Behaviour Research and Therapy 116 (2019) 69–79

Contents lists available at ScienceDirect

Behaviour Research and Therapy

journal homepage: www.elsevier.com/locate/brat

Open trial of a personalized modular treatment for mood and anxiety

Aaron J. Fisher[a,*], Hannah G. Bosley[a], Katya C. Fernandez[b], Jonathan W. Reeves[a], Peter D. Soyster[a], Allison E. Diamond[a], Jonathan Barkin[c]

[a] University of California, Berkeley, USA
[b] Center for Creative Leadership, Greensboro, NC, USA
[c] San Francisco Bay Area Center for Cognitive Therapy, USA

ARTICLE INFO

Keywords:
Idiography
N-of-1
Depression
Anxiety
Psychotherapy
Precision medicine
Personalized treatment

ABSTRACT

Psychosocial treatments for mood and anxiety disorders are generally effective, however, a number of treated individuals fail to demonstrate clinically-significant change. Consistent with the decades-old aim to identify 'what works for *whom,*' personalized and precision treatments have become a recent area of interest in medicine and psychology. The present study followed the recommendations of Fisher (2015) to employ a personalized modular model of cognitive-behavioral therapy. Employing the algorithms provided by Fernandez, Fisher, and Chi (2017), the present study collected intensive repeated measures data prior to therapy in order to perform person-specific factor analysis and dynamic factor modeling. The results of these analyses were then used to generated personalized modular treatment plans on a person-by-person basis. Thirty-two participants completed therapy. The average number of sessions was 10.38. Hedges *g's* for the Hamilton Rating Scale for Depression (HRSD) and Hamilton Anxiety Rating Scale (HARS) were 2.33 and 1.62, respectively. The change per unit time was $g = .24$/session for the HRSD and $g = 0.17$/session for the HARS. The current open trial provides promising data in support of personalization, modularization, and idiographic research paradigms.

1. Introduction

Although psychosocial treatments for psychiatric conditions are generally efficacious (c.f. Cuijpers et al., 2013; Cuijpers et al., 2014; Cuijpers, van Straten, Andersson, & van Oppen, 2008), even empirically-supported treatments that are considered to be the gold standard of care for common disorders often fail to produce desired improvement. For example, a meta-analysis of treatment studies by Hofmann and Smits (2008) showed that for generalized anxiety disorder (GAD) – one of the more prevalent psychiatric problems (Wittchen, 2002) – the odds of responding to cognitive-behavioral interventions are not significantly different from the odds of responding to a placebo. Historically, the fields of psychology and psychiatry have primarily focused on the 'what' component of Paul's (1967) famous question – "*What treatment, by whom, is most effective for this individual with that specific problem, and under which set of circumstances?*" – conducting 'horse race' controlled trials (CTs) and randomized controlled trials (RCTs) that pit one macroscopic intervention against another. Lost is the potential granularity of the specific problems, sets of circumstances and, perhaps most importantly, the individual. However, more recently there has been an explosion of interest in the information at the individual level, and a bourgeoning interest in personalized and precision interventions.

Personalized and precision medicine have become prominent areas of interest in recent years. Former United States President, Barack Obama announced a precision medicine initiative in 2015 (Ashley, 2015), the European Union established the European Alliance for Personalized Medicine in 2012 (Dzau, Ginsburg, Van Nuys, Agus, & Goldman, 2015), and both the Food and Drug Administration and the National Institutes of Health (NIH) have made calls for increased attention to personal-level predictors of morbidity and treatment response (Hamburg & Collins, 2010). Within the NIH, institutes such as the National Institute of Mental Health (NIMH) have included personalization as central parts of their strategic plans (Insel, 2009). In addition, calls for personalized and person-level research paradigms have recently been found in Science (Iyengar, Altman, Troyanskya, & FitzGerald, 2015), Nature (Schork, 2015), the Lancet (Dzau et al., 2015), and the Journal of the American Medical Association (Ashley, 2015; Khoury & Evans, 2015), to name a few. This explosion of interest in precision medicine has been proposed to be the result of biotechnological advances and cost reductions in genomic sequencing (Khoury & Evans, 2015), because "biomedical technology now allows a deeper understanding of many diseases," (Ashley, 2015, p. 2119).

* Corresponding author.
E-mail address: afisher@berkeley.edu (A.J. Fisher).

https://doi.org/10.1016/j.brat.2019.01.010
Received 1 May 2018; Received in revised form 5 January 2019; Accepted 28 January 2019
Available online 21 February 2019
0005-7967/

Yet, it should be noted that biomedical technology, per se, is not required for a deeper or more complex understanding of many diseases and syndromes. This is likely especially true for psychiatric conditions, for which the degree of genomic contribution remains unclear (Iyengar et al., 2015). Whereas behavioral genetics have demonstrated that psychiatric disorders are heritable constructs, identifying their genetic sources through linkage analysis or genome-wide association studies has not yielded actionable results to date (Burmeister, McInnis, & Zöllner, 2008). We argue that the behavioral level of analysis remains a potent and fertile avenue for understanding, defining, measuring, and *deepening* our descriptions and predictions of human phenomenology and behavior. Although it has been argued that psychiatric disorders are *brain* disorders (Insel & Cuthbert, 2015), they nevertheless manifest as behavioral syndromes. Examining the mood and anxiety disorders of the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5; American Psychiatric Association, 2013) reveals clinical criteria predominantly related to thoughts, actions, feelings, and interpersonal functioning. Moreover, the combinatory nature of the Kraepelinian system in the DSM-5 produces high degrees of heterogeneity in the specific componentry of a given diagnosis from person to person (Fisher, 2015; Galatzer-Levy & Bryant, 2013). Given the overwhelming degree of heterogeneity – including 636,020 different combinations for a diagnosis of posttraumatic stress disorder – it is reasonable to expect that additional specificity and quantitative predictability would be found by utilizing research paradigms that quantify individuals (i.e. idiography).

Calls for idiographic research are not new (c.f. Lamiell, 1981), and have been substantiated on theoretical (Mischel, 1973), mathematical (Molenaar, 2004), and empirical (Fisher, Medaglia, & Jeronimus, 2018) grounds. To the latter, Fisher and colleagues examined data from six independent sites in the United States and the Netherlands and found that the covariance among aggregated (i.e. nomothetic) data did not reflect the covariance of the constituent individuals in each data set. This result was consistent for psychiatric, psychophysiological, and general population survey data. Consistent with this, Fisher, Reeves, Lawyer, Medaglia, and Rubel (2017) examined the contemporaneous and time-lagged network structure of 40 participants with DSM-5 mood and anxiety disorders on a person-by-person basis. These authors found a high degree of heterogeneity in individual network structures, temporal dynamics, and centrality – the degree to which one symptom influences others.

As other have observed (Molenaar, 2004), the aggregation of data across participants obscures information about the constituent individuals in the sample. Patterns of covariation at baseline, response trajectories during treatment, and attempts to determine 'what works for whom' (Paul, 1967), are undermined by the statistical limitation that aggregated analyses reveal only what can be known about the group and do not reflect or reveal individual outcomes. For this reason, there have been a number of calls for idiographic methods (Fisher, Newman, & Molenaar, 2011) and N-of-1 studies (Barlow & Nock, 2009; Schork, 2015). Calling traditional, nomothetic clinical trials 'imprecision medicine,' Schork (2015) has argued that single person treatment studies will be vital to understanding which individuals respond to specific interventions. Aggregating large numbers of N-of-1 trials, he argues, will provide a better means with which to identify subpopulations and match treatments to individuals. However, few authors have put forward paradigms for carrying out such work.

Fisher (2015) recently proposed such a paradigm, providing a framework by which researchers can measure individual-level psychiatric symptoms on a moment-to-moment basis, and leverage these measurements to reveal idiosyncratic syndrome structures and dynamics. Specifically, Fisher proposed that researchers should endeavor to (a) collect intensive repeated measures data *in vivo*, (b) factor analyze the data to reveal idiographic syndrome structures, and (c) employ dynamic analyses to identify the influence of symptom structures on a moment-to-moment basis. One advantage of the proposed methodology is that it does not require or presuppose formal diagnostic classifications such as those provided by the DSM-5. Instead, syndromes are generated via idiographic analysis, with the variance-covariance among each individual's symptoms dictating the constructs in play. However, Fisher (2015) did not provided a manner by which these methods would be applied to the construction and delivery of personalized treatments. For this reason, Fernandez, Fisher, and Chi (2017) developed systematic procedures for translating person-specific data analyses to personalized treatment delivery. These authors created methodologies based on clinician-based and machine-based selection paradigms, for use with idiographic factor analysis and dynamic factor modeling (Molenaar, 1985). This approach was recently extended to be applied to network modeling approaches as well (Rubel, Fisher, Husen, & Lutz, In Press).

The current study followed the paradigm established by Fisher (2015) to conduct a trial of personalized, modular cognitive-behavioral therapy (CBT). To that end, we collected data multiple times per day for ~30 days in order to generate person-specific contemporaneous factor models and time-lagged dynamic factor models. We then employed the methods provided by Fernandez et al. (2017) to construct personalized treatment plans on a person-by-person basis. Consistent with the proposal by Schork (2015), we conceptualized the current study as an aggregation of multiple N-of-1 studies. To that end, in the sections to follow, we provide both idiographic and aggregated outcome data. The present study utilized the Unified Protocol for Transdiagnostic Treatment of Emotional Disorders (UP; Barlow et al., 2011) because of its modular design and demonstrated efficacy across anxiety and mood disorders (e.g., Boswell, 2013). We recruited individuals with presenting symptoms consistent with DSM-5 GAD and major depressive disorder (MDD). These putative disorder categories are noteworthy for their high degree of co-occurrence and shared diagnostic features (Kessler, Chiu, Demler, & Walters, 2005). They have been proposed to derive from a shared, underlying source of neuroticism (Barlow, Sauer-Zavala, Carl, Bullis, & Ellard, 2014) or negative affectivity (Brown, Chorpita, & Barlow, 1998), and some have argued that they should be classified together as disorders of distress (Watson, 2005). Consistent with Fernandez et al. (2017), treatment modules were linked to presenting symptoms via a logical matrix. Thus, treatment planning was not based on putative disorder categories, but on the idiosyncratic presentation of symptoms for each individual.

We hypothesized that the personalized nature of the study design would yield two observable benefits, an efficacious treatment response and a more efficient treatment delivery. That is, we believed that the treatment would promote clinically-significant and reliable change in symptom severity, and that these changes could take place over as few sessions as possible. Thus, the term efficiency is meant to connote a comparable to superior degree of symptom change over a smaller number of sessions.

Because the current study was an uncontrolled open trial with no comparison condition, it was not possible to control for therapeutic factors that are common to all psychotherapies (such as expectancy), nor would it be possible to conclusively demonstrate the efficacy of the proposed model above and beyond an established gold-standard treatment. Thus, it should be emphasized at the outset that any inferences drawn from these data are strictly preliminary.

Nevertheless, it may be possible to estimate the efficacy and efficiency of the intervention via available benchmarks – to the degree that a given benchmark reflects the underlying psychopathology and treatment approach employed in the present study. A recent meta-analysis by Johnsen and Friborg (2015) documented all available, English-language CTs and RCTs that used the Hamilton Rating Scale for Depression (HRSD) as a primary outcome measure, and reported the average effect and number of sessions across studies. Although ostensibly a depression measure, the HRSD reflects the shared topology of GAD and MDD, and can be employed to measure mixed anxious-depressive distress. To wit, of the 17 items in the HRSD, four have direct overlap with the clinical

criteria for GAD (sleep disturbance, restlessness, apprehension and worry, and fatigue), and an additional symptom, somatic anxiety, that reflects adrenergically-mediated fear and arousal. Moreover, although the benchmark provided by Johnsen and Friborg may be considered exclusive to depression, it should be noted that (a) these authors used the more general classification of depressive disorder, rather than MDD, and included studies that classified depression based on HRSD and Beck Depression Inventory scores, and (b) multiple studies in the meta-analysis included additional primary and secondary comorbidities. Finally, effect sizes for diagnostically heterogeneous and comorbid populations tend to be lower than homogeneous depressed samples (Johnsen & Friborg, 2015). Thus, using the HRSD benchmark provided by these authors is likely a conservative comparison for the transdiagnostic sample utilized in the present study.

Across 31 CTs and RCTs, the average pre-post Hedge's *g* for the HRSD was 1.72, over an average of 14 sessions (Johnsen & Friborg, 2015). We hypothesized that our personalized modular therapy would demonstrate equivalent to superior overall effectiveness (i.e. a $g \geq 1.72$). Moreover, we hypothesized that our personalization would generate superior gains per unit time. Thus, we hypothesized that we would return a *g*/session greater than 0.12. Finally, we measured participant progress at six-months following the conclusion of treatment. We hypothesized that post-treatment gains would be maintained over this period.

2. Method

2.1. Participants

Individuals with symptomatic experiences consistent with possible diagnoses of GAD and MDD were recruited from the greater Bay Area community via flyers, referrals, and Internet advertisements. After passing a brief telephone screening interview, 174 potential participants were invited to in-person appointment at the first author's laboratory at the University of California, Berkeley during which they completed a structured clinical interview for diagnostic assessment. Inclusion criteria were primary diagnosis of GAD or MDD, age of 18–65 years, and a web-enabled mobile phone. In addition, given that the main outcome measure for the current study was the HRSD, a minimum score of 7 on the HRSD was required for inclusion, as this reflects the minimum clinically-significant change (Cusin, Yang, Yeung, & Fava, 2009; Frank et al., 1991). One participant who met clinical criteria for GAD was excluded due to a subthreshold score on the HRSD (P113).

Exclusion criteria were any history of psychosis or mania, concurrent treatment or cognitive–behavioral treatment within the past 12 months, and as-needed medication. A total of 57 individuals (33%) met inclusion criteria for the current study. Of these, 40 began treatment after three declined to participate, 10 withdrew or provided insufficient data during the survey period, and an additional four participants withdrew after completing surveys, but before initiating therapy. Seven participants withdrew from the study during treatment, and one participant failed to complete a post-treatment assessment, leaving 32 participants who completed a full course of treatment and post-treatment assessment. Fig. 1 provides the CONSORT diagram,including samples sizes for each stage of the study.

Interrater reliability was calculated based on video recordings of the structured clinical interviews. Two videos (Participants 7 and 37) were lost in the course of the study—one overwritten and one not adequately captured—leaving 55 participants available for reliability ratings. Following the recommendation of McHugh (2012), we calculated both percent agreement and Cohen's kappa to rate diagnostic interrater reliability. The inclusion criteria, GAD and MDD, returned kappa values of 0.68 and 0.84, and percent agreement of 95% and 92%, respectively, with two mismatches for GAD and three mismatches for MDD. Percent agreement and kappa values for secondary diagnoses were SAD (92%, $\kappa = 0.79$), specific phobia (92%, $\kappa = 0.63$), panic disorder (95%, $\kappa = 0.64$), agoraphobia (97%, $\kappa = 0.79$), posttraumatic stress disorder (97%, $\kappa = 0.84$), and alcohol use disorder (100%, $\kappa = 1$).

Of the 32 treatment completers, 14 met for a current primary diagnosis of GAD, 10 met for a current primary diagnosis of MDD, and 8 met for co-primary diagnoses of both GAD and MDD. 16 of these participants met for at least one current comorbid disorder other than GAD or MDD: Social anxiety disorder (SAD), $n = 12$; specific phobia, $n = 4$; panic disorder, $n = 1$; agoraphobia, $n = 3$; persistent depressive disorder, $n = 3$; post-traumatic stress disorder, $n = 2$. The mean overall clinician ratings for the sample of treatment completers on the HRSD and the Hamilton Anxiety Rating Scale (HARS) were 13.81 (*SD* 3.78) and 15.81 (*SD* 6.94), respectively. For this group, mean self-reported ratings of depression and anxiety symptoms via the DASS-D and DASS-A were 22.81 (SD 9.22) and 12.88 (SD 7.68) respectively.

2.2. Measures

2.2.1. Anxiety and related disorders interview schedule for DSM–5 (ADIS-5; Brown & Barlow, 2014)

The ADIS-5 is a semi-structured clinical interview that is designed to diagnose current anxiety, mood, and related disorders according to new DSM-5 criteria. This updated version of the ADIS-5 builds upon previous versions (the ADIS, ADIS-R, and ADIS-IV for DSM-III, DSM−III−R, and DSM-IV, respectively), which had well-established reliability. The ADIS-IV demonstrates good-to-excellent interrater reliability for DSM-IV disorders (*kappa* ranging from 0.67 to 0.86, with the exception of dysthymia, *kappa* = .31).

2.2.2. Hamilton Anxiety Rating Scale (HARS; Hamilton, 1959)

The HARS assesses severity of anxious symptomatology. This 14-item clinician administered scale provides a severity rating of each overarching symptom cluster on a scale from 0 (*not present*) to 4 (*very severe*). Internal consistency is excellent (0.92; Kobak, Reynolds, & Greist, 1993). Retest reliability for the HARS was very good (intraclass correlation coefficient 0.86) across 2 days and interrater reliability ranged from an intraclass correlation coefficient of 0.74–0.96 (Bruss, Gruenberg, Goldstein, & Barber, 1994). Construct validity has also been demonstrated in clinical samples (Beck & Steer, 1991).

2.2.3. Hamilton Rating Scale for Depression (HRSD; Hamilton, 1960)

The HRSD was developed to assess the severity of depressive symptomatology. This 13-item clinician administered scale provides a rating of severity of each overarching symptom cluster on a scale from 0 (*not present*) to 4 (*very severe/incapacitating*). Internal consistency of the HRSD ranges from adequate to good (0.73–0.81; Moras, Di Nardo, & Barlow, 1992; Steer, Beck, Riskind, & Brown, 1987). Interrater reliabilities of the HRSD total score range from 0.78 to 0.82 (Moras et al., 1992; Steer et al., 1987). HRSD scores correlate significantly with self-report measures of depression in clinical samples (Steer, McElroy, & Beck, 1983).

2.2.4. Experience sampling survey

For each survey, participants rated their experience of each item over the preceding hours using a 0–100 visual analog slider with the anchors *not at all* and *as much as possible* for the 0 and 100 positions, respectively. Surveys contained the extant symptoms of the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* criteria for GAD and MDD (*down and depressed, hopeless, loss of interest or pleasure, worthless or guilty, worried, restless, irritable, difficulty concentrating, muscle tension, fatigued*), as well as an additional 11 items measuring positive affect (*positive, energetic, enthusiastic, and content*), negative affect (*angry* and *afraid*), rumination (*dwelled on the past*), behavioral avoidance (*avoided people, avoided activities, and procrastinated*), and reassurance seeking (*sought reassurance*).

2.3. Procedures

2.3.1. Clinical interview

Interested individuals contacted the Idiographic Dynamics Laboratory at the University of California, Berkeley, and trained research assistants administered a brief telephone screening interview to assess study eligibility. Those who passed the initial screening interview were invited to the laboratory for an in-person structured clinical interview. The ADIS-5 was administered by graduate students in clinical psychology supervised by a doctoral level clinical psychologist. At this appointment, clinician ratings of symptom severity were obtained via the HRSD and HARS, and participants also completed a battery of self-report symptom measures including the DASS.

2.3.2. Ecological momentary assessment (EMA)

After enrolling in the study, participants' mobile phone numbers were entered into a secure web-based survey system. The survey system prompted participants to answer survey questions four times per day during waking hours (tailored to participants' self-reported wake up times); participants received surveys approximately every 4 h, the exact time of the ping was randomized within a 30-min window. Each survey prompt was received as a text message containing a hyperlink to a web-based survey. Every time pings were sent to participants, the back-end of the system recorded a time stamp, whether the participant completed the survey or not. Participants completed surveys for a minimum of 30 days. The mean number of observations across the 44 participants who completed the survey period was 110.7 (SD = 11.2), with a minimum of 87 and a maximum of 140. For the sample of 32 treatment completers, the mean number of observations was 111.2 (SD = 11.3).

2.3.3. Data preparation and analysis

Data collection during the EMA period yielded a multivariate time series for each participant, which were each then subjected to P-technique factor analysis and dynamic factor modeling as delineated in Fisher and Boswell (2016).

2.3.4. P-technique factor analysis

A two-stage approach was taken wherein we first used exploratory factor analysis via the Psych package in R (Revelle, 2013) to reveal the latent structure in each person's symptom experience. An iterative approach for determining the number of factors was employed whereby a one-factor model was initially tested and assessed for model fit, followed by two-, three-, and four-factor models (when indicated). A final model was retained when an acceptable fit was indicated by the (a) chi-square goodness-of-fit statistic, (b) root mean square error of approximation and (c) standardized root mean square residual (for a review of these criteria and their respective optimal cutoffs, see Hu & Bentler, 1999). Factor loadings less than |0.30| were removed before generating factor scores.

2.3.5. Dynamic factor modeling

Once the factor solution was confirmed, the matrix of raw input data was multiplied by a weighting matrix of factor loadings from the confirmatory solution. Because the data were unevenly sampled due to the overnight interval (with intervals of ~4, ~4, ~4, and ~12 h) a cubic spline interpolation was applied to the factor-scored time series, yielding a time series with evenly sampled 6-h intervals. These interpolated time series were then each duplicated and lagged by one observation, resulting in a set of time-lagged factors (time $t - 1$) and time-forward factors (time t). FIML estimation was used to analyze the raw data in LISREL. For each individual, an initial model was run with all contemporaneous factor correlations and autoregressions. Cross-lagged parameters were then revealed on an iterative basis using the Lagrange multiplier test (Chou & Bentler, 1990).

2.3.6. Module selection

Compared to diagnosis-specific treatment manuals, transdiagnostic protocols offer benefits such as targeting common underlying dimensions of mood and anxiety disorders, as well as promoting flexibility and personalization in treatment delivery (Barlow et al., 2014; Brown & Barlow, 2009). In the present study, we chose to utilize the UP (Barlow et al., 2011) because of its modular design and established efficacy across anxiety and mood disorders (e.g. Boswell, 2013). To maximize the UP's potential for personalization, in the current study we treated the UP as a "menu" of treatment modules, each designed to treat a targeted subset of symptom domains. Only modules that addressed the predominant symptom dimensions within each individual were selected for treatment delivery and therapeutic elements deemed to be unnecessary or irrelevant to the client's symptom structure were removed. In addition, the order in which modules were delivered was determined by the dynamic relationships identified by person-specific dynamic factor models. That is, interventions for symptoms shown to drive the behavior of other symptoms were preferentially delivered earlier in therapy.

As described in detail in Fisher and Boswell (2016) and Fernandez, Fisher, & Chi (2017), we utilized a two-phase approach to the development of a treatment selection algorithm. It should be noted at the outset that the original intent of this approach was to establish a set of treatment selection procedures under manual, human-delivered conditions (phase 1) that could then be digitized and automated (phase 2). Thus, the explicit assumption was that the automation would simply codify and standardize the treatment selection process, and neither increase nor decrease treatment performance. However, the two procedures provided differential performance across participants (see Results). Further attention is given to this issue in the Discussion.

2.3.7. Expert panel treatment selection

In the first phase, an expert panel of clinicians comprising the principal investigator (first author), a postdoctoral researcher (third author), a postdoctoral clinician, and practicing clinician from the Bay Area community (seventh author) met to review the results of the P-technique and dynamic models for each participant. A personalized treatment plan was then developed on the basis of symptom predominance, as established via the explanatory power of identified factors, both within (P-technique) and across time (dynamic modeling). The factors accounting for the largest share of the variance in a person's symptom structure were considered to best represent the latent structure and organization of the individual's presenting problems. Considering the cross-lagged effects as identified by the dynamic models, factors that exhibited predictive time-lagged effects on other factors were thought to represent the greatest potential for intervention because of their downstream effect on other symptoms later in time. Thus, these factors were given preferential consideration.

2.3.8. Automated selection

In phase 2, our group developed the Dynamic Assessment and Treatment Algorithm (DATA; Fernandez et al., 2017), in order to quantify the contribution of each individual symptom or behavior as a function of the factor to which each symptom belonged. For comprehensive details of this algorithm, we refer readers to Fernandez et al. (2017). Briefly, a set of weighting equations were created to (a) calculate the predominance among factors within time as reflected by the percent variance accounted for among presenting symptoms and behaviors, (b) the predominance of factors across time as reflected by the percentage of predictive (lagged regression) variance accounted for by each factor, and (c) assign an item score for each symptom or behavior based on the relative strength of association (i.e., standardized factor loading) between the item and its associated factor. The final step (d) weighted the item scores by their mean levels over the pretherapy period. DATA assigns a factor score to each factor derived from Steps (a) and (b) and then an item score to each item based on Steps (c) and

(d). Module selection and ordering was then based on a logical matrix, matching modules to items—with the highest scoring module delivered first, followed by successively lower scoring modules.

2.3.9. Allocation to automated versus expert panel selection

Patients were allocated sequentially and no randomization was employed. The first 20 participants to complete pre-therapy data collection were designated to receive expert-panel selections. Of these, one (participant 2) provided insufficient data. Thus, the first 19 treatment initiators received expert-panel treatment selections. The next 15 participants were designated to receive DATA-based treatment selections, however, one of these participants (participant 168) was excluded due to a history of psychosis. Thus, 14 participants were allocated to receive DATA-based treatment selections. The remaining participants (participants 202–244) were assigned treatment plans based on expert-panel selections.

2.3.10. Treatment length

Treatment length was set by the selected modules and their respective session recommendations, per the UP. That is, the number and order of modules was determined by algorithmic selection, as described above, and the number of sessions corresponded to the recommended allocation of sessions per module. Modules 1, 3, 4, 6, 7, and 8 were each accompanied by a single chapter in the UP patient workbook and took a minimum of one session. Modules 2 and 5 were each accompanied by two chapters in the patient workbook and required at least two sessions. Therapists were allowed to administer additional sessions with permission from the first author, however, they were not permitted to shorten the length of treatment (e.g. condense modules 2 or 5 into a single session). The first module delivered was the most often extended, as acclimatizing and introductory material (e.g. explanation of the cognitive-behavioral model, goal-setting, patient's statement of presenting problems) was allowed to take up to (but not exceeding) one complete session.

2.3.11. Treatment delivery

Once a treatment plan was determined, treatment was administered by the treatment team described above. Clients presented to the lab on a weekly basis for 50-min treatment sessions. In the initial session, clients' idiosyncratic results from the P-technique and dynamic factor modeling analyses were presented and explained by the therapist, and the resulting treatment plan was discussed with the client. For subsequent sessions, treatment modules of the UP were delivered in the order prescribed by the client's personalized treatment plan. Treatment was delivered by doctoral level clinical psychologists and advanced graduate students in clinical psychology supervised by doctoral-level clinicians.

2.3.12. Follow-up assessments

Within days of completing treatment, participants were invited back to the lab for a follow-up assessment during which 1) they completed a self-report measures and 2) trained graduate students again administered a diagnostic structured clinical interview and provided clinician ratings via the HRSD and HARS to assess change in diagnosis and symptom experience. Six months following the date of treatment completion, participants were invited to one final follow-up appointment during which the same assessment battery was administered.

3. Results

3.1. Baseline characteristics

Table 1 presents the baseline characteristics for all 57 participants meeting inclusion criteria. No baseline differences were found in HRSD scores between those who initiated therapy (N = 40) and those who dropped out before therapy began (N = 17; *beta* = −1.06, *SE* = 1.40, $t = -0.76$, $p = .45$), nor between those who completed (N = 32) and dropped out (N = 8) of therapy (*beta* = −2.19, *SE* = 1.58, $t = -1.38$, $p = .18$). Likewise, there were no significant differences between those who initiated therapy and those who dropped out before therapy (*beta* = −0.25, *SE* = 1.90, $t = -0.13$, $p = .90$), nor between treatment completers and dropouts (*beta* = −2.81, *SE* = 2.70, $t = -1.04$, $p = .31$) on the HARS.

3.2. P-technique, dynamic factor models, and module selection

The complete P-technique results for all 44 individuals who completed surveys are available online at the Open Science Framework at: https://osf.io/8vjcq/. Likewise, the 44 dynamic factor models for these individuals. Consistent with Fisher et al. (2017), the lagged regression (i.e. beta) matrix from each dynamic factor model was run through *qgraph* (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012) package in R, in order to visualize the models as networks. In this format, autoregressions are presented as self-loops; positive relationships are represented by green lines, and negative relationships are represented by red lines. The relative strength of relationship is reflected in the thickness of each line.

Table 2 presents the module orders for all 40 participants who initiated therapy (dropouts are denoted by *). There was no difference between the number of modules assigned to treatment completers and dropouts (*beta* = −0.38, *SE* = 0.30, $t = -1.25$, $p = .22$). The minimum number of modules assigned was four, the maximum seven, the median and mode were six modules. As can be seen from Table 2, module order exhibited a great deal of heterogeneity from participant to participant. The most common initial module was module 2, related to psychoeducation and tracking of emotional experiences. This module was prescribed first for 16 of the 32 treatment completers and five of the eight dropouts.

3.3. Treatment outcomes

Hedges's *g* effects were calculated for the change from pre-to post-therapy in order to be comparable to results from Johnsen and Friborg (2015). Hedge's *g* and Cohen's *d* differ only in the method employed for pooling standard deviations (utilizing *n-1* versus *n*, respectively). Two sets of analyses were carried out, (1) a completer analysis for those who completed therapy and were evaluated via post-treatment assessment, and (2) an intent-to-treat analysis, including all participants who progressed to the treatment stage of the study. Intent-to-treat (ITT) has been recommended as standard practice for the analysis of RCT data, in order to maintain randomization when comparing treatment arms (Yelland et al., 2015). However, participants with missing outcome data can be excluded when such exclusions do not affect randomization (Fergusson, Aaron, Guyatt, & Hébert, 2002). Given the open trial design employed in the present study, ITT is unnecessary for maintaining randomization and likely to lead to overly-conservative estimates of treatment effect (Abraha et al., 2015). Nevertheless, it has been recommended that clinical trials provide separate reports for participants with complete and incomplete data (Alshurafa et al., 2012). Thus, in the following, we report results for the 32 treatment completers and the 40 individuals who progressed from the assessment stage of the study to the treatment stage. For ITT analyses, the baseline measurement was carried forward as a post-treatment score, yielding a Hedge's *g* of zero.

3.3.1. Completer analysis

All raw baseline and post-treatment data for the HRSD and HARS are provided in Table 3. The average change in depression was 8.03 points (*SD* = 3.60) and the average change in anxiety was 9.22 points (*SD* = 6.35). The average Hedge's *g's* for depression and anxiety were −2.33 (95% CI [-1.97, −2.73]) and −1.62 (95% CI [-1.18, −2.05]), respectively. The average number of sessions delivered in the study was 10.38 (min = 4, max = 14, mode = 9), and the average Hedge's *g* per

Table 1
Participant baseline characteristics.

ID	Sex	Age	Ethnicity	Primary Diagnoses	Comorbidities	HRSD	HARS
001	Female	28	Latino	MDD, GAD	Panic	23	27
002	Male	43	Black	MDD	SAD, GAD	10	13
003	Male	29	White	MDD, GAD		16	15
004	Female	32	Latino	GAD		16	33
006	Male	26	White	MDD, GAD	SAD	13	13
007	Female	33	Black	MDD, GAD	Agor, SAD, Spec Phob	11	17
008	Female	23	Asian American	MDD, GAD	PTSD, Body Dys	19	15
009	Female	25	Other	GAD, SAD	Spec Phob	17	9
010	Male	33	Asian American	MDD, GAD	SAD	22	22
012	Female	36	Latino	GAD, Agor		9	13
013	Male	26	White	MDD, GAD	SAD	14	19
014	Male	22	Latino	MDD		10	12
019	Female	30	Asian American	MDD	SAD	10	10
021	Male	59	Other	GAD	SAD	15	16
023	Female	64	White	GAD		8	7
025	Male	31	White	GAD, SAD		15	14
030	Male	35	White	MDD	Tobacco Use	9	9
032	Male	62	White	GAD	Spec Phob	12	18
033	Female	28	White	GAD	Agor, SAD, OCD	8	14
037	Female	28	Latino	GAD, SAD	Illness Anxiety, Spec Phob	12	23
040	Female	29	White	GAD	Agor, SAD, MDD, Spec Phob	21	41
048	Male	57	Asian American	MDD, GAD	SAD, Spec Phob	14	17
068	Female	42	White	GAD		11	14
072	Female	38	Asian American	MDD	GAD	15	13
074	Female	56	White	MDD		12	10
075	Female	27	Asian American	GAD		18	23
100	Male	31	White	GAD	PTSD	7	14
111	Female	23	Asian American	GAD	Panic, SAD, PTSD	18	15
115	Female	42	White	MDD, GAD	SAD	18	19
117	Male	59	White	MDD, GAD		12	18
127	Male	29	Latino	GAD	SAD	9	13
137	Male	45	Asian American	MDD		16	15
138	Female	24	Asian American	GAD	Spec Phob	4	12
139	Female	62	White	MDD	GAD	14	12
145	Female	47	Other	GAD	SAD, PTSD	21	30
160	Male	50	White	GAD	PDD	13	11
163	Female	58	Asian American	MDD	GAD, PDD	16	16
169	Male	29	White	MDD		13	15
176	Female	26	Latino	GAD	MDD, Spec Phob, Panic, PTSD	19	26
198	Female	38	Asian American	GAD	SAD, MDD, PDD	11	11
200	Male	28	Asian American	MDD	GAD, Alcohol Use, Spec Phob	21	16
201	Male	51	Latino	MDD, SAD		28	22
202	Female	34	White	GAD	PDD	10	11
203	Female	21	Asian American	MDD, GAD	SAD	18	20
204	Female	57	White	GAD		12	16
206	Female	39	Other	GAD, SAD		11	16
214	Female	26	Black	MDD	PDD	15	15
215	Female	31	Black	GAD		17	23
217	Female	31	White	GAD	MDD	17	14
219	Female	23	Asian American	GAD	MDD	21	27
220	Male	64	White	MDD	GAD	14	13
223	Male	56	White	MDD	GAD	21	12
234	Female	64	White	MDD, GAD	SAD	21	21
237	Male	NA	Asian American	GAD, SAD	Panic	11	17
238	Female	20	Other	MDD	GAD	14	14
242	Male	24	Latino	GAD	PDD	18	15
244	Female	21	Asian American	MDD		12	8

Agor = agoraphobia; GAD = generalized anxiety disorder; MDD = major depressive disorder; PDD = persistent depressive disorder; PTSD = posttraumatic stress disorder; SAD = social anxiety disorder; Spec Phob = specific phobia.

unit time was 0.24 for the HRSD and 0.17 for the HARS. Thus, the present study returned a Hedge's *g* 35% greater than the meta-analytic average provided by Johnsen and Friborg ($g = 1.72$). Moreover, the effect per unit time of $g = 0.24$/session was exactly twice the effect per unit time reported for the 31 CT and RCT studies.

3.3.2. Intent to treat analysis

Analysis of the ITT sample returned an average change in depression of 6.43 points ($SD = 4.57$) and an average change in anxiety of 7.38 points ($SD = 6.78$). The average Hedge's *g's* for depression and anxiety were −1.86 (95% CI [-1.48, −2.24]) and −1.29 (95% CI [-0.90, −1.69]), respectively. Thus, under ITT assumptions the present study returned a Hedge's *g* only 8% greater than the Johnsen and Friborg benchmark. No effects per unit time could be calculated for those with missing treatment data.

3.4. Six-month follow-up

Twenty-one of the 32 treatment completers presented for follow-up assessments six months after completing treatment. No significant differences were observed between the 21 who completed follow-up assessments and 11 who did not in treatment response of the HRSD

Table 2
Module orders for all 40 treatment initiators. Treatment dropouts denoted with *.

ID	Module Order						
3	2	3	4	6	8		
4	2	7	5	4	8		
6	2	3	5	4	7	6	8
7	1	2	3	7	4	5	8
8*	2	5	4	7	6		
9	2	3	4	5	8		
10*	2	3	5	7	4	8	
12	5	2	3	4	8		
13	2	3	5	7	4	8	
14	1	2	3	4	5	8	
19	2	3	4	6	8		
21*	2	3	4	6			
23	2	3	4	5	8		
25*	2	3	4	5	7		
33*	6	2	3	4	8		
37*	1	2	3	4	5	7	
40	2	3	4	6	5	7	8
48	4	2	3	5	7	6	8
68	2	3	4	6	8		
72	2	3	4	5	7	8	
74	3	2	4	5	8		
75	1	2	4	3	5	7	6
100	4	3	2	5			
111	2	6	3	4	5	7	
115	2	4	3	5	6		
117	3	2	4	6	5	7	
127	2	4	3	5	7	6	
137*	5	3	7	2	6	4	
139	2	4	5	3	7		
145*	2	4	3	5	6	7	
160	2	5	7	3	4	6	
163	5	7	3	2	4	6	
169	3	2	4	5	7	6	
202	1	4	2	3	5	7	
203	4	2	3	6	5	7	
206	4	2	3	6	5	7	
219	1	2	3	4	6	5	
220	5	7	3	2	4	8	
223	5	6	7	4	2	3	
244	2	3	4	5	8		

Note: Module 1 = Motivation Enhancement for Treatment Engagement; Module 2 = Psychoeducation and Tracking of Emotional Experiences; Module 3 = Emotion Awareness Training; Module 4 = Cognitive Appraisal and Reappraisal; Module 5 = Emotion Avoidance and Emotion-Driven Behaviors; Module 6 = Awareness and Tolerance of Physical Sensations; Module 7 = Interoceptive and Situation-Based Emotion Exposures; Module 8 = Relapse Prevention.

(*beta* = −0.46, *SE* = 1.36, *t* = −0.34, *p* = .74) or HARS (*beta* = 0.91, *SE* = 2.40, *t* = 0.38, *p* = .71). Hedges's *g* effects were calculated for the change from pre-treatment to six-month follow-up. The average change in depression was 8.29 points (*SD* = 6.44) and the average change in anxiety was 9.19 points (*SD* = 6.64). The average Hedge's *g's* for depression and anxiety were −1.91 (*SD* = 1.57) and −1.34 (*SD* = 1.03), respectively. The average Hedge's *g* per unit time was 0.21 for the HRSD and 0.14 for the HARS.

3.5. Linear mixed-effect model of treatment response

We also examined the aggregated response trajectory via linear mixed-effect regression, in order to examine the shape of change in the sample over the treatment and follow-up period. We used a piecewise model in order to generate separate coefficients for the treatment and follow-up periods. Time was scaled in months.

3.5.1. Completer analysis

For treatment completers, both the HRSD (*beta* = −2.15, *SE* = 0.19, *t* = −11.11, *p* < .001) and HARS (*beta* = −2.43, *SE* = 0.29, *t* = −8.36, *p* < .001), exhibited a significant negative slope during the treatment period and non-significant slopes during follow-up (*beta* = −0.03, *SE* = 0.11, *t* = −0.32, *p* = .80; and *beta* = −0.04 *SE* = 0.14, *t* = −0.26, *p* = .80, respectively), indicating that – as a group – our sample exhibited significant decreases in anxious and depressive symptoms that were maintained over six months. Fig. 2 presents the raw data and predicted response trajectories for the HRSD and HARS over the study period.

3.5.2. Intent-to-treat analysis

Analyses of the ITT sample returned consistent, albeit somewhat attenuated results. Both the HRSD (*beta* = −1.77, *SE* = 0.20, *t* = −8.95, *p* < .001) and HARS (*beta* = −1.98, *SE* = 0.30, *t* = −6.64, *p* < .001) again exhibited significant negative slopes during the treatment period and non-significant slopes during follow-up (*beta* = −0.005, *SE* = 0.12, *t* = −0.04, *p* = .97; and *beta* = −0.20 *SE* = 0.14, *t* = −1.44, *p* = .16, respectively).

3.6. Comparison of expert panel and DATA

3.6.1. Completer analysis

As noted above, DATA (Fernandez et al., 2017) was created in order to codify the selection procedures employed in the present study into an automated system. We were thus interested in the degree to which patients with expert panel-selected treatments fared compared to those with DATA-selected treatments. Three sets of comparisons were conducted via ordinary least squares regression: comparison of changes in HRSD and HARS from baseline to post-treatment, comparison of changes in Hamilton scales per unit time, and comparison of changes in HRSD and HARS from baseline to six-month follow-up. One comparison, for change in HRSD from baseline to post-treatment, was significant (*beta* = 3.25, *SE* = 1.20, *t* = 2.72, *p* = .01), indicating that the subsample with expert panel-selected treatments decreased 3 points more on the HRSD than the subsample with DATA-selected treatments. However, this difference was non-significant when examined per unit time (*beta* = 0.08, *SE* = 0.04, *t* = 1.92, *p* = .06), and no differences were found for raw change in the HARS (*beta* = 1.82, *SE* = 2.33, *t* = 0.78, *p* = .44), nor in the HARS per unit time (*beta* = 0.003, *SE* = 0.045, *t* = 0.06, *p* = .95). Finally, there were no differences observed for change in the HRSD (*beta* = 0.64, *SE* = 3.05, *t* = 0.21, *p* = .84) or HARS (*beta* = 0.93, *SE* = 3.15, *t* = 0.30, *p* = .77) from baseline to six-month follow-up.

3.6.2. Intent-to-treat analysis

Only one set of analyses was repeated for the ITT sample, the comparison of changes in HRSD and HARS from baseline to post-treatment. Comparisons of change per unit time, and change from baseline to six-month follow-up were not examined, given that no effects per unit time could be calculated for those with missing treatment data, and the degree of attrition from post-treatment to six-month follow-up would require imputation of both completer *and* dropout data. Of note, under ITT assumptions, the comparison of change in HRSD from baseline to post-treatment, was non-significant (*beta* = 1.97, *SE* = 1.50, *t* = 1.32, *p* = .20), reflecting no difference in outcome between the subsample with expert panel-selected treatments and the subsample with DATA-selected treatments. Similarly, no difference was found for change in the HARS (*beta* = 0.69, *SE* = 2.37, *t* = 0.30, *p* = .76).

3.7. Clinical significance and reliable change in completer sample

Finally, we examined the degree to which treatment completers in the present study exhibited clinically significant change and reliable change (Jacobson & Truax, 1991). For the former, we used the threshold established by Rush et al. (2003) – a reduction of 27% on the HRSD – to determine minimum clinically-significant change on the

Table 3
Baseline and post-treatment data for 32 treatment completers.

ID	HRSD.B	HRSD.P	HRSD_Δ	HARS.B	HARS.P	HARS_Δ	HRSD g	HARS g	Avg. g	# Sessions	HRSD g/time	HARS g/time	Avg. g/time
P003	16	3	−13	15	4	−11	−3.76	−1.93	−2.85	9	−0.42	−0.21	−0.32
P004	16	7	−9	33	13	−20	−2.60	−3.51	−3.06	12	−0.22	−0.29	−0.26
P006	13	8	−5	13	6	−7	−1.45	−1.23	−1.34	11	−0.13	−0.11	−0.12
P007	11	3	−8	17	4	−13	−2.32	−2.28	−2.30	14	−0.17	−0.16	−0.16
P009	17	7	−10	9	11	2	−2.89	0.35	−1.27	12	−0.24	0.03	−0.11
P012	9	0	−9	13	1	−12	−2.60	−2.11	−2.36	8	−0.33	−0.26	−0.30
P013	14	3	−11	19	6	−13	−3.18	−2.28	−2.73	10	−0.32	−0.23	−0.27
P014	10	3	−7	12	6	−6	−2.03	−1.05	−1.54	9	−0.23	−0.12	−0.17
P019	10	5	−5	10	3	−7	−1.45	−1.23	−1.34	7	−0.21	−0.18	−0.19
P023	8	1	−7	7	2	−5	−2.03	−0.88	−1.45	10	−0.20	−0.09	−0.15
P040	21	8	−13	41	9	−32	−3.76	−5.62	−4.69	13	−0.29	−0.43	−0.36
P048	14	6	−8	17	8	−9	−2.32	−1.58	−1.95	11	−0.21	−0.14	−0.18
P068	11	6	−5	14	12	−2	−1.45	−0.35	−0.90	9	−0.16	−0.04	−0.10
P072	15	6	−9	13	4	−9	−2.60	−1.58	−2.09	8	−0.33	−0.20	−0.26
P074	12	8	−4	10	11	1	−1.16	0.18	−0.49	10	−0.12	0.02	−0.05
P075	18	11	−7	23	18	−5	−2.03	−0.88	−1.45	8	−0.25	−0.11	−0.18
P100	7	2	−5	14	1	−13	−1.45	−2.28	−1.86	4	−0.36	−0.57	−0.47
P111	18	11	−7	15	8	−7	−2.03	−1.23	−1.63	13	−0.16	−0.09	−0.13
P115	18	8	−10	19	8	−11	−2.89	−1.93	−2.41	9	−0.32	−0.21	−0.27
P117	12	9	−3	18	7	−11	−0.87	−1.93	−1.40	8	−0.11	−0.24	−0.18
P127	9	4	−5	13	5	−8	−1.45	−1.40	−1.43	12	−0.12	−0.12	−0.12
P139	14	9	−5	12	9	−3	−1.45	−0.53	−0.99	14	−0.10	−0.04	−0.07
P160	13	6	−7	11	3	−8	−2.03	−1.40	−1.71	13	−0.16	−0.11	−0.13
P163	16	10	−6	16	5	−11	−1.74	−1.93	−1.83	14	−0.12	−0.14	−0.13
P169	13	9	−4	15	3	−12	−1.16	−2.11	−1.63	12	−0.10	−0.18	−0.14
P202	10	4	−6	11	7	−4	−1.74	−0.70	−1.22	12	−0.14	−0.06	−0.10
P203	18	5	−13	20	10	−10	−3.76	−1.76	−2.76	10	−0.38	−0.18	−0.28
P206	11	2	−9	16	4	−12	−2.60	−2.11	−2.36	9	−0.29	−0.23	−0.26
P219	21	5	−16	27	9	−18	−4.63	−3.16	−3.89	10	−0.46	−0.32	−0.39
P220	14	10	−4	13	9	−4	−1.16	−0.70	−0.93	12	−0.10	−0.06	−0.08
P223	21	3	−18	12	2	−10	−5.21	−1.76	−3.48	9	−0.58	−0.20	−0.39
P244	12	3	−9	8	3	−5	−2.60	−0.88	−1.74	10	−0.26	−0.09	−0.17
Average	**13.81**	**5.78**	**−8.03**	**15.81**	**6.59**	**−9.22**	**−2.33**	**−1.62**	**−1.97**	**10.38**	**−0.24**	**−0.17**	**−0.20**

HRSD = Hamilton rating scale for depression; HARS = Hamilton rating scale for anxiety; g = Hedges *g*; Avg. = average; B = baseline; P = post-treatment; Δ = change from baseline to post-treatment;/time = per unit time (sessions).

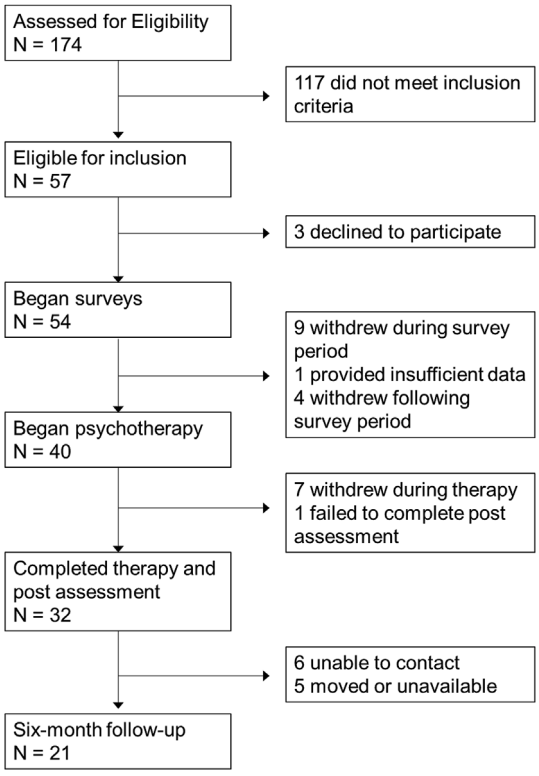

Fig. 1. CONSORT diagram.

HRSD. One participant (117) exhibited a change in HRSD of only 25%. The remaining 97% (n = 31) of the participants who completed treatment exhibited change in the HRSD that was at least *minimally* clinically significant. Cusin et al. (2009) have argued that a higher threshold of 50% reduction on the HRSD should be employed. This threshold was met by 72% of the treatment completers (n = 23) in the present study. We next calculated reliable change using the reliability for the HRSD provided by Cusin and colleagues (alpha = .83). The standard deviation for the change in HRSD was 3.60, yielding a reliable change denominator of 1.48 (where the denominator = 3.60 x $\sqrt{1 - .83}$). All 32 participants exceeded the 1.96 threshold established by Jacobson and Truax (1991). The minimum reliable change was 2.02, with a maximum of 12.14 and a mean of 5.42.

4. Discussion

The present study employed the paradigm proposed by Fisher (2015) for personalized modular CBT. Individuals with symptomatic experiences consistent with DSM-5 GAD and MDD completed surveys four-times-per-day for approximately 30 days prior to treatment. These data were subjected to within-person factor analysis (i.e. P-technique) to determine the idiosyncratic structure of their mood and anxiety pathology. These factor solutions were then used to generate factor scores for the individual time series which were then analyzed via dynamic factor modeling in order to assess the degree to which factors predicted each other from moment to moment. The algorithm provided by Fernandez et al. (2017) was then applied to the resulting output in order to construct personalized treatment plans for each individual. Finally, a personalized selection and order of UP modules was delivered.

Overall, participants responded well to the treatment. The average

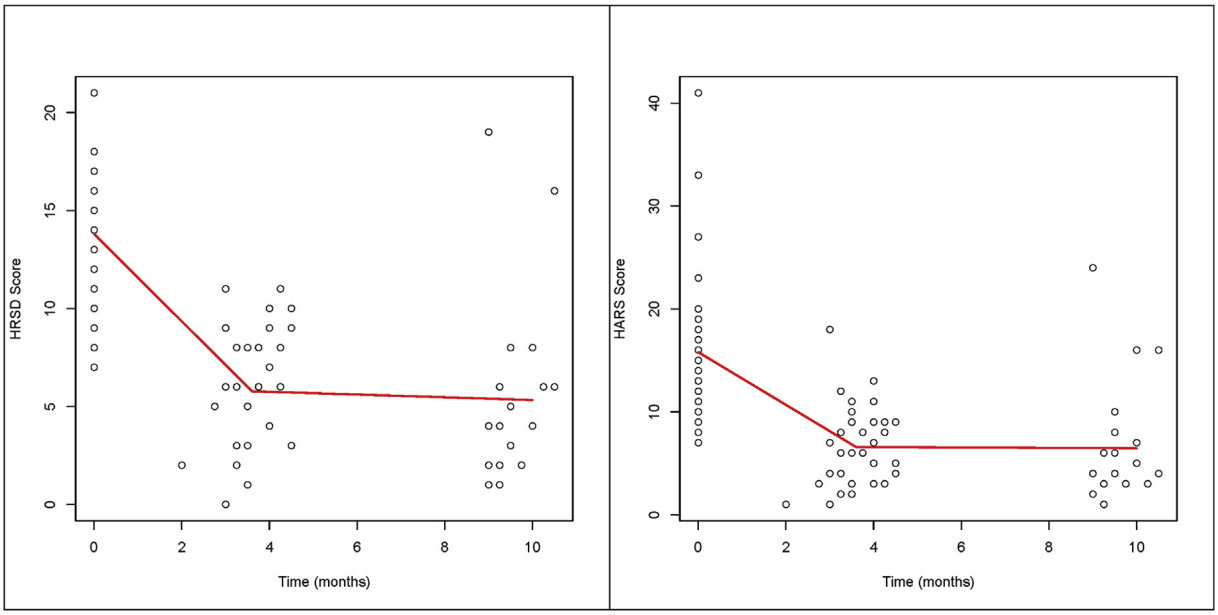

Fig. 2. Response trajectories for HRSD (left) and HARS (right) over treatment and follow-up period.

change in the HRSD – the primary outcome measure for the study – was 8.03 points, with an average Hedge's *g* effect size of 2.33, when using data from treatment completers only and 7.38 points, with an average Hedge's *g* effect size of 1.86 for the ITT sample. Given that the present study was an uncontrolled open trial, we were interested in evaluating the effectiveness of the current trial against the benchmark HRSD effect provided by Johnsen and Friborg (2015). These authors analyzed 31 CTs and RCTs for depression from 1977 to 2014, finding an average Hedge's *g* of 1.72. Thus, results from both the completer and ITT samples demonstrated treatment responses greater than this meta-analytic average. In addition, Fisher (2015) proposed that personalized modular therapies held the promise of making therapy more efficient, by eschewing unnecessary treatment components and potentially front-loading the most efficacious modules, thereby helping patients to get better faster. To this end, we were also interested in examining the effect of treatment per unit time (i.e. session). With an average of 10.38 sessions per participant, the HRSD effect per session was $g = 0.24$/session for treatment completers – exactly twice the effect per session reported by Johnsen and Friborg. Effects for the HARS were more modest, but still strong. The average change in the HARS was 9.22 points, with an average Hedge's *g* effect size of 1.62 and an effect of $g = 0.17$/session.

Twenty-one of the 32 treatment completers returned for six-month follow-up assessments. Treatment effects remained robust at six-months, with a Hedge's *g* of 1.91 for the HRSD and a *g* of 1.34 for the HARS. Examining the trajectory of change during the treatment and follow-up periods via linear mixed-effect regression revealed a change in the HRSD of 2.15 points per month, on average, during treatment and a change in the HARS of 2.43 points per month, on average, during treatment. Both the HRSD and HARS exhibited non-significant slopes during the follow-up period.

4.1. Examining differences between treatment-selection approaches

The present study employed two treatment selection procedures for personalizing modular CBT, one conducted by an expert panel and one performed by an automated algorithm. Both approaches followed the same general rubric, with automation intended to digitize and automate the expert panel selection procedures. However, analyses revealed a significant difference in treatment outcome between the two procedures, with the expert panel performing 3.25 points better than the algorithm on the HRSD. This difference was observed in the completer data but not in the ITT data, nevertheless, it bears some further inspection and explication. Unfortunately, a serious limitation of the present study – and obstacle to drawing inferences about the differential performance of the two treatment selection approaches – is that we did not randomize participants into treatment selection procedures, using a simple sequential procedure instead. Thus, the following should be considered hypotheses for further evaluation.

Examining points of divergence between the two approaches reveals three potentially important differences that may have impacted the efficacy of selected treatments. These were, (i) the interpretation and incorporation of secondary symptoms in p-technique factors, (ii) the weighting of cross-lagged prediction paths, and (iii) the inclusion versus exclusion of mean values. To the first point, the expert panel employed a thematic approach to incorporating p-technique factors. That is, each factor was given a semantic label, and symptom hierarchies tended to be weighted toward the archetypal or defining symptoms of each factor. For instance, the first factor for participant 004 was labeled 'Avoidance,' and contained factor loadings of 0.75 and 0.83 for avoiding activities and procrastination, respectively. However, this factor also explained 26% of the variance in fatigue (loading = 0.51), yet this symptom was de-emphasized, and the module used to address physical sensations such as fatigue (module 6), was selected as the final module. DATA, in turn, would have returned relatively similar item scores for avoiding activities (10.4) and fatigue (9.0). Although the expectation was that DATA would sharpen the estimation of symptom hierarchies by not omitting small but crucial details or overemphasizing semantically salient features, it may be that, instead, greater efficacy is derived from a gestalt perspective that better summarizes the nature of the latent variable.

The DATA calculation includes a component called the 'raw item score,' which is the product of the explanatory power of the p-technique factor (percentage of contemporaneous symptom variance accounted for) and the explanatory power of the dynamic factor (percentage of lagged symptom variance accounted for). Relative to autoregressions, cross-lagged effects were typically small in magnitude. Thus, their influence on treatment planning tended to be tertiary to the explanatory power of the p-technique model and the autoregressions in the dynamic model. In contrast, the expert panel tended to treat cross-lagged effects as either present or absent, with factors exerting cross-lagged influence given precedence as primary treatment targets. It was assumed that

DATA would correct for this imprecise, potentially biased application of cross-lagged effects. However, it may be that signal detection issues attenuated the magnitude of cross-lagged predictions, and relatively small coefficient values belied stronger relationships. Future research should attempt to identify optimal sampling frequencies and other potential measurement features that may better detect cross-lagged relationships.

Finally, it should be noted that the expert panel selection procedure was exclusively based on covariance relationships and did not incorporate mean levels. Although the majority of the DATA calculations regarded variance-covariance information, the final score was weighted by item means – simply the average severity for each item. In determining symptom predominance and modular treatment selection, it may be that the correlational relationships between symptoms provide more treatment-relevant information than hierarchies of level or severity. That is, it may be more important to understand the degree of influence of a given symptom than the perceived severity or intensity of that symptom.

4.2. Conclusion

The present study is the first to use pre-therapy multivariate time series data to generate prospective treatment plans. Consistent with Schork's (2015) proposal, we viewed the present study as an accumulation of 32 separate N-of-1 studies. Aggregation was applied to the 32 samples in order to make generalizable conclusions about the effectiveness of the applied approach *generally*. However, we believe that the strength of the current approach is that it does not require a large sample to make quantitatively rigorous conclusions about symptom severity, idiosyncratic syndrome structures, or treatment response.

Broadly, our findings illustrate that data-driven personalization of psychosocial interventions can increase the efficiency with which gains are made by helping patients to get better faster. Increased efficiency may aid in the solution of large-scale problems that persist even within empirically-derived psychosocial interventions. First, individuals tend to drop out of treatment at higher rates when they perceive a lack of progress (Hunsley, Aubry, Verstervelt, & Vito, 1999), so methods that yield more rapid change may aid in retention. Further, the high burden on the mental healthcare system is both a local and a global issue, for which the World Health Organization (2013) and others have issued calls for action. Thus, the ability of the present methods to provide clinically significant change over fewer sessions may allay the burden on the mental healthcare system so that more individuals can receive necessary psychosocial intervention. In the future, a RCT design could be employed to further establish and understand the benefits of the current approach. Systematically varying and randomizing to differing treatment length and treatment selection approaches could better-isolate those factors that increased efficacy per-unit-time in the present study.

4.3. Limitations

The present study had three notable weaknesses. First, it was an uncontrolled open trial. Despite calls from our group and others for idiographic research paradigms, the RCT remains the gold-standard for demonstrating efficacy. These goals are not incompatible. The idiographic approach employed in the current research could be tested against standard nomothetic treatment strategies, with one personalized arm and one standardized arm. Additionally, controlled comparisons between treatment lengths, data collection paradigms, and analytic strategies could all help to determine the specific mechanisms by which the current approach may elicit an incremental treatment response. Second, the current study had a relatively small sample size. Although the data collected *per person* was substantial, the between-subject samples of N = 32 for completer analyses and N = 40 for ITT analyses likely limit the generalizability of these findings. Taken together with the uncontrolled nature of the trial, the present findings should be considered preliminary, and thus require further investigation and replication.

Finally, it should be noted that the current study employed a latent factor approach to conceptualizing and modeling psychopathology. Recently, a number of researchers have proposed that the latent factor model is not an appropriate choice for psychiatric syndromes, as these models infer the existence of a latent disease entity and require conditional independence in the factor indicators (Schmittmann et al., 2013). The network theory of psychopathology has been proposed as an alternative, possibly superior model (Borsboom, 2017). Work in our group has applied network models to idiographic mood and anxiety data, with promising results (Fisher et al., 2017). In addition, the algorithms provided by Fernandez et al. (2017) have recently been extended to network model methodology (Rubel et al., In Press). Thus, future work should likely test the comparative efficacy of treatment selection based on factor analytic versus network analytic methods.

References

Abraha, I., Cherubini, A., Cozzolino, F., De Florio, R., Luchetta, M. L., Rimland, J. M., ... Montedori, A. (2015). Deviation from intention to treat analysis in randomised trials and treatment effect estimates: meta-epidemiological study. *BMJ, 350*, h2445. https://doi.org/10.1136/bmj.h2445.

Alshurafa, M., Briel, M., Akl, E. A., Haines, T., Moayyedi, P., Gentles, S. J., ... Guyatt, G. H. (2012). Inconsistent definitions for intention-to-treat in relation to missing outcome data: Systematic review of the methods literature. *PLoS One, 7*(11), e49163. https://doi.org/10.1371/journal.pone.0049163.

American Psychiatric Association (2013). *The diagnostic and statistical manual of mental disorders: DSM, Vol. 5*. bookpointUS.

Ashley, E. A. (2015). The precision medicine initiative: A new national effort. *Journal of the American Medical Association, 313*(21), 2119–2120.

Barlow, D. H., Farchione, T. J., Fairholme, C. P., Ellard, K. K., Boisseau, C. L., Allen, L. B., et al. (2011). *Unified protocol for transdiagnostic treatment of emotional disorders: Therapist guide.* New York, NY: Oxford University Press.

Barlow, D. H., & Nock, M. K. (2009). Why can't we be more idiographic in our research? *Perspectives on Psychological Science, 4*(1), 19–21.

Barlow, D. H., Sauer-Zavala, S., Carl, J. R., Bullis, J. R., & Ellard, K. K. (2014). The nature, diagnosis, and treatment of neuroticism: Back to the future. *Clinical Psychological Science, 2*(3), 344–365. https://doi.org/10.1177/2167702613505532.

Beck, A. T., & Steer, R. A. (1991). Relationship between the Beck anxiety inventory and the Hamilton anxiety rating scale with anxious outpatients. *Journal of Anxiety Disorders, 5*(3), 213–223. https://doi.org/10.1016/0887-6185(91)90002-B.

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry, 16*(1), 5–13. https://doi.org/10.1002/wps.20375.

Boswell, J. F. (2013). Intervention strategies and clinical process in transdiagnostic cognitive–behavioral therapy. *Psychotherapy, 50*(3), 381.

Brown, T. A., & Barlow, D. H. (2009). A proposal for a dimensional classification system based on the shared features of the DSM-IV anxiety and mood disorders: Implications for assessment and treatment. *Psychological Assessment, 21*(3), 256–271.

Brown, T. A., & Barlow, D. H. (2014). *Anxiety and related disorders interview schedule for DSM-5: Clinician manual.* Oxford University Press.

Brown, T. A., Chorpita, B. F., & Barlow, D. H. (1998). Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. *Journal of Abnormal Psychology, 107*(2), 179–192. https://doi.org/10.1037/0021-843X.107.2.179.

Bruss, G. S., Gruenberg, A. M., Goldstein, R. D., & Barber, J. P. (1994). Hamilton anxiety rating scale interview guide: Joint interview and test-retest methods for interrater reliability. *Psychiatry Research, 53*(2), 191–202. https://doi.org/10.1016/0165-1781(94)90110-4.

Burmeister, M., McInnis, M. G., & Zöllner, S. (2008). Psychiatric genetics: Progress amid controversy. *Nature Reviews Genetics, 9*(7), 527.

Chou, C. P., & Bentler, P. M. (1990). Model modification in covariance structure modeling: A comparison among likelihood ratio, Lagrange multiplier, and wald tests. *Multivariate Behavioral Research, 25*(1), 115–136. https://doi.org/10.1207/s15327906mbr2501_13.

Cuijpers, P., Berking, M., Andersson, G., Quigley, L., Kleiboer, A., & Dobson, K. S. (2013). A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments. *Canadian Journal of Psychiatry, 58*(7), 376–385.

Cuijpers, P., Sijbrandij, M., Koole, S., Huibers, M., Berking, M., & Andersson, G. (2014). Psychological treatment of generalized anxiety disorder: A meta-analysis. *Clinical Psychology Review, 34*(2), 130–140. https://doi.org/10.1016/j.cpr.2014.01.002.

Cuijpers, P., van Straten, A., Andersson, G., & van Oppen, P. (2008). Psychotherapy for depression in adults: A meta-analysis of comparative outcome studies. *Journal of Consulting and Clinical Psychology, 76*(6), 909.

Cusin, C., Yang, H., Yeung, A., & Fava, M. (2009). Rating scales for depression. *Handbook of clinical rating scales and assessment in psychiatry and mental health* (pp. 7–35). Springer.

Dzau, V. J., Ginsburg, G. S., Van Nuys, K., Agus, D., & Goldman, D. (2015). Aligning incentives to fulfil the promise of personalised medicine. *The Lancet, 385*(9982),

2118–2119.
Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software, 1*(4), 2012. https://doi.org/10.18637/jss.v048.i04.
Fergusson, D., Aaron, S. D., Guyatt, G., & Hébert, P. (2002). Post-randomisation exclusions: The intention to treat principle and excluding patients from analysis. *BMJ British Medical Journal, 325*(7365), 652–654.
Fernandez, K. C., Fisher, A. J., & Chi, C. (2017). Development and initial implementation of the dynamic assessment treatment algorithm (DATA). *PLoS One, 12*(6), e0178806. https://doi.org/10.1371/journal.pone.0178806.
Fisher, A. J. (2015). Toward a dynamic model of psychological assessment: Implications for personalized care. *Journal of Consulting and Clinical Psychology, 83*(4), 825–836. https://doi.org/10.1037/ccp0000026.
Fisher, A. J., & Boswell, J. F. (2016). Enhancing the personalization of psychotherapy with dynamic assessment and modeling. *Assessment, 23*(4), 496–506. https://doi.org/10.1177/1073191116638735.
Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). *Lack of group-to-individual generalizability is a threat to human subjects research.* Proceedings of the National Academy of Sciences.
Fisher, A. J., Newman, M. G., & Molenaar, P. C. M. (2011). A quantitative method for the analysis of nomothetic relationships between idiographic structures: Dynamic patterns create attractor states for sustained posttreatment change. *Journal of Consulting and Clinical Psychology, 79*(4), 552–563. https://doi.org/10.1037/a0024069.
Fisher, A. J., Reeves, J. W., Lawyer, G., Medaglia, J. D., & Rubel, J. A. (2017). Exploring the idiographic dynamics of mood and anxiety via network analysis. *Journal of Abnormal Psychology, 126*(8), 1044–1056. https://doi.org/10.1037/abn0000311.
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.
Galatzer-Levy, I. R., & Bryant, R. A. (2013). 636,120 ways to have posttraumatic stress disorder. *Perspectives on Psychological Science, 8*(6), 651–662. https://doi.org/10.1177/1745691613504115.
Hamburg, M. A., & Collins, F. S. (2010). The path to personalized medicine. *New England Journal of Medicine, 363*(4), 301–304. https://doi.org/10.1056/NEJMp1006304.
Hamilton, M. (1959). The assessment of anxiety states by rating. *British Journal of Medical Psychology, 32*(1), 50–55.
Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery & Psychiatry, 23*(1), 56.
Hofmann, S. G., & Smits, J. A. (2008). Cognitive-behavioral therapy for adult anxiety disorders: A meta-analysis of randomized placebo-controlled trials. *Journal of Clinical Psychiatry, 69*(4), 621.
Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.
Hunsley, J., Aubry, T. D., Verstervelt, C. M., & Vito, D. (1999). Comparing therapist and client perspectives on reasons for psychotherapy termination. *Psychotherapy: Theory, Research, Practice, Training, 36*(4), 380.
Insel, T. R. (2009). Translating scientific opportunity into public health impact: A strategic plan for research on mental illness. *Archives of General Psychiatry, 66*(2), 128–133. https://doi.org/10.1001/archgenpsychiatry.2008.540.
Insel, T. R., & Cuthbert, B. N. (2015). Brain disorders? Precisely. *Science, 348*(6234), 499–500.
Iyengar, R., Altman, R. B., Troyanskya, O., & FitzGerald, G. A. (2015). Personalization in practice. *Science, 350*(6258), 282–283.
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.
Johnsen, T. J., & Friborg, O. (2015). The effects of cognitive behavioral therapy as an anti-depressive treatment is falling: A meta-analysis. *Psychological Bulletin, 141*(4), 747–768. https://doi.org/10.1037/bul0000015.
Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the national comorbidity survey replication. *Archives of General Psychiatry, 62*(6), 617–627.
Khoury, M. J., & Evans, J. P. (2015). A public health perspective on a national precision medicine cohort: Balancing long-term knowledge generation with early health benefit. *Journal of the American Medical Association, 313*(21), 2117–2118.
Kobak, K. A., Reynolds, W. M., & Greist, J. H. (1993). Development and validation of a computer-administered version of the Hamilton Rating Scale. *Psychological Assessment, 5*(4), 487.
Lamiell, J. T. (1981). Toward an idiothetic psychology of personality. *American Psychologist, 36*(3), 276–289. https://doi.org/10.1037/0003-066X.36.3.276.
McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282.
Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review, 80*(4), 252–283.
Molenaar, P. C. M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika, 50*(2), 181–202.
Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives, 2*(4), 201–218.
Moras, K., Di Nardo, P. A., & Barlow, D. H. (1992). Distinguishing anxiety and depression: Reexamination of the reconstructed Hamilton scales. *Psychological Assessment, 4*(2), 224.
Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology, 31*(2), 109.
Revelle, W. (2013). *psych: Procedures for psychological, psychometric, and personality research.* Northwestern University.
Rubel, J. A., Fisher, A. J., Husen, K., & Lutz, W. (in press). Translating person-specific network models into personalized treatments: Development and demonstration of the dynamic assessment treatment algorithm for individual networks (DATA-IN). Psychotherapy and psychosomatics.
Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., ... Manber, R. (2003). The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry, 54*(5), 573–583.
Schmittmann, V. D., Cramer, A. O. J., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology, 31*(1), 43–53. https://doi.org/10.1016/j.newideapsych.2011.02.007.
Schork, N. J. (2015). Personalized medicine: Time for one-person trials. *Nature, 520*(7549), 609–611.
Steer, R. A., Beck, A. T., Riskind, J. H., & Brown, G. (1987). Relationships between the Beck depression inventory and the Hamilton psychiatric rating scale for depression in depressed outpatients. *Journal of Psychopathology and Behavioral Assessment, 9*(3), 327–339.
Steer, R. A., McElroy, M. G., & Beck, A. T. (1983). Correlates of self-reported and clinically assessed depression in outpatient alcoholics. *Journal of Clinical Psychology, 39*(1), 144–149.
Watson, D. (2005). Rethinking the mood and anxiety disorders: A quantitative hierarchical model for DSM-V. *Journal of Abnormal Psychology. Special Issue: Toward a Dimensionally Based Taxonomy of Psychopathology, 114*(4), 522–536.
Wittchen, H. U. (2002). Generalized anxiety disorder: Prevalence, burden, and cost to society. *Depression and Anxiety, 16*(4), 162–171.
World Health Organization (2013). *Investing in mental health: Evidence for action.* Retrieved from http://www.who.int/mental_health/publications/financing/investing_in_mh_2013/en/.
Yelland, L. N., Sullivan, T. R., Voysey, M., Lee, K. J., Cook, J. A., & Forbes, A. B. (2015). Applying the intention-to-treat principle in practice: Guidance on handling randomisation errors. *Clinical Trials, 12*(4), 418–423. https://doi.org/10.1177/1740774515588097.