:670-674 [FREE Full text] [doi: 10.1136/bmj.319.7211.670]

8. Bell ML, Fairclough DL. Practical and statistical issues in missing data for longitudinal patient-reported outcomes. Stat Methods Med Res 2014 Oct;23(5):440-459. [doi: 10.1177/0962280213476378] [Medline: 23427225]
9. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;7(2):147-177. [doi: 10.1037/1082-989x.7.2.147]
10. Rubin DB. Multiple imputation after 18+ Years. J Am Stat Assoc 1996 Jun;91(434):473-489. [doi: 10.1080/01621459.1996.10476908]
11. Little RJ. Modeling the drop-out mechanism in repeated-measures studies. J Am Stat Assoc 1995 Sep;90(431):1112-1121. [doi: 10.1080/01621459.1995.10476615]
12. Enders CK. Applied Missing Data Analysis (Methodology in the Social Sciences), 1st edition. New York, NY: The Guilford Press; 2010.
13. Robins J, Rotnitzky A, Scharfstein DO. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Statistical Models in Epidemiology, the Environment, and Clinical Trials. New York, NY: Springer; 2000:1-94.
14. Mallinckrodt CH. Preventing and Treating Missing data in Longitudinal Clinical Trials: A Practical Guide. New York, NY: Cambridge University Press; 2013.
15. Little RJ, Rubin DB. Statistical Analysis With Missing Data, 2nd Edition. New Jersey: Wiley-Interscience; 2014.
16. Christensen H, Mackinnon A. The law of attrition revisited. J Med Internet Res 2006 Sep 29;8(3):e20 [FREE Full text] [doi: 10.2196/jmir.8.3.e20]
17. Donkin L, Hickie IB, Christensen H, Naismith SL, Neal B, Cockayne NL, et al. Rethinking the dose-response relationship between usage and outcome in an online intervention for depression: randomized controlled trial. J Med Internet Res 2013 Oct 17;15(10):e231 [FREE Full text] [doi: 10.2196/jmir.2771] [Medline: 24135213]
18. Melville KM, Casey LM, Kavanagh DJ. Dropout from internet-based treatment for psychological disorders. Br J Clin Psychol 2010 Nov;49(Pt 4):455-471. [doi: 10.1348/014466509X472138] [Medline: 19799804]
19. Hedman E, Ljótsson B, Blom K, El Alaoui S, Kraepelien M, Rück C, et al. Telephone versus internet administration of self-report measures of social anxiety, depressive symptoms, and insomnia: psychometric evaluation of a method to reduce the impact of missing data. J Med Internet Res 2013 Oct 18;15(10):e229 [FREE Full text] [doi: 10.2196/jmir.2818] [Medline: 24140566]
20. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clin Trials 2004 Sep;1(4):368-376. [doi: 10.1191/1740774504cn032oa] [Medline: 16279275]
21. Titov N, Dear BF, Johnston L, Lorian C, Zou J, Wootton B, et al. Improving adherence and clinical outcomes in self-guided internet treatment for anxiety and depression: randomised controlled trial. PLoS One 2013 Jul;8(7):e62873 [FREE Full text] [doi: 10.1371/journal.pone.0062873] [Medline: 23843932]
22. Dear BF, Staples LG, Terides MD, Karin E, Zou J, Johnston L, et al. Transdiagnostic versus disorder-specific and clinician-guided versus self-guided internet-delivered treatment for generalized anxiety disorder and comorbid disorders: a randomized controlled trial. J Anxiety Disord 2015;36:63-77 [FREE Full text] [doi: 10.1016/j.janxdis.2015.09.003]
23. Titov N, Dear BF, Staples LG, Terides MD, Karin E, Sheehan J, et al. Disorder-specific versus transdiagnostic and clinician-guided versus self-guided treatment for major depressive disorder and comorbid anxiety disorders: a randomized controlled trial. J Anxiety Disord 2015 Oct;35:88-102 [FREE Full text] [doi: 10.1016/j.janxdis.2015.08.002] [Medline: 26422822]
24. Titov N, Dear BF, Staples LG, Bennett-Levy J, Klein B, Rapee RM, et al. MindSpot clinic: an accessible, efficient, and effective online treatment service for anxiety and depression. Psychiatr Serv 2015 Oct;66(10):1043-1050. [doi: 10.1176/appi.ps.201400477] [Medline: 26130001]
25. Kroenke K, Spitzer RL, Williams JB. The PHQ-9. Validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x]
26. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
27. Ecentreclinic. Welcome to the eCentreClinic URL: https://www.ecentreclinic.org/ [accessed 2018-02-23] [WebCite Cache ID 6xRTeITXq]
28. Clark DM. Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: the IAPT experience. Int Rev Psychiatry 2011 Aug;23(4):318-327 [FREE Full text] [doi: 10.3109/09540261.2011.606803] [Medline: 22026487]
29. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. CMAJ 2012 Feb 21;184(3):E191-E196 [FREE Full text] [doi: 10.1503/cmaj.110829] [Medline: 22184363]
30. Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. New York: Springer Science & Business Media; 2015:9425.
31. Nagelkerke NJ. A note on a general definition of the coefficient of determination. Biometrika 1991;78(3):691-692. [doi: 10.2307/2337038]

32. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73(1):13-22. [doi: 10.1093/biomet/73.1.13]
33. Diggle P, Liang KY, Zeger SL. Analysis of Longitudinal Data. Oxford, UK: Oxford University Press; 1994.
34. Hubbard AE, Ahern J, Fleischer NL, Van der Laan M, Lippman SA, Jewell N, et al. To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. Epidemiology 2010 Jul;21(4):467-474. [doi: 10.1097/EDE.0b013e3181caeb90] [Medline: 20220526]
35. Greene CJ, Morland LA, Durkalski VL, Frueh BC. Noninferiority and equivalence designs: issues and implications for mental health research. J Trauma Stress 2008 Oct;21(5):433-439 [FREE Full text] [doi: 10.1002/jts.20367] [Medline: 18956449]
36. IBM. 2013. IBM SPSS Statistics V22.0 URL: https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=an&subtype=ca&appname=gpateam&supplier=897&letternum=ENUS213-309 [accessed 2018-02-22] [WebCite Cache ID 6xPrtZhpM]
37. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
38. Alfonsson S, Olsson E, Hursti T. Motivation and treatment credibility predicts dropout, treatment adherence, and clinical outcomes in an internet-based cognitive behavioral relaxation program: a randomized controlled trial. J Med Internet Res 2016 Mar 8;18(3):e52. [doi: 10.2196/jmir.5352]
39. DiMatteo MR, Lepper HS, Croghan TW. Depression is a risk factor for noncompliance with medical treatment: meta-analysis of the effects of anxiety and depression on patient adherence. Arch Intern Med 2000 Jul 24;160(14):2101-2107. [Medline: 10904452]
40. Bower P, Kontopantelis E, Sutton A, Kendrick T, Richards DA, Gilbody S, et al. Influence of initial severity of depression on effectiveness of low intensity interventions: meta-analysis of individual patient data. BMJ 2013;346:f540. [doi: 10.1136/bmj.f540]
41. Streiner DL. Missing data and the trouble with LOCF. Evid Based Ment Health 2008 Feb;11(1):3-5. [doi: 10.1136/ebmh.11.1.3-a] [Medline: 18223040]
42. Molnar FJ, Man-Son-Hing M, Hutton B, Fergusson D. Have last-observation-carried-forward analyses caused us to favour more toxic dementia therapies over less toxic alternatives? A systematic review. Open Med 2009;3(2):e31-e50.
43. Gilbody S, Littlewood E, Hewitt C, Brierley G, Tharmanathan P, Araya R, et al. Computerised cognitive behaviour therapy (cCBT) as treatment for depression in primary care (REEACT trial): large scale pragmatic randomised controlled trial. BMJ 2015 Nov 11;351:h5627. [doi: 10.1136/bmj.h5627]

Abbreviations

BOCF: baseline observation carried forward
CBT: cognitive behavioral therapy
GEE: generalized estimating equation
GAD-7: Generalized Anxiety Disorder Scale 7-item
iCBT: Internet cognitive behavioral therapy
LOCF: last observation carried forward
MAR: missing at random
MCAR: missing completely at random
MNAR: missing not at random
MUM: Macquarie University Web-based Model
RCT: randomized controlled trial
PHQ-9: Patient Health Questionnaire 9-item

Edited by G Eysenbach; submitted 06.07.17; peer-reviewed by B Meyer, B Clough, H Baumeister, J Apolinário-Hagen; comments to author 19.09.17; revised version received 14.11.17; accepted 22.12.17; published 19.04.18

Please cite as:
Karin E, Dear BF, Heller GZ, Crane MF, Titov N
"Wish You Were Here": Examining Characteristics, Outcomes, and Statistical Solutions for Missing Cases in Web-Based Psychotherapeutic Trials
JMIR Ment Health 2018;5(2):e22
URL: http://mental.jmir.org/2018/2/e22/
doi:10.2196/mental.8363
PMID:29674311

Chapter 5

Examining Characteristics, Outcomes, and Statistical Solutions for Missing Cases in Web-Based Psychotherapeutic Trials – a replication and extension (Study 4)

This chapter describes a study that replicated and extended Study 3 and which aimed to identify features of missing cases and the type of statistical assumption that can be made about the outcomes of missing psychotherapy cases. The study employed a similar methodology conducted in Study 3 but using data from psychotherapy routine care. The study also explored the ability to approximate clinical outcomes using several symptom scales

including those measuring anxiety, psychological distress and depressive symptoms. The study aims to replicate finding from the first study, about the trajectory of missing cases as a broader phenomena. In this way, the study aims to gauge the opportunities for achieving measurement that is more suited to the features of missing cases as a phenomena (gauging internal validity), as well as the application of these features across clinical contexts (gauging external validity).

Publication status

This chapter has been submitted for publication, to the *Journal of Psychotherapy Research* and is currently under review (TPSR-2019-0172).

Author contribution:

Mr Eyal Karin designed, analysed, and wrote the study. Dr Monique Crane, Associate Professor Blake F. Dear, and Dr Rony Kayrouz provided the dataset, assisted with the refinement of the manuscript, and helped frame the methodological content for a clinical audience. Professor Nick Titov oversaw the conception of the project and the drafting of the manuscript.

Abstract

Background: Missing cases are common in psychotherapy trials, and challenge the ability to evaluate the effects of treatment. Whilst common, little is known about the characteristics of missing cases, their likely clinical outcomes, or the suitability of different replacement methodologies.

Objective: To explore the characteristics of missing cases, their likely treatment outcomes, and the ability of different statistical models to accurately approximate missing post-treatment data.

Methods: A sample of web-based cognitive behavioural therapy participants in routine care (*n*=6701) was used to identify predictors of missing cases probability, and predictors that moderated clinical outcomes, such as psychological distress, anxiety and depressive symptoms. These variables were then incorporated into a range of statistical models that approximate missing cases replacement outcomes, with the results compared through sensitivity and cross-validation analyses.

Results: Lower treatment adherence and increased symptoms at pre-treatment were identified as the dominant predictors of missing cases, as well as the rate of symptom change. Statistical replacement methods that overlooked theses prominent features underestimated missing cases outcomes by as much as 40%.

Conclusions: Missing cases experienced treatment outcomes that were distinct from the remaining observed sample. By overlooking the features of missing cases, clinical measurement, and the evaluation of treatment can be compromised.

STUDY 4 - CHARACTERISTICS AND SOLUTIONS FOR MISSING CASES

Introduction

The ability to accurately evaluate psychotherapy depends on the measurements conducted during and after interventions. Unfortunately, some participants are unable to complete such measurements, and become so-called *missing cases,* thus threatening the validity of conclusions of such trials. Missing cases are frequently reported in psychotherapy trials (Fernandez, Salem, Swift, & Ramtahal, 2015; Christensen, Griffiths & Farrer, 2009; Karyotaki, Kleiboer, Smit, Turner, Pastor, Andersson,…, & Cuijpers, 2015; Waller & Gilbody, 2009), and pose a risk to the validity of clinical evidence (DeSouza, Legedza, & Sankoh, 2009; Little, D'Agostino, Cohen, Dickersin, Emerson, Farrar, ... & Stern, 2012; Karin, Dear, Heller, Crane & Titov, 2018). Overlooking the causes and outcomes of missing cases can lead to systematic measurement bias and misrepresentation of treatment outcomes, and therefore risks compromising the validity of clinical research (Bell & Fairclough, 2014; Rubin, 1976; Little, et al., 2012). For this reason, missing cases are considered an important part of the measurement process of clinical evidence.

Although the importance of handling missing cases is well understood (Lang & Little, 2018; Little et al., 2012), accounting for the outcomes of missing cases is a challenging task as researchers can never verify if the replacement values they generate accurately captures the outcomes of patients. Thus, researchers must rely on statistical approximation and the assumption that any replacement outcomes are suitable (DeSouza, et al., 2009; Mealli & Rubin, 2015; Schafer & Graham, 2002).

A key requirement for handling missing data is to ensure the outcomes of missing cases are represented within statistical analyses (Mealli & Rubin, 2015), and this usually involves using a statistical solution to generate replacement values for missing cases. To determine that a statistical solution for missing cases is suitable, researchers rely on statistical methods that

explore the characteristics of missing cases, and whether these features could also be associated with distinct clinical outcomes. This is typically achieved through analyses that identify variables that predict both the probability of participants becoming missing, as well as a participants' clinical outcome (Karin, Dear, Heller, Crane & Titov, 2018; Little & Rubin, 2014; Mallinckrodt, 2013; Mealli & Rubin, 2015). Identifying such variables enables researchers to generate replacement scores that are likely to capture the outcomes of treatment for missing cases (Lang & Little, 2018; Little & Rubin, 2014; Blackwell, Honaker & King, 2017). For example, if an increased age is associated with a decreased probability of becoming a missing case, and an increased rate of symptom change, a statistical model that can adjust for participant's age would create replacement outcomes that are more accurate and representative of the effects of treatment than models that overlook age. In statistical terms, variables that both predict the likelihood of missing cases and the outcome of missing cases are termed the mechanism of non-ignorable missing cases (Rubin, 1976, Little, 1995; Rubin & Little, 2014).

Although the replacement values for missing cases using adjusted statistical models has been in use for decades (Lang & Little, 2018; Mealli & Rubin, 2015; Schafer & Graham, 2002), limited research is available to determine the characteristics psychotherapy missing cases, or identify non-ignorable mechanisms of missingness that may influence the measurement of psychotherapy outcomes (Alfonsson, Olsson & Hursti, 2016; Karyotaki et al., 2015). Consequently, little is known about the ability of researchers to approximate the outcomes for missing cases in psychotherapy studies or the impact of missing cases on psychotherapy outcome evaluation. This gap in methodological research, may result from (1) the limited knowledge about missing cases, and the patient features that may generalise across clinical trials (Karyotaki et al., 2015), and (2) the scarcity of large and comparable treatment samples that are statistically powered to explore non-ignorable mechanisms of missingness.

Preliminary evidence from web-based cognitive behavioural therapy trials suggests that common patient variables, such as treatment adherence and baseline depressive symptom severity, dominantly predicted both the likelihood of cases to become missing, as well as moderating the clinical effect (Karin, et al., 2018; Karyotaki et al., 2015). These findings suggested missing cases are not comparable to the patients that provide their data following treatment, and that missing cases are characterised by lower treatment adherence and high baseline symptoms. Consequentially, these variables represent non-ignorable mechanisms of missing data mechanisms and an important consideration in estimating the outcomes of missing cases in web-based psychotherapy. Without accounting for these variables, web-based psychotherapy researchers risk overlooking a systematic pattern of poorer treatment outcomes for missing cases, and may generate estimates of treatment effects that are unrepresentative and unrealistic. However, the available literature concerning missing cases in web-based psychotherapy is limited to a single study that focused on depression symptoms and employed data from a highly controlled clinical trial with good participant retention rates (Karin, et al., 2018). Thus, more research replicating this work and employing different samples and a broader range of outcomes is needed before conclusions can be drawn concerning the characteristics of missing cases in web-based psychotherapy, their impacts on clinical outcomes and the optimal approaches for handing missing cases.

The present study

The primary aim of this study was to extend on earlier work and explore the characteristics of missing cases, their possible clinical outcomes following psychotherapy, and compare different statistical methods for estimating missing outcomes. To do this, the current study employs a large routine care sample from an established digital mental health service offering web-based psychotherapy clinic (n = 6701), and explored outcomes of missing cases in the context of a broader range of clinical outcome domains, such as depression, anxiety and

psychological distress. In line with previous preliminary research concerning the features of missing cases in psychotherapy (Alfonsson, et al., 2016; Fernandez et al., 2015; Karin et al., 2018; Karyotaki et al., 2015), it was hypothesised that lower treatment adherence and increased baseline depressive symptoms will predict both increased missing cases likelihood as well as higher post-treatment depressive symptom outcomes (H1). Consequently, statistical models that account for these features (i.e., treatment adherence, baseline symptoms) will result in higher post-treatment symptom replacement scores compared to statistical models that overlook these features and operate under a missing completely at random assumption (H2).

Method

The sample

The present study employed the clinical participant data from an Australian national digital mental health service, the MindSpot Clinic (mindspot.org.au). All participants provided informed consent for their de-identified data to be used in evaluation and quality improvement activities, and approval for such activities was provided by the Macquarie University Human Research Ethics Committee. Additional information about the sample, the effectiveness of the internet-delivered cognitive behavior therapy (iCBT), course content and delivery protocols can be found elsewhere (Titov et al., 2017; Nielssen, Dear, Staples, Dear, Ryan, Purtell, & Titov, 2015). It is important to note that web-based psychotherapy presents a unique opportunity for investigating missing cases and their trajectories due primarily to the highly standardized nature of the clinical engagement and outcome measurement procedures. By reducing the variability associated with treatment delivery the resulting variance in treatment outcomes can be attributed more to the individual differences of cases rather than the delivery of treatment. Thus, the treatment standardization represents a unique opportunity to measure missing cases outcomes with increased internal validity.

The total sample consisted of 6701 participants who initiated treatment, 64% of whom provided data at post-treatment (*n*=4271), and 36% of the sample missing at post-treatment (*n*=2430). The symptom change estimates were captured by measuring symptoms before treatment and after the eight-week course. This sample comprises treatment-seeking individuals registering for treatment within a time window of 30 months.

The sample of 6701 was randomly allocated into five subgroups, each including over 1340 participants at pre-treatment, and over 840 complete measurements at post-treatment. Table 1 collates the samples demographic information, including chi-square test statistics obtained from statistical tests that checked for successful randomisation within each stratum of the sample.

Table 1 - Randomisation of cross validation samples and participant characteristics

Sample available at pre-treatment		Available sample at post treatment	Randomisation test
Total sample (*n*=6701)		*n*=4271 (64%)	
Replication sample 1 (*n*=1341)		*n*=842 (64%)	
Replication sample 2 (*n*=1340)		*n*=846 (64%)	$\chi2$ =3.768, p = 0.438
Replication sample 3 (*n*=1340)		*n*=843 (64%)	
Replication sample 4 (*n*=1340)		*n*=846 (64%)	
Replication sample 5 (*n*=1340)		*n*=848 (64%)	
Variables considered	Mean (SD)	Count (% of total)	Randomisation test
Average age (SD)	37.57 (10.9)		$\chi2$ =3.768, p = 0.438
Completed 1/5 modules		513 (8%)	
Completed 2/5 modules		715 (11%)	
Completed 3/5 modules		718 (11%)	$\chi2$ =7.533, p = 0.962
Completed 4/5 modules		653 (10%)	
Completed 5/5 modules		4102 (61%)	
In a relationship		4458 (67%)	$\chi2$ =0.546, p = 0.969
Employment (employed)		4908 (73%)	$\chi2$ =0.755, p = 0.944
Education (Tertiary)		3239 (49%)	$\chi2$ =3.952, p = 0.413
Gender (Female)		4866 (73%)	$\chi2$ =6.803, p = 0.147
Comorbidity (GAD-7 ≤ 8 and PHQ-9 ≤ 10)		3437 (51%)	$\chi2$ =2.976, p = 0.562

PHQ-9 - Patient Health Questionnaire -9 Item; GAD-7 – Generalized Anxiety Disorder-7-Item Scale

Intervention

The Wellbeing Course was developed at the eCentreClinic, Macquarie University, in Sydney, Australia (Dear et al., 2015; 2016; Fogliati et al., 2016; Titov, Dear, Staples, Terides

et al., 2015). The course is designed for patients experiencing depression and or anxiety and contains five lessons. The lessons are released gradually over 8 weeks, covering: (1) the cognitive behavioural model and symptom identification; (2) thought monitoring and challenging; (3) de-arousal strategies and pleasant activity scheduling; (4) graduated exposure; and (5) relapse prevention. Patients can also download lesson summaries, patient stories as well as additional resources (e.g., sleep, problem-solving, communication). Each of the lessons provided homework assignments to assist participants in learning and applying the skills described in the lessons. Participants are strongly encouraged to practice the skills taught within the course daily and to gradually adopt them into their everyday lives.

Measures

The primary outcome measures for this study comprised several standardised symptom scales. These scales and their psychometric properties are presented below in Table 2.

Table 2 - Assessed psychometric properties of outcome measures

Scale	Primary pathology measured	Cited origin	Range	Cut-off indicative of the clinical range**	Internal consistency (Cronbach's α)	Intra class correlation coefficient (Two-way random, single score)*
Patient Health Questionnaire – 9 Item Scale (PHQ-9)	Depression	Kroenke, Spitzer, & Williams, 2001	0-27	10	0.848	0.716
Generalized Anxiety Disorder 7-Item Scale (GAD-7)	Anxiety	Spitzer, Kroenke, Williams, & Lowe, 2006	0-21	8	0.849	0.744
The Kessler 10-Item Scale	Psychological distress	Kessler, Andrews, Colpe, Hiripi, Mroczek, Normand, Walters & Zaslavsky, 2002	10-50 (denoted 0-40)	20 (denoted 10)	0.830	0.710

*Estimates identified by comparing patient scores during assessment intake and then again at the point of pre-treatment scores 4-8 weeks later. **Proposed cut-offs by the authors of the original papers; PHQ-9 - Patient health questionnaire, nine-item scale; GAD-7 – Generalised anxiety disorder scale, seven-item scale; K-10 - The Kessler 10-Item Scale

The following measures were also included as possible independent variables/ predictors that may characterise missing cases and their clinical trajectories through treatment.

Comorbidity

Individuals were considered to have comorbidity if they demonstrated scores of both anxiety and depression above pre-determined clinical thresholds (GAD-7 ≥8 & PHQ-9≥10 at baseline) (Johansson, Carlbring, Heedman, Paxling, & Andersson, 2013; Johnston, Titov, Andrews, Dear, & Spence, 2013).

Demographic measures

Age in years at the start of treatment, relationship status, pre-treatment symptom scores, pre-treatment anxiety scores, and educational attainment were considered as predictor variables of both treatment outcomes and missing cases. The frequencies of participants within categories of educational attainment, relationship status, treatment adherence, and gender, are presented in Table 1.

Treatment adherence

Treatment adherence was measured as the minimal, but a continued progression of participants through the intended five modules of the course; consistent with common definitions of treatment adherence in eHealth interventions (Sieverink, Kelders, ., & van Gemert-Pijnen, 2017). Increased adherence was measured by (1) minimal login to the assigned secured website, and (2) access the lesson modules that were incrementally released as a part of the Wellbeing Course syllabus.

Analytical plan

Identifying predictors of missing cases and the rate of clinical change

The characterisation of missing cases and the approximation of their likely outcomes was operationalised with three steps using SPSS (IBM) version 25 and a dedicated *R* software (R Core Team 2014). The first step aimed to identify and explore the relative importance of

variables that predicted missing cases probability. Initially, all possible predictors of missing cases were tested through separate logistic regression models. In these logistic regression models, the missing case status of the patient at post-treatment was the binary dependent variable. Following a series of univariate models, a stepwise variable selection analysis was used to identify a multivariate but parsimonious list of predictors for missing cases. This was done by considering all available predictors in a "saturated" binary logistic model, including treatment adherence, baseline depression score, baseline anxiety score, and demographic variables such as gender, age, employment status, educational attainment, and relationship status. Following the variable selection strategy detailed by Harrell (2015) and others (Diggle & Kenward, 1994; Rochon, 1999), predictors that increased the probability of becoming a missing case were retained in a final model as significant and dominant predictors of missing cases probability. Each possible predictor of missing cases was assessed for statistical significance at an adjusted, more conservative, *p*-value of 0.01 or less. In addition, the ability of each predictor to account for the probability variance of missing cases likelihood was represented with the *Nagelkerke R* Square statistic. This statistic illustrates the predictive contribution of each variable, and the variance it can account for in comparison to a model with no predictors (Nagelkerke, 1991).

Longitudinal statistical models were also employed to test the ability of baseline and treatment variables to moderate the rate of symptom change. Together, these models sought to identify variables that jointly predicted missing cases and increased/decreased rate of symptom change; that is, to identify mechanisms of missing cases. Longitudinal predictors of symptom change were examined through generalized estimated equation models (GEE; Liang & Zenger, 1986; Hubbard, Ahern, Fleischer, Van der Laan, Lippman, Jewell, ... & Satariano, 2010) that included a time covariate, each of the predictors as a main effect, and a time by predictor interaction. The moderation of symptom change following treatment was tested by examining

the time by covariate interaction. All models included a gamma scale, an unstructured pattern of within-subjects' correlation matrix, and log link function to account for positive skewness and the proportional pattern of symptom change from baseline (Karin, Dear, Heller, Gandy & Titov, 2018). In addition, these models were tested with the overall sample and within each of the five sub-samples for cross-validation purposes.

Power analyses

A power analysis was conducted for both the GEE longitudinal models of symptom change, and the binary logistic regression models of missing cases probability at post-treatment (Cook, Julious, Sones, Hampson, Hewitt, Berlin, ... & Wilson, 2018), using a dedicated R software (R Core Team 2014) package '*longpower'* (Donohue & Eland, 2013). The '*longpower'* package observed statistical parameters from pilot GEE models, such as the rate of change over time, the variance of symptom scores at each time point, and within-subject correlation. This information was then used to determine the minimal differences to the rate of longitudinal change (moderation of longitudinal change) that could be refuted as false negatives. The pilot data used to determine the overall rate of change was replication sample 1 (of 5; n=1341) and the differences from the overall rate of symptom change, or missing cases likelihood, were calculated as a relative difference ($exp\beta$) from the overall rate of change. These power analyses determine whether non-significant tests of symptom change variance, or missing cases probability, are genuine non-significant results, or whether certain non-significant results could be masked by the size of the sample. Separate power estimates were created for the GEE models of symptom change and the binary logistic regression models of missing cases probability and all analyses also specified the probability of power at 80%, and Type I error of 0.05. The resulting power estimates are further described in the result section.

Comparison of different missing cases outcome approximation models

Approximated missing cases replacement scores were generated using stratified longitudinal GEE models. Models differed from one another by the inclusion of different covariates and a covariate by time interaction term. In this way, different models approximated different adjusted outcomes. For example, by including gender, and a time-by-gender interaction term, the approximated missing cases outcomes simulated by the GEE model take into consideration the gender of the individual missing cases, and their likely clinical outcomes as a male or a female. These various models are considered to produce model-based replacement values that account and adjust for different missing cases mechanisms. The adjustment of each model by different covariate would imply that different models make different assumptions about missing cases and adjust for their outcomes through the consideration of different mechanisms. In statistical terms, the conditional adjustment of missing cases outcomes by different influence is often referred to as the replacement of missing cases under a conditional missing at random assumption (MAR)(Mealli & Rubin, 2015).

The accuracy of various adjusted models were interpreted as either overestimating, underestimating, or being equivalent to models that overlook the features of missing cases. Specifically, if the mean confidence interval from an adjusted model was within the mean confidence interval of an unadjusted model, evidence of statistical equivalence was concluded (Greene, Morland, Durkalski, & Frueh, 2008). If the confidence interval of the mean replacement scores was outside the mean of the scores from unadjusted models cases, the models were considered to approximate distinct (statistically significant) symptom outcomes.

Missing cases were also replaced through the last observation carried forward (LOCF), and baseline observation carried forward (BOCF) missing cases replacement methods. These methods represent a commonly used approach for handling missing data without the use of statistical models (Bell et al., 2014). The contrast of these methods to the completer’s analysis

enables a contrast of model-based and non-model based missing cases replacement methodology.

Results

Predictors of missing cases and the rate of clinical change

Results from the logistic regression models, testing for predictors of missing cases at post-treatment, are presented in Table 3.1.

Table 3.1 Univariate model of the total sample *(n*=6701*)*

Probability of missing values at post-treatment [95%CI]	*p-value*	*Time* Predictor odds ratio*	Variance explained (R^2)	*RRI* % missing [95%CI]**
Sample average	*<0.001*	***0.566***	**--**	***36% [35%, 37%]***
Demographic				
Age (% per year)	***<0.001***	**0.967**	**3.8%**	**-1% [-1.1%, -1.1%]**
Gender: Female	**0.003**	**1.188**	**0.2%**	**37% [36%, 39%]**
Male				**33% [31%, 35%]**
Employment status: At least some employment	0.618	0.972	0.0%	36% [35%, 37%]
Otherwise				37% [34%, 39%]
Relationship status: In a relationship	**0.014**	**0.876**	**0.1%**	**35% [34%, 37%]**
Otherwise				**38% [36%, 40%]**
Education level: Tertiary education	***<0.001***	**0.736**	**0.7%**	**32% [31%, 34%]**
Otherwise				**40% [38%, 41%]**
Initial Severity				
Baseline anxiety symptoms (% per GAD-7 point)	*<0.001*	1.024	0.5%	0.7% [0.7%, 0.72%]
Baseline depression symptoms (% per PHQ-9 point)	***<0.001***	**1.037**	**1.4%**	**1.4% [1.4%, 1.44%]**
Baseline psychological distress (% per K-10 point)	***<0.001***	**1.033**	**1.9%**	**1.1% [1.1%, 1.08%]**
Comorbidity at baseline: (PHQ-9≥10 & GAD-7≥8)	*<0.001*	0.718	0.9%	40% [38%, 42%]
None				32% [31%, 34%]
Treatment adherence				
Completed all modules	***<0.001***		**60.3%**	**10% [9%, 11%]**
Completed (4 of 5)		**9.104**		**49% [45%, 53%]**
Completed (3 of 5)		**33.715**		**78% [75%, 81%]**
Completed (2 of 5)		**106.010**		**92% [90%, 94%]**
Completed (1 of 5)		**162.104**		**95% [92%, 96%]**

*RRI – relative risk increment; PHQ-9 - Patient health questionnaire, nine-item scale; GAD-7 – Generalised anxiety disorder scale, seven-item scale; K-10 - The Kessler 10-Item Scale

The binary models indicated that increased psychological distress ($Wald\ \chi^2 = 70.090$, $p < 0.001$), increased baseline depressive symptoms ($Wald\ \chi^2 = 152.4$, $p < 0.001$), decreased treatment adherence ($Wald\ \chi^2 = 2247.443$, $p < 0.001$) and decreased age ($Wald\ \chi^2 = 183.139$, $p < 0.001$) were significant predictors of increased missing cases probability. Together these

variables predicted 60.8% of the variance observed out of the total probability variance for becoming missing at post-treatment (Nagelkerke R square = 60.8%). Additional significant predictors of missing cases were also identified, including relationship status, educational attainment, and comorbidity. However, these variables accounted for a substantially lower (R^2<0.005) explained variance compared to models with treatment adherence.

The effect of increased baseline severity demonstrated that for every one additional PHQ-9 point at baseline, the probability of a participant to become a missing case at post-treatment increased by 2%, or 0.7% as a relative risk (e.g. 0.7% of 36%). Similarly, the effect of a one-point increase in psychological distress symptoms (K-10) at baseline increased the odds of an individual to become missing by 1.6%, or 0.56% as a measure of relative risk.

The participant's age seemed to predict the reduced probability of missing cases, with each additional year of age reducing the probability odds of becoming a missing case reduced by 3.3%, or 1.2% of the total probability (relative risk). However, from the range of screened predictors, treatment adherence as a single variable accounted for the absolute majority of the probability variance. Specifically, 60.3% of the total 60.8% probability variance of missing cases was explained by the number of lessons completed during treatment. Treatment adherence was identified as the dominant predictor of missing cases, where participants who completed the entire program had only a 10% probability of becoming missing at post-treatment. In contrast, participants who completed only one lesson were over 95% likely to present as missing post-treatment cases.

An interaction between depressive baseline severity and treatment adherence was also explored and found to be non-significant ($Wald\ \chi^2_{Treatment\ adherence*Time}$ = 2.162, p=.706); as was an age by treatment adherence interaction ($Wald\ \chi^2_{Age*Time}$ = 4.883, p=.300). The non-significant interactions imply that predictors such as PHQ-9 baseline severity, age and

treatment adherence were distinct main effects of missing cases probability. These predictors were also significant across the replication samples, showing replicability and consistency in all five samples. The influence of variables such as age, PHQ-9 and K-10 baseline symptoms and treatment adherence on the likelihood of missingness were replicated with minimal differences in each of the five randomised sub-samples.

Table 3.2 collates the overall estimates of different missing cases predictors and the replication of these results within each of the five randomised sub-samples.

Table 3.2 Univariate Model of the total sample *(n=6701)*

Probability estimate of missing values at post-treatment in replication subsamples [95%CI]

	***RRI Rep 1 (n=1341)**	***RRI Rep 2 (n=1340)**	***RRI Rep 3 (n=1340)**	***RRI Rep 4 (n=1340)**	***RRI Rep 5 (n=1340)**
Sample average	***36% [34%, 39%]***	***36% [34%, 39%]***	***36% [34%, 39%]***	***36% [34%, 39%]***	***36% [34%, 39%]***
Demographic					
Age (% per year)	**-1.1% [-1.5%, -0.7%]**	**-1.1% [-1.5%, -0.7%]**	**-1.1% [-1.5%, -0.7%]**	**-1.1% [-1.5%, -0.7%]**	**-1.4% [-1.8%, -1%]**
Gender: Female	37% [34%, 41%]	36% [34%, 40%]	38% [35%, 41%]	37% [34%, 40%]	37% [34%, 40%]
Male	33% [29%, 38%]	35% [31%, 41%]	31% [27%, 36%]	34% [29%, 39%]	33% [28%, 38%]
Employment status: At least some employment	36% [33%, 39%]	35% [32%, 38%]	37% [34%, 40%]	36% [33%, 39%]	36% [34%, 40%]
Otherwise	37% [32%, 42%]	40% [35%, 45%]	35% [31%, 40%]	36% [32%, 42%]	35% [30%, 40%]
Relationship status: In a relationship	**35% [32%, 38%]**	**34% [31%, 37%]**	**37% [34%, 40%]**	**35% [32%, 38%]**	**35% [32%, 38%]**
Otherwise	**38% [33%, 42%]**	**41% [36%, 45%]**	**35% [30%, 39%]**	**39% [35%, 44%]**	**38% [34%, 43%]**
Education level: Tertiary education	32% [29%, 36%]	31% [27%, 34%]	35% [31%, 39%]	32% [28%, 35%]	33% [30%, 37%]
Otherwise	40% [36%, 44%]	41% [37%, 45%]	37% [34%, 41%]	40% [37%, 44%]	39% [36%, 43%]
Initial Severity					
Baseline anxiety symptoms (% per GAD-7 point)	1.1% [0.3%, 1.9%]	0.7% [-0.1%, 1.5%]	1.4% [0.6%, 2.2%]	0.4% [-0.4%, 1.2%]	0.7% [-0.1%, 1.5%]
Baseline depression symptoms (% per PHQ-9 point)	**1.4% [0.7%, 2.1%]**	**1.4% [0.7%, 2.1%]**	**1.1% [0.4%, 1.8%]**	**1.4% [0.7%, 2.2%]**	**1.4% [0.7%, 2.1%]**
Baseline psychological distress (% per K-10 point)	**1.1% [0.5%, 1.6%]**	**1.1% [0.5%, 1.6%]**	**1.4% [0.9%, 2%]**	**1.1% [0.5%, 1.6%]**	**1.1% [0.5%, 1.6%]**
Comorbidity at baseline: (PHQ-9≥10 & GAD-7≥8)	40% [36%, 44%]	40% [36%, 44%]	41% [37%, 44%]	40% [36%, 44%]	40% [36%, 43%]
No comorbidity	33% [29%, 36%]	33% [29%, 36%]	32% [28%, 35%]	32% [29%, 36%]	33% [30%, 37%]
Treatment adherence					
Completed all modules	**9% [7%, 11%]**	**9% [8%, 12%]**	**11% [9%, 13%]**	**10% [8%, 12%]**	**9% [7%, 11%]**
Completed (4 of 5)	**57% [48%, 66%]**	**47% [39%, 55%]**	**48% [39%, 56%]**	**51% [42%, 59%]**	**45% [36%, 53%]**
Completed (3 of 5)	**77% [70%, 83%]**	**82% [75%, 88%]**	**72% [64%, 79%]**	**79% [72%, 85%]**	**81% [73%, 86%]**
Completed (2 of 5)	**90% [84%, 94%]**	**91% [85%, 95%]**	**91% [86%, 95%]**	**95% [90%, 98%]**	**93% [87%, 96%]**
Completed (1 of 5)	**95% [88%, 98%]**	**96% [90%, 99%]**	**95% [89%, 98%]**	**94% [87%, 97%]**	**93% [86%, 96%]**

*RRI – relative risk increment; Rep – randomised subsample for cross-validation purposes; PHQ-9 - Patient health questionnaire, nine-item scale; GAD-7 – Generalised anxiety disorder scale, seven-item scale; K-10 - The Kessler 10-Item Scale

Power analyses of missing cases probability models

Post hoc power analyses of the missing cases models illustrated the five replication sub-samples were powered to refute false negative effects that were as little as 10% of the overall sample probability of missing cases (36%). For example, sample 1 (*n*=1341), was powered to refute false-negative predictors that moderated the missing cases probability rate by 3.6% or more (10% of 36%). Refuting non-significant tests of predictors that were smaller than 3.6% required a sample larger than the sample available (*n*=1341). The power to refute non-significant results can be illustrated with the test of the gender predictor in Table 3.2, where men's missing cases were estimated as 33% and women at 37%. The difference between men and women was not statistically significant, and the current sample was large enough to refute this difference as a genuine non-significant (true-negative) result with the statistical power of at least 80%.

Predictors of the rate of clinical improvement

Variables that moderated the rate of symptom improvement were also tested, to determine whether similar variables identified to predict missingness also moderated the rate of symptom change over time. The coefficients statistics in Table 3.3 to 3.5, illustrate the symptom change moderation for each of the three symptom outcomes different variables, with depressive symptoms (Table 3.3), anxiety symptoms (Table 3.4) and psychological distress symptoms (Table 3.5).

Table 3.3 Longitudinal estimates of average depressive (PHQ-9) symptom moderation

	Moderation of the rate of PHQ-9 (depressive) symptom change		
Sample average	***p-value***	***Time*Predictor coefficient (exp(ß))***	**Symptom change rate [95%CI]**
	<0.001	***0.521***	**48% [47%, 49%]**
Demographic			
Age (% per year)	0.119	0.998	-0.2% [-0.4%, 0%]
Gender: Female	0.176	0.967	48% [47%, 50%]
Male			47% [44%, 49%]
Employment status: At least some employment	0.022	0.946	49% [47%, 50%]
Otherwise			46% [43%, 48%]

Relationship status: In a relationship	***<0.001***	**0.893**	**50% [48%, 52%]**
Otherwise			**44% [42%, 46%]**
Education level: Tertiary education	0.818	0.995	48% [46%, 50%]
Otherwise			48% [46%, 50%]
Initial Severity			
Baseline anxiety symptoms (% per GAD-7 point)	<0.001	1.003	0.3% [-0.1%, 0.7%]
Baseline depression symptoms (% per PHQ-9 point)	<0.001	0.988	-1.2% [-1.6%, -0.9%]
Baseline psychological distress (% per K-10 point)	<0.001	1.003	0.3% [0%, 0.6%]
Comorbidity at baseline: (PHQ-9≥10 & GAD-7≥8)	0.006	1.051	36% [34%, 37%]
No comorbidity			39% [37%, 41%]
Treatment adherence			
Completed all lesson modules	**<0.001**		**49% [48%, 51%]**
Completed (4 of 5)		**0.874**	**42% [37%, 47%]**
Completed (3 of 5)		**0.779**	**35% [28%, 42%]**
Completed (2 of 5)		**0.75**	**33% [20%, 45%]**
Completed (1 of 5)		**0.711**	**29% [13%, 45%]**

PHQ-9 - Patient health questionnaire, nine-item scale; GAD-7 – Generalised anxiety disorder scale, seven-item scale; K-10 - The Kessler 10-Item Scale; all estimated were derived from GEE models and their marginal means.

Table 3.3 illustrates that post-treatment depressive symptoms were moderated by treatment adherence, all three baseline symptom levels, and relationship status; all presenting with significant predictor by time interactions. Thus, increases in baseline symptom severity, increased treatment adherence, and relationship status significantly increased the rate of depressive symptom improvement in therapy.

Significant predictors of the rate of anxiety symptoms change were similarly identified. Specifically, increased baseline anxiety symptoms, increased treatment adherence, and the relationship status in treatment seemed to increase the rate of symptom change. The results of anxiety moderators are presented in Table 3.4.

Table 3.4 Longitudinal estimates of average anxiety (GAD-7) symptom moderation

	Moderation of the rate of GAD-7 (anxiety) symptom change		
Sample average	***p-value***	***Change coefficient (exp(β))***	**Symptom change rate [95%CI]**
	<0.001	***0.519***	**48% [47%, 49%]**
Demographic			
Age (% per year)	0.624	0.999	-0.1% [-0.3%, 0.2%]
Gender: Female	0.287	0.975	48% [47%, 50%]
Male			47% [45%, 49%]
Employment status: At least some employment	0.046	0.952	49% [47%, 50%]
Otherwise			46% [44%, 49%]
Relationship status: In a relationship	***<0.001***	**0.887**	**50% [49%, 52%]**
Otherwise			**44% [41%, 46%]**

Education level: Tertiary education	0.456	0.984	48% [47%, 50%]
Otherwise			48% [46%, 49%]
Initial Severity			
Baseline anxiety symptoms (% per GAD-7 point)	*<0.001*	0.976	-2.4% [-2.9%, -2%]
Baseline depression symptoms (% per PHQ-9 point)	0.617	1.001	0.1% [-0.3%, 0.5%]
Baseline psychological distress (% per K-10 point)	0.304	1.002	0.2% [-0.1%, 0.5%]
Comorbidity at baseline: (PHQ-9≥10 & GAD-7≥8)	0.086	0.963	49% [47%, 50%]
No comorbidity			47% [45%, 49%]
Treatment adherence			
Completed all modules	***<0.001***		**49% [48%, 51%]**
Completed (4 of 5)		**0.82**	**43% [38%, 48%]**
Completed (3 of 5)		**0.699**	**35% [28%, 42%]**
Completed (2 of 5)		**0.694**	**38% [27%, 49%]**
Completed (1 of 5)		**0.686**	**40% [27%, 53%]**

PHQ-9 - Patient health questionnaire, nine-item scale; GAD-7 – Generalised anxiety disorder scale, seven-item scale; K-10 - The Kessler 10-Item Scale; all estimated were derived from GEE models and their marginal means.

Similar analyses exploring moderators of general psychological distress (K-10) yielded the same pattern, with results presented in Table 3.4, showing treatment adherence, baseline severity, and relationship status to significantly moderate change in psychological distress.

Table 3.5 Longitudinal estimates of average psychological distress (K-10) symptom moderation

	Moderation of the rate of K-10 (psychological distress) symptom change		
Sample average	***p-value***	***Time*Predictor Change coefficient (exp(β))***	**Symptom change rate [95%CI]**
	<0.001	***0.63***	**37% [36%, 38%]**
Demographic			
Age (% per year)	0.638	1	0% [-0.2%, 0.1%]
Gender: Female	0.287	0.975	48% [47%, 50%]
Male			47% [45%, 49%]
Employment status: At least some employment	0.005	0.946	38% [36%, 40%]
Otherwise			34% [32%, 37%]
Relationship status: In a relationship	***<0.001***	**0.892**	**39% [38%, 41%]**
Otherwise			**32% [30%, 35%]**
Education level: Tertiary education	0.789	1.005	37% [35%, 39%]
Otherwise			37% [35%, 39%]
Initial Severity			
Baseline anxiety symptoms (% per GAD-7 point)	0.009	1.005	0.5% [0.1%, 0.8%]
Baseline depression symptoms (% per PHQ-9 point)	0.003	1.005	0.5% [0.2%, 0.8%]
Baseline psychological distress (% per K-10 point)	*<0.001*	0.994	-0.6% [-0.9%, -0.4%]
Comorbidity at baseline: (PHQ-9≥10 & GAD-7≥8)	0.079	0.962	49% [47%, 50%]

No comorbidity			47% [45%, 49%]
Treatment adherence			
Completed all modules	***<0.001***		**38% [37%, 39%]**
Completed (4 of 5)		**0.881**	**34% [29%, 39%]**
Completed (3 of 5)		**0.77**	**27% [19%, 34%]**
Completed (2 of 5)		**0.763**	**30% [19%, 41%]**
Completed (1 of 5)		**0.644**	**18% [2%, 34%]**

PHQ-9 - Patient health questionnaire, nine-item scale; GAD-7 – Generalised anxiety disorder scale, seven-item scale; K-10 - The Kessler 10-Item Scale; all estimated were derived from GEE models and their marginal means.

Power analyses of symptom change rate models

Post hoc power analyses of the GEE symptom change models demonstrated that each of the five replication sub-samples was adequately powered to determine non-significant predictors, that moderated the rate of symptom change by as little 12% of the total depression symptom change effect (5.7% of 48%). Within the anxiety symptom change models, the sample was powered to refute non-significant predictors that moderated 12% of the total reduction of anxiety symptom reduction (5.7% of 48%), and 13% of the total psychological distress symptom reduction (4.4% of 37%). Refuting predictor effects that were smaller than 5.7% (PHQ-9/GAD-7) and 4.4% (K-10) required a sample that was larger than the 842 participants available in each of the sub-samples.

Identified mechanisms of non-ignorable missing cases

The predictors of treatment adherence, baseline symptoms, and to a lesser extent relationship status, demonstrated an association with both the likelihood of missing data at post-treatment and the rate of symptom change over time. These results indicate that treatment adherence, and to lesser extent baseline symptoms, formed non-ignorable mechanisms of missing cases.

The association of treatment adherence and baseline symptoms with both clinical improvement and risk of presenting as missing cases are illustrated in Figure 1 (missing cases probability at post-treatment and symptom change, associated with program adherence), and Figure 2 (missing cases and symptom change trends associated with depressive symptom baseline severity and depressive symptom outcomes). These figures illustrate how the probability of missing cases is likely to increase for those individuals who also experience higher depressive symptoms at the end of the treatment period (8 weeks); as a result of low treatment adherence (Figure 1) and increased baseline symptoms (Figure 2).

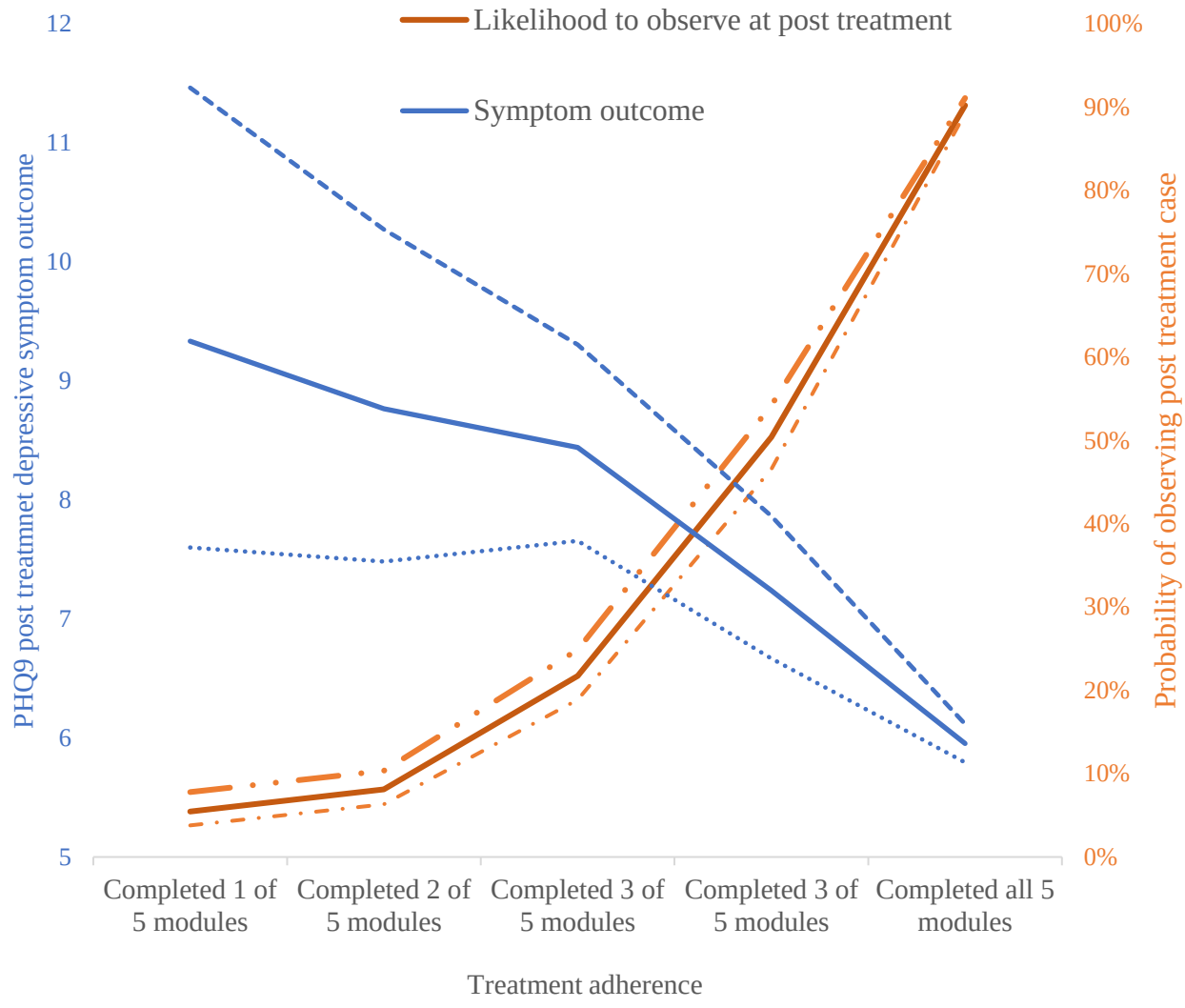


Figure 1 - Missing cases and treatment outcome trends associated with treatment adherence; 95% confidence interval drawn around each effect in dotted lines.

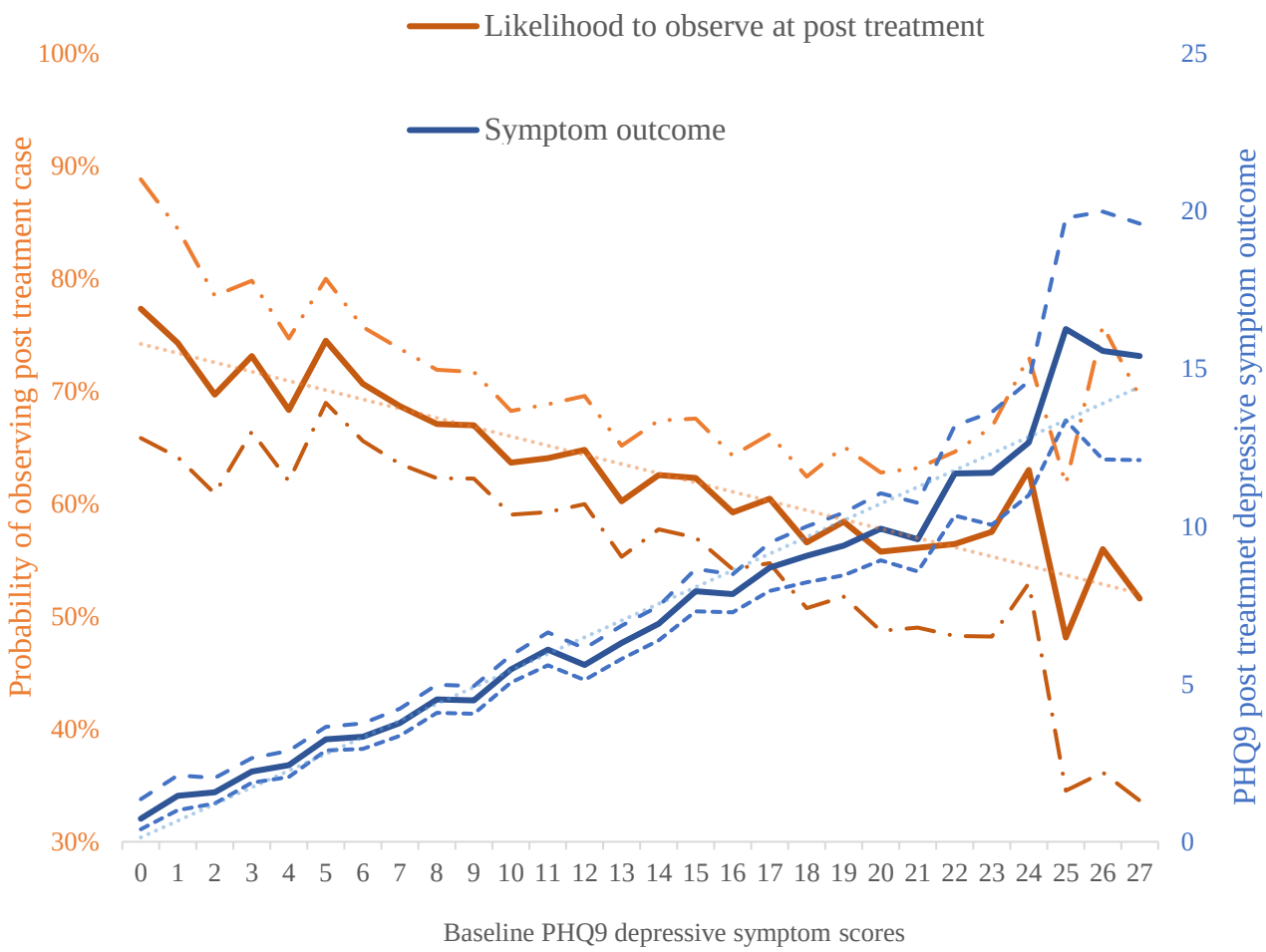


Figure 2 - Missing cases and treatment outcomes trends associated with depressive symptoms baseline severity; 95% confidence interval drawn around each effect in dotted lines.

Comparison of replacement outcomes from different statistical models

In this step, the statistical approximation of replacement symptom outcomes were compared across three different statistical models: (1) models that adjust for the predictors that form missing cases mechanisms (e.g., treatment adherence), (2) models that only adjust for time (completer's analysis), and (3) models that adjust for predictors that are not considered to form missing cases mechanism (e.g. Gender, Age, education). These models differ from one another by the inclusion of different covariates that adjust the projected outcomes of missing cases.

Tables 4.1 to 4.3 presents the approximated mean PHQ-9, GAD-7 and K-10 scores, and confidence intervals for replacement scores, for the various models.

Table 4. 1 predicted PHQ-9 outcomes generated with different replacement models – compared to average post-treatment model estimate (MCAR)

	Mean predicted post-treatment score [95% CI]	Relative to completers' case analysis [95% CI]	The conclusion drawn about replacement approach
Scores from missing cases prior to treatment	13.09 [12.8, 13.34]	--	--
(MCAR) Completer case analysis	6.3 [6.2, 6.5]	--	--
Models adjusted for predictors that do not form missing cases mechanisms			
(MAR) Age	6.3 [6.3, 6.3]	1% [1% 1%]	statistical equivalence to MCAR
(MAR) Gender	6.3 [6.3, 6.3]	0% [0% 0%]	statistical equivalence to MCAR
(MAR) Employment status	6.3 [6.3, 6.3]	0% [0% 1%]	statistical equivalence to MCAR
(MAR) Relationship status	6.3 [6.3, 6.4]	1% [0% 1%]	statistical equivalence to MCAR
(MAR) Education level	6.3 [6.3, 6.4]	1% [1% 1%]	statistical equivalence to MCAR
Models adjusted for predictors that form non-ignorable missing cases mechanisms (missingness & PHQ-9 outcomes)			
(MAR) Baseline anxiety symptoms	6.5 [6.4, 6.6]	3% [2% 4%]	Significant increase above MCAR
(MAR) Baseline depressive symptoms	6.9 [6.8, 7.1]	10% [8% 12%]	Significant increase above MCAR
(MAR) Baseline psychological distress	6.9 [6.8, 7]	10% [8% 12%]	Significant increase above MCAR
(MAR) Comorbidity (PHQ-9≥10 & GAD-7≥8)	6.6 [6.5, 6.6]	4% [3% 6%]	Significant increase above MCAR
(MAR) Treatment adherence	8.1 [8.1, 8.2]	29% [29% 30%]	Significant increase above MCAR
(MAR) Treatment adherence & baseline symptoms	8.8 [8.6, 8.9]	39% [36% 42%]	Significant increase above MCAR
LOCF	10.4 [10.2, 10.7]	65% [62% 69%]	Significant increase above MCAR
BOCF	13.1 [12.8, 13.3]	108% [104% 112%]	Significant increase above MCAR

**Relative to predicted MCAR scores at follow-up and not post-treatment; MAR – Missing at random; MCAR- Missing completely at random (GEE model with time Coefficient only); LOCF – Last observation carried forward; BOCF – Baseline observation carried forward.

Table 4. 2 predicted K-10 outcomes generated with different replacement models – compared to average post-treatment model estimate (MCAR)

	Mean predicted post-treatment score [95%CI]	Relative to completers' treatment effect MCAR [95%CI]	The conclusion drawn about replacement approach
Scores from missing cases prior to treatment	19.44 [19.1, 19.8]	--	--
(MCAR) Completer's analysis	11.4 [11.1, 11.6]	--	--
Models adjusted for predictors that do not form missing cases mechanisms			
(MAR) Age	11.4 [11.4, 11.4]	1% [1% 1%]	statistical equivalence to MCAR
(MAR) Gender	11.3 [11.3, 11.4]	0% [0% 0%]	statistical equivalence to MCAR
(MAR) Employment status	11.3 [11.3, 11.4]	0% [0% 0%]	statistical equivalence to MCAR
(MAR) Relationship status	11.4 [11.4, 11.4]	1% [0% 1%]	statistical equivalence to MCAR
(MAR) Education level	11.4 [11.4, 11.4]	0% [0% 1%]	statistical equivalence to MCAR
Models adjusted for predictors that form non-ignorable missing cases mechanisms (missingness & K-10 outcomes)			
(MAR) Baseline anxiety symptoms	12.4 [12.2, 12.7]	10% [8% 12%]	Significant increase above MCAR
(MAR) Baseline depressive symptoms	12.2 [12, 12.4]	7% [6% 9%]	Significant increase above MCAR
(MAR) Baseline psychological distress	11.7 [11.5, 11.8]	3% [2% 4%]	Significant increase above MCAR
(MAR) Comorbidity (PHQ-9≥10 & GAD-7≥8)	11.8 [11.6, 11.9]	4% [3% 5%]	Significant increase above MCAR
(MAR) Treatment adherence	13.7 [13.7, 13.8]	21% [21% 22%]	Significant increase above MCAR
(MAR) Treatment adherence & Baseline symptoms	14.6 [14.3, 14.9]	29% [26% 31%]	Significant increase above MCAR
LOCF	17.8 [17.5, 18.2]	56% [54%, 59%]	Significant increase above MCAR
BOCF	19.4 [19.1, 19.8]	71% [68%, 74%]	Significant increase above MCAR

**Relative to predicted MCAR scores at follow-up and not post-treatment; ; MAR – Missing at random; MCAR- Missing completely at random (GEE model with time Coefficient only). LOCF – Last observation carried forward; BOCF – Baseline observation carried forward.

Table 4. 3 predicted GAD-7 outcomes generated with different replacement models – compared to average post-treatment model estimate (MCAR)

	Mean predicted post-treatment score [95%CI]	Relative to completers' only [95%CI]	The conclusion drawn about replacement approach
Scores from missing cases prior to treatment	11.45 [11.2, 11.7]	--	--
(MCAR) Completer's analysis	5.7 [5.6, 5.8]	--	--
Models adjusted for predictors that do not form missing cases mechanisms			
(MAR) Age	5.8 [5.8, 5.8]	2% [1% 2%]	statistical equivalence to MCAR
(MAR) Gender	5.7 [5.7, 5.7]	0% [0% 0%]	statistical equivalence to MCAR
(MAR) Employment status	5.7 [5.7, 5.7]	0% [0% 0%]	statistical equivalence to MCAR
(MAR) Relationship status	5.7 [5.7, 5.7]	0% [0% 1%]	statistical equivalence to MCAR
(MAR) Education level	5.7 [5.7, 5.7]	1% [1% 1%]	statistical equivalence to MCAR
Models adjusted for predictors that form non-ignorable missing cases mechanisms (missingness & GAD-7 outcomes)			
(MAR) Baseline anxiety symptoms	6 [5.9, 6.1]	5% [3% 7%]	Significant increase above MCAR
(MAR) Baseline depressive symptoms	6 [5.9, 6.1]	6% [4% 7%]	Significant increase above MCAR
(MAR) Baseline psychological distress	6.1 [6, 6.2]	7% [6% 9%]	Significant increase above MCAR
(MAR) Comorbidity (PHQ-9≥10 & GAD-7≥8)	5.9 [5.8, 6]	4% [3% 5%]	Significant increase above MCAR
(MAR) Treatment adherence	6.8 [6.8, 6.8]	19% [19% 20%]	Significant increase above MCAR
(MAR) Treatment adherence & baseline symptoms	7.1 [6.9, 7.2]	24% [22% 27%]	Significant increase above MCAR
LOCF	9.1 [8.8, 9.3]	60% [56% 63%]	Significant increase above MCAR
BOCF	11.5 [11.2, 11.7]	102% [98% 105%]	Significant increase above MCAR

**Relative to predicted MCAR scores at follow-up and not post-treatment; ; MAR – Missing at random; MCAR- Missing completely at random (GEE model with time Coefficient only). LOCF – Last observation carried forward; BOCF – Baseline observation carried forward.

Tables 4.1 to 4.3 illustrate that the statistical models that adjust their approximation of missing cases outcomes by the prominent characteristics of missing cases (mechanism of nonignorable missingness) resulted in higher symptom outcomes. For example, the PHQ-9 predicted estimate for missing cases, from the model adjusted for treatment adherence was 29% higher than the outcomes from a completer's analysis (Table 4.1). Similarly, the adjusted model, adjusting for both baseline and treatment adherence resulted in missing cases replacement outcomes that are 39% higher than the average treatment effect. In contrast, the application of models that adjust missing cases replacement scores by covariates that *only* predict missing cases (e.g. age), or the rate of symptom change (e.g. relationship status) did not result in missing cases symptom approximation that were different than average (non-adjusted models).

The influence of non-ignorable mechanisms of missing cases were repeated in Tables 4.2 (GAD-7) and Tables 4.3 (K-10). Specifically, by accounting for the role of low treatment adherence on missing cases, the projected symptom scores for missing cases increased by 20%. When the role of baseline symptom severity was also considered, the predicted missing cases outcomes increased even further to nearly 30% above the average symptom outcome score. In contrast, models that adjust their predicted outcome by variables that do not jointly predict missing cases and symptom change, have resulted in near-identical outcomes to the completers' treatment outcomes.

In addition, last observation carried forward (LOCF) and baseline observation carried forward (BOCF) replacement methodologies were compared to outcomes generated under a completer's analysis. The results in tables 4.1-4.3 also illustrate that under the BOCF and LOCF methodologies, replacement scores for missing cases were higher and significantly more conservative by twofold (102% - 148% higher), when compared with the statistical approximation of outcomes under a completer's analysis.

Discussion

The objective of this study was to better understand the characteristics of missing cases in psychotherapy, and compare methods for estimating the symptom outcomes of missing cases in psychotherapy; exemplified in this study with the investigation of likely missing cases outcomes on depressive symptoms scales (PHQ-9), anxiety symptoms scales (GAD-7), and psychological distress scales (K-10).

The first hypothesis, postulating that treatment adherence and baseline symptoms would predict missing cases probability and moderate symptom outcomes, was supported. Notably, treatment adherence explained the majority of the missing cases probability variance at post-treatment (R^2 <60%). Specifically, those participants who completed all of the intervention were also those cases who were 90% likely to provide symptom data at post-treatment. In contrast, those participants who completed a single lesson module were only 5% likely to provide data at post-treatment. This pattern was replicated consistently within all five cross-validation samples. As well as a predictor of missing cases probability, treatment adherence also moderated the rate of symptom improvement for depression, anxiety and distress. The combined associations between greater treatment adherence and a rapid decrease in missing data likelihood and with an increased rate of symptom improvement, forming a key non-ignorable missing cases mechanism. Specifically, an individual's level of treatment adherence changed the rate of symptom change by more than two-fold for psychological distress (K-10; 18% symptom reduction for low adherence vs. 38% for high adherence; a 2.1:1 ratio), depressive symptoms change by a ratio of 1:1.68 (PHQ-9; 29% vs 49% symptom reduction), and moderated the change rate of anxiety symptoms (GAD-7) by up to forty percent (35% vs 49% symptom reduction; 1.4:1 ratio). Given this, the outcomes for missing cases were

non-comparable to the remaining sample and required a statistical approximation that reflected their unique characteristics.

The identification of treatment adherence, missing cases, and clinical outcomes, as related concepts, is consistent with longstanding methodological thinking that links all three (Diggle & Kenward, 1994; Hollis & Campbell, 1999; Rochon, 1999). However, the importance of integrating these concepts into a process of measuring treatment outcomes has, to date, rarely been conducted in psychotherapy research, nor in missing data studies. Rather, treatment adherence, missing cases attrition, and clinical outcomes have been defined as distinct outcomes (Cavanagh, 2010; Sieverink, et al., 2017), and empirically explored as parallel outcomes in meta-analyses on dropout (Karyotaki et al., 2015), or in predictor papers on dropout (Alfonsson, et al. 2016) that overlook the effect of the these features on the missing cases replacement and treatment evaluation.

The current study indicated that treatment adherence jointly predicted the probability of missing cases as well as the amount of symptom change as a result of treatment; a finding consistent with previous work (Karin et al., 2018). In turn, these features suggest that the outcomes of missing cases are likely to have distinctly worse treatment outcomes, that would be overlooked without the adjustment for the rate of treatment adherence and the severity of a patient's symptoms at baseline. Researchers seeking to produce accurate and representative outcome estimates for missing cases should account for the relationship between treatment adherence, the likelihood of cases to become missing, and the role of adherence in driving clinical improvement.

In line with these findings, the key recommendation of the study concerns the measurement of treatment adherence, and its importance for approximating outcomes in clinical trials and routine care. The measurement of treatment adherence could enable

researchers to both better understand the characteristics of missing cases, and more accurately approximate their missing outcomes. From an overview perspective on the methodology used in psychotherapy, a norm to minimally measure and report the incremental rate or progress through a prescribed treatment protocol, or dosage where relevant, could enable researchers to explore mechanisms of missing cases, improve the understanding of the impact of treatment, and approximate the statistical solutions of their outcomes.

The current research is critical in light of broader findings within the clinical trial literature, suggesting that missing cases patterns are mostly overlooked (Bell et al., 2014). At the same time, missing cases in psychotherapy research are common and often reflect a substantial portion of psychotherapy research samples (Fernandez et al., 2015; Waller & Gilbody, 2009). To our knowledge, limited research is available to identify non-ignorable mechanisms of missingness in psychotherapy, their effect on clinical outcomes, or the suitability of different statistical methods to handle missing cases.

The second aim of the study was to explore the suitability of different statistical solutions for replacing the outcomes of missing cases and identify methodological opportunities for psychotherapy researchers. From the range of patient characteristics, two types of models were identified: (1) models that included the key non-ignorable mechanisms of treatment adherence and (2) models that included alternative less dominant predictors (age, gender, education). For example, the analyses of psychotherapy patient characteristics demonstrated that higher psychological distress symptoms at baseline, higher depressive symptoms at baseline, or relatively younger age, also predicted the increased probability of missing cases at post-treatment. Although these predictors of missing cases probability are consistent with findings identified in previous studies (Alfonsson, et al. 2016; Karyotaki et al., 2015), the results of this study, demonstrated that age, gender, and baseline symptoms are limited in their ability to account for the variance in missing cases (R^2 <5%) or account for the

outcomes of missing cases. Consequently, models that considered age, gender, and baseline symptoms were limited in their ability to approximate the probability of missing cases or their likely distinct symptom outcomes. In contrast, models that factored in treatment adherence far outweighed other competing explanations for missing cases (e.g., age, education), and for this reason, were considered to approximate outcomes that reflect the prominent features of missing cases. In this way, the study results supported the second hypothesis, postulating that models that adjust for treatment adherence and baseline severity would be more representative of the outcomes of missing cases.

The second hypothesis also reflects the key finding of the study, demonstrating that the variable of treatment adherence replicated as the single dominant mechanism of non-ignorable missing cases across clinical trials and routine care contexts. Together, this result supports the proposed recommendation to use treatment adherence as a key mechanism of missing cases, and as an adjustment variable in the process of approximating missing cases outcomes.

Limitations and future directions

The findings of this study must be considered in light of several key limitations. First, and foremost, the demonstration of missing cases, their characteristics, their outcomes, and the suitability of replacing missing cases through adjusted models can only be considered preliminary and at this time, as relevant to one specific treatment model (iCBT; Titov et al., 2016). Given that missing cases estimates vary between treatments (Bell et al., 2014; Christensen et al., 2009), it is possible that the patterns, predictors and outcomes of missing cases also vary between treatment models. Although this sample employed extensive cross-validation efforts, the trajectories of missing cases identified in this sample should be considered preliminary and experimental, pending on replications. Replication of these findings across different treatment paradigms could affirm the generalisability of treatment

adherence as a key non-ignorable mechanism of missing cases, and the importance of treatment adherence for psychotherapy missing cases more broadly.

In addition, this study did not comprehensively explore alternative predictors that may characterise the trajectories of missing cases, nor does it exhaust the theoretical causes that may explain missing cases. Other important participant variables not explored in the current study could also play a role in explaining why certain cases become missing and how their outcomes are affected as a result. For example, the presence of a major depression diagnosis (DiMatteo, Lepper & Croghan, 2000), perception of treatment credibility (Fernández-Álvarez, Díaz-García, González-Robles, Baños, García-Palacios & Botella, 2017), or motivation (Alfonsson et al., 2016) may also lead to different rates of treatment adherence, and better capture the expected trajectory of participants in psychotherapy. For this reason, future studies may consider a more direct or more sophisticated measurement of participant engagement, such as motivation and time spent engaged with treatment.

Although not a limitation of the current study, it is important to note that the ability to use statistical replacement models that are adjusted by treatment adherence and baseline symptoms may not be realistic in datasets involving small samples (Cook, Hislop, Altman, Fayers, Briggs, Ramsay, ... & Ford, 2015; Hilgers, Roes & Stallard, 2016). For example, many psychotherapy trials contain small samples of fifty patients or less and, may be underpowered to detect the associations found in the current study and would have insufficient variance to model missing cases outcomes using adjusted models. For this reason, more parsimonious solutions that account for treatment adherence should be considered. For example, the application of LOCF for cases which do not complete treatment could be coupled with the replacement values from unadjusted models for cases who complete treatment in full. This type of hybrid solution may result in a less statistically demanding procedure, but balance overly conservative LOCF statistics with overly liberal unadjusted model approximation. Such

solutions are beyond the scope of this study, but represent an important direction for future research given that many psychotherapy trials involve small samples (Fernandez et al., 2015).

In summary, this study aimed to explore the characteristics of missing cases, the possible clinical outcomes of missing cases following web-based psychotherapy, and the suitability of different strategies for accounting for the outcomes of missing cases in psychotherapy trials. The findings of the current study suggest that: (1) missing cases are associated with lower treatment adherence, (2) the clinical trajectories of missing cases are not likely to be similar to the average surveyed participant, and (3) overlooking the non-ignorable mechanisms of missing cases is likely to result in erroneous replacement of missing cases outcomes. Importantly, this pattern of findings was replicated in the current study using five, large, randomised subsamples, and three different symptom domains. Together, the findings of this suggest that researchers need to consider how they account for the outcomes of missing cases in psychotherapy trials where non-ignorable missing cases mechanisms are likely to occur likely.

References

Alfonsson, S., Olsson, E., & Hursti, T. (2016). Motivation and treatment credibility predicts dropout, treatment adherence, and clinical outcomes in an internet-based cognitive behavioral relaxation program: a randomized controlled trial. Journal of medical Internet research, 18(3), e52.

Altman, D. G., & Simera, I. (2016). A history of the evolution of guidelines for reporting medical research: the long road to the EQUATOR Network. Journal of the Royal Society of Medicine, 109(2), 67-77.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. American Psychologist, 73(1), 3.

Bell, M. L., & Fairclough, D. L. (2014). Practical and statistical issues in missing data for longitudinal patient-reported outcomes. Statistical methods in medical research, 23(5), 440-459.

Bell, M. L., Fiero, M., Horton, N. J., & Hsu, C. H. (2014). Handling missing data in RCTs; a review of the top medical journals. BMC medical research methodology, 14(1), 118.

Blackwell, M., Honaker, J., & King, G. (2017). A unified approach to measurement error and missing data: overview and applications. Sociological Methods & Research, 46(3), 303-341.

Christensen, H., Griffiths, K. M., & Farrer, L. (2009). Adherence in internet interventions for anxiety and depression: systematic review. Journal of medical Internet research, 11(2), e13.

Cook, J. A., Hislop, J., Altman, D. G., Fayers, P., Briggs, A. H., Ramsay, C. R., ... & Ford, I. (2015). Specifying the target difference in the primary outcome for a randomised controlled trial: guidance for researchers. Trials, 16(1), 12.

DeSouza, C. M., Legedza, A. T., & Sankoh, A. J. (2009). An overview of practical approaches for handling missing data in clinical trials. Journal of biopharmaceutical statistics, 19(6), 1055-1073.

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. Journal of the Royal Statistical Society: Series C (Applied Statistics), 43(1), 49-73.

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.

Donohue, M. C., & Edland, S. D. (2013). longpower: Power and sample size calculators for longitudinal data. *R package version*, 1-0.

Eysenbach, G. (2005). The law of attrition. Journal of medical Internet research, 7(1), e11.

Fernandez, E., Salem, D., Swift, J. K., & Ramtahal, N. (2015). Meta-analysis of dropout from cognitive behavioral therapy: Magnitude, timing, and moderators. Journal of Consulting and Clinical Psychology, 83(6), 1108.

Fernández-Álvarez, Javier, Amanda Díaz-García, Alberto González-Robles, R. Baños, Azucena García-Palacios, and Cristina Botella. "Dropping out of a transdiagnostic online intervention: A qualitative analysis of client's experiences." Internet interventions 10 (2017): 29-38.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. Annual review of psychology, 60, 549-576.

Greene, C. J., Morland, L. A., Durkalski, V. L., & Frueh, B. C. (2008). Noninferiority and equivalence designs: issues and implications for mental health research. Journal of traumatic stress, 21(5), 433-439.

Hilgers, R. D., Roes, K., & Stallard, N. (2016). Directions for new developments on statistical design and analysis of small population group trials. *Orphanet journal of rare diseases*, *11*(1), 78.

Hollis, S., & Campbell, F. (1999). What is meant by intention to treat analysis? Survey of published randomised controlled trials. Bmj, 319(7211), 670-674.

Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Lippman, S. A., Jewell, N., ... & Satariano, W. A. (2010). To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. Epidemiology, 21(4), 467-474.

Johansson, R., Carlbring, P., Heedman, Å., Paxling, B., & Andersson, G. (2013). Depression, anxiety and their comorbidity in the Swedish general population: point prevalence and the effect on health-related quality of life. *PeerJ*, *1*, e98.

Johnston, L., Titov, N., Andrews, G., Dear, B. F., & Spence, J. (2013). Comorbidity and internet-delivered transdiagnostic cognitive behavioural therapy for anxiety disorders. Cognitive Behaviour Therapy, 42(3), 180-192.

Karin, E., Dear, B. F., Heller, G. Z., Crane, M. F., & Titov, N. (2018). "Wish You Were Here": Examining Characteristics, Outcomes, and Statistical Solutions for Missing Cases in Web-Based Psychotherapeutic Trials. JMIR mental health, 5(2), e22.

Karin, E., Dear, B. F., Heller, G. Z., Crane, M. F., & Titov, N. (2018). "Wish You Were Here": Examining Characteristics, Outcomes, and Statistical Solutions for Missing Cases in Web-Based Psychotherapeutic Trials. JMIR mental health, 5(2), e22.

Karin, E., Dear, B. F., Heller, G. Z., Gandy, M., & Titov, N. (2018). Measurement of symptom change following web-based psychotherapy: Statistical characteristics and analytical methods for measuring and interpreting change. JMIR mental health, 5(3), e10200.

Karyotaki, E., Kleiboer, A., Smit, F., Turner, D. T., Pastor, A. M., Andersson, G., ... & Christensen, H. (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.

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S. L., ... & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychological medicine, 32(6), 959-976.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. Journal of general internal medicine, 16(9), 606-613.

Lang, K. M., & Little, T. D. (2018). Principled missing data treatments. *Prevention Science*, *19*(3), 284-294.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73(1), 13-22.

Little, R. J., & Rubin, D. B. (2014). Statistical analysis with missing data. John Wiley & Sons.

Little, R. J., D'agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., ... & Neaton, J. D. (2012). The prevention and treatment of missing data in clinical trials. New England Journal of Medicine, 367(14), 1355-1360.

Mallinckrodt, C. H. (2013). Preventing and treating missing data in longitudinal clinical trials: a practical guide. Cambridge University Press.

Mealli, F., & Rubin, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. Biometrika, 102(4), 995-1000.

Nagelkerke NJ. A note on a general definition of the coefficient of determination. Biometrika 1991;78(3):691-692.

Nielssen, O., Dear, B. F., Staples, L. G., Dear, R., Ryan, K., Purtell, C., & Titov, N. (2015). Procedures for risk management and a review of crisis referrals from the MindSpot Clinic, a national service for the remote assessment and treatment of anxiety and depression. BMC psychiatry, 15(1), 304.

R Core Team (2014). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rochon, J. (1999). Issues in adjusting for covariates arising post randomization in clinical trials. Drug Inform. J. 33:1219–1228.

Rubin, D. B. (1976), "Inference and Missing Data," Biometrika, 63, 581-592.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. Psychological methods, 7(2), 147. http://dx.doi.org/10.1037/1082-989x.7.2.147

Sieverink, F., Kelders, S. M., & van Gemert-Pijnen, J. E. (2017). Clarifying the concept of adherence to eHealth technology: systematic review on when usage becomes adherence. Journal of medical Internet research, 19(12), e402.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. Archives of internal medicine, 166(10), 1092-1097.

Titov, N., Dear, B. F., Staples, L. G., Bennett-Levy, J., Klein, B., Rapee, R. M., ... & Nielssen, O. B. (2017). The first 30 months of the MindSpot Clinic: Evaluation of a national e-mental health service against project objectives. Australian & New Zealand Journal of Psychiatry, 51(12), 1227-1239.

Waller, R., & Gilbody, S. (2009). Barriers to the uptake of computerized cognitive behavioural therapy: a systematic review of the quantitative and qualitative evidence. Psychological medicine, 39(5), 705-712.

Chapter 6

Classification of symptom improvement in psychotherapy: A new proposed approach for classifying minimal treatment-related response (Min-TR) (Study 5)

This chapter concerns a third and incremental step in the process of measuring and interpreting psychotherapy evidence, that is, the conversion of continuous symptom scores into interpretable individual outcome categories. The chapter describes a pilot study, proposing a novel method for classifying the impact of treatment by accounting for the change that also occurs under conditions where treatment is not available (non-specific symptom change). The contribution of the study aims to differentiate and classify a new category of individual symptom change, the change that is minimal but specific to treatment; or considered inversely, classifying those individuals who are not impacted by treatment. The broader aim of the paper is to establish a new methodology for characterizing treatment specific and non-specific symptom change through treatment.

Publication status

This chapter has been submitted for publication to the *Journal of Consulting and Clinical Psychology* and is currently under review (CCP-2019-1704).

Author contribution:

Mr Eyal Karin designed, analysed, and wrote the study.

Professor Gillian Heller advised about the choice of software (OptimalCutpoint) and oversaw the analysis. Dr Monique Crane, Associate Professor Blake F. Dear, and Professor Olav Nielssen provided the dataset, assisted with the refinement of the manuscript and helped frame the methodological content for a clinical audience. Professor Nick Titov oversaw the conception of the project and the drafting of the manuscript.

Abstract

Background: Classifying symptom change into interpretable categories is an important step in assessing the effect of psychotherapy in clinical trials. The currently used methods such as the Reliable Change Index (RCI) have improved our ability to classify and interpret change from baseline, but do not describe the trajectory of symptoms through treatment, or distinguish it from the change that can also occur for other reasons, for example, spontaneous remission.

Objectives: We employed a novel classification approach, the Minimal Treatment-Related Response (Min-TR) analysis, in order to differentiate and classify the symptom change that is likely to occur under treatment conditions, and unlikely to occur while on a waitlist, placebo treatment, and other control conditions.

Method: We examined data from trials of an internet-delivered cognitive behavioural therapy, containing standardized scores on scales of anxiety and depression. The participant's allocation to treatment or waitlist control was used as a benchmark to classify the symptom change associated with treatment and non-treatment conditions, and was combined with a novel application of discriminatory analyses.

Results: A 25% improvement in symptoms optimally differentiated and classified individuals who underwent treatment (sensitivity 80%) from those whose symptoms changed while not receiving treatment (specificity 80%). The Min-TR cut-off optimally classified outcomes as relating to treatment and non-treatment, where alternative methods resulted in unspecific classification.

***Conclusions*:** Min-TR may offer a way to (1) represent the quantity of symptom change that optimally differentiates change due to treatment from change due to other factors, and (2) compare methods for measuring treatment specific change using a flexible, non-parametric test.

Keywords

Treatment outcome classification, symptom change, psychotherapy measurement,

Statistical Methodology, treatment evaluation

Introduction

Psychotherapy trials depend on statistical methods that are valid, in the sense that they accurately report the effect of clinical change due to treatment, and reliable in the sense that they can be consistently reproduced. The ability to effectively measure and classify individuals who change in their symptoms as a result of treatment is fundamental to the ability to evaluate treatment (Kazdin, 2008; Ogles, Lunnen & Boneesteel, 2001; Wise, 2004), compare the efficacy of different treatments (Gyani, Shafran, Layard & Clark, 2011; Lambert & Ogles, 2009; Wyrwich, K. W., Norquist, J. M., Lenderking, W. R., Acaster, S., & Industry Advisory Committee of International Society for Quality of Life Research (ISOQOL. (2013)), identify moderators of clinical change (Lin, Huang, Simon, & Liu, 2016; Panagiotakopoulos, Lyras, Livaditis, Sgarbas, Anastassopoulos, & Lymberopoulos, 2010; Wise, Streiner & Gallop, 2016), and identify treatment-related adverse outcomes (Costello, Swendsen, Rose & Dierker, 2008; Kraemer, Noda & O'Hara, 2004; Rozental, Andersson, Boettcher, Ebert, Cuijpers, Knaevelsrud, Ljótsson, Kaldo, Titov, & Carlbring, 2014). These objectives require methods of measurement and classification that convert observed clinical change into interpretable and clinically meaningful categories.

To classify treatment-related change, a longstanding practice in psychotherapy research has been to convert continuous symptom scores into simplified categories that convey either a positive change due to treatment, or no change, or even clinical deterioration (Frank, Prien, Jarrett, Keller, Kupfer, Lavori, ... , Weissman & 1991; Jacobson, Roberts, Berns & McGlinchey, 1999; Ogles et al., 2001). The dichotomisation of symptom scales can result in the loss of measurement sensitivity (Altman, Lausen, Sauerbrei & Schumacher, 1994; Royston, Altman & Sauerbrei, 2006). However, converting those scores into categorical outcomes has several benefits, including (1) allowing the ready interpretation of change following treatment (Kazdin, 1999; Zweig & Campbell, 1993), (2) identifying patients who experience distinct

patterns of change such as deterioration, non-response, improvement and remission while in treatment (Castellani et al., 2016; Costello et al., 2008; Kraemer, et al., 2004), and (3) enhancing the ability to identify subgroups of interest in circumstances where the average trends are unsuitable (Castellani et al., 2016; Kievit, Frankenhuis, Waldorp & Borsboom, 2013; Lin et al., 2016; Zweig & Campbell, 1993).

Methods for dichotomising symptoms in categories in psychotherapy

The most common method for converting continuous symptom scores into categories in psychotherapy research is the *Reliable Change Index* (RCI) (Jacobson & Truax, 1991). The dominance of the RCI is reflected in consensus statements about measurement (e.g. Rozental et al. 2014), treatment evaluation methodology literature (e.g. Delgadillo, McMillan, Leach, Lucock, Gilbody & Wood, 2014), and treatment evaluation studies (e.g. Clark, 2011; Grant, Hotopf, Breen, Cleare, Grey, Hepgul, ... , Young, 2014; Gyani et al., 2011; Hofmann, Asnaani, Vonk, Sawyer & Fang, 2012; King, 2011). Under the RCI, the symptom scores at which the scales are dichotomised can be determined before treatment has commenced. The RCI threshold can be calculated from the variance in pre-treatment symptom scores and the test retest-reliability of a symptom scale (Jacobson & Truax, 1991). The calculated RCI cut-off then represents a symptom change threshold that individuals must exceed to be classified as having "changed significantly" from the pre-treatment symptom scores of the group as a whole (Jacobson & Truax, 1991). This threshold represents the statistically significant departure from the range of baseline symptoms (95%), where individuals who demonstrate a symptom change that is greater than the RCI cut-off are interpreted as having experienced a statistically significant shift in their symptoms. In this way, individuals can be classified as having demonstrated an "improvement," "deterioration," or "recovery" depending on their scores in symptom scales administered after treatment. For example, in studies of psychotherapy for

depression, researchers , such as Gyani and colleagues (Gyani et al., 2011), have calculated that a 5 point change in scores on the Patient Health Questionnaire-9 Item (PHQ-9; Kroenke, Spitzer, & Williams, 2001) can be used to classify participants as 'improved', or 'recovered' when coupled with minimal post-treatment symptoms.

A second approach for dichotomising symptom change into meaningful and interpretable categories is the use of discriminatory analyses (Gao, Calhoun & Sui, 2018; Kennedy & Ceniti, 2018; Kessler, van Loo, Wardenaar, Bossarte, Brenner, Cai, ... & Nierenberg, 2016), which evaluate the ability of different cut-offs to optimally predict a known clinical outcome with no subjective input from the user (López-Ratón, Rodríguez-Álvarez, Cadarso-Suárez, Gude-Sampedro, 2014; Zweig & Campbell, 1993). Authors such as McMillan and colleagues (2010), Levis and colleagues (Levis, Benedetti & Thombs, 2019) and Kessler and colleagues (2016) have used discriminatory analyses to evaluate changes in symptom scores for their ability to predict recovery from an episode of major depression (MDE). Through these statistical algorithms, the change in symptom scores (e.g., such as 5 or more points of improvement on the PHQ-9) that can most accurately predict a change in MDE diagnosis is then selected as the cut-off threshold for dichotomising symptom change into categories.

A limitation in the use of current dichotomisation methods

A major limitation in the current methods for classifying change is that they are unable to account for the symptom change arising from non-treatment effects, such as waitlists, placebo treatments or control groups involved in active interventions. The RCI can classify individuals who change in their symptoms into categories of statistically significant improvement, non-change or deterioration, on the condition that individuals have a change in symptoms that is larger than the calculated standardised error (Jacobson & Truax, 1991).

However, even statistically significant change from baseline can be due to non-treatment effects, such as regression to the mean (Bland & Altman,1994), the effect of being promised treatment, or other non-treatment effects (Andrews, 2001; Meister, Jansen, Härter, Nestoriuc, & Kriston, 2017), and the RCI is not able to distinguish between the kind of symptom change that can occur without treatment and the change that is specific to treatment (Hsu, 1989; Hiller et al., 2012, Jacobson et al., 1999).

Similarly, discriminatory analytic methods that rely on the presence or absence of clinical diagnosis cannot differentiate between the type of change that is specific to treatment (i.e., the treatment effect) and the change that is unrelated to treatment (e.g., spontaneous remission). In brief, the ability to identify a symptom cut-off that conveys the change on a clinical diagnosis of depression (e.g., 5 points; Hiller et al., 2012; McMillan et al., 2010) is informative, but the degree to which this change is specific to treatment is not known. Without the ability to measure and take into account the symptom change that can also occur without treatment, it is not possible to determine which cases should be classified as responding to the effects of treatment, and which cases should be classified as experiencing change that is unrelated to treatment.

Arguably, psychotherapy research would benefit from the ability to differentiate the rate of symptom change that individuals experience in treatment from that which can also happen in non-treatment conditions. In the same way that randomised control trials (RCT) are designed to control for the symptom change that is non-specific to treatment (Bothwell, Greene, Podolsky, & Jones, 2016), the classification of treatment-related change could enable researchers to evaluate the effect of treatment in new ways, such as the investigation of groups who are unresponsive to the effects of treatment. To achieve this classification, however, researchers would need to explore the characteristics of nonspecific treatment change and seek

a potential threshold that would capture the differences between treatment and control; rather than just capturing a statistically significant change from baseline symptoms.

The Minimal Treatment-Related Response (Min-TR): a new proposed way of classifying treatment response

This study explores a technique we describe as the Minimal Treatment-related Response (Min-TR) which is a proposed method of separating change due to treatment, from the change that could also associated with other causes. The method combines discriminatory analyses and the known dichotomisation from randomising individuals to treatment and control groups in treatment trials. The resulting discriminatory analyses are used to evaluate the predictive accuracy of a range of possible cut-offs against a known binary outcome (Duda et al., 2012; Gallop, Crits-Christoph, Muenz & Tu, 2003; López-Ratón, et al., 2014; Vickers, & Elkin, 2006;. These discriminatory analyses can then be used to evaluate the symptom outcome cut-offs such as pre-post percentage differences in symptom scores, remaining post-treatment scores and pre-post difference scores in order to identify the optimal cut-off for each potential outcome and for different symptom outcomes. Moreover, discriminatory analyses can flexibly test and compare various predictors against the same binary outcome, such as the area under the curve (AUC), and sensitivity and specificity of predicting the same binary outcome (López-Ratón, et al., 2014; Vickers, & Elkin, 2006; Zweig & Campbell, 1993) because they share the same accuracy metrics. Further, the cut-offs identified through discriminatory analyses do not require independent variables to comply with statistical assumptions such as the linearity of change or even a parametric association between an independent variable and the binary outcome (López-Ratón et al., 2014), which allows the comparison of different types of symptom cut-offs under one statistical method.

The Min-TR method makes use of the known allocation of randomised individuals to either the treatment or control conditions in RCTs. The binary treatment allocation can be seen as a variable that either represents the experience of treatment or the experience of conditions that are unrelated to treatment. The combination of these features enables the identification of change that is specific and non-specific to treatment in a way that is similar to traditional RCT analyses. In traditional RCT designs, the randomised condition allocation is used as an independent variable for identifying the average rate of symptom improvement between treatment and control conditions. The Min-TR also draws on the association between these conditions and symptom change. However, the randomised allocation is entered as the dependent variable, and symptom change as the independent variable. As a consequence, the interpretation of the analysis changes with the binary outcome, and instead of the average symptom change associated with the allocation to a treatment condition, the Min-TR seeks to identify the amount of symptom change that can distinguish symptom changes due to treatment from those that were not. The resulting binary analysis can then be used to identify a symptom cut-off that (1) represents the quantity of symptom change that optimally differentiates treatment change from non-treatment change, (2) describes the extent to which cases in the treatment and control condition can be separated with statistical accuracy, and (3) compare different ways of measuring treatment-specific and nonspecific change through a flexible, non-parametric, binary test.

The aims of the present study

The primary aim of this study was to pilot a new method for classifying the minimal symptom change that can be attributed to treatment, rather than any change that can also occur under non-treatment conditions. To do this, a large sample of participants (n = 1096) was employed from previously published wait-list controlled trials of a validated internet-delivered cognitive behavioural therapy (iCBT) for treatment of anxiety and depression. Further aims

were (1) to compare the utility of the Min-TR in measuring the rate of symptom change compared with classification using the RCI and change in clinical diagnosis, and (2) to compare the classification performance and scientific implications of using each method.

Method

The sample

The sample used in this study (n=1096) combined clinical data from three previously published RCTs that were designed to evaluate an Internet-delivered cognitive behaviour therapy intervention (iCBT) for change in symptoms of depression (PHQ-9; Kroenke et al., 2001) and anxiety (GAD-7; Spitzer et al., 2006) (Dear, Staples, Terides, Karin, Zou, Johnston, ... & Titov, 2015; Dear, Zou, Ali, Lorian, Johnston, Sheehan, ... & Titov, 2015; Titov, Dear, Ali, Zou, Lorian, Johnston, ., ... & Fogliati, 2015; Titov, Dear, Johnston, Lorian, Zou, Wootton, ... & Rapee, 2013; Titov, Dear, Staples, Terides, Karin, Sheehan, ... & McEvoy, 2015. The samples included 96 participants allocated to a waitlist control condition for the eight weeks of treatment, during which there was no contact or intervention.

The participants were randomly allocated from consecutive eligible applicants to the same clinic (www.ecentreclinic.org) over a 24 month period between 2012 and 2014, and offered near-identical treatment programs (The Macquarie University Model; Titov, et al., 2015). Precautionary longitudinal mixed model testing for similarities across all three treatment trials demonstrated that the baseline symptoms for the various treatment interventions were homogenous (*PHQ-9;* F_{group}=1.92, p=0.147)(*GAD-7;* F_{group}=3.159, p=.043), as were the rate of change within each of the three groups (*PHQ-9,* $F_{groupByTime}$ = 1.875, p=0.154)(*GAD-7,* $F_{groupByTime}$ = 4.89, p = .087), and as a result, the three samples were collapsed into a single treatment sample to test the Min-TR method. In the combined sample the treatment group had an average reduction in symptoms of depression on the PHQ-9 of 5.94 points, which was

significantly greater than the combined control group, that only improved by half a point (0.50) during the 8 weeks on the waitlist ($F_{\text{groupByTime}} = 82.43, p < .001$). The change in anxiety symptoms as measured by the GAD-7 was similar, with a reduction of 5.43 points, compared with a drop of 0.63 points in the control group ($F_{\text{groupByTime}} = 55.79, p < .001$).

Brinley plots (Blampied, 2017) were used to further illustrate sample characteristics, such as symptom severity, symptom change, and the variance of scores in each time point, for each condition and outcome. The pre-post PHQ-9 symptoms are presented in Figures 1 (Treatment) and 2 (Control), and the GAD-7 pre-post symptoms are presented in Figures 3 (Treatment) and 4 (Control). These Figures illustrate the minimal pre-post symptom change and variance in the waitlist condition (Figures 2 and 4), and a mixture of treatment-related symptom responses in both the outcome of depression symptoms (Figure 1) and anxiety symptoms (Figure 3).

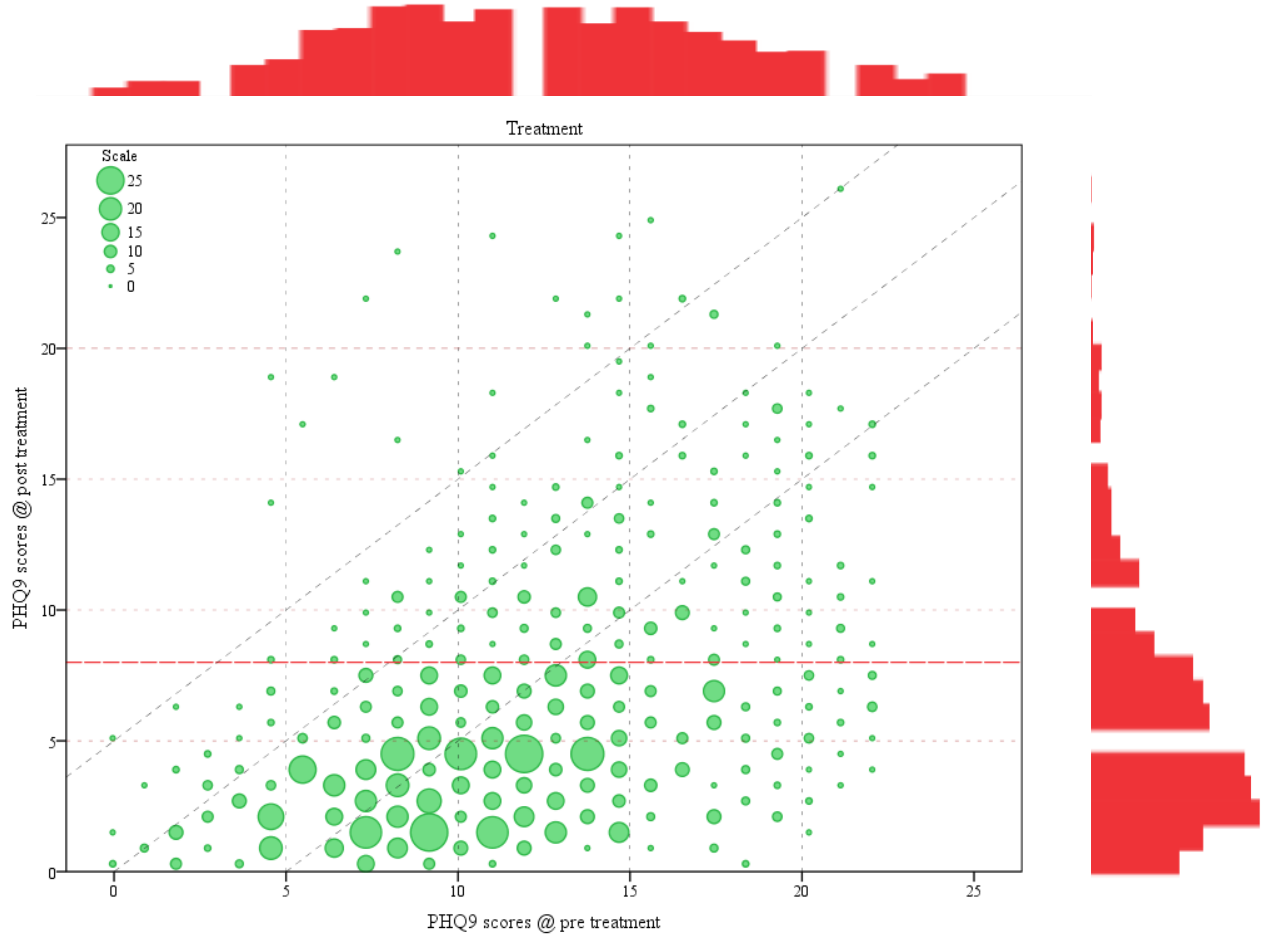


Figure 1: Brinley plot of PHQ-9 symptom of individuals in the treatment condition. The dotted red line denote a "clinical" level of remaining symptoms at post-treatment (Krenoke et al., 2001). Red bars denote the dispersions (percentage proportion) of participants along the pre-treatment (horizontal axis) and post-treatment (vertical) PHQ-9 score axis.

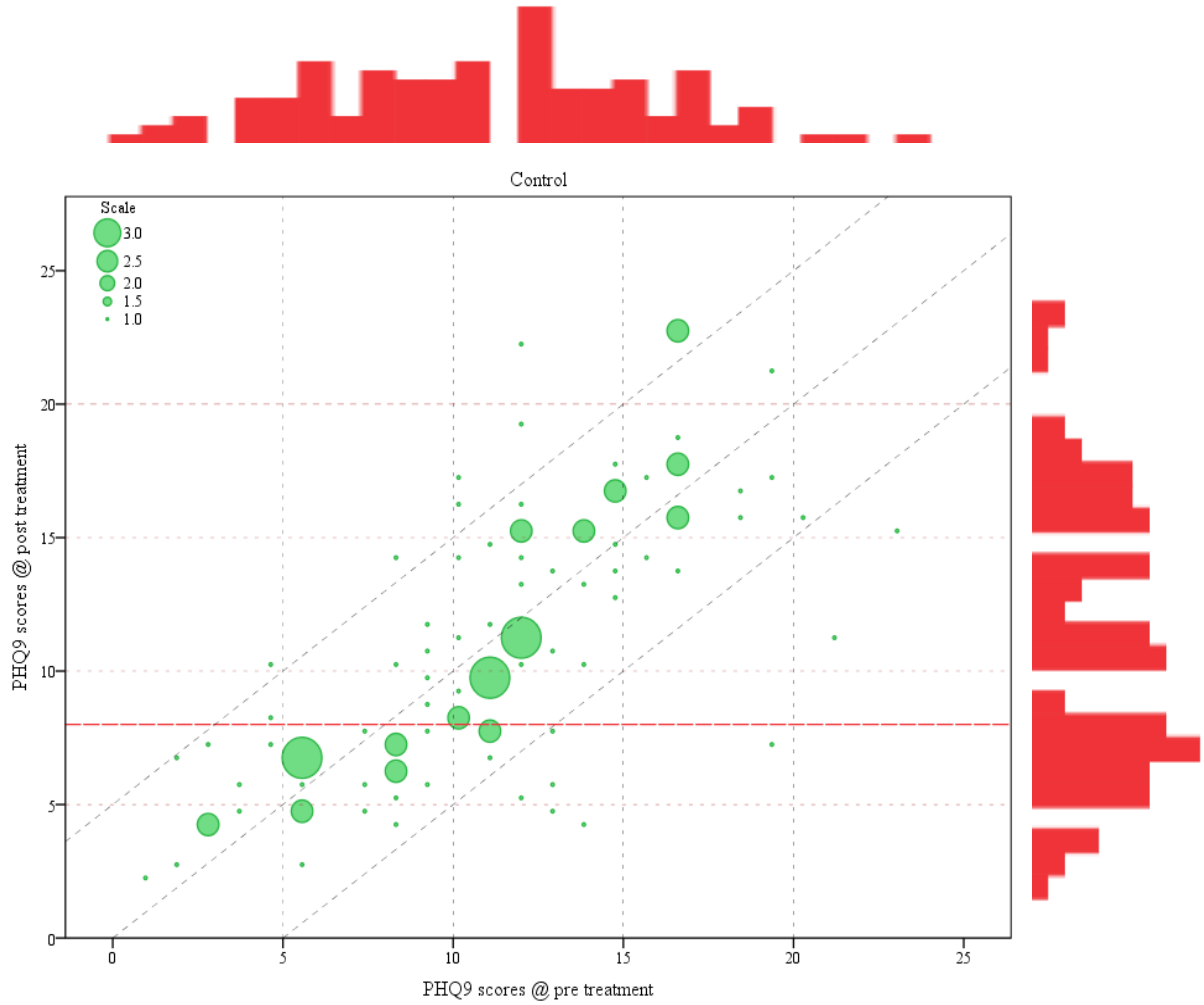


Figure 2: Brinley plot of PHQ-9 symptom of individuals in the control (waitlist) condition. The dotted red line denotes a "clinical" level of remaining symptoms at post-treatment (Krenoke et al., 2001). Red bars denote the dispersions (percentage proportion) of participants along the pre-treatment (horizontal axis) and post-treatment (vertical) PHQ-9 score axis.

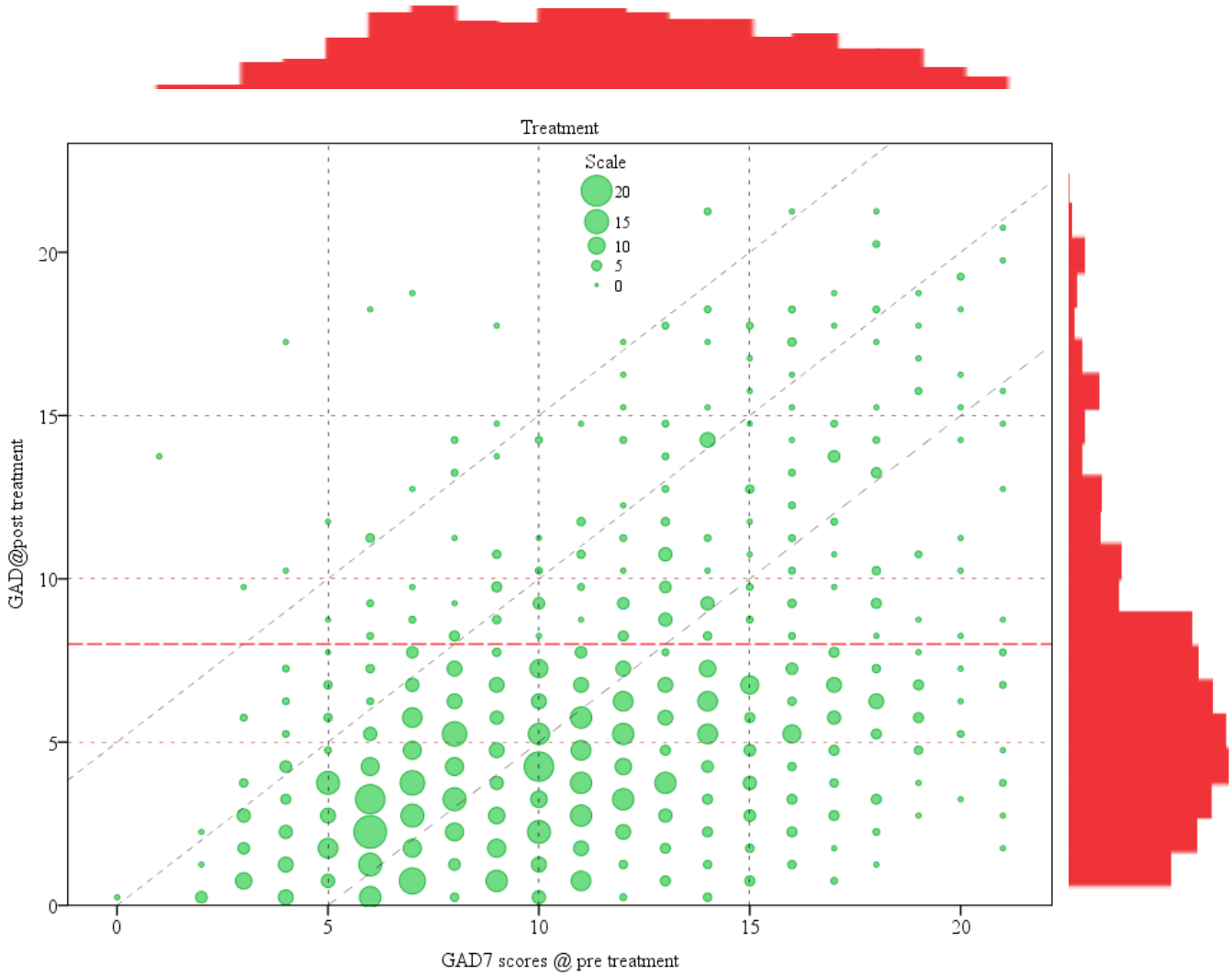


Figure 3: Brinley plot of GAD-7 symptom of individuals in the Treatment condition. The dotted red line denotes a "clinical" level of remaining symptoms at post-treatment (Spitzer et al., 2006). Red bars denote the dispersions (percentage proportion) of participants along the pre-treatment (horizontal axis) and post-treatment (vertical) GAD-7 score axis.

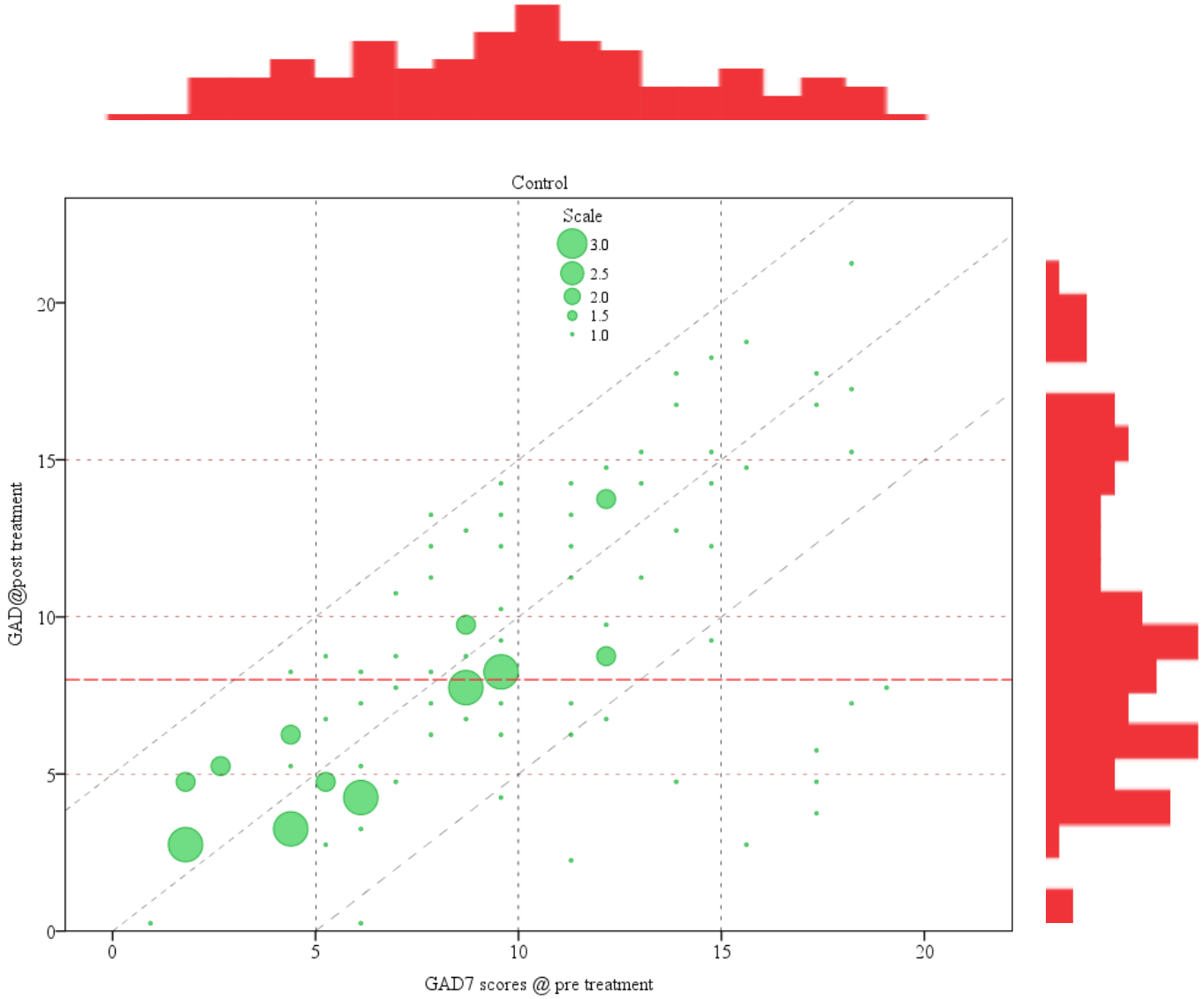


Figure 4: Brinley plot of GAD-7 symptom of individuals in the control (waitlist) condition. The dotted red line denotes a "clinical" level of remaining symptoms at post-treatment (Spitzer et al., 2006). Red bars denote the dispersions (percentage proportion) of participants along the pre-treatment (horizontal axis) and post-treatment (vertical) GAD-7 score axis.

In two of the three treatment samples (Titov et al. 2015; Dear et al., 2015), or 628 of the 1096 participants, clinical interviews were administered by telephone before and after treatment to test the presence of a DSM-IV TR diagnosis (4th ed., text rev.; DSM-IV-TR; American Psychiatric Association, 2000) of either major depressive episode (MDE), or generalised anxiety disorder (GAD). Of the 628 participants interviewed in this way, 372 (59%)

were diagnosed with MDE at pre-treatment, 590 met the diagnostic criteria for GAD (93%), and 345 individuals met the diagnostic criteria for both disorders (56%). These cases were selected as subgroups for discriminatory analysis to determine the symptom change associated with a change in a clinical diagnosis so that the cut-offs associated with a clinical diagnosis change could be compared to the Min-TR cut-off.

More detailed demographic and symptom information about the treatment and control samples are presented in Table 1.

Table 1: Demographic and symptom features of collated treatment and control samples.

Demographic and symptom features	Control sample	Collated treatment sample	Treatment sub-sample with MDE	Treatment sub-sample with GAD
	(*n*=96)	(*n*=1098)	(*n*=372)	(*n*=443)
Gender proportions				
Male	53% (51)	30% (330)	28% (105)	30% (133)
Age (Mean, SD) in years				
	56.3 (13.0)	52.8 (14.2)	43.7 (11.7)	43.3 (11.4)
Sample proportions with Marital status				
married/de facto	47% (45)	65% (713)	58% (216)	65% (288)
single/never married	22% (21)	10% (109)	27% (102)	23% (102)
separated/divorced/widowed	31% (30)	25% (274)	15% (54)	12% (53)
Employment status				
Employed	51% (49)	58% (636)	69% (259)	73% (323)
Education level attained				
High school	41% (39)	16% (176)	15% (56)	14% (62)
Vocational	25% (24)	28% (307)	24% (134)	18% (80)
Degree	35% (37)	56% (615)	54% (202)	68% (301)
Pre-treatment symptom levels				
PHQ-9 (Mean, SD)	10.95 (4.7)	11.73 (4.8)	14.8 (3.8)	12.5 (4.7)
GAD-7 (Mean, SD)	9.45 (4.5)	10.9 (4.5)	12.24 (4.6)	12.2 (4.4)
Post-treatment symptom levels				
PHQ-9 (Mean, SD)	10.9 (4.7)	5.59 (4.6)	6.71 (5.2)	5.86 (4.8)
GAD-7 (Mean, SD)	8.83 (5.4)	5.47 (4.3)	5.58 (4.8)	5.59 (4.6)
Clinical diagnosis at post-treatment				
Major depression diagnosis	--	--	16.2% (47)	12% (54)
Generalised anxiety diagnosis	--	--	26.1% (88)	37% (163)

GAD - generalised anxiety disorder; MDE - major depressive episode; PHQ-9 - Patient Health Questionnaire -9 Item; GAD-7 – Generalized Anxiety Disorder-7-Item Scale.

Measures

The symptom scales used in the study were the PHQ-9 (Kroenke et al., 2001) and GAD-7 (Spitzer et al., 2006), which reflect the DSM-IV TR diagnostic criteria for MDE and generalised anxiety disorder, respectively, and are widely used in clinical trials and evaluations (e.g., Choi et al., 2014; Clark, 2011; Titov et al., 2015). The PHQ-9 and GAD-7 have been shown to have high internal consistency and to be sensitive to the presence and change of clinical and subclinical depression and anxiety diagnoses. The psychometric properties of the scales derived from the combined sample are presented in Table 2.

Table 2*:* Psychometric properties of the PHQ-9 and GAD-7 in the present sample.

Scale	Range of scores	Internal consistency (Cronbach's α)	Intraclass correlation coefficient*
PHQ-9	0-27	.857	.75
GAD-7	0-21	.832	.713

* Estimate is based on a two-way random, single score analysis of items from assessment to pre-treatment
PHQ-9 - Patient Health Questionnaire -9 Item; GAD-7 – Generalized Anxiety Disorder-7-Item Scale

Design and Analytical plan

The first step sought to dichotomise symptom scores associated with RCI, clinical diagnosis change, and the Min-TR. The second step sought to compare the classification performance of the three methods. First, to identify the RCI cut-off, the steps outlined by Jacobson and Truax (1991) were followed. A detailed description of the steps and estimates used in the RCI calculation are presented in the supplementary material (to-be) published on-line (A). To identify the symptom cut-off that represents a change in MDE or GAD clinical diagnoses, a series of discriminatory analyses were conducted through the open-source R statistical software (version 3.2.1; R Core Team, 2016) and a dedicated software package that evaluates continuous markers for diagnostic tests, the OptimalCutpoints package (López-

Ratón, Rodríguez-Álvarez, Cadarso-Suárez, & Gude-Sampedro, 2014). This software performs discriminatory analyses that weigh and identify those cut-points that optimally predict a diagnostic test.

The predictors entered were various symptom score calculations, with the binary outcome of the change in diagnosis identified by clinical interview used as the diagnostic reference point. Three predictor variables were entered, representing the three common ways to use symptom scale scores for evaluating treatment (Hiller et al., 2012; Levis et al., 2019), (1) a percentage change score of symptoms from baseline to post-treatment, (2) the pre-post symptom difference score, and (3) a measure of the remaining symptoms at post-treatment. For the prediction of GAD diagnosis change, GAD-7 scores were used. For the prediction of MDE change, PHQ-9 scores were used.

To establish the Min-TR cut-off, the discriminatory analysis procedure was repeated using the same 3 symptom score calculations. In these analyses, the allocation to either a treatment or control group was entered as the binary outcome.

Within each of the discriminatory analyses employed, the specificity and sensitivity of each possible symptom score cut-offs were also calculated. Sensitivity and specificity were also estimated through Diagnostic Likelihood Ratios, representing the ratio of correct (positive) and incorrect (negative) predictions (Deek & Altman, 2004). A Youden's J statistic was included as a measure that collates sensitivity and specificity into an overall prediction effectiveness statistic (Schisterman, Perkins, Liu, & Bondell, 2005), with scores ranging from 0, indicating completely inaccurate predictions, and minimal specificity and sensitivity, to 1, indicating completely accurate model prediction, with corresponding high specificity and sensitivity.

The area under the curve metric (AUC) was used to evaluate the overall diagnostic accuracy associated with the use of pre-post difference scores, pre-post percentage change

scores, or residual symptom scores after treatment for predicting either the clinical diagnosis change or the treatment allocation (MinTR). In this way, the three ways of using symptom outcomes to predict treatment-specific effects can be compared for their overall diagnostic accuracy, and the optimal way of measuring diagnosis change. The AUC statistic assumed equal weighting between false positives and false negatives, assigned to offset the imbalance in the outcome variable between the 96 waitlisted participants and 1096 participants offered treatment, to simulate an analysis that randomly redraws cases from either the treatment or control conditions (Hand, 2009). The optimal cut-offs were identified for each of the PHQ-9 and GAD-7 cut-offs separately. Exemplar Optimal Cutpoints code for the Min-TR analysis is included in the Supplementary material (B).

In the second step, the cut-offs associated with the RCI, and improvement in GAD and MDE diagnoses and symptom scores were evaluated as clinical outcomes that are specific, or non-specific to treatment. In these analyses, the cut-offs identified under the RCI or clinical diagnosis change were evaluated as predictors of whether participants were in the treatment or control groups. The cut-offs determined by the RCI and clinical diagnoses methods were entered into a Min-TR discriminatory analysis and evaluated for their predictive accuracy of allocation of participants as being in treatment or wait-list control. In this way, the degree to which methods such as the RCI produce outcomes that are specific or non-specific to treatment was evaluated. The differences in the way different symptom dichotomization methods classify treatment specific outcomes was also demonstrated graphically, through receiver operator curves and through histograms that outline specificity and sensitivity for each symptom cut-off. These illustrations aim to demonstrate where the differences between the methods occur, and their specificity as outcomes that occur in treatment and outside of treatment.

Results

Step 1 – Identifying the symptom dichotomisation cut-offs of different methods

In the first set of analyses, the symptom cut-offs associated with the RCI and clinical diagnosis change were determined. The RCI calculation, representing the threshold of change from baseline symptoms, resulted in a threshold of 5-point pre-post improvement score (Supplementary Material A). This 5-point or more cut-off was identified for both the treatment participants and control group after they completed treatment and for both the PHQ-9 and GAD-7 scales. A series of discriminatory analyses predicting the MDE or GAD diagnosis change were then conducted. The resulting cut-offs for classifying MDE or GAD diagnosis change are presented in Table 3. From the range of possible pre-post difference scores, a cut-off of 6 or more points was identified to optimally predict MDE improvement. From the range of pre-post percentage change scores, a cut-off of 44% was identified as an optimal classifier of MDE improvement. From the range of remaining residual scores, eight or less post-treatment PHQ-9 points were identified. From the three possible cut-offs, the optimal classifier of MDE improvement was the percentage improvement of 44%, which had the highest overall sensitivity, specificity, and Youden's J.

Table 3*:* The predictive performance of PHQ-9 and GAD-7 cut-offs scores for the classification of MDE/GAD diagnosis change

Cut-off identified	Sensitivity	Specificity	Positive likelihood ratio (change)	Negative likelihood (no change)	Youden's J	AUC
MDE diagnosis change cut-offs identified						
PHQ-9 pre-post % change ≥ 44%	80% [74%, 85%]	80% [65%, 90%]	4	4	0.6	88% [82%, 93%]
PHQ-9 pre-post change score ≥ 6	75% [61%, 81%]	73% [57%, 86%]	2.78	2.92	0.48	83% [77%, 90%]
PHQ-9 at post treatment ≤ 8	80% [65%, 91%]	76% [70%, 81%]	3.33	3.8	0.56	88% [83%, 93%]
GAD diagnosis change cut-offs identified						
GAD-7 pre-post % change ≥ 58%	62% [55%, 68%]	62% [52%, 71%]	1.63	1.63	0.24	68% [62%, 74%]
GAD-7 pre-post change score ≥ 7	50% [43%, 57%]	57% [47%, 66%]	1.16	1.14	0.07	60% [53%, 67%]
GAD-7 at post treatment ≤ 6	62% [53%, 72%]	71% [65%, 77%]	2.14	1.87	0.33	73% [67%, 79%]

Positive likelihood ratio = sensitivity/(1-specificity) - The ratio of the probability of a positive (or negative) test results in the patients with the disorder to the probability of the same test result in the patients without the disorder; Negative likelihood ratio = specificity /(1-sensitivity). The ratio of the odds of a positive test result in patients with the disorder compared to the odds of the same test result in patients without disease; Youden's J statistic = (sensitivity) + (specificity) – 1. Values range between (low diagnostic accuracy) and 1 (high diagnostic accuracy); GAD - generalised anxiety disorder; MDE - major depressive episode. PHQ-9 - Patient Health Questionnaire -9 Item; GAD-7 – Generalized Anxiety Disorder-7-Item Scale. AUC – area under the curve.

The discriminatory analyses, aiming to identify a symptom cut-off for GAD diagnosis improvement, also resulted in three possible symptom cut-offs, and are presented in Table 3. From the range of pre-post GAD-7 difference scores, a cut-off of 7 or more points was identified to optimally predict GAD diagnosis change. From the range of percentage change scores, a cut-off of 58% was identified as an optimal percentage score cut-off score. From the range of residual scores, a score of 6 post-treatment points or less was selected, which also had the overall optimal predictive performance from the three types of GAD-7 cut-offs. Notably, however, the prediction of MDE change was more accurately predicted with PHQ-9 scores (Optimal Youden's J = .6) than the change of GAD with GAD-7 scores (Optimal Youden's J = .33).

Discriminatory analyses of MinTR and diagnosis change

To identify the symptom cut-offs associated with the Min-TR change, the second series of discriminatory analyses were conducted. The remaining post-treatment (1), pre-post difference scores (2), and pre-post percentage difference scores (3) were entered as predictors in this series of discriminatory analyses. The allocation to treatment or control group were entered as a binary dependent variable. The resulting symptom cut-offs that optimally predicted the known treatment allocation of individuals are presented in Table 4, along with relative accuracy metrics.

Table 4*:* The predictive performance of PHQ-9 and GAD-7 cut-off scores for the classification of Min-TR and MDE/GAD diagnosis change

Cut-off identified	Sensitivity	Specificity	Positive Likelihood ratio (Tx condition)	Negative Likelihood ratio (non-Tx condition)	Youden's J	AUC
Min-TR PHQ-9 cut-offs identified						
PHQ-9 pre-post % change ≥ 25%	80% [77%, 82%]	80% [71%, 88%]	4	4	0.6	86% [83%, 90%]
PHQ-9 pre-post change score ≥ 3	82% [72%,89%]	77% [75%,80%]	3.57	4.28	0.59	83% [80%, 87%]
PHQ-9 at post treatment ≤ 8	67% [57%, 77%]	76% [72%, 77%]	2.79	2.3	0.43	80% [77%, 84%]
Min-TR GAD-7 cut-offs identified						
GAD-7 pre-post % change ≥ 25%	75% [72%,77%]	76% [66%,84%]	3.13	3.04	0.51	80% [75%, 85%]
GAD-7 pre-post change score ≥ 3	73% [70%,75%]	78% [68%,86%]	3.32	2.89	0.51	78% [73%, 83%]
GAD-7 at post treatment ≤ 7	64% [54%,74%]	68% [66%,71%]	2	1.89	0.32	72% [67%, 77%]

Positive likelihood ratio = sensitivity/(1-specificity) - The ratio of the probability of a positive (or negative) test results in the patients with disorder to the probability of the same test result in the patients without the disorder; Negative likelihood ratio = specificity /(1-sensitivity). The ratio of the odds of a positive test result in patients with disorder compared to the odds of the same test result in patients without disease; Youden's J statistic = (sensitivity) + (specificity) – 1. Values range between (low diagnostic accuracy) and 1 (high diagnostic accuracy). PHQ-9 - Patient Health Questionnaire -9 Item; GAD-7 – Generalized Anxiety Disorder-7-Item Scale. AUC – area under the curve.

Min-TR using pre-post 25% improvement. The results in Table 4 show that out of the range of possible cut-offs, the optimal cut-off identified by the Min-TR identified was a pre-post improvement in scores of 25% for both the GAD-7 and PHQ-9. A pre-post PHQ-9 change of 25% or more correctly separated cases into treatments and controls by an accuracy ratio of 4:1. In other words, the 25% cut-off point captured the change observed in 80% of cases in treatment and 80% of the cases observed in the control condition. Similarly, a pre-post change in GAD-7 symptoms of 25% had an accuracy ratio of 3:1 for those in treatment and the controls.

Min-TR using 3 pre-post symptom change score. The optimal cut-offs within the range of symptom change scores (pre-post treatment difference) was a 3-point cut-off on both the PHQ-9 and GAD-7. This cut-off had slightly lower in predictive accuracy when compared to the percentage change scores, but was still highly predictive.

Min-TR using post-treatment scores.

When compared to the estimates of change in symptom scores and percentage change scores, the cut-off generated by using residual post-treatment scores did not differentiate between those who received treatment and the controls. From the range of possible post-treatment PHQ-9 scores, a cut-off of eight points or less was identified as a best-case scenario. This cut-off implies that cases with a score of eight PHQ-9 points or more, following eight weeks, would be classified as belonging to the control condition. This cut-off was associated with a predictive accuracy of 2.79 correct predictions for treatment, and an even lower prediction ratio of 2.0 for the GAD-7 with the optimal cut-off of 7 points or less.

Together, these results suggest that the measurement of percentage change scores optimally differentiates treatment from controls, with reasonable accuracy, and can be used as a standardized way to evaluate predictive accuracy regardless of the symptom scale, or symptom score cut-off for the PHQ-9 and GAD-7.

The ability of the Min-TR to differentiate treatment from controls is illustrated in Figures 5.1, 5.2 and 5.3. The histograms that overlay the dispersions of symptom scores in treatment and control along with sensitivity and specificity predictive accuracy from the Min-TR analysis. Figures 5.1 to 5.3 illustrate the pre-post difference scores cut-off (Figure 5.1), pre-post percentage change outcomes (Figure 5.2), and remaining symptoms at post-treatment (Figure 5.3). In each Figure, the cut-off that best differentiates between the distributions of cases in the treatment (red bars) and waitlist (green bar) condition is marked with a dotted line. The dotted-red drop-line indicates the optimal cut-off on the x-axis.

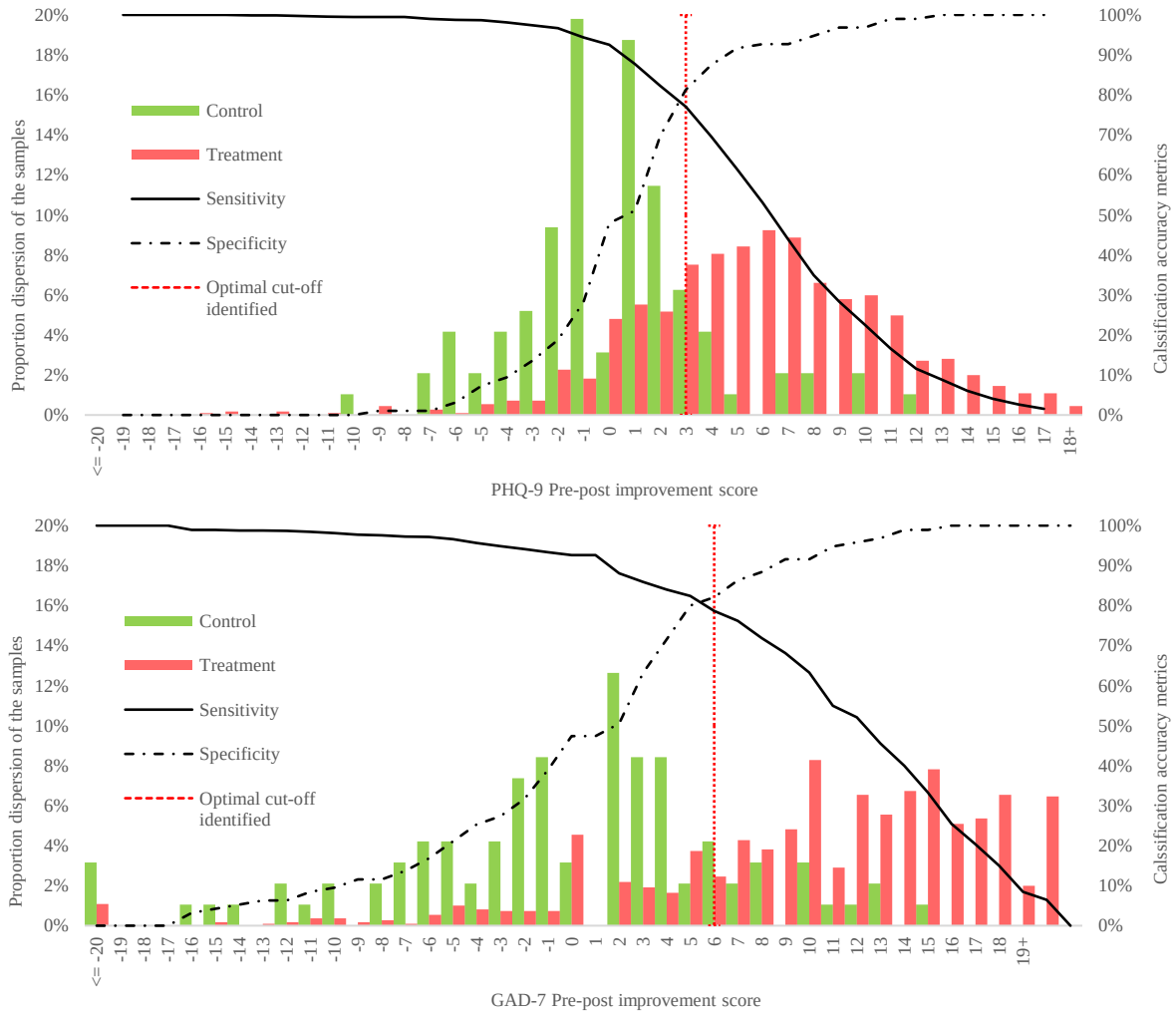


Figure 5.1*:* The frequency of difference change scores in treatment and control - presenting depressive PHQ-9 symptom outcomes (top), and anxiety GAD-7 symptom outcome (bottom)

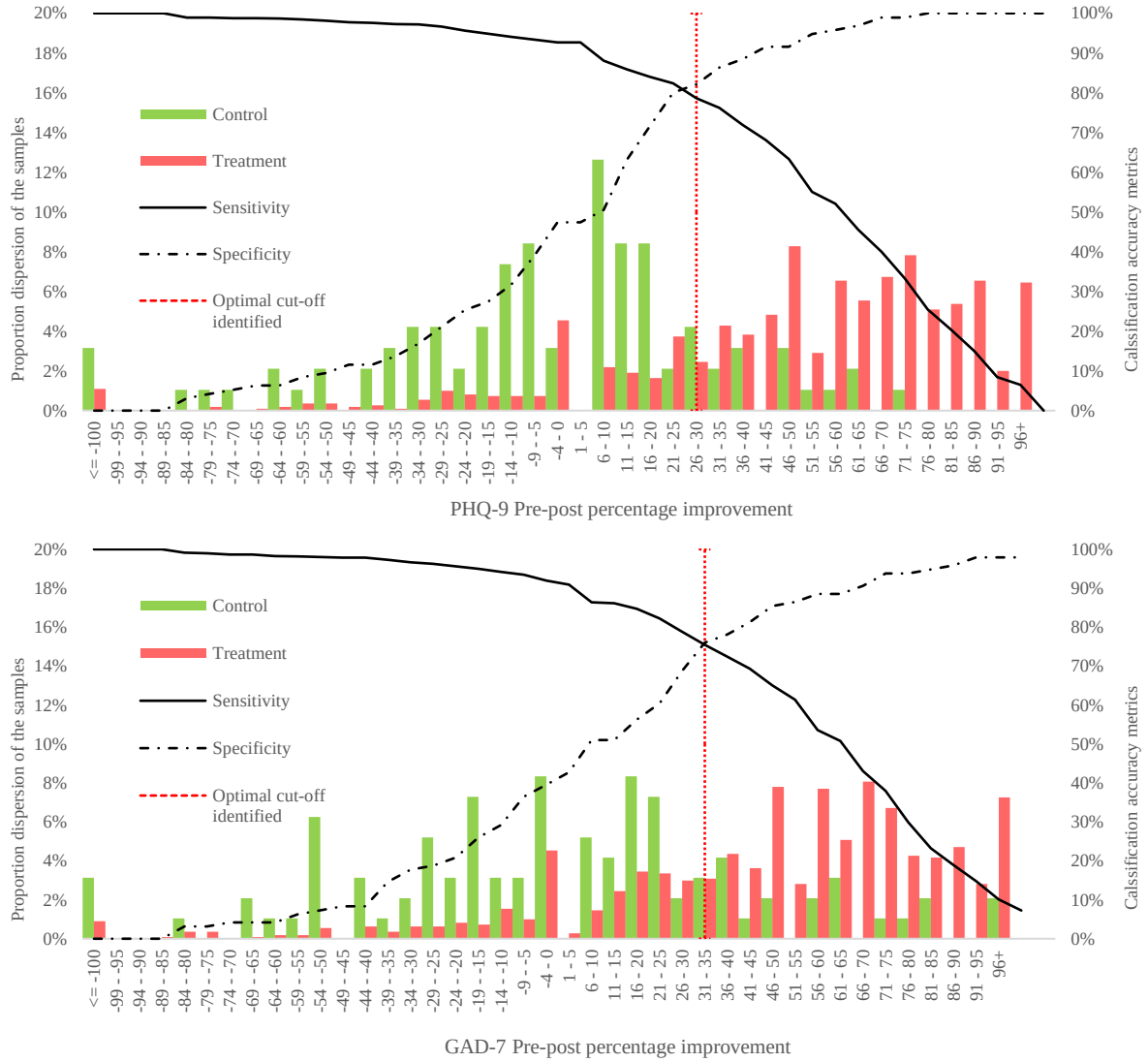


Figure 5.2: The frequency of pre-post percentage difference change scores in treatment and control; presenting depressive PHQ-9 symptom outcomes (top), and anxiety GAD-7 symptom outcome (bottom).

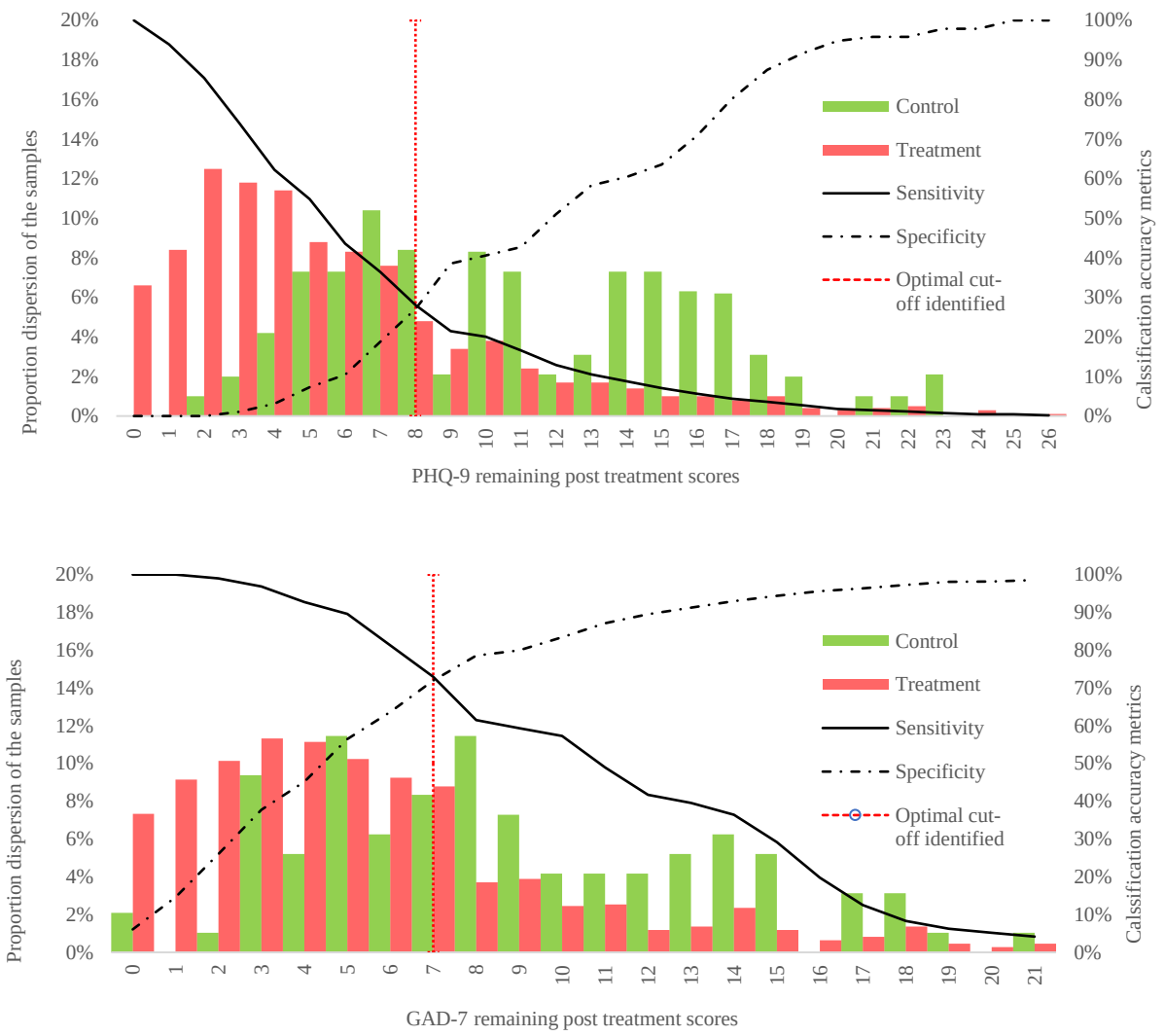


Figure 5.3: The frequency of pre-post percentage difference change scores in treatment and control; presenting depressive PHQ-9 symptom outcomes (top), and anxiety GAD-7 symptom outcome (bottom).

Figures 5.1 to 5.3 illustrate the ability of the Min-TR to identify the point that differentiates the two distributions. In Figures, 5.1 – 5.3 the cut-offs are marked with dotted red vertical lines, at the point where sensitivity and specificity are at equilibrium; which is also the point where differentiation accuracy is at a maximum.

Step 2 - Comparison of RCI cut-offs and MDE/GAD diagnoses change as treatment-specific effects

To further compare the Min-TR method to more traditional symptom dichotomisation methods, eight discriminatory analyses were performed. These analyses used the cut-offs, identified under the methods of RCI and clinical diagnosis change, to predict whether cases had received treatment or control allocation. In this way, traditional methods for dichotomising symptoms outcomes were evaluated for their ability to identify treatment specific clinical outcomes from changes due to other factors.

Each of these analyses included one of the following predictors: (1) the identified PHQ-9 RCI 5-point cut-offs, (2) the identified GAD-7 RCI 5-point cut-offs, (3) the MDE diagnosis change PHQ-9 pre-post change score ≥ 6, (4) MDE diagnosis change PHQ-9 at post-treatment ≥ 8, (5) MDE diagnosis change PHQ-9 pre-post % change ≥ 44%, (6) GAD diagnosis change pre-post change score ≥ 7, (7) GAD diagnosis change at post-treatment ≥ 6, (8) GAD diagnosis change pre-post % change ≥ 58%. PHQ-9. The results of these discriminatory analyses are presented in Table 5.

Table 5*:* Testing the RCI and MDE diagnosis cut-offs as treatment-specific effects. Comparing cut-offs from different methodologies as treatment-specific effects

	Sensitivity (Accurate treatment)	Specificity (Accurate non-treatment)	Positive Likelihood ratio (Tx allocation)	Negative Likelihood (non-Tx allocation)	Youden's J
PHQ-9 cut-offs under different dichotomisation methods					
Optimal Min-TR cut-off identified (% change ≥ 25%)	80% [77%, 82%]	80% [71%, 88%]	4	4	0.6
RCI (PHQ-9 pre-post change score ≥ 5)	60% [57%, 62%]	91% [80%, 100%]	6.67	2.28	0.51
MDE diagnosis change (PHQ-9 pre-post change score ≥ 6)	56% [53%, 59%]	91% [80%,100%]	6.22	2.07	0.47
MDE diagnosis change (PHQ-9 at post treatment ≤ 8)	68% [57%, 77%]	75% [73%, 78%]	2.72	2.34	0.43
MDE diagnosis change (PHQ-9 pre-post % change ≥ 44%)	65% [62%, 67%]	92% [84%, 96%]	8.13	2.63	0.57
GAD-7 cut-offs under different dichotomisation methods					
Optimal Min-TR cut-off identified (% change ≥ 25%)	75% [72%,77%]	76% [66%,84%]	3.13	3.04	0.51
RCI (GAD-7 pre-post change score ≥5 GAD-7)	54% [51%, 57%]	86% [78%, 93%]	6.67	2.28	0.51
GAD diagnosis change (GAD-7 pre-post change score ≥ 7)	37% [34%, 40%]	92% [84%, 96%]	4.63	1.46	0.29
GAD diagnosis change (GAD-7 at post treatment ≤ 6)	65% [54%, 74%]	69% [66%, 71%]	2.10	1.97	0.34
GAD diagnosis change (GAD-7 pre-post % change ≥ 58%)	48% [45%, 51%]	90% [82%, 95%]	4.80	1.73	0.38

Positive likelihood ratio = sensitivity/(1-specificity) - The ratio of the probability of a positive (or negative) test results in the patients with the disorder to the probability of the same test result in the patients without the disorder; Negative likelihood ratio = specificity /(1-sensitivity). The ratio of the odds of a positive test result in patients with disorder compared to the odds of the same test result in patients without disease; Youden's J statistic = (sensitivity) + (specificity) – 1. Values range between (low diagnostic accuracy) and 1 (high diagnostic accuracy). PHQ-9 - Patient Health Questionnaire -9 Item; GAD-7 – Generalized Anxiety Disorder-7-Item Scale. AUC – area under the curve.

Table 5 shows that cut-offs identified from the discriminatory analyses of MDE and GAD diagnosis change, nor the RCI, optimally differentiate between cases in the treatment and control conditions. From the Table, the RCI and MDE and GAD diagnosis change cut-offs are associated with high specificity (correctly classifying cases in control), but reduced sensitivity scores (inability to accurately differentiate cases in treatment), that ranged from 50% (random chance, indicating poor predictive accuracy) to 70% (moderate predictive accuracy). In other words, the cut-offs associated with methods such as the RCI were effective for classifying change that does not occur in controls (90%), but RCI overlooked cases that changed by a rate that was higher than the control condition, but lower than the 5-point threshold.

These results indicate that the Min-TR, RCI, and clinical diagnosis change analyses, resulted in (1) different dichotomisation thresholds, (2) different proportions of the treatment and control samples, classified as 'improved', and (3) different predictive accuracy to differentiate between cases that changed as a result of treatment or nonspecific conditions.

The differences between the RCI, Min-TR and diagnoses change cut-offs are illustrated graphically with the use of a receiver operator curve plot; Figure 6. This figure compares the sensitivity and specificity of different cut-offs and their ability to differentiate between the treatment and control conditions; approximating treatment specific and nonspecific symptom change.

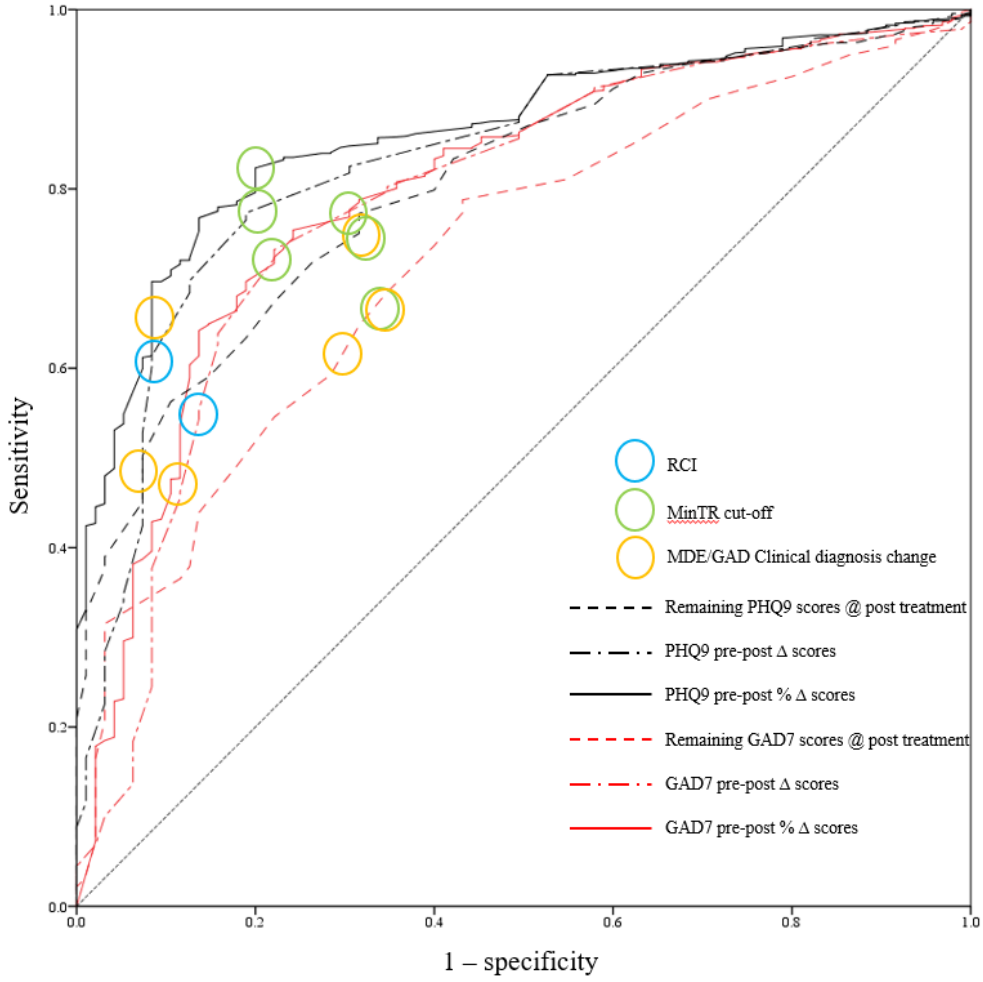


Figure 6*:* Receiver operator curve comparing the ability of different methods to classify outcomes that either specific or nonspecific to treatment accuracy. Lines represent different uses of PHQ-9 and GAD-7 outcome calculation. The circles represent the identified cut-offs under each of the RCI (blue circles), Min-TR (green circles) or MDE/GAD change (Yellow circles).

Figure 6 illustrates that, of the possible ways to measure treatment-related symptom change, the measurement of percentage change scores (unbroken lines) seems to be the most effective at differentiating the effects that occurred in treatment, and not in controls for both anxiety (GAD-7) and depression (PHQ-9) outcomes. Further, specific cut-offs such as the RCI (blue circles) can be seen to produce predictive outcomes that are specific to control (high specificity) but insensitive to the type of change that occurs in treatment (low sensitivity). Similarly to the RCI, the outcomes associated with the dichotomisation of symptoms that represent an MDE or GAD diagnosis also resulted in high specificity but poor sensitivity. In contrast to the RCI, the Min-TR method can be seen to optimally differentiate between

treatment and control, resulting in the most accurate (perpendicular) points on each of the PHQ-9 or GAD-7 symptom score curves.

Discussion

The primary aim of this study was to use a large sample, from RCTS of iCBT for anxiety and depression, to test a novel statistical approach against standard approaches to classifying symptom that occur with treatment, from the change that can also occur without treatment. The Min-TR allowed the comparison of multiple types of symptom outcomes (e.g., percentage change, remaining post-treatment scores) in order to identify the threshold that can best represent outcomes that are specific to treatment. From the range of options, the Min-TR method identified a 25% pre-post improvement cut-off as the optimal threshold for differentiating the minimal change that is specific to treatment, and which is uncommon in the control groups for both the PHQ-9 and GAD-7 symptom outcomes. The methods that apply a pre-determined calculated threshold, such as the RCI, resulted in a high ability to specify the cases of the control group (90%) but also resulted in a near-random prediction of cases in treatment (50% sensitivity). In contrast, the Min-TR cut-off point illustrated that by reducing the specificity rate from 90% to 80% (equivalent to a Type I error of 20%), the ability to predict the cases in treatment (sensitivity) improved in accuracy from 50% to 80% (a reduction in Type II error from 50% to 20%). In other words, by changing the cut-off from a reduction in score of 5 points to a 25% improvement, the Min-TR resulted in a net increase to the ability to differentiate the symptom changes of the treatment and control groups, and accurately predict up to four of five cases in either the treatment or waitlist control groups.

The Min-TR method is different from other approaches in three important ways. First, to classify cases that improve as a result of treatment, the Min-TR directly contrasts the outcomes associated with treatment against the outcomes observed under waitlist conditions,

and offers the possibility of using direct observation to separate non-specific symptom change from the change associated with treatment. There is limited research reporting on the classification of non-specific symptom change, and the characteristics of non-specific symptom change are rarely considered in psychotherapy methodology studies (Hiller et al., 2012; McMillan et al., 2010) or reported in meta-analyses about the effects of treatment (Meister et al., 2017). The Min-TR could offer an accurate alternative method to the RCI and change in clinical diagnosis to establish where symptoms might be changing in a way that is unrelated to treatment.

Second, the Min-TR method demonstrated the ability to consider a range of symptom scores without the need to meet statistical assumptions, such as the linearity of change, kurtosis, skewness or modality of any symptom score distribution. The RCI rely on a linear scale, and strict statistical assumptions (e.g., normal distribution) and results in a single cut-off (Jacobson et al., 1999). In contrast, the Min-TR employs discriminatory analysis that applied a non-parametric cut-off analysis. In this way, the Min-TR searches through linear, nonlinear, and non-parametric ways to differentiate the effects of treatment from the effects of control.

Third, the Min-TR method could provide an opportunity for psychotherapy researchers to have another distinctive way of reporting clinical change, the *minimal treatment-related response*. Using the RCI, participants in treatment were required to demonstrate a change of either 5 points on the PHQ-9 or GAD-7 in order to be classified as improved, which meant that nearly 40% of individuals were considered as non-responders to treatment. Using the Min-TR, the threshold for identifying treatment-related change was lowered to 3 points, with only 25% of participants classified as non-responders, with a further 15% classified as changing by symptom margins that were minimal, but treatment-related. This gap between minimal change, and more moderate change represents an additional category of clinical outcome, termed *minor clinical improvement*. As a result, rather than classifying the 15% of individuals who were

under the RCI threshold as non-responders, they may now be categorised as having had experienced *a minimal treatment-related change*. This minimal treatment-related response category could represent an intermediate category alongside the categories of deterioration, non-response and remission often reported in treatment evaluation studies. The benefit of using a distinct category of minimal treatment-related response may prove useful for research concerning the measurement of partial symptom response (Zimmerman, 2003; Zhou, Li, Pei, Gao, & Kong, 2016), or research seeking to identify and further examine participants who are not showing at least a minimal response to treatment. It also enables researchers not to overlook treatment response that may be important for patients.

Limitations and future research

Several limitations should be considered about both the use of treatment allocation as a binary outcome within discriminatory analyses, and the identification of the resulting Min-TR cut-offs. First and foremost, the ability of the Min-TR method to distinctly classify treatment-related change from nonspecific change is reliant on clear differences between the symptom change individuals experience under control and treatment conditions. This method was tested using a sample in which the majority of cases in treatment and control had different outcomes. For this reason, the Min-TR was accurate and effective at differentiating the symptom outcomes of treatment and control. However, in other research contexts, such as placebo conditions or active control, the symptom outcomes of treatment and control conditions might be less pronounced (Barber et al., 2012; Dimidjian et al., 2006). In these circumstances, the ability of the MinTR to differentiate between cases in treatment and control could be poor, and methods such as the RCI may offer a clearer distinction between treatment responders and non-responders. For this reason, the Min-TR methodology should be considered as a preliminary concept that should be further explored in other clinical settings. Future research seeking to explore minimal but treatment-related response may consider re-

examining samples from other studies using the Min-TR, in particular, RCTs employing active and placebo control conditions. Without replication, the robustness of both the Min-TR methodology and the specific use of the three PHQ-9 point cut-off cannot be confirmed.

Second, it is also important to note that this study did not employ additional variables that could verify the occurrence of an actual improvement from the perspective of patients and whether this was reflected in the identification of minimal treatment-related response. For example, the extent of the minimal improvement group might have been verified by ratings of satisfaction, quality of life, and also functional performance, such as days spent out of normal role. Similarly to other classification methods, such as the minimal clinical importance difference (Jaeschke, Singer & Guyatt, 1989), the use of self-rated experience of change could be used to further verify whether the minimal symptom improvement is representative of a positive experience in treatment (Wu, Liu, Tanadini, Lammertse, Blight ,Kramer ,... & Fawcett, 2015). Future research aiming to verify the concept of minimal treatment-related response could compare the ability of the Min-TR, RCI, and MDE to also identify the experience of treatment and functional improvement from the perspective of patients.

Conclusion

The findings of this study suggest that minimal symptom change that is likely to occur following treatment, over and above any nonspecific change, can be identified using the Min-TR method. The ability to identify minimal treatment-related response has significant potential for psychotherapy research and is promising in that it may improve the validity and reliability of psychotherapy research. Future research is needed to test the validity and reliability of the Min-TR method in different contexts with a range of different outcome measures and control conditions.

References

Altman, D. G., Lausen, B., Sauerbrei, W., & Schumacher, M. (1994). Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *JNCI: Journal of the National Cancer Institute*, *86*(11), 829-835.

American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders (4th ed., text rev.). doi:10.1176/appi.books.9780890423349.

Bamber D (1975). "The Area above the Ordinal Dominance Graph and the Area below the Receiver Operating Graph." Journal of Mathematical Psychology, 12, 387–415.

Barber, J. P., Barrett, M. S., Gallop, R., Rynn, M. A., & Rickels, K. (2012). Short-term dynamic psychotherapy versus pharmacotherapy for major depressive disorder: a randomized, placebo-controlled trial. *The Journal of clinical psychiatry*, *73*(1), 66-73.

Clark, D. M. (2011). Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: the IAPT experience. *International review of psychiatry*, *23*(4), 318-327.

Dear, B. F., Staples, L. G., Terides, M. D., Karin, E., Zou, J., Johnston, L., ... & Titov, N. (2015). Transdiagnostic versus disorder-specific and clinician-guided versus self-guided internet-delivered treatment for generalized anxiety disorder and comorbid disorders: a randomized controlled trial. *Journal of anxiety disorders*, *36*, 63-77.

Dear, B. F., Zou, J. B., Ali, S., Lorian, C. N., Johnston, L., Sheehan, J., ... & Titov, N. (2015). Clinical and cost-effectiveness of therapist-guided internet-delivered cognitive behavior therapy for older adults with symptoms of anxiety: a randomized controlled trial. *Behavior Therapy*, *46*(2), 206-217.

Delgadillo, J., McMillan, D., Leach, C., Lucock, M., Gilbody, S., & Wood, N. (2014). Benchmarking routine psychological services: a discussion of challenges and methods. *Behavioural and cognitive psychotherapy*, *42*(1), 16-30.

Delong ER, Delong DM, Clarke-Pearson DL (1988). "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach." Biometrics, 44, 837–845.

DeRubeis, R. J., Hollon, S. D., Amsterdam, J. D., Shelton, R. C., Young, P. R., Salomon, R. M., ... & Gallop, R. (2005). Cognitive therapy vs medications in the treatment of moderate to severe depression. *Archives of general psychiatry*, *62*(4), 409-416.

Dimidjian, S., Hollon, S. D., Dobson, K. S., Schmaling, K. B., Kohlenberg, R. J., Addis, M. E., ... & Atkins, D. C. (2006). Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the acute treatment of adults with major depression. *Journal of consulting and clinical psychology*, *74*(4), 658.

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Ebert, D. D., Donkin, L., Andersson, G., Andrews, G., Berger, T., Carlbring, P., ... & Kleiboer, A. (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), 2679-2693.

Frank, E., Prien, R. F., Jarrett, R. B., Keller, M. B., Kupfer, D. J., Lavori, P. W., ... & Weissman, M. M. (1991). Conceptualization and rationale for consensus definitions of terms in major depressive disorder: remission, recovery, relapse, and recurrence. *Archives of general psychiatry*, *48*(9), 851-855.

Gallop, R. J., Crits-Christoph, P., Muenz, L. R., & Tu, X. M. (2003). Determination and interpretation of the optimal operating point for ROC curves derived through generalized linear models. *Understanding statistics*, *2*(4), 219-242.

Hageman, W. J., & Arrindell, W. A. (1993). A further refinement of the reliable change (RC) index by improving the pre-post difference score: Introducing RC ID. *Behaviour research and therapy*, *31*(7), 693-700.

Hageman, W. J., & Arrindell, W. A. (1999). Clinically significant and practical! Enhancing precision does make a difference. Reply to McGlinchey and Jacobson, Hsu, and Speer. Behaviour Research and Therapy, 37(12), 1219-1233. PII: S0005-7967(99)00036-4

Hiller, W., Schindler, A. C., & Lambert, M. J. (2012). Defining response and remission in psychotherapy research: A comparison of the RCI and the method of percent improvement. *Psychotherapy Research*, *22*(1), 1-11.

Hayes, A. M., Laurenceau, J. P., Feldman, G., Strauss, J. L., & Cardaciotto, L. (2007). Change is not always linear: The study of nonlinear and discontinuous patterns of change in psychotherapy. *Clinical psychology review*, *27*(6), 715-723.

Grant, N., Hotopf, M., Breen, G., Cleare, A., Grey, N., Hepgul, N., ... & Young, A. H. (2014). Predicting outcome following psychological therapy in IAPT (PROMPT): a naturalistic project protocol. *BMC psychiatry*, *14*(1), 170.

Gyani, A., Shafran, R., Layard, R., & Clark, D. M. (2013). Enhancing recovery rates: lessons from year one of IAPT. *Behaviour Research and Therapy*, 51(9), 597-606.

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: description, application, and alternatives. *Journal of consulting and clinical psychology*, *67*(3), 300.

Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status: ascertaining the minimal clinically important difference. *Controlled clinical trials*, *10*(4), 407-415.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of consulting and clinical psychology*, *59*(1), 12.

López-Ratón, M., Cadarso-Suárez, C., Molanes-López, E. M., & Letón, E. (2016). Confidence intervals for the symmetry point: an optimal cutpoint in continuous diagnostic tests. *Pharmaceutical statistics*, *15*(2), 178-192.

López-Ratón, M., Rodríguez-Álvarez, M. X., Cadarso-Suárez, C., & Gude-Sampedro, F. (2014). OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests. *Journal of statistical software*, *61*(8), 1-36.

Löwe, B., Unützer, J., Callahan, C. M., Perkins, A. J., & Kroenke, K. (2004). Monitoring depression treatment outcomes with the patient health questionnaire-9. Medical care, 42(12), 1194-1201.

Kazdin, A. E. (1999). The meanings and measurement of clinical significance. doi.org/10.1037/0022-006X.67.3.332

Kraemer, H. C., Noda, A., & O'Hara, R. (2004). Categorical versus dimensional approaches to diagnosis: methodological challenges. *Journal of Psychiatric Research*, *38*(1), 17-25.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The phq-9. *Journal of general internal medicine*, *16*(9), 606-613.

McMillan, D., Gilbody, S., & Richards, D. (2010). Defining successful treatment outcome in depression using the PHQ-9: a comparison of methods. *Journal of affective disorders*, *127*(1), 122-129.

Metz, C. E. (1978, October). Basic principles of ROC analysis. In *Seminars in nuclear medicine* (Vol. 8, No. 4, pp. 283-298). WB Saunders.

Panagiotakopoulos, T. C., Lyras, D. P., Livaditis, M., Sgarbas, K. N., Anastassopoulos, G. C., & Lymberopoulos, D. K. (2010). A contextual data mining approach toward assisting the treatment of anxiety disorders. *IEEE Transactions on Information Technology in Biomedicine*, *14*(3), 567-581.

Pepe, M. S. (2003). The statistical evaluation of medical tests for classification and prediction. Medicine. ISBN 9780198509844

R Core Team (2014). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in medicine*, *25*(1), 127-141.

Rozental, A., Andersson, G., Boettcher, J., Ebert, D. D., Cuijpers, P., Knaevelsrud, C., ... & Carlbring, P. (2014). Consensus statement on defining and measuring negative effects of Internet interventions. *Internet interventions*, *1*(1), 12-19.

Titov, N., Dear, B. F., Ali, S., Zou, J. B., Lorian, C. N., Johnston, L., ... & Fogliati, V. J. (2015). Clinical and cost-effectiveness of therapist-guided internet-delivered cognitive behavior therapy for older adults with symptoms of depression: a randomized controlled trial. *Behavior therapy*, *46*(2), 193-205.

Titov, N., Dear, B. F., Johnston, L., Lorian, C., Zou, J., Wootton, B., ... & Rapee, R. M. (2013). Improving adherence and clinical outcomes in self-guided internet treatment for anxiety and depression: randomised controlled trial. *PLoS One*, *8*(7), e62873.

Titov, N., Dear, B. F., Staples, L. G., Terides, M. D., Karin, E., Sheehan, J., ... & McEvoy, P. M. (2015). Disorder-specific versus transdiagnostic and clinician-guided versus self-guided treatment for major depressive disorder and comorbid anxiety disorders: a randomized controlled trial. *Journal of Anxiety Disorders*, *35*, 88-102.

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. Medical Decision Making, 26(6), 565-574.

Wise, E. A. (2004). Methods for analyzing psychotherapy outcomes: A review of clinical significance, reliable change, and recommendations for future directions. *Journal of personality assessment*, *82*(1), 50-59.

World Health Organization, Composite international diagnostic interview, version 1.0, World Health Organization, Geneva (1990) Wu, X., Liu, J., Tanadini, L. G., Lammertse, D. P., Blight, A. R., Kramer, J. L., ... & Fawcett, J. (2015). Challenges for defining minimal clinically important difference (MCID) after spinal cord injury. *Spinal Cord*, *53*(2), 84.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32-35.

Zimmerman, M. (2003). What should the standard of care for psychiatric diagnostic evaluations be?. *The Journal of nervous and mental disease*, *191*(5), 281-286.

Zhou, T., Li, X., Pei, Y., Gao, J., & Kong, J. (2016). Internet-based cognitive behavioural therapy for subthreshold depression: a systematic review and meta-analysis. *BMC psychiatry*, *16*(1), 356.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, *39*(4), 561-577.

Supplementary material A

The RCI calculation steps

The RCI cut-off was determined following the calculation steps outlined by Jacobson and Truax (1991), using estimates of variance and change from the treatment sample (*n*=1096). The RCI symptom cut-off was calculated following the Jacobson and Truax (1991) formula, a symptom cut of five-points or more was identified. The calculation details are presented below.

$$RC = \frac{X_{PreTx} - X_{PostTx}}{\sqrt{2\left(S_x\sqrt{(1} - R_{xx}\right)^2}}$$

S_x Standard deviation of scores at pre-treatment

R_{xx} = Test retest reliability of the measure, determined prior to treatment;
correlation of two time points prior to treatment
(e.g reliability of scores under control conditions; Jacobson & Truax, 1991)

X_{PreTx} = Individual score at pre-treatment

X_{PostTx} = Individual score at post treatment

RC = Standardized reliable change; score greater than 1.96 are considered significant

The RCI calculation for PHQ-9 change identified the cut-off at five points.

S.D of scores at pre-treatment = 4.82

Test retest reliability of the measure = 0.84 (Kroenke et al., 2001)

$$\text{Symptom cut-off } = 1.96 * \sqrt{2 * \left(4.82\sqrt{(1} - 0.84)\right)^2} = 5.344 \approx 5 \text{ PHQ} - 9 \text{ points}$$

The RCI calculation for GAD-7 change similarly identified the cut-off at five points.

S.D of scores at pre-treatment = 4.55

Test retest reliability of the measure = 0.83 (Spitzer et al., 2006)

$$\text{Symptom cut-off } = 1.96 * \sqrt{2 * \left(4.55\sqrt{(1} - 0.83)\right)^2} = 5.20 \approx 5 \text{ GAD} - 7 \text{ points}$$

Supplementary material B

The basic Min-TR code needed to apply within the OptimalCutpoints package

> NameYourRObject <- optimal.cutpoints.default(X = "PHQ-9_difference_score", status = "TxAllocation", tag.healthy = "1", methods = c("SpEqualSe"), data = YourDataSetName, categorical.cov = "PossibleSubgroups", control = control.cutpoints(), ci.fit = TRUE)

Arguments explained
NameYourRObject – The element that will house the analysis information

X = "PHQ-9_difference_score" – This feature specifies what test variable (predictor) should be used to classify treatment allocation. In this example, PHQ-9_difference_score is a variable collating the difference in an individual's score from pre to post-treatment. status = "TxAllocation" – The known randomised treatment allocation of individuals to either a treatment or waitlist/placebo condition. It is suggested that in this binary variable, zero will be coded as control, and one as the treatment.

tag.healthy = "1" – Indicating the targeted outcome for classification.
methods = c("SpEqualSe") – Weighing the influence of sensitivity and specificity. Several options are available here (See Lopez-Raton et al., 2014 for more detail). In this example, equal weighting was given to specificity and sensitivity, and this option.
data = YourDataSetName – The dataset uploaded to the R environment.

categorical.cov = "PossibleSubgroups"- This feature is akin to a split file command. Where the discriminatory analysis will be run within subgroups. If the entire sample is considered, this variable should have a column with a single value, i.e. 1 for all cases. If researchers seek to investigate cut-points within subgroups such as gender, or baseline severity bands, the categorical.cov feature should be specified.

control = control.cutpoints() - Used to set various parameters controlling the optimal-cutpoint selection process. No criterion were applied in this analysis. See Lopez-Raton et al., 2014 for more detail

ci.fit = TRUE – Estimation of confidence intervals
Reviewing results
> summary(NameYourRObjectHere)

For additional software information see
Lopez-Raton, M., Rodriguez-Alvarez, M.X, Cadarso-Suarez, C. and Gude-Sampedro, F. (2014). OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. Journal of Statistical Software 61(8), 1–36. URL http://www.jstatsoft.org/v61/i08/.
And
https://cran.r-project.org/web/packages/OptimalCutpoints/OptimalCutpoints.pdf

Chapter 7

General Discussion

General summary of the research problems, gaps and aims of this thesis

The General Introduction to this thesis highlighted the complexity of selecting appropriate outcome measurement techniques, statistical methods, and metrics for psychotherapy research and treatment evaluation. It also highlighted a key dilemma faced by clinical researchers in their efforts to generate clinical evidence where, on the one hand, researchers need to capture the unique and specific (ideographic) features of their data, while also wanting to maximise the comparability and generalisability (nomothetic features) of their clinical evidence with existing evidence. That is to say, whether they are aware of it or not, clinical researchers face a decision along an ideographic-nomothetic axiom when selecting between research methods and metrics (Robinson, 2011). On the nomothetic end of the axiom are metrics and methods that maximise the comparability and generalisability of clinical evidence. However, these methods can come at the cost of the ability to capture features that are unique and specific to a context (idiosyncratic measurement), and by overlooking these features the accuracy and its validity are reduced. At the idiosyncratic end of the axiom are methods and metrics that capture all of the unique and specific features (idiographic measurement), maximising validity, but at the cost of comparability and generalisability. Thus, validity (idiosyncratic measurement) and generalisability (nomothetic measurement) can form competing priorities for clinical researchers in psychotherapy. The General Introduction also noted that there is a much stronger focus on validity than generalisability within the statistical literature, but a much stronger focus on generalisability than validity within the clinical literature.

This preference for generalisability among clinical researchers seems to have led certain statistical methods and metrics to become dominant across the psychotherapy literature. For example, linear measurement methods and metrics such as Cohen's *d*, the handling of missing cases through LOCF or MCAR replacement approaches, and classification of clinical outcomes through the RCI methodology, were identified as dominant in psychotherapy; as noted in the Introduction's review of recent literature concerning internet-delivered psychotherapy for anxiety and depression (Table 2) and elsewhere (Bell, Olivier, & King, 2013; Clarke, 2007; Laken, 2013). Unfortunately, as was also noted in the Introduction, very little published research has explored the validity of these measures and metrics

in the psychotherapy context; a gap that is reflected in the psychology literature (Baldwin Fellingham & Baldwin, 2016; Lambert & Ogles, 2009; Silberzahn, Uhlmann, Martin, Anselmi, Aust, Awtrey & Carlsson, 2018), through the lack of reported statistical assumption testing reviewed amongst the practices clinical researchers employ in psychotherapy trials (Table 2), and in methodological reviews of the psychotherapy research literature more broadly (Harris, Reeder & Hyun; 2011; Miettunen, Nieminen, & Isohanni, 2002; Nieminen & Kaur, 2019; Nieminen, Virtanen & Vähänikkilä, 2017). The limited research available about the features of psychotherapy data, or the suitability of the current predominant metrics to capture the impact of psychotherapy was identified as a major research gap; with the potential to impact the quality of evidence regarding psychotherapy.

To start to address this gap this thesis employed a novel approach. It first explored in detail the core features of anxiety and depressive symptoms across different symptom scales and how these changed over time with and without psychotherapy (e.g. functions of symptom change, statistical distributions, treatment-related change, non-specific symptom change). These features were then used to examine the relative validity of different methods and metrics that could be used in evaluating psychotherapy outcomes, including those that are currently widely employed. This approach intended to optimise both the validity and generalisability of the research methods and metrics used within the literature by identifying those methods and metrics that best reflected the features of psychotherapy data concerning anxiety and depression. In taking this approach, the research studies of the thesis intended to point to new methodological opportunities for measuring and interpreting clinical evidence with an increased balance of validity and generalisability.

This aim of maximising both validity and generalisability is somewhat unconventional from the perspective of both statistical literature and practice (Baldwin et al., 2016), and clinical literature and practice (Lambert & Ogles, 2009). Nevertheless, the studies of the thesis aimed to contribute to the psychotherapy literature with a series of studies engaging with three core domains of clinical evidence generation. Studies 1 and 2 explored the different statistical features of clinical change in symptoms of anxiety and depression as a result of psychotherapy. Studies 3 and 4 explored different approaches for the handling of missing cases in psychotherapy research for anxiety and depression. Study 5 explored

methods for the identification and classification of clinical change in symptoms of anxiety and depression as a result of treatment. Together, the studies of the thesis aimed to inform clinical researchers about the features of psychotherapy data concerning anxiety and depression, and the methods and metrics that should be considered in the multistep process of generating clinical evidence. These studies and their findings are discussed below.

Overview of findings and new knowledge gained

The measurement of symptom change as a proportional function

Studies 1 and 2 examined the statistical features and assumptions that characterise anxiety and depression symptom scores, and the function of symptom change that patients exhibited over the course of treatment. Both studies strongly indicated that a proportional function of symptom change, and symptom score distributional skewness, was common across several widely used symptom scales (the PHQ-9, GAD-7, and K-10) and that these patterns were found among both research trial data and routine care data. Consequently, methods that reflected the features and assumptions of proportional change, and distributional skew, were identified to enhance the validity of evidence across contexts. Specifically, the use of methods that reflected the proportionality of symptom change led to a more accurate and representative measurement of patient outcomes through treatment, when compared to a linear function of change. However, Study 1 was limited in that it relied on data from research trials and only examined the features of change in depression symptoms.

Study 2 replicated the methods used in Study 1 and demonstrated that the results generalised to other symptom domains, namely anxiety, general psychological distress, as well as depression. It also indicated that the same pattern of findings generalised across therapy contexts, specifically, by observing the replication of the results from clinical trials in routine clinical care samples. In combination, these two studies suggested that methods and metrics that recognise the distributional skewness of symptom data and the proportional nature of symptom change were likely to result in substantial gains in measurement accuracy and conclusion validity, that is, compared with existing methods and metrics used across the literature. In this way, the proportional measurement methods and

metrics explored were considered to strike a more optimal balance between validity and generalisability than currently used methods and metrics.

Together, as a more abstract learned conclusion, Studies 1 and 2 demonstrated that the symptom change individuals experience in psychotherapy can be more accurately be described as a proportional function of symptom change. Because this feature was observed across several contexts such as symptom scales and treatment contexts (albeit preliminary), the research of these studies suggests that researchers can measure, interpret and compare the symptom change in psychotherapy with increased validity and generalisability. That is, on the condition that researchers employ those measurement methods that reflect the proportionality of symptom change from scores that preceded treatment. These studies also demonstrated that the use of methods that overlook the proportionality of symptoms can result in validity threats and markedly less accurate prediction of outcomes.

Measuring and identifying the outcomes of missing cases

Studies 3 and 4 explored the statistical features and assumptions that can be made about missing cases in psychotherapy data. Study 3 demonstrated that missing cases were strongly associated with reduced levels of treatment adherence, and that this feature impacted both the symptom outcomes of patients as well as their likelihood of becoming missing at post-treatment. In statistical terms, the results of Study 3 provided evidence that missing cases followed a *conditional* missing at random assumption, that is, cases were missing at randomly within groups of individuals that underwent different levels of treatment adherence. A comparison between the measurement accuracy of replacement methods that consider this feature demonstrated considerably improved accuracy in predicting the outcomes of missing cases (~30-40%), compared with methods that overlooked this feature. In this way, methods that account for the features of missing cases, namely treatment adherence, appeared to increase the validity of the replacement scores. However, this study was again limited in that it only focused on symptoms of depression, and employed data from research trials.

Study 4 replicated and extended the methods of Study 3 by also examining symptoms of anxiety and general psychological distress, using a large dataset from routine care. Study 4 found treatment

adherence was the single most important variable associated with missing cases. Thus, with the addition of Study 4, this overall finding was replicated across different symptom scales and across both research trials and routine care samples. Together, Studies 3 and 4 identified that missing cases were not comparable in their outcomes to cases that provided outcome data (non-missing) following treatment.

Together, as a conclusion, Studies 3 and 4 demonstrated that in order to accurately predict the trajectories of missing cases and therefore draw valid conclusions about the effectiveness of treatments, researchers should consider a set of specific variables, such as treatment adherence. In these studies, the nuanced variable of reduced treatment adherence was demonstrated to be the single biggest feature of missing cases, and this feature had clear implications for the ability to measure and interpret the outcome for missing cases. Researchers on their part could measure, interpret and compare the outcomes of missing cases in treatment if they also considered the degree to which patients adhere to the protocol of treatment. These studies also demonstrated that the use of methods that overlook the reduced adherence of missing cases can result in validity threats and markedly less accurate prediction of outcomes.

Measuring and interpreting the minimal but treatment specific impact of treatment on individuals

Study 5 investigated a novel proposed classification approach designed to identify and quantify the minimal amount of change in symptoms of anxiety and depression associated with treatment, called Minimal Treatment-Related change (Min-TR). This approach used the change associated with treatment, compared with change observed in control conditions, to controls for statistical features such as regression to the mean, and the proportionality of symptom change. This methodological approach provided a new tool to investigate the specificity and sensitivity of different measurement approaches, for example, comparing linear change scores, proportional change and residual post-treatment symptom outcomes. In brief, Study 5 highlighted that researchers' ability to identify and quantify treatment-related change must consider the features of nonspecific symptom change and the proportionality of symptom change, and highlighted the Min-TR as a viable approach for validly identifying and quantifying the treatment-related change.

Together, as a learned conclusion, Study 5 demonstrated that the degree of nonspecific symptom change (~±30%) can be informative for deducing the classification of individual-level outcome to treatment that is specific or nonspecific to the treatment condition. Further, this approach illustrated the importance of addressing features such as the proportionality of change in order to increase the accuracy of deducing treatment-specific and nonspecific symptom change. These studies also demonstrated that the use of methods that overlook the rate of non-specific symptom change, or the proportionality of symptoms, can result in less sensitive and less accurate classification of the outcome of treatment; being the minimal treatment-specific response (Min-TR), or even the change on a clinical diagnosis status.

Comparison of findings with the existing methodology literature in psychotherapy

The result of the five studies comprising this thesis suggests that the validity and generalisability of clinical evidence and psychotherapy evaluation rely heavily on the choice of methods and metrics used. Together, the five studies demonstrated the importance of identifying the statistical features that characterise the data sets and then selecting measurement methods that reflect those features.

The results from the five studies also suggest that the dominant methods and metrics currently used in psychotherapy research (Cohen's *d*, *RCI*, *MCAR*) may overlook important statistical features and assumptions that are common to symptom data and important for the context of psychotherapy. As demonstrated in each of the studies, overlooking these statistical features of the data resulted in avoidable measurement error, measurement bias, and in some circumstances, even potentially faulty conclusions about symptom change and the efficacy of treatment. For example, by overlooking the proportionality of symptom change, samples with severe baseline symptoms obtain Cohen's *d* effects sizes that were larger than the average treatment effect, even when the sample was drawn from a control group who did not receive treatment (Study 1). Moreover, by overlooking the characteristics of missing cases and applying MCAR or LOCF methods, the estimation of missing cases over-estimated the symptom change individuals experienced in treatment by as much as 39% (Studies 3, 4). Similarly, by

overlooking the symptom change that can occur without treatment (e.g., due to regression to the mean) and applying the widely-used RCI method, the classification of individuals who *responded* to treatment was not specific enough to reflect the effects of treatment (Study 5). Specifically, participants in both treatment and control groups were classified as having made an improvement using the RCI approach. In combination, the five studies of the thesis highlighted the critical need to select methods and metrics that consider the features of symptom data, and the prospect of these features to optimise the validity and generalisability of any resulting analyses.

The results of the studies in the thesis are consistent with findings reported in the broader methodological literature, but are novel and challenge some of the prominent practices in psychotherapy research. Within the broader methodological literature the need to consider the features and assumptions that present within a dataset is a common and well-understood practice. This is reflected in multiple clinical research streams such as scale development research (Mokkink, Terwee, Patrick, Alonso, Stratford, Knol, ., ... & De Vet, 2010; Angst, 2011), clinical data mining (Bellazzi & Zupan, 2008; Herland, Khoshgoftaar, & Wald, 2014), epidemiology (Brakenhoff, Mitroiu, Keogh, Moons, Groenwold & van Smeden, 2018; Keogh & White, 2014) and medical data modelling (Diggle, 2015; Obermeyer & Emanuel, 2016) and even psychopathology studies (Ebrahim, Sohani, Montoya, Agarwal, Thorlund ,Mills & Ioannidis, 2014; Field & Wilcox, 2017). This is because the pre-screening of datasets for ideographic features is known to guide researchers towards the selection of methods that better fit the data, which in turn increases accuracy and validity. This established practice was supported by the results from all five studies, where measurement models, missing cases handling and classification accuracy, were all improved by the selection of measurement methods and metrics that matched the features of the data. These features were identified at the point of pre-screening data.

Although the notion of data pre-screening and statistical assumption testing is a widely accepted part of the statistical methodology literature, the novel contribution of this thesis was the ability to provide preliminary evidence pointing to a specific set of data features that are highly relevant to treatment-related change. For example, although the need to identify the function of symptom change through data screening might be well understood in the statistical literature (Ng & Cribbie, 2017), within

the psychotherapy literature the exploration of the function of symptom change is rarely practiced or researched (de Beurs, Barendregt, de Heer, van Duijn, Goeree, Kloos, ., ... & Merks, 2016). It is also important to recognise that the data features (e.g. proportional function of change) identified in this thesis are also not commonly discussed within those areas of psychotherapy that rely on advanced and extensive statistical modelling. For example, methodological discussion papers on comparative meta-analyses (Barth, Munder, Gerger, Nüesch, Trelle, ,Znoj, ... & Cuijpers, 2016) and meta-analysis methodology (Grove, Zald, Lebow, Snitz, ,& Nelson ,2000; Sanders & Hunsley, 2018), do not consider the role of proportional symptom change, or the impact of missing cases, and therefore overlook the biases these issues may create. In a similar way, researchers can overlook the proportional function of symptom change by transforming the symptom outcome variables through log-linear functions (creating a multiplicative/proportional function of change) and then undermine the importance of baseline symptoms by reporting minimal correlations between baseline symptoms scores and the magnitude of change (Liu & Maxwell, 2019; Paul, Andlauer, Czamara, Hoehn, Lucae, Pütz, ,... & Sämann, 2019). Furthermore, the proportionality of symptom change is rarely considered within papers about advanced statistical modelling for psychotherapy, including those focused on artificial intelligence algorithms and learning machines (Boman, Abdesslem, Forsell, Gillblad, ,Görnerup ,Isacsson., ... & Kaldo, 2019) or technical papers concerning psychotherapy prediction models (Safinianaini, Boström, & Kaldo, 2019). Whilst these papers consider a great range of technical topics, these and others (Cañete-Massé, Peró-Cebollero, Gudayol-Ferré, & Guàrdia-Olmos, 2018; Flood, Page, & Hooke, 2018) do not currently consider the importance of examining the fundamental features of symptom data or symptom change.

The above examples highlight that the identification of data features, such as proportional symptom change, are not trivial and can be overlooked by even the most experienced researchers and in studies involving advanced statistical knowledge and analytics. The examples also demonstrate that the function of symptom change must be understood as an inherent feature of psychotherapy data and therefore considered in the process of measuring and interpreting clinical evidence. For this reason, whilst the practice of data pre-screening is common and somewhat simple practice in statistical research,

the ability to identify generalisable features of symptoms and symptom change from psychotherapy and draw links to the methodological literature, is novel and potentially helpful.

The studies of this thesis also raise several considerations and recommendations that are not common within the statistical literature, which emphasises the selection of methods by context, that is, ideographic measurement. As noted in the introduction, currently, methodological guidelines for clinical analysis (e.g., Lang & Altman, 2013; Baldwin et al., 2016; Field & Wilcox, 2017), for measurement (e.g., Delucchi & Bostrom, 2004), or for the handling of missing cases (e.g., Blankers, Koeter & Schippers, 2010; Little, 1995), recommend the investigation of statistical assumptions and selection of method based on the presenting data features. Such authors, however, tend to assume that researchers will prioritise those methods that fit the specific (ideographic) features of the data, rather than other potential priorities, such as the ability to compare outcomes with existing studies (generalisability), or other methodological practices that are shared in the field as a standard. However, at the same time, the need and preference for generalisability, and shared, comparable metrics (nomothetic measurement) cannot be overlooked (Clarke, 2007; Gottfredson, Cook, Gardner, Gorman-Smith, Howe, Sandler, & Zafft, 2015), especially when the majority of researchers chose widely methods such as Cohen's *d* and RCI (Lakens, 2013; Lambert & Ogles, 2009). Thus, as outlined in the General Introduction, there is a critical need for accessible research that explores methods and metrics that maximise both validity and generalisability and can help to guide clinical researchers in the methods and metrics they employ. Emphasising idiographic methods and metrics that maximise validity, as the statistical literature does, represents only one part of a bigger methodological challenge for clinical researchers, who strive to compare their results and interpret their results in the context of the existing clinical literature.

Implications for treatment efficacy and clinical evidence

The results of the five studies imply several points about the dominant methods used to measure, interpret and compare clinical evidence, and in particular the use of Cohen's *d* effect sizes, RCI classification of individual effects and MCAR, LOCF missing cases approaches. First, each of the

five studies illustrates that the current use of these dominant methods and metrics can lead to biased results and misleading conclusions about the efficacy of treatment. The degree of bias is, of course, dependent on both the choices made by researchers and the features of the data. For example, in Studies 1 and 2, the measurement error associated with the selection of an unsuitable linear function of change resulted in prediction error that varied between 29% (K-10) and 39% (PHQ-9). Further, the bias associated with different measurement choices was found to vary as a result of initial symptom severity of the sample. For example, Cohen's *d* effect sizes artificially increased or decreased the estimate of treatment efficacy by up to three times, even when the same treatment was being evaluated. From Studies 1 and 2, we learnt that this error in measurement and interpretation can be circumvented by applying a measurement method that considered the feature of proportional symptom change. Similarly, the results of the studies of the thesis indicate that trials with high degree of missing cases are likely to artificially increase estimates of the efficacy of treatment as missing cases tend to obtain worse outcomes than the average person providing data following treatment. However, as Studies 3 and 4 illustrate, this can be circumvented with the addition of minimal but specific methodological steps, and the specific consideration of treatment adherence in the handling of missing cases.

Second, the results of the five studies also suggested that the process for measuring and interpreting evidence in clinical trials must be considered as an incremental process with distinct but interacting components; as outlined by analytical guidelines in frameworks such as the JARS and CONSORT. For example, the measurement accuracy gains from Studies 1 and 2 can be used to improve researcher's ability to estimate the outcomes of missing cases. In turn, the optimal measurement of symptom change and handling of missing cases then incrementally improves researchers' ability to identify the individual outcomes of treatment. Thus, while each component is distinct and important in itself, in combination they represent interacting components of the research process that can contribute to or detract from the generation of valid clinical evidence measurement, that is, depending on the methodological decisions made. Importantly, the studies of this thesis also demonstrate how bias can be introduced into research findings via poor methodological decisions at each step of the research process, whilst still following the recommendations outlined by and being compliant with JARS and

CONSORT frameworks. Thus, by placing the onus of methodological choices on clinical researchers, such frameworks are arguably insufficient for preventing measurement bias. Rather, specific research and guidance about the selection of methods are needed, in order to identify the features of psychotherapy, assumptions and approaches that can mitigate bias.

Third, the ability to measure the outcomes of treatment in a way that more accurately reflects the trajectory of patients holds promise for a range of clinical research streams other than treatment evaluation. These may include research topics such as treatment cost-effectiveness, evaluations of psychological processes, risk profiling of individuals in treatment, and research seeking to compare the efficacy of treatments. Research in these areas relies on the valid and generalisable measurement of treatment efficacy, and the clear labelling of outcomes such as patient remission or risk. For example, research concerning cost of a treatment and cost-effective ratio of successful outcomes relies on researchers' ability to agree on a magnitude of clinical change that is clinically meaningful (Murray, Hekler, Andersson, Collins, Doherty, Hollis ... & Wyatt, 2016; Ross, Zivin & Maixner, 2018) and this outcome in turn forms the dependent variable associated with costs and other economic outcomes. However, numerous areas of clinical research are reliant on researcher's ability to accurately identify clinical change, including research focused on the risk profiling of patients (Delgadillo, Moreea & Lutz, 2016; Karyotaki, Kemmeren, Riper, Twisk, Hoogendoorn, Kleiboer, ... & Littlewood, 2018), the profiling of patients who benefit the most from treatment (Castellani, Rajaram, Gunn, ., & Griffiths, 2016; Driessen, Abbass, Barber, Gibbons, Dekker, Fokkema, ... & Town, 2018; Gunn, Elliott, Densley, Middleton, Ambresin, Dowrick, ., ... & Griffiths, 2013), as well as comparative studies of treatment efficacy (Leichsenring, Abbass, Driessen, Hilsenroth, Luyten, Rabung, & Steinert, 2018). The abler clinical researchers are to accurately identify clinical outcomes, the greater their ability to research treatment mechanisms, predictors of outcomes, and the profiles of patients. For example, whilst currently there are few studies to compare how measurement issues identified in this thesis affect cost-effectiveness research, it is logical that the associated reductions in measurement error would improve the ability to detect and understand the cost-effectiveness of psychotherapy (Murray et al., 2016).

Limitations and recommended future directions

The findings and conclusions drawn from each of the five studies must be considered alongside important limitations, several of which were discussed in the individual studies, and are revisited here as recurrent themes across the five studies. The themes of these limitations include (1) the preliminary and limited range of clinical contexts explored, and the need to replicate the results of the studies across additional contexts; (2) the limited ability to translate the measurement validity into clinical validity, and the need for further verification that link statistics to the experience of patients in treatment; and (3) the challenge to achieve reform in psychotherapy evidence. Each of these will be discussed as a topic and coupled with proposed future research.

The preliminary and limited range of clinical contexts explored, and the need to replicate the results of the studies across addition psychotherapy contexts

As noted in each of the introductions and related studies, the methods applied in this thesis represent a subset of a large range of possible methods of measurement and analysis of outcomes, and of psychotherapy contexts. For example, in the current thesis, the examination of clinical change was restricted to changes on self-reported symptom scales. Further to this, the review and critique of measurement practice and clinical evidence were also limited to self-reported symptom scales. Specifically, clinical evidence was narrowly defined as the integration of standardized symptom scales, statistical methods (e.g. effect sizes) that together form the dominant methods for evaluating the impact of psychotherapy. This limited definition for the measurement of change was further limited to the exploration of preliminary and exemplary contexts that included the use of three specific symptoms scales (GAD7, PHQ9, and K10) and the clinical context of iCBT under clinical trials and routine care.

Together, the limited range of psychotherapy contexts (e.g., internet-delivered CBT, symptoms of anxiety and depression) explored in the thesis means that the measurement solutions offered in this thesis are preliminary, and currently relevant to a narrow range of contexts. Without the additional exploration of the features of psychotherapy data, across additional scales and treatment contexts including different samples and types of treatment, the relevance of results and proposed measurement

solutions for psychotherapy research more broadly are uncertain. That is, it is currently unclear to what extent features such as proportional symptom change, missing cases mechanisms, and the classification of outcomes are representative and generalisable to other psychotherapy contexts. Future research on the data features of symptom change might seek to replicate the studies on measurement function (Study 1 and 2), the patterns of missing cases (Study 3 and 4), and the classification of outcomes in additional psychotherapy contexts (Study 5). Replications and elaboration should be conducted in contexts that include traditional face-to-face psychological services, other symptom domains, and use data from active control samples and pharmaceutical placebos, and with different symptom scales.

The relevance of the methods and metrics examined in this thesis for different samples, types of treatments or other symptom scales could be investigated through meta-analytic studies. For example, a meta-analysis about psychotherapy efficacy could be conducted to include multiple estimates of clinical change about the same treatment and across multiple types of symptom scales. Multiple estimates of clinical change for the *same treatment* would provide the necessary data to confirm the occurrence of proportional change under different scales and explore the variance associated with the use of linear and proportional functions of symptom change. Alternatively, a meta-analysis comparing different treatments and samples for the *same symptom scale* (e.g. multiple studies that employ PHQ-9) could be used to explore the occurrence of linear or proportional symptom change across different types of treatment and samples, whilst holding other measurement factors constant. Further, if several clinical samples were to be combined into a dataset for individual participant meta-analysis (Karyotaki, et al., 2018; Riley, Lambert, & Abo-Zaid, 2010) patterns of missing cases, and the sensitivity of the Min-TR cut-offs could also be investigated. Such research could shed light on the generalisability of data features, methods and metrics, which was a key aim of this thesis. For example, if datasets concerning the same types of treatments, such as telephone supported iCBT were aggregated into a dataset for an intra-individual meta-analysis, it would be possible to quantify the measurement variance of scales, when the same treatment is measured through different symptom scales. Thus, one avenue for future research, with significant potential, is to start to combine datasets from across multiple

contexts in order to more comprehensively explore the kinds of methodological issues examined within this thesis.

It is also important to consider the findings of this thesis as preliminary because the five studies examined only a limited range of statistical methods and assumptions for operationalising the measurement of psychotherapy effects and clinical evidence. As preliminary studies, this thesis did not have the scope to compare additional and important methods that could be relevant for psychotherapy data. For example, in this thesis the function of symptom change (Studies 1 and 2) considered two options (linear and multiplicative) where other options such as a zero-inflated, two-part models (Ferrer, Conger & Robins, 2016), non-parametric semi-parametric modelling such as generalised additive models (GAM)(Ng & Cribbie, 2017), may result in at least comparable, if not more accurate, estimates of change. Further, it is important to research how existing and relatively simple measurement solutions, such as a logarithmic-transformation of the scale (Liu & Maxwell, 2019) may enable the use of linear analysis within a context where symptoms change proportionally. For example, solutions such as scale transformation may be useful for researchers who rely on methods such as linear regression and are not in a position to apply alternative methods such as generalised linear modelling. Similarly, within the missing cases studies, the exploration of missing cases prediction was operationalised through a single model imputation, and did not consider other solutions such as multiple imputations (Li, Stuart & Allison, 2015; Sterne, White, Carlin, Spratt, Royston, Kenward, ... & Carpenter, 2009), pattern mixture models (Graham, 2009) or other variables associated with missingness that may provide effective solutions for identifying and correcting for bias arising from missing cases. Thus, future research exploring the performance of alternate missing cases solutions might identify other mechanisms of missing cases and more effective ways to account for the outcomes of missing cases with minimal bias.

Translation of clinical measurement into clinical validity and the need for clinical verification

A second limitation across the five studies is the inability to demonstrate whether the gains in measurement accuracy from the methods and metrics proposed actually translate into gains in clinical validity. Specifically, within the studies, the ability to evaluate the accuracy of outcome measurement

relied on statistical metrics such as AIC and AUC that reflect model measurement variance (σ^2). The studies also relied on the use of clinical 'anchors' such as the contrasts of waitlist and treatment groups (Studies 1 and 4), different levels of treatment adherence (Studies 3 and 4), or symptom change with clinical diagnosis (Study 5). Although these '*anchors'* (e.g., waitlist control groups, low treatment adherence, clinical diagnosis) are important for the creation of valid clinical evidence, these *anchors* are limited in their ability to convey anything about the actual experience of individuals in treatment, or the acceptability of the conclusions drawn by clinicians. For example, in order to conclude that a 50% reduction in symptoms represents a similar clinical effect across moderate and severely symptomatic individuals, additional qualitative research is needed to verify the experience of a treatment effect and the difference in the wellbeing of patients (Studies 1 and 2). Similarly, the ability to conclude that a 25% cut-off can represent a minimal treatment-related change (Study 5) requires additional research to establish whether this reflects a change from the perspectives of both patients and clinicians. Without verification, the ability to translate any advances in methods and metrics into patient experiences is limited, and therefore the clinical validity and utility of new methods and metrics also remain uncertain.

The absence of a direct relationship between statistics and clinical conclusions has been raised as a critical issue by various authors (e.g., Boers, Kirwan, Wells, Beaton, Gossec, d'Agostino, ... & March, 2014; Jaeschke, Singer & Guyatt, 1989; King, 2011; Thompson, 2002). For example, several authors have argued that the measurement of statistically significant change does not necessarily translate to a change that is clinically significant from the perspective of patients. In fact, they have noted that, at times, statistically significant changes may not even be detectable by patients or clinicians (King, 2011; Thompson, 2002). Thus, rather than to rely on statistical metrics, it is important to consider alternative outcomes, such as functional performance, treatment satisfaction, measures of quality of life or other core measures of wellbeing. Future research seeking to bridge this gap between measurement accuracy and clinical significance could include feedback measures about the experience of patients in treatment, and in this way evaluate the improvement in statistical prediction as a clinically relevant issue. Future research seeking to verify the clinical validity of different measurement methods could

also associate the estimates of symptom change from different measurement techniques with self-report qualitative feedback measures. For example, if participants were asked about their levels of treatment satisfaction at mid-treatment, this information could be used to profile the individuals who become missing cases at post-treatment and their possible motivation for not persisting. Similarly, measures such as treatment satisfaction could be used to model linear or proportional symptom change and compare the 50% reduction in symptoms across patients with mild, moderate and severe symptoms at baseline. This type of research would add another level of verification to clinical evidence and could increase the ability to optimise clinical conclusions from measurements.

The challenge to achieve reform and disseminate new practices for measuring and evaluating psychotherapy outcomes

An important and outstanding set of issues relates to the challenge of promoting the uptake of alternative methods of measurement and evaluation; an issue that is described here as an additional limitation but also represents a broader challenge for the field. The studies of this thesis aimed to identify the features and methods that could lead to both more valid and generalizable evidence ***if*** the methods were widely adopted. However, whilst the measurement validity and statistical accuracy of studies may be improved with the adoption of the solutions identified in this thesis, their benefits rely on the degree to which they are adopted by other researchers as a shared standard. In other words, without the broad adoption of the methods and solutions proposed, the ability to achieve improvements is limited; ironically, because different researchers would opt for differing measurement methodology. For example, a clinical study that adopts the recommended measurement of clinical change through proportional change metrics and the use of Min-TR may end up with estimates that are incomparable to existing studies employing other methods and metrics, such as Cohen's *d* and RCI. Critically, as demonstrated in all five of the studies, researchers who adopt the proposed alternatives can, on the one hand, increase the statistical rigour of their results. However, on the other, they can end up with more conservative estimates than those studies that overlook the features of psychotherapy data for anxiety and depression. Thus, the decision to adopt new methods and metrics places researchers in a bind, where the choice of new methods and metrics may result in evidence that shows weaker treatment effects and

initially incomparable effects. Together, these issues may dissuade researchers from initially adopting the methods proposed in this thesis, even if these methods are justified from a statistical point of view.

It is also important to note that over the past 30 years there have been numerous attempts to challenge the use of the current and dominant methods and metrics, such as Cohen's *d* effect sizes (Kelley & Preacher, 2012), methods of managing missing cases (Sterne, White, Carlin, Spratt, Royston, Kenward, ... & Carpenter, 2009), and the classification of clinical outcomes (King, 2011). Some of the efforts to change these dominant methods and metrics have come from the authors of the original metrics themselves. For example Cohen, the author of the effect size (1992; 2016), Jacobson, the author of the RCI (1999), and Rubin, the author of missing cases imputation method (Rubin, 1976; Little & Rubin, 2014), all described shortcomings of their methods and argued against their use when they do not fit the features of the data. However, these methods have become dominant, even in the face of critics and alternatives. As reviewed by Sharpe (2013), Cohen (2017), and Cummings (2014) there is an understandable reluctance to abandon easily applied and commonly used methods and metrics. At the same time, however, as demonstrated in this thesis, changes in measurement and evaluation practice offer significant advantages. Moreover, in light of the preliminary findings of this thesis, the efforts to optimise the methods and metrics used in generating psychotherapy evidence must continue.

Several research efforts could be taken to promote the adoption of new, shared, methods and metrics for psychotherapy research. These could include a series of studies that verify and promote the value of measurement advances in some of the different domains of psychotherapy research. For example, researchers concerned with the identification of moderators or mediators of clinical change, health economic analyses of clinical efficacy, or data mining studies about clinical profiles, may be particularly receptive to the methodology that can reduce measurement error and increase the validity of the resultant analysis. In addition, as mentioned previously, the replication of this thesis' results in different contexts could encourage clinical researchers to adopt or at least consider reporting the methods and metrics proposed in this thesis. For example, researchers in the fields of chronic pain management (Dear, Titov, Perry, Johnston, Wootton, Terides, ... & Hudson, 2013), psycho-oncology (Hopko, Bell, Armento, Robertson ,Mullane, Wolf, & Lejuez, 2008), or paediatric (Rapee, Lyneham,

Wuthrich, Chatterton, Hudson, Kangas, & Mihalopoulos, 2017) and geriatric mental health and its treatment (Hollon, Jarrett, Nierenberg, Thase, Trivedi & Rush, 2005), all rely on the accurate measurement of symptoms and symptom change. Whilst these areas are often considered as separate fields within psychotherapy with dedicated journals and distinct readership, the adoption of new methods and metrics could lead to advances in other fields that rely on the measurement of clinical change and promote awareness of the recommendations proposed in this thesis. For this reason, a possible future direction would be to replicate the research about the measurement in symptoms across a range of contexts where psychotherapy is employed. In this way, the findings and recommendations from the thesis could be more broadly tested and, pending on the generalisability of the findings, could promote a more cross-disciplinary discussion about the nature and measurement of clinical symptom change.

Furthermore, as mentioned in Study 1, a practical way to bridge the gap between emerging and future measurement practices is to promote dual reporting of results under both methods in clinical evidence. For example, clinical trials evaluating new treatments, and even meta-analytic studies focused on treatment efficacy, could report clinical change using co-reported metrics or include supplementary sensitivity analyses that compare different ways to measure change. Together with the cross-context replication mentioned above, this could form a dissemination pathway to achieve more suitable methods for measuring the efficacy of psychotherapy, and in this way result in more valid and generalisable clinical evidence.

Concluding remarks

In these concluding remarks, three points are restated as overarching learning points from the thesis as a whole. First, it is important to acknowledge that the measurement of clinical evidence in psychotherapy remains a complex, diverse, and multifaceted challenge, and this has been the case since the inception of the field. The aim of this thesis, as a body of work, was to contribute to improving our ability to measure and evaluate treatment outcomes. However, despite the efforts to replicate findings across different symptom scales and treatment contexts, the findings and suggestions within each of the

studies are preliminary and associated with clear limitations, uncertainty and contextual considerations that require future research.

Second, this thesis demonstrated the importance of considering separate but incremental methodological steps that form the measurement and evaluation of psychotherapy evidence; including the selection of measurement models, approaches for managing missing cases, and the identification and classification of clinical change. Similarly, the efforts in the thesis to bring together clinical and statistical considerations show promise for optimising both validity and generalisability.

Third, at the heart of this methodological research are patients who are seeking effective treatment. The research of this thesis aimed to find the most suitable ways to measure and evaluate the outcomes of psychotherapy for anxiety and depression, and in this way progress the state of clinical research and science. Although statistical analysis and measurement methodology for treatment evaluation may seem detached from actual patients, the ability to accurately measure the treatment outcomes of patients and empower clinical researchers to make valid conclusions about treatment response is critical. The need to optimise outcome measurement is ever-present for clinical researchers who fundamentally rely on accurate and interpretable clinical evidence. This promise of greater clarity about the measurement of patient outcomes has been the core driver for this thesis' research and needs to be a core focus of research moving forward.

References

Angst, F. (2011). The new COSMIN guidelines confront traditional concepts of responsiveness. BMC medical research methodology, 11(1), 152.

Baldwin, S. A., Fellingham, G. W., & Baldwin, A. S. (2016). Statistical models for multilevel skewed physical activity data in health research and behavioral medicine. Health Psychology, 35(6), 552.

Barth, J., Munder, T., Gerger, H., Nüesch, E., Trelle, S., Znoj, H., ... & Cuijpers, P. (2016). Comparative efficacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis. Focus, 14(2), 229-243.

Bell, M. L., Olivier, J., & King, M. T. (2013). Scientific rigour in psycho-oncology trials: why and how to avoid common statistical errors. Psycho-Oncology, 22(3), 499-505.

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. International journal of medical informatics, 77(2), 81-97.

de Beurs, E., Barendregt, M., de Heer, A., van Duijn, E., Goeree, B., Kloos, M., ... & Merks, A. (2016). Comparing methods to denote treatment outcome in clinical research and benchmarking mental health care. Clinical psychology & psychotherapy, 23(4), 308-318.

Blankers, M., Koeter, J.M., Schippers, M.G., 2010. Missing data approaches in eHealth research: simulation study and a tutorial for nonmathematically inclined researchers. J Med Internet Res 12, e54.

Boers, M., Kirwan, J. R., Wells, G., Beaton, D., Gossec, L., d'Agostino, M. A., ... & March, L. (2014). Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. Journal of clinical epidemiology, 67(7), 745-753.

Boman, M., Abdesslem, F. B., Forsell, E., Gillblad, D., Görnerup, O., Isacsson, N., ... & Kaldo, V. (2019). Learning machines in Internet-delivered psychological treatment. Progress in Artificial Intelligence, 1-11.

Brakenhoff, T. B., Mitroiu, M., Keogh, R. H., Moons, K. G., Groenwold, R. H., & van Smeden, M. (2018). Measurement error is often neglected in medical literature: a systematic review. Journal of clinical epidemiology, 98, 89-97.

Cañete-Massé, C., Peró-Cebollero, M., Gudayol-Ferré, E., & Guàrdia-Olmos, J. (2018). Longitudinal estimation of the clinically significant change in the treatment of Major Depression Disorder. Frontiers in psychology, 9, 1406.

Castellani, B., Rajaram, R., Gunn, J., & Griffiths, F. (2016). Cases, clusters, densities: Modeling the nonlinear dynamics of complex health trajectories. Complexity, 21(S1), 160-180.

Cohen, B. (2017). Why the resistance to statistical innovations? A comment on Sharpe (2013). Psychological Methods, 22(1), 204-210.

Cohen, J. (1992). A power primer. Psychological bulletin, 112(1), 155.

Cohen, J. (2016). The earth is round (p<. 05). In What if there were no significance tests? (pp. 69-82). Routledge.

Dear, B. F., Titov, N., Perry, K. N., Johnston, L., Wootton, B. M., Terides, M. D., ... & Hudson, J. L. (2013). The Pain Course: a randomised controlled trial of a clinician-guided Internet-delivered cognitive behaviour therapy program for managing chronic pain and emotional well-being. PAIN®, 154(6), 942-950.

Delgadillo, J., Moreea, O., & Lutz, W. (2016). Different people respond differently to therapy: a demonstration using patient profiling and risk stratification. Behaviour Research and Therapy, 79, 15-22.

Delucchi, K. L., & Bostrom, A. (2004). Methods for analysis of skewed data distributions in psychiatric clinical studies: working with many zero values. American Journal of Psychiatry, 161(7), 1159-1168.

Diggle, P. J. (2015). Statistics: a data science for the 21st century. Journal of the Royal Statistical Society: Series A (Statistics in Society), 178(4), 793-813.

Ebrahim, S., Sohani, Z. N., Montoya, L., Agarwal, A., Thorlund, K., Mills, E. J., & Ioannidis, J. P. (2014). Reanalyses of randomized clinical trial data. Jama, 312(10), 1024-1032.

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: a primer for clinical psychology and experimental psychopathology researchers. Behaviour Research and Therapy.

Flood, N., Page, A., & Hooke, G. (2018). A comparison between the clinical significance and growth mixture modelling early change methods at predicting negative outcomes. Psychotherapy Research, 1-12.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. Annual review of psychology, 60, 549-576.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. Psychological assessment, 12(1), 19.

Gunn, J., Elliott, P., Densley, K., Middleton, A., Ambresin, G., Dowrick, C., ... & Griffiths, F. (2013). A trajectory-based approach to understand the factors associated with persistent depressive symptoms in primary care. Journal of affective disorders, 148(2), 338-346.

Harris, A., Reeder, R., & Hyun, J. (2011). Survey of editors and reviewers of high-impact psychology journals: statistical and research design problems in submitted manuscripts. The Journal of psychology, 145(3), 195-209.

Herland, M., Khoshgoftaar, T. M., & Wald, R. (2014). A review of data mining using big data in health informatics. Journal of Big data, 1(1), 2.

Hollon, S. D., Jarrett, R. B., Nierenberg, A. A., Thase, M. E., Trivedi, M., & Rush, A. J. (2005). Psychotherapy and medication in the treatment of adult and geriatric depression: which monotherapy or combined treatment?. The Journal of clinical psychiatry.

Hopko, D. R., Bell, J. L., Armento, M., Robertson, S., Mullane, C., Wolf, N., & Lejuez, C. W. (2008). Cognitive-behavior therapy for depressed cancer patients in a medical care setting. Behavior therapy, 39(2), 126-136.

Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. Control Clin Trials 1989; 10: 407–415.

Karyotaki, E., Kemmeren, L., Riper, H., Twisk, J., Hoogendoorn, A., Kleiboer, A., ... & Littlewood, E. (2018). Is self-guided internet-based cognitive behavioural therapy (iCBT) harmful? An individual participant data meta-analysis. Psychological medicine, 48(15), 2456-2466.

Kelley, K., & Preacher, K. J. (2012). On effect size. Psychological Methods, 17(2), 137-152. http://dx.doi.org/10.1037/a0028086 - need for research on effect sizes

Keogh, R. H., & White, I. R. (2014). A toolkit for measurement error correction, with a focus on nutritional epidemiology. Statistics in medicine, 33(12), 2137-2155.

King, M. T. (2011). A point of minimal important difference (MID): a critique of terminology and methods. Expert review of pharmacoeconomics & outcomes research, 11(2), 171-184.

Lang, T. A., & Altman, D. G. (2013). Basic statistical reporting for articles published in biomedical journals: the "Statistical Analyses and Methods in the Published Literature" or the SAMPL Guidelines". Handbook, European Association of Science Editors, 256, 256.

Leichsenring, F., Abbass, A., Hilsenroth, J., Luyten, P., Munder, T., Rabung, S., & Steinert, C. (2018). "Gold standards", plurality and monocultures: the need for diversity in psychotherapy. Frontiers in psychiatry, 9, 159.

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. Journal of the American Statistical Association, 90(431), 1112-1121. DOI: 10.1080/01621459.1995.10476615

Little, R. J., & Rubin, D. B. (2014). Statistical analysis with missing data. John Wiley & Sons.

Liu, Q., & Maxwell, S. E. (2019). Multiplicative treatment effects in randomized pretest-posttest experimental designs. Psychological methods.

Miettunen, J., Nieminen, P., & Isohanni, M. (2002). Statistical methodology in general psychiatric journals. Nordic journal of psychiatry, 56(3), 223-228.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., ... & De Vet, H. C. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research, 19(4), 539-549.

Murray, E., Hekler, E. B., Andersson, G., Collins, L. M., Doherty, A., Hollis, C., ... & Wyatt, J. C. (2016). Evaluating digital health interventions: key questions and approaches.

Nieminen, P., & Kaur, J. (2019). Reporting of data analysis methods in psychiatric journals: Trends from 1996 to 2018. International journal of methods in psychiatric research, e1784-e1784.

Nieminen, P., Virtanen, J. I., & Vähänikkilä, H. (2017). An instrument to assess the statistical intensity of medical research papers. PloS one, 12(10), e0186882.

Ng, V. K., & Cribbie, R. A. (2017). Using the Gamma Generalized Linear Model for modeling continuous, skewed and heteroscedastic outcomes in psychology. Current Psychology, 36(2), 225-235.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. The New England journal of medicine, 375(13), 1216.

Rapee, R. M., Lyneham, H. J., Wuthrich, V., Chatterton, M. L., Hudson, J. L., Kangas, M., & Mihalopoulos, C. (2017). Comparison of stepped care delivery against a single, empirically validated cognitive-behavioral therapy program for youth with anxiety: A randomized clinical trial. Journal of the American Academy of Child & Adolescent Psychiatry, 56(10), 841-848.

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: rationale, conduct, and reporting. Bmj, 340, c221.

Rubin, D. B. (1976), "Inference and Missing Data," Biometrika, 63, 581-592.

Sanders, S. G., & Hunsley, J. (2018). The new Caucus-race: Methodological considerations for meta-analyses of psychotherapy outcome. Canadian Psychology/psychologie canadienne, 59(4), 387.

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. Psychological Methods, 18(4), 572-582.

Safinianaini, N., Boström, H., & Kaldo, V. (2019, June). Gated Hidden Markov Models for Early Prediction of Outcome of Internet-Based Cognitive Behavioral Therapy. In Conference on Artificial Intelligence in Medicine in Europe (pp. 160-169). Springer, Cham.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj, 338, b2393.

Thompson, B. (2002). "Statistical,""practical," and "clinical": How many kinds of significance do counselors need to consider?. Journal of Counseling & Development, 80(1), 64-71.

Appendix B – longitudinal model syntax, exemplifying the use of generalized estimation equation model for predicting post-treatment outcomes

Exemplar Linear models that overlook the role of baseline scores.

* denotes explanation commentary, which are not a part of the syntax codes.

```
* Generalized Estimating Equations.
GENLIN PHQ9_in_long_Stacked_data_format BY Time (ORDER=ASCENDING)  /MODEL Time
INTERCEPT=YES DISTRIBUTION=GAMMA LINK=LOG  /CRITERIA METHOD=FISHER(1)
SCALE=PEARSON MAXITERATIONS=100 MAXSTEPHALVING=5     PCONVERGE=1E-006(ABSOLUTE)
SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95 LIKELIHOOD=FULL  /REPEATED
SUBJECT=id_long WITHINSUBJECT=Time SORT=YES CORRTYPE=UNSTRUCTURED ADJUSTCORR=YES
COVB=ROBUST MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1  /MISSING
CLASSMISSING=EXCLUDE  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION
(EXPONENTIATED) WORKINGCORR  /SAVE MEANPRED(PHQ9_GEE_Linear_Cov).

Compute Resid_PHQ9_GEE_Linear_change = PHQ9_long - PHQ9_GEE_Linear_Cov.

*aggregating the respective model measurement error (residual errors for each
  case).

*Generalized Estimating Equations.
GENLIN GAD7_in_long_Stacked_data_format (ORDER=ASCENDING)  /MODEL Time
INTERCEPT=YES DISTRIBUTION=GAMMA LINK=LOG  /CRITERIA METHOD=FISHER(1)
SCALE=PEARSON MAXITERATIONS=100 MAXSTEPHALVING=5     PCONVERGE=1E-006(ABSOLUTE)
SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95 LIKELIHOOD=FULL  /REPEATED
SUBJECT=id_long WITHINSUBJECT=Time SORT=YES CORRTYPE=UNSTRUCTURED DJUSTCORR=YES
COVB=ROBUST MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1  /MISSING
CLASSMISSING=EXCLUDE  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION
(EXPONENTIATED) WORKINGCORR  /SAVE MEANPRED(GAD7_GEE_Linear_Cov).


Compute Resid_GEE_Linear_change = GAD7_long - GAD7_GEE_Linear_Cov.

*aggregating the respective model measurement error (residual errors for each
  case).

*Generalized Estimating Equations.
GENLIN K10_in_long_Stacked_data_format BY Time (ORDER=ASCENDING)  /MODEL Time
INTERCEPT=YES DISTRIBUTION=GAMMA LINK=LOG  /CRITERIA METHOD=FISHER(1)
SCALE=PEARSON MAXITERATIONS=100 MAXSTEPHALVING=5     PCONVERGE=1E-006(ABSOLUTE)
SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95 LIKELIHOOD=FULL  /REPEATED
SUBJECT=id_long WITHINSUBJECT=Time SORT=YES CORRTYPE=UNSTRUCTURED DJUSTCORR=YES
COVB=ROBUST MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1  /MISSING
CLASSMISSING=EXCLUDE  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION
(EXPONENTIATED) ORKINGCORR  /SAVE MEANPRED(K10_GEE_Linear_Cov).


Compute Resid_K10_GEE_Linear_change = K10_long - K10_GEE_Linear_Cov.

*aggregating the respective model measurement error (residual errors for each case).
```

*Exemplar multiplicative models that make adjustment for baseline scores.

*Generalized Estimating Equations.

GENLIN PHQ9_in_long_Stacked_data_format BY Time (ORDER=ASCENDING) WITH PHQ9_cov

/MODEL Time PHQ9_cov Time*PHQ9_cov INTERCEPT=YES

DISTRIBUTION=GAMMA LINK=LOG

/CRITERIA METHOD=FISHER(1) SCALE=PEARSON MAXITERATIONS=100 MAXSTEPHALVING=5 PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95 LIKELIHOOD=FULL /REPEATED SUBJECT=id_long WITHINSUBJECT=Time SORT=YES CORRTYPE=UNSTRUCTURED DJUSTCORR=YES COVB=ROBUST MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1 /MISSING CLASSMISSING=EXCLUDE /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION (EXPONENTIATED) WORKINGCORR / SAVE MEANPRED(PHQ9_GEE_gamma_Cov).

Compute Resid_PHQ9_GEE_Multiplicative = PHQ9_in_long_Stacked_data_format - PHQ9_GEE_gamma_Cov.

*aggregating the respective model measurement error (residual errors for each case).

*Generalized Estimating Equations.

GENLIN GAD7_in_long_Stacked_data_format (ORDER=ASCENDING) WITH GAD7_cov /MODEL Time GAD7_cov Time*GAD7_cov INTERCEPT=YES DISTRIBUTION=GAMMA LINK=LOG /CRITERIA METHOD=FISHER(1) SCALE=PEARSON MAXITERATIONS=100 MAXSTEPHALVING=5 PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95 LIKELIHOOD=FULL /REPEATED SUBJECT=id_long WITHINSUBJECT=Time SORT=YES CORRTYPE=UNSTRUCTURED ADJUSTCORR=YES COVB=ROBUST MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1 /MISSING CLASSMISSING=EXCLUDE /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION (EXPONENTIATED) WORKINGCORR /SAVE MEANPRED(GAD7_GEE_gamma_Cov).

Compute Resid_GAD7_GEE_Multiplicative = GAD7_in_long_Stacked_data_format - GAD7_GEE_gamma_Cov.

*aggregating the respective model measurement error (residual errors for each case).

*Generalized Estimating Equations.

GENLIN K10_in_long_Stacked_data_format BY Time (ORDER=ASCENDING) /MODEL Time INTERCEPT=YES DISTRIBUTION=GAMMA LINK=LOG /CRITERIA METHOD=FISHER(1) SCALE=PEARSON MAXITERATIONS=100 MAXSTEPHALVING=5 PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95 LIKELIHOOD=FULL /REPEATED SUBJECT=id_long ITHINSUBJECT=Time SORT=YES CORRTYPE=UNSTRUCTURED ADJUSTCORR=YES COVB=ROBUST MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1 /MISSING CLASSMISSING=EXCLUDE /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION (EXPONENTIATED) WORKINGCORR /SAVE MEANPRED(K10_GEE_linear_Cov).

Compute Resid_K10_GEE_linear = K10_in_long_Stacked_data_format - K10_GEE_linear_Cov.

*aggregating the respective model measurement error (residual errors for each case).

*graphical exemples of the residual distribution

```
USE ALL.COMPUTE filter_$=(Time = 2).VARIABLE LABELS filter_$ 'Time = post-Treatment (FILTER)'.VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'. FORMATS filter_$ (f1.0). FILTER BY filter_$. EXECUTE.

* Chart Builder.

GGRAPH

  /GRAPHDATASET NAME="graphdataset" VARIABLES=PHQ9_long Resid_PHQ9_GEE_Loglink_Cov MISSING=LISTWISE     REPORTMISSING=NO  /GRAPHSPEC SOURCE=INLINE.BEGIN GPL  SOURCE: s=userSource(id("graphdataset"))  DATA: PHQ9_long=col(source(s), name("PHQ9_long"))  DATA: Resid_PHQ9_GEE_Loglink_Cov=col(source(s), name("Resid_PHQ9_GEE_Loglink_Cov"))  GUIDE: axis(dim(1), label("PHQ9_long"))  GUIDE: axis(dim(2), label("Resid_PHQ9_GEE_Loglink_Cov"))  ELEMENT: interval(position(PHQ9_long*Resid_PHQ9_GEE_Loglink_Cov), shape.interior(shape.square))END GPL.
```