Personalized Prescriptions of Therapeutic Skills from Patient Characteristics: An Ecological Momentary Assessment Approach

Christian A. Webb, Ph.D.[1], Marie Forgeard, Ph.D.[1,2], Elana S. Israel, B.A.[1], Nathaniel Lovell-Smith, B.A.[1], Courtney Beard, Ph.D. [1], Thröstur Björgvinsson, Ph.D., ABPP[1]

[1] Harvard Medical School – McLean Hospital, Boston, MA
[2] Department of Clinical Psychology, William James College, Newton, MA

Corresponding author:
Christian A. Webb, Ph.D.
Harvard Medical School & McLean Hospital
115 Mill Street, Belmont, MA 02478.
Email: cwebb@mclean.harvard.edu

Funding

Funding for the study was provided by the Behavioral Health Partial Program and McLean Hospital. The first author (Webb) was partially supported by R01 MH116969, the Tommy Fuss Fund and a NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation. The opinions and assertions contained in this article should not be construed as reflecting the views of the sponsors.

Abstract

Objective: Rather than relying on a single psychotherapeutic orientation, most clinicians draw from a range of therapeutic approaches to treat their clients. To date, no data-driven approach exists for personalized predictions of which skill domain would be most therapeutically beneficial for a given patient. The present study combined ecological momentary assessment (EMA) and machine learning to test a data-driven approach for predicting patient-specific skill-outcome associations.

Method: Fifty (M_{age}= 37 years old, 54% female, 84% White) adults received training in Behavioral Therapy (BT) and Dialectical Behavior Therapy (DBT) skills within a behavioral health partial hospital program. Following discharge, patients received four EMA surveys per day for two weeks (total observations = 2,036) assessing use of therapeutic skills and positive/negative affect (PA/NA). Clinical and demographic characteristics were submitted to elastic net regularization to predict, via cross-validation, patient-specific associations between use of BT vs. DBT skills and level of PA/NA.

Results: Cross-validated accuracy was 81% (sensitivity=93%; specificity=63%) in predicting whether a patient would exhibit a stronger association between use of BT vs. DBT skills and PA level. Predictors of positive DBT skills-PA associations included higher levels of non-suicidal self-injury and sleep disturbance, whereas predictors of positive BT skills-PA relations included higher emotional lability and anxiety disorder comorbidity, and lower psychomotor retardation/agitation and worthlessness/guilt. Corresponding models with NA yielded no predictors.

Conclusions: Findings from this initial proof-of-concept study highlight the potential of data-driven approaches to inform personalized prescriptions of which skill domains may be most therapeutically beneficial for a given patient.

Keywords: cognitive behavioral therapy; personalized; ecological momentary assessment; affect

Public Health Significance Statement

Therapists must decide which skills (e.g., behavioral activation, cognitive restructuring, mindfulness) are expected to be most beneficial for a given client initiating therapy. This study demonstrates the potential of a data-driven approach – based on pre-treatment clinical and demographic characteristics – to inform which skills are predicted to be most therapeutically beneficial for a given client.

Personalized Prescriptions of Therapeutic Skills from Patient Characteristics: An Ecological Momentary Assessment Approach

A range of empirically supported psychotherapeutic (e.g., cognitive behavioral therapy [CBT], interpersonal therapy [IPT]) and pharmacological (e.g., selective serotonin reuptake inhibitors [SSRIs]) treatments are available for individuals suffering from depression. Randomized clinical trials (RCTs) comparing the relative efficacy of these different interventions provide information about *average* differences between treatments. However, certain individuals may be better suited to one treatment vs. another (e.g., by virtue of their unique demographic, clinical or neurobiological characteristics). In an effort to develop data-driven approaches for treatment matching, there have been a growing number of studies leveraging advances in machine learning approaches to predict optimal treatment assignment (Cohen & DeRubeis, 2018; DeRubeis, 2019). For example, several recent studies have utilized the *personalized advantage index* (PAI) (DeRubeis et al., 2014), a machine learning algorithm informed by a patient's pretreatment characteristics, to predict whether a given depressed patient is better suited to CBT vs. an SSRI (DeRubeis et al., 2014), CBT vs. IPT (Huibers et al., 2015), CBT vs. psychodynamic therapy (Cohen et al., 2019), SSRI vs. placebo (Webb et al., 2018), and supportive expressive therapy vs. SSRI vs. placebo (Zilcha-Mano et al., 2016).

Prior studies estimated the relative efficacy of whole treatment packages (e.g., CBT vs. IPT) for individual patients. In other words, they were designed to predict whether a given depressed individual will exhibit a better outcome in treatment A vs. treatment B. Critically, the majority of community clinicians do *not* rely on a single theoretical orientation but draw from more than one therapeutic school (e.g., teaching clients both behavioral therapy (BT) and DBT skills) (Cook et al., 2010; DiGiorgio et al., 2010; Michalak et al., 2018; Norcross et al., 2019).

Indeed, surveys indicate that only approximately 2-12% of therapists adhere to a single therapeutic orientation (Cook et al., 2010; Hollanders & McLeod, 1999; Norcross et al., 2019). In current clinical practice, deciding which therapeutic approach or skill domain to emphasize for a given patient is based on individual case conceptualizations and a trial-and-error approach. No data-driven approach exists for personalized predictions of which skill domain would be most therapeutically beneficial for a given patient.

To address this gap, the present study combined ecological momentary assessment (EMA) and machine learning to test a data-driven approach to predicting patient-specific skill-affect associations for depressed and anxious patients receiving treatment in a naturalistic psychiatric hospital unit teaching BT and DBT skills. EMA allows for a dense (i.e., repeated daily surveys) and ecologically valid assessment of the use of skills and positive/negative affect (PA and NA) in the daily lives of patients. With regards to predictors of skill-affect associations, any single predictor variable will account for limited variance. Multivariable machine learning approaches, such as elastic net regularization (ENR), can accommodate the combination of relatively large sets of predictors in an effort to account for maximal variance in outcome (in the present study, skill-affect associations). ENR is a highly utilized variant of conventional regression, which has several features well-suited to this study. First, ENR, in contrast to conventional (ordinary least squares) linear regression, can handle a relatively large number of predictors and multicollinearity among them (Friedman et al., 2010; Zou & Hastie, 2005). Second, ENR has been shown to yield superior out-of-sample predictive performance (i.e., increases the generalizability of the model to new individuals) relative to linear regression (Webb et al., 2020; Zou & Hastie, 2005). Finally, another desirable feature of ENR is that it is less complex and more easily interpretable than other machine learning approaches, many of

which have “black box” properties (e.g., non-linear associations and high-order interactions may underlie the predictive performance of random forest) (see Supplement for additional details on ENR).

In the present study, patient clinical and demographic characteristics were submitted to ENR to predict, via cross-validation, patient-specific associations between use of BT skills vs. DBT skills and PA/NA (assessed via EMA). We hypothesize that ENR-derived personalized predictions of skill-affect relations (in particular, the ability to predict which skill domain, BT or DBT, is more strongly associated with improved affect for each individual) will be significantly more accurate than chance. To the extent that such a predictive effort is successful (i.e., as quantified by prediction accuracy, sensitivity and specificity), a resulting model could inform personalized prescriptions of which skill domain would have the highest likelihood of being emotionally beneficial for a given patient.

Method

Participants

Fifty participants were recruited between December 2016 and July 2017 by research assistants during their treatment at a behavioral health partial hospital program (PHP) in the [*omitted for blinded submission*], as part of a larger study examining daily predictors of affect following discharge from partial hospitalization [*omitted for blinded submission]*. All PHP patients were eligible for the study except for those with active mania or psychosis; in addition, some could not be included due to practical constraints (early discharge from PHP, acuity requiring clinical care be prioritized over research activities, and erratic program attendance). All participants completed a 30-minute training session with the research assistant prior to discharge to review study procedures as well as download and practice answering surveys on the

smartphone app used in the present study. Participants were 36.82 years old on average (*SD* = 14.42, range = 19-70). **Table 1** provides detailed information about our sample's demographic characteristics.

Participants reported an average PHQ-9 score of 6.98 (*SD* = 4.54, range = 0-20) and an average GAD-7 score of 4.77 (*SD* = 3.78, range = 0-18) on the first day of this study, suggesting mild residual depressive and anxiety symptoms the day after discharge from the PHP. The most common current diagnosis (established by a structured diagnostic interview, as described below) in this sample was a major depressive episode in the context of Major Depressive Disorder (74% of the sample), followed by Generalized Anxiety Disorder (66%) and Social Anxiety Disorder (40%). Twenty-four percent of the sample had scores on the Mclean screening instrument for borderline personality disorder (MSI-BPD) (Zanarini, Vujanovic, Parachini, Boulanger, Frankenburg, & Hennen, 2003) which were above the cut-off (total score ≥ 7) commonly used to suggest a possible BPD diagnosis. Most participants (76%) met criteria for 2 or more disorders. **Table 2** provides detailed information about our sample's diagnostic characteristics. Participants in this study had completed an average of 10.26 treatment days (*SD* = 1.24) at the PHP before discharge.

The set of clinical and demographic predictors (**Table 3**) submitted to our ENR models were selected based on empirical, theoretical and practical considerations. First, to our knowledge, no prior study has tested predictors of skill-affect associations. Thus, there were no directly relevant prior studies to inform variable selection (i.e., those patient characteristics that have previously been shown to predict skill-affect associations). However, there are prior relevant studies testing predictors of either skill use (Strunk et al., 2014; Webb et al., 2012) or treatment outcome (For a review, see Kessler et al., 2017). For example, we included variables

previously shown to predict depression treatment outcome (for a review, see Kessler et al., 2017), including within the same partial hospital setting as the present study (Webb et al., 2020), such as depression severity, anxiety severity, suicidality, anhedonia, sleep difficulties, fatigue, concentration difficulties, relationship difficulties, BPD symptoms, prior psychiatric treatment, age, and education level. Second, and following discussions with clinical staff in the PHP, we included theoretically relevant variables that may plausibly predict BT skill-affect and/or DBT skill-affect associations (e.g., BPD symptoms, suicidality, treatment history). Finally, given the practical constraints of conducting research in a naturalistic clinical setting, assessments had to be relatively brief in order not to interfere with clinical care.

Setting

Patients in the PHP receive individual and group therapy in addition to working with a psychiatrist and case manager who coordinate an individualized treatment plan. The PHP provides treatment for patients who experience a range of disorders, the most common being mood and anxiety disorders. Treatment within the PHP focuses on teaching patients core CBT skills (including behavioral activation, exposure, and cognitive restructuring) (Beck et al., 1979), as well as DBT skills (including emotion regulation, distress tolerance, interpersonal effectiveness, and mindfulness) (Linehan, 1993) via daily group therapy and individual therapy. Patients attend the PHP daily (Monday-Friday; 9am – 3pm) and receive approximately 5 hours of clinical services each day [*omitted for blinded submission*].

Materials and Procedures

The Partners Healthcare Institutional Review Board (IRB) assessed and sanctioned all procedures and materials used for this research. Participants began completing daily self-report measures for current study upon discharge from the PHP. The survey period lasted 14 days with

participants receiving EMA surveys four times per day. All participants submitted their responses via a smartphone app installed onto their phones. The app, Metricwire, permits researchers to build and distribute surveys that are HIPAA-compliant. Participants earned $20/week for completing any survey, with an additional $30 per week for completing ≥ 80% of surveys (up to $100 in total). Informed written consent was acquired from all participants to use both their daily measures and the clinical data associated with their time at the PHP.

EMA Surveys. Participants completed EMA surveys four times a day for 14 days, at semi-random intervals between 10am and 8pm (separated by at least 2 hours). Participants had up to an hour to answer each survey. The average interval between completed surveys within each day was 2 hours and 54 minutes.

Positive and Negative Affect. At each EMA survey, participants were instructed to rate how they felt immediately prior to receiving the notification using 6 positive affect (PA) and 6 negative affect (NA) adjectives rated on a 7-point Likert scale from 1 (*not at all*) to 7 (*extremely*). These adjectives were similar to those from existing self-report measures (Watson et al., 1988) and previous EMA studies of momentary affect (Bylsma & Rottenberg, 2011; Forbes et al., 2009; Peeters et al., 2010). However, we ensured that an equal number of activated (i.e., "excited," "energized," "active") and deactivated (i.e., "calm," "peaceful," "relaxed") PA states, as well as activated (i.e., "nervous," "frustrated," "angry") and deactivated ("bored," "sad," "tired") NA states, were represented (see Supplement for list of items). Given the nested data structure (i.e., repeated EMA assessment nested within patients), we used a multilevel approach to assess the reliability of the brief positive and negative affect (PA and NA) scales used in our EMA surveys (as described by Geldhof et al., 2014; Nezlek, 2007; Nezlek & Gable, 2001). Estimates of between-person reliability (PA = .95, NA = .91) as well as within-person reliability

(PA = .75, NA = .71) showed that these scales demonstrated adequate internal consistency in the present sample.

Skills Use. At each EMA survey, participants were presented with a list of skills. Participants were then asked to identify which, if any, of the skills they had used *since the last survey administration*. All skills were drawn from the core of the PHP treatment, and thus no skills were considered novel. For the purposes of the present study, skills were categorized as either BT skills (i.e., behavioral activation, behavioral scheduling, exposure) or DBT skills (i.e., mindfulness, distress tolerance, emotion regulation, interpersonal effectiveness), and each skill had a corresponding description offered in the app as a reminder to participants (see Supplement). An identifying/challenging negative automatic thoughts item was excluded from this study due to limited variance and endorsement by patients (e.g., 10 participants had a standard deviation of 0 for this variable). This is consistent with the short-term nature of treatment within the PHP and the emphasis on behavioral skills as opposed to relatively more complex cognitive restructuring skills, which are more challenging to learn and effectively deploy. Seven participants had no variability in either PA, NA, BT skills or DBT skills, and were thus removed from the analyses given that skill-PA/NA correlations could not be computed these patients.

Depression and anxiety symptoms. At discharge, participants completed the Patient Health Questionnaire (PHQ-9) (Kroenke & Spitzer, 2002) to assess depression symptoms and the 7-item Generalized Anxiety Disorder Questionnaire (GAD-7) (Spitzer et al., 2006) to assess anxiety symptoms, two widely used measures with established reliability and validity. For both, participants indicated the frequency with which they experienced specific symptoms over the

past 24 hours on a 4-point Likert-type scale (from 0 = not at all to 3 = nearly all the time). The PHQ-9 and GAD-7 had high internal consistency in this study (α = .89 and α = .91 respectively).

Behavior and Symptom Identification Scale (BASIS-24) (Cameron et al., 2007). The BASIS-24 consists of 24 items and assesses symptoms over the past week. It consists of 6 subscales: (1) Depression/Functioning ("Feel sad or depressed?"), (2) Interpersonal Problems ("Get along with people in your family?"), (3) Self-Harm ("Think about ending your life?"), (4) Emotional Lability ("Have mood swings?"), (5) Psychosis ("Hear voices or see things?"), and (6) Substance Abuse/Dependence ("Did you have an urge to drink alcohol or take street drugs?"). Participants are asked to rate items on a 5-point Likert type rating scale from 0 *(none of the time)* to 4 *(all of the time),* with higher scores indicating worse functioning. Subscale scores range from 0-8 (self-harm) to 0-24 (depression/functioning) and total scores reflect overall functioning. The BASIS-24 has previously been shown to demonstrate good reliability, validity, and responsiveness to change as measure of mental well-being and functioning (Cameron et al., 2007). Internal consistency for the total score at discharge in this sample was high (α = .88).

Mclean screening instrument for borderline personality disorder (MSI-BPD) (Zanarini, Vujanovic, Parachini, Boulanger, Frankenburg, & Hennen, 2003). The MSI-BPD is a brief 10-item measure screening for borderline personality disorder traits in a yes/no format. It has demonstrated adequate psychometric properties (Zanarini et al., 2003), and internal consistency was adequate in the present sample (α = .73).

Diagnostic interview. We obtained diagnostic information from a structured interview completed on the second day of treatment at the PHP, the Mini International Neuropsychiatric Interview for DSM-IV (MINI) (Sheehan et al., 1998). This structured interview was administered by doctoral interns or practicum students trained and supervised by a postdoctoral

fellow. The MINI has strong reliability and validity in relation to the Structured Clinical Interview for DSM-IV (Sheehan et al., 1998), and MINI diagnoses correlate with those made by program psychiatrists in this PHP (Kertz et al., 2012). MINI training included two, two-hour group didactic meetings, as well as independent time reviewing administration manuals, listening to training audio recordings, rating test interviews, and completing mock interviews. Diagnosticians were required to pass a final training interview with their supervisor. To prevent drift, all diagnosticians met as a group monthly with the supervisor to review procedures and discuss cases. Twice per year, diagnosticians all rated the same interview and discussed diagnoses during this monthly meeting.

Analytic Strategy

A schematic of the analysis pipeline is presented in **Figure 1**. First, two within-subject correlations were computed for every patient representing their *observed* correlation between self-reported use of BT skills since the last survey and current PA, and between use of DBT skills since the last survey and current PA. Next, we generated two *predicted* within-subject correlations reflecting the expected relation between use of BT skills and PA, and between use of DBT skills and PA. The latter predictions were generated via ENR (glmnet package, alpha = 0.5; Friedman et al., 2010) with 24 clinical and demographic predictors (see **Table 3** for list of variables). Two ENR models were run: predicting the BT skills-PA correlation and DBT skills-PA correlations, separately. Tuning of ENR's lambda parameter was performed via the glmnet package (cv.glmnet), and the optimal value was selected (via minimum cross-validated error criterion). In an effort to minimize overfitting, the ENR models were run via 10-fold cross-validation (CV). For each of the ten folds, models were trained on 9/10th of the data (from the other 9 folds) and skills-PA predictions were generated for that held out fold (which consisted of

the other 1/10th of the training sample). Importantly, the CV procedure ensures that the predictions of BT-PA and DBT-PA correlations for all patients are generated from models that are constructed without using their own data. To generate stable estimates, the 10-fold CV ENR models were repeated 1000 times. Thus, for every patient, 1000 pairs of predicted within-subject BT skills-PA and DBT skills-PA correlations were generated. Next, for every patient, we computed the difference between the mean predicted BT skills-PA correlation and DBT skills-PA correlation, such that a positive value indicated that our model predicted that this patient would exhibit a numerically larger association between use of BT skills and levels of PA, relative to use of DBT skills and PA. This difference score was then translated into a binary variable reflecting whether a given patient was predicted to have either (1) a stronger association between use of BT skills and PA or (2) between DBT skills and PA. A corresponding binary score was computed from the *observed* BT-skills PA vs. DBT skills PA correlations. Finally, we compared these predicted and observed values (via accuracy, sensitivity and specificity; ROCit package (Khan & Brandenburger, 2019) and ROCR package (Sing et al., 2005)) to answer the question: How accurately can we predict whether a given patient will have a stronger association between BT skills vs. DBT skills and level of PA. The above steps were repeated for models predicting skills-NA associations. However, no predictors emerged in these corresponding models (see Supplemental Results). Thus, only the skills-PA results are presented below.

Results

Between-patient and within-patient variability in skills and affect

Intraclass correlation coefficients (ICCs) were .57 and .55 for PA and NA, respectively, indicating that just under half (43-45%) of the variance in affect was due to variability *within* patients over time. The ICCs for BT and DBT skills were 22% and 31%, respectively, indicating

that the majority (69-78%) of the variance in use of skills was within-patient variability rather than attributable to average between-patient differences. In sum, there was substantial within-individual variability in affect and skills over time, which we modeled below.

Predicting patient-specific skill-affect associations

Figure 2 displays the distribution of observed within-patient correlations between BT skills and PA (red), and between DBT skills and PA (green). Although the mean BT skills-PA (*r* = .14; sd = .17; range = -.23 – .58) and DBT skills-PA (*r* = .04; sd = .20; range= -.51 - .39) associations were small, the distributions revealed substantial variability across patients in the magnitude (and direction) of these relationships. Predictors of a more positive DBT skills-PA relation included presence of non-suicidal self-injury (past month) and greater sleep disturbance; whereas predictors of positive BT skills-PA relations included higher emotional lability and anxiety disorder comorbidity, and lower psychomotor retardation/agitation and worthlessness/guilt (**Table 4**). As noted above, the 10-fold CV ENR models were repeated 1000 times. Thus, for every patient, 1000 pairs of predicted within-subject BT skills-PA and DBT skills-PA correlations were generated and averaged. The within-person standard deviation (SD) was .04 and .03 for BT skills – PA and DBT skills – PA *r* values, respectively.

The area under the curve (AUC) for the predictions was 0.80 (**Figure 3**). **Supplemental Figure 1** plots model accuracy at different cutoffs (computed as mean predicted BT skills-PA correlation - mean predicted DBT skills-PA correlation). As displayed, maximum accuracy (81.4% [95% CI 66.6 – 91.6%; sensitivity = 92.6%; specificity = 62.5%) was achieved at a cutoff value of .069. This accuracy was significantly greater than the no-information rate (Acc > NIR, p = .007). More specifically, at the above cutoff, the model correctly classified 35/43 patients (i.e., accuracy of 81.4%); whereas 8 patients were misclassified (i.e., incorrectly

predicted to have a stronger relation between BT skills and PA relative to DBT skills and PA, or vice versa). Examined as continuous variables, the correlation between the predicted (i.e., difference between the mean predicted BT skills-PA correlation and DBT skills-PA correlation) and observed (i.e., difference between the observed BT skills-PA correlation and DBT skills-PA correlation) difference variables was $r = 0.30$ ($p = 0.051$).

Discussion

Inspired by the potential of "personalized medicine" approaches to mental health treatment, a growing number of studies have leveraged machine learning algorithms in an effort to predict optimal treatment assignment for individual patients (Cohen & DeRubeis, 2018). To our knowledge, this is the first study to adapt such an approach to predict personalized skill-affect associations. Informed by patient clinical and demographic characteristics, a cross-validated ENR model achieved an accuracy of 81% (sensitivity = 92.6%; specificity = 62.5%) for predicting whether a given patient would exhibit a stronger association between use of BT skills vs. DBT skills and PA. Importantly, skill use and affect were assessed in the daily lives of patients via EMA. If replicated in a larger sample, such findings could inform personalized prescriptions for which skill domain is expected to have the greatest emotional benefit for a given patient. It is also worth noting that there was a positive (although nonsignificant) correlation (r = .30) between the difference in predicted BT skills-PA correlations vs. DBT skills-PA correlation and the observed (i.e., actual) differences in these two correlations. Assuming these findings are replicated, skill recommendations could be reserved for those with relatively larger predicted differences in skill-affect associations (e.g., if the DBT skills PA correlation is predicted to be meaningfully, rather than trivially, larger than the BT skills-PA correlation, and vice versa).

Ultimately, a prospective test is needed in which therapist-patient dyads are randomly assigned to algorithm-guided skills recommendation (i.e., therapists receive feedback from the algorithm indicating which skill domain is predicted be most therapeutically beneficial for their patient) vs. treatment as usual (i.e., the therapist receives no information from the algorithm regarding optimal skills assignment). To the extent that patients within the former condition experience better treatment outcomes than those in the latter condition, results would support the therapeutic benefits of algorithm-based personalized skills prescriptions. A range of important questions would need to be considered in such a trial. For example, for the therapists in the algorithm-guided condition, what percentage of the time are they adhering to the algorithm recommendation (vs. trusting their own clinical judgment regarding what skills to prescribe, which may or may not be consistent with the algorithm recommendation)? For the therapists in the treatment as usual condition, what percentage of the time is their clinical recommendation (uninformed by any algorithm) in fact consistent with the algorithm-derived recommendation? What should be the format, duration and treatment setting for the skills training? In each condition, what is the extent of skill acquisition for patients, and is that moderated by clinical characteristics (e.g., some patients struggle to acquire certain therapeutic skills, for a variety of reasons)?

The pattern of predictors that emerged is informative and partially consistent with psychotherapeutic models. For example, individuals engaging in non-suicidal self-injury (NSSI) exhibited stronger associations between DBT skills (but not BT skills) and levels of PA, relative to those without NSSI. DBT is a treatment specifically designed for individuals with intense affective and behavioral dysregulation, commonly characterized by NSSI (King et al., 2019). These findings suggest that individuals with a current history of NSSI are more likely to derive

affective benefits from engaging in DBT skills. Similarly, those with an anxiety disorder had a stronger relation between use of BT skills (i.e., behavioral activation/scheduling and exposure) and PA, relative to those without an anxiety disorder. This latter pattern of findings may not be surprising given the core therapeutic role of exposure in anxiety treatment. However, other skills-affect predictors that emerged are less clearly interpretable (e.g., greater sleep difficulties being associated with stronger DBT skills-PA associations).

Mean skill-affect associations were small (e.g., mean BT skills-PA r = .14; SD = .17). Sample (e.g., hospitalized patients with relatively severe levels of psychopathology and comorbidity, including personality pathology), setting (e.g., brief partial hospital treatment) and methodological features (e.g., assessment of a limited subset of skills, sampling affect at only 4 moments per day) may have contributed to the small mean associations. In particular, a broad array of factors (e.g., current activity, interpersonal context, recent or anticipated stressors), beyond recent use of BT or DBT skills, collectively contribute to current affect. Accordingly, multivariable models incorporating a combination of relevant predictors are needed to increase the variance accounted for in PA and NA. Moreover, other skill domains, not assessed in the present study, may be more predictive of affect. In addition, the present study focused on short-term PA and NA. A short-term increase in PA may not translate to long-term therapeutic benefit. For example, certain maladaptive behaviors (e.g., substance use, risky behaviors) can improve affect in the short-term, but be associated with longer-term negative outcomes. Conversely, exposure exercises may increase NA (and decrease PA) in the short-term, but be associated with longer-term therapeutic benefits (including lower mean levels of NA and higher PA). The present study did not consider other relevant outcomes beyond momentary affect (e.g., changes in specific clinical symptoms or functional impairment). Finally, we focused on momentary

affect assessed via EMA. Future studies could consider other EMA designs (e.g., event contingent sampling of affect when skills are used).

Although mean correlations were small, it is important to highlight that there was substantial variability across patients in these skill-affect associations (e.g., BT skills-PA range: r = -.23 to 0.58; see distribution in **Figure 2**). Thus, for some patients, there is a weak (or even inverse) association between use of BT skills and levels of PA; whereas for others there is a larger association. In addition, patients also differed substantially in the *difference* between BT skills-PA and DBT skills-PA relations (i.e., mean absolute difference r = .19; range: r = 0 to 0.90). For approximately a quarter of patients (24%) the difference was greater than r = .25. Ultimately, this model, and any subsequent adaptations, is likely only applicable to a certain subset of patients. In particular, the predicted skills-PA difference scores (i.e., difference between the predicted BT skills-PA correlation and DBT skills-PA correlation for every patient) were positively correlated (r = 0.30) with the corresponding observed values (i.e., difference between the observed BT skills-PA correlation and DBT skills-PA correlation for every patient). In other words, a larger predicted difference between the BT skills-PA correlation and DBT skills-PA correlation is associated with a larger observed (i.e., actual) difference between these associations. If the model predicts a small difference between these associations, a patient may be unlikely to benefit from using one skill set vs another. However, those patients with a larger predicted difference may be more likely to benefit from prioritizing the skill set (BT or DBT) prescribed by the model.

A key benefit of EMA relative to traditional laboratory-administered self-report questionnaires is the higher frequency of assessments, as well as minimizing risk of recall bias. However, a critical consideration in designing EMA studies is the time lag between surveys. In

the present study, the mean interval between completed surveys within each day was just under 3 hours. If the time course of the causal relation between use of specific skills (e.g., behavioral activation) and their effect on PA/NA is briefer (e.g., on the order of seconds or minutes) then EMA surveys spaced several hours apart, on average, would be too long to reliably capture these effects. For example, how soon after engaging in a scheduled pleasurable activity (i.e., behavioral activation) would one expect affect to improve (e.g., on the order of seconds), and how long would this effect be detectable after the activity ends (e.g., a few minutes or hours)? Undoubtedly, this relation would be moderated by a range of variables (e.g., the nature of the activity, intervening cognitions). In contrast, other skill-outcome relations may have a longer time lag (e.g., the effect of sleep hygiene skills on improved sleep). In sum, the frequency and time lag between EMA surveys should be based on a careful consideration of likely causal time course, as well as important practical considerations (e.g., not overburdening participants with too many surveys per day, which may negatively influence EMA compliance and the validity of responses) (Vachon et al., 2019).

It is also important to note that BT and DBT are related, overlapping approaches. In the present study, the DBT skills variable represented the 4 core skill domains (i.e., mindfulness, distress tolerance, emotion regulation, interpersonal effectiveness) taught in DBT (Linehan, 2014). In contrast, the BT skills variable was a composite of 3 skills (behavioral activation, behavioral scheduling and exposure). There are significant differences, but also overlap, between these two skill subsets. For example, mindfulness, distress tolerance and interpersonal effectiveness are not represented in the BT skills variable. However, the DBT skill of “emotion regulation” is a broad category (e.g., encompassing training in several specific skills, including problem solving, opposite action, checking the facts/cognitive reappraisal strategies,

accumulating positive emotions) (Linehan, 2014), which includes strategies that overlap with BT skills (e.g., opposite action and behavioral activation, accumulating positive emotions and behavioral scheduling).

There are several notable limitations. First, findings were based on a small sample size (although we had over 2,000 skill-affect observations in total given repeated assessments), in particular with regards to implementing cross-validation. Findings will need to be replicated in a significantly larger sample size, and with a more diverse set of participants (the present sample was 84% White). We consider this manuscript as presenting an initial empirical demonstration or "proof of concept" study, a first step in testing whether patient clinical and demographics characteristics can be used to predict patient-specific skill-affect associations. Second, reverse causation cannot be ruled out. For example, our model identified patients who exhibited a relatively strong association between self-reported use BT skills since the last EMA survey, and current levels of PA. For this subset of individuals, rather than use of BT skills *causing* increased PA, it may be that elevated PA influences self-reported or actual use of BT skills. In addition, it is important to note that a corresponding NA model yielded no significant predictors. Of relevance, a related recently published study (Rubel et al., 2019) attempted to predict patient-specific therapeutic alliance-outcome associations from baseline patient characteristics, but the resulting model proved unsuccessful. In contrast to the ENR approach used in the present study, the authors used a Boruta algorithm (a variant of random forest) for variable selection coupled with a nearest neighbor approach to generate predicted alliance-outcome correlations. However, in the latter study, predicted alliance-outcome correlations were not associated with the observed values. The extent to which the modeling approach used influences results is unclear. Moreover, rather examining the relatively broad BT and DBT skill domains, an alternative study design

could examine specific skill-affect associations *within* a treatment modality (e.g., cognitive skills-outcome vs. behavioral skills-outcome associations within CBT; Webb et al., 2019). Finally, we relied on patient-reported assessments of skills. Objective observer-rated measures of skills have been developed (Strunk et al., 2007; Webb et al., 2012) and could be used in future studies. In summary, the present study represents a promising initial demonstration of the potential of a data-driven approach to inform personalized prescriptions of which skill domains may be most therapeutically beneficial for a given patient. A future study is needed using a prospective design in which therapists are randomly assigned to algorithm-guided skills recommendations vs. treatment as usual.

Acknowledgments

[Omitted for blinded submission]

References

Beck, A. T., Rush, J. A., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression*. Guilford Press.

Bylsma, L. M., & Rottenberg, J. (2011). Uncovering the dynamics of emotion regulation and dysfunction in daily life with ecological momentary assessment. In *Emotion regulation and well-being* (pp. 225–244). Springer.

Cameron, I. M., Cunningham, L., Crawford, J. R., Eagles, J. M., Eisen, S. V., Lawton, K., Naji, S. A., & Hamilton, R. J. (2007). Psychometric properties of the BASIS-24© (Behaviour and Symptom Identification Scale–Revised) Mental Health Outcome Measure. *International Journal of Psychiatry in Clinical Practice*, *11*(1), 36–43. https://doi.org/10.1080/13651500600885531

Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment Selection in Depression. *Annual Review of Clinical Psychology*, *14*(1), 209–236. https://doi.org/10.1146/annurev-clinpsy-050817-084746

Cohen, Z. D., Kim, T. T., Van, H. L., Dekker, J. J. M., & Driessen, E. (2019). A demonstration of a multi-method variable selection approach for treatment selection: Recommending cognitive–behavioral versus psychodynamic therapy for mild to moderate adult depression. *Psychotherapy Research*, *0*(0), 1–14. https://doi.org/10.1080/10503307.2018.1563312

Cook, J. M., Biyanova, T., Elhai, J., Schnurr, P. P., & Coyne, J. C. (2010). What do psychotherapists really do in practice? An internet study of over 2,000 practicioners. *Psychotherapy (Chicago, Ill.)*, *47*(2), 260–267. https://doi.org/10.1037/a0019788

DeRubeis, R. J. (2019). The history, current status, and possible future of precision mental health. *Behaviour Research and Therapy*, *123*, 103506. https://doi.org/10.1016/j.brat.2019.103506

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: Translating Research on Prediction into Individualized Treatment Recommendations. A Demonstration. *PLOS ONE*, *9*(1), e83875. https://doi.org/10.1371/journal.pone.0083875

DiGiorgio, K. E., Glass, C. R., & Arnkoff, D. B. (2010). Therapists' Use of DBT: A Survey Study of Clinical Practice. *Cognitive and Behavioral Practice*, *17*(2), 213–221. https://doi.org/10.1016/j.cbpra.2009.06.003

Forbes, E. E., Hariri, A. R., Martin, S. L., Silk, J. S., Moyles, D. L., Fisher, P. M., Brown, S. M., Ryan, N. D., Birmaher, B., Axelson, D. A., & Dahl, R. E. (2009). Altered Striatal Activation Predicting Real-World Positive Affect in Adolescent Major Depressive Disorder. *American Journal of Psychiatry*, *166*(1), 64–73. https://doi.org/10.1176/appi.ajp.2008.07081336

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*(1), 72.

Hollanders, H., & McLeod, J. (1999). Theoretical orientation and reported practice: A survey of eclecticism among counsellors in Britain. *British Journal of Guidance & Counselling*, *27*(3), 405–414. https://doi.org/10.1080/03069889908256280

Huibers, M. J. H., Cohen, Z. D., Lemmens, L. H. J. M., Arntz, A., Peeters, F. P. M. L., Cuijpers, P., & DeRubeis, R. J. (2015). Predicting Optimal Outcomes in Cognitive Therapy or Interpersonal Psychotherapy for Depressed Individuals Using the Personalized Advantage Index Approach. *PLOS ONE*, *10*(11), e0140771. https://doi.org/10.1371/journal.pone.0140771

Kertz, S. J., Bigda-Peyton, J. S., Rosmarin, D. H., & Björgvinsson, T. (2012). The importance of worry across diagnostic presentations: Prevalence, severity and associated symptoms in a partial hospital setting. *Journal of Anxiety Disorders*, *26*(1), 126–133. https://doi.org/10.1016/j.janxdis.2011.10.005

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Ebert, D. D., de Jonge, P., Nierenberg, A. A., Rosellini, A. J., Sampson, N. A., Schoevers, R. A., Wilcox, M. A., & Zaslavsky, A. M. (2017). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiology and Psychiatric Sciences*, *26*(1), 22–36. https://doi.org/10.1017/S2045796016000020

Khan, M. R. A., & Brandenburger, T. (2019). *ROCit: Performance Assessment of Binary Classifier with Visualization* (1.1.1) [Computer software]. https://CRAN.R-project.org/package=ROCit

King, A. M., Rizvi, S. L., & Selby, E. A. (2019). Emotional experiences of clients with borderline personality disorder in dialectical behavior therapy: An empirical investigation of in-session affect. *Personality Disorders: Theory, Research, and Treatment*, *10*(5), 468.

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals*, *32*(9), 509–515. https://doi.org/10.3928/0048-5713-20020901-06

Kuhn, M. (2008). Caret package. *Journal of Statistical Software*, *28*(5), 1–26.

Linehan, M. M. (1993). *Cognitive-Behavioral Treatment of Borderline Personality Disorder* (1 edition). The Guilford Press.

Michalak, J., Steinhaus, K., & Heidenreich, T. (2018). (How) Do Therapists Use Mindfulness in Their Clinical Work? A Study on the Implementation of Mindfulness Interventions. *Mindfulness*. https://doi.org/10.1007/s12671-018-0929-9

Nezlek, J. B. (2007). A multilevel framework for understanding relationships among traits, states, situations and behaviours. *European Journal of Personality*, *21*(6), 789–810. https://doi.org/10.1002/per.640

Nezlek, J. B., & Gable, S. L. (2001). Depression as a Moderator of Relationships between Positive Daily Events and Day-to-Day Psychological Adjustment. *Personality and Social Psychology Bulletin*, *27*(12), 1692–1704. https://doi.org/10.1177/01461672012712012

Norcross, J. C., Beutler, L. E., & Goldfried, M. R. (2019). Cognitive-Behavioral Therapy and Psychotherapy Integration. *Handbook of Cognitive-Behavioral Therapies*, 318.

Peeters, F., Berkhof, J., Rottenberg, J., & Nicolson, N. A. (2010). Ambulatory emotional reactivity to negative daily life events predicts remission from major depressive disorder. *Behaviour Research and Therapy*, *48*(8), 754–760. https://doi.org/10.1016/j.brat.2010.04.008

Posner, K., Brown, G. K., Stanley, B., Brent, D. A., Yershova, K. V., Oquendo, M. A., Currier, G. W., Melvin, G. A., Greenhill, L., Shen, S., & Mann, J. J. (2011). The Columbia–Suicide Severity Rating Scale: Initial Validity and Internal Consistency Findings From Three Multisite Studies With Adolescents and Adults. *The American Journal of Psychiatry*, *168*(12), 1266–1277. https://doi.org/10.1176/appi.ajp.2011.10111704

Rubel, J. A., Zilcha-Mano, S., Giesemann, J., Prinz, J., & Lutz, W. (2019). Predicting personalized process-outcome associations in psychotherapy using machine learning approaches—A demonstration. *Psychotherapy Research*, 1–10.

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., & Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of Clinical Psychiatry*, *59 Suppl 20*, 22-33;quiz 34-57.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, *21*(20), 3940–3941.

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Strunk, D. R., DeRubeis, R. J., Chiu, A. W., & Alvarez, J. (2007). Patients' competence in and performance of cognitive therapy skills: Relation to the reduction of relapse risk following treatment for depression. *Journal of Consulting and Clinical Psychology*, *75*(4), 523–530. https://doi.org/10.1037/0022-006X.75.4.523

Strunk, D. R., Hollars, S. N., Adler, A. D., Goldstein, L. A., & Braun, J. D. (2014). Assessing Patients' Cognitive Therapy Skills: Initial Evaluation of the Competencies of Cognitive Therapy Scale. *Cognitive Therapy and Research*, *38*(5), 559–569. https://doi.org/10.1007/s10608-014-9617-9

Vachon, H., Viechtbauer, W., Rintala, A., & Myin-Germeys, I. (2019). Compliance and Retention With the Experience Sampling Method Over the Continuum of Severe Mental

Disorders: Meta-Analysis and Recommendations. *Journal of Medical Internet Research*, *21*(12), e14475. https://doi.org/10.2196/14475

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063.

Webb, C. A., Cohen, Z. D., Beard, C., Forgeard, M., Peckham, A. D., & Björgvinsson, T. (2020). Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: A comparison of machine learning approaches. *Journal of Consulting and Clinical Psychology*, *88*(1), 25–38. https://doi.org/10.1037/ccp0000451

Webb, C. A., DeRubeis, R. J., Dimidjian, S., Hollon, S. D., Amsterdam, J. D., & Shelton, R. C. (2012). Predictors of patient cognitive therapy skills and symptom change in two randomized clinical trials: The role of therapist adherence and the therapeutic alliance. *Journal of Consulting and Clinical Psychology*, *80*(3), 373–381. https://doi.org/10.1037/a0027663

Webb, C. A., Trivedi, M. H., Cohen, Z. D., Dillon, D. G., Fournier, J. C., Goer, F., Fava, M., McGrath, P. J., Weissman, M., Parsey, R., Adams, P., Trombello, J. M., Cooper, C., Deldin, P., Oquendo, M. A., McInnis, M. G., Huys, Q., Bruder, G., Kurian, B. T., … Pizzagalli, D. A. (2018). Personalized prediction of antidepressant v. placebo response: Evidence from the EMBARC study. *Psychological Medicine*, 1–10. https://doi.org/10.1017/S0033291718001708

Zanarini, M. C., Vujanovic, A. A., Parachini, E. A., Boulanger, J. L., Frankenburg, F. R., & Hennen, J. (2003). A screening measure for BPD: The McLean Screening Instrument for

Borderline Personality Disorder (MSI-BPD). *Journal of Personality Disorders*, *17*(6), 568–573.

Zilcha-Mano, S., Keefe, J., Chiu, H., Rubin, A., Barrett, M. S., & Barber, J. P. (2016). Reducing Dropout in Treatment for Depression: Translating Dropout Predictors Into Individualized Treatment Recommendations. *The Journal of Clinical Psychiatry*, *77*(12), e1584–e1590. https://doi.org/10.4088/JCP.15m10081

Table 1.

Demographic Characteristics for the Sample (N = 50).

	N	**(%)**
Gender		
Female	27	(54)
Male	22	(44)
Non-binary or Genderfluid	1	(2)
Ethnicity/Race		
Asian	6	(12)
Multiracial	2	(4)
White	42	(84)
Sexual Orientation		
Bisexual	7	(14)
Gay / Lesbian	4	(8)
Heterosexual/Straight	38	(76)
Not listed	1	(2)
Education		
Some College	17	(34)
College Graduate	19	(38)
Post-college Education	14	(28)
Employment		
Not Employed	20	(40)
Employed Full-time	9	(18)
Employed Part-time	21	(42)
Marital Status		
Never Married	31	(62)
Married or Living with Partner	15	(30)
Separated/ Divorced or Widowed	4	(8)
Psychiatric Hospitalization (last 6 mos)	19	(38)
Age (*M*, *SD*)	36.82	(14.42)

Table 2.

Clinical Characteristics for the Sample (N = 50)

Clinical Characteristics	***N***	**(%)**
Current Diagnosis		
MDE within MDD	37	(74)
MDE within Bipolar Disorder	5	(10)
Generalized Anxiety Disorder	33	(66)
Social Anxiety Disorder	20	(40)
Obsessive-Compulsive Disorder	14	(28)
Panic Disorder	9	(18)
Post-Traumatic Stress Disorder	7	(14)
Alcohol Dependence or Abuse	7	(14)
Psychotic Disorder	0	(0)
Number of Diagnoses		
0	3	(6)
1	9	(18)
2	13	(26)
3	11	(22)
4	8	(16)
5	5	(10)
6	0	(0)
7	1	(2)

Notes. MDE = Major Depressive Episode; MDD = Major Depressive Disorder. Diagnoses were established using the Mini International Neuropsychiatric Interview for DSM-IV-TR (MINI; Sheehan et al., 1998).

Table 3. *Baseline predictors submitted to models, including estimates of internal consistency for multi-item scales.*

Predictors Submitted to Elastic Net Models	
PHQ-9 Total (α = .88)	BASIS-24 Substance Abuse (α = .62)
PHQ-9 Interest/Pleasure	BASIS-24 Relationships (α = .74)
PHQ-9 Depressed Mood	BASIS-24 Psychosis (α = .57)
PHQ-9 Insomnia/Hypersomnia	MSI-BPD Total (α = .74)
PHQ-9 Fatigue	Suicidality in the past month (CSSRS)
PHQ-9 Appetite	Non-suicidal self-injury in the past month (CSSRS)
PHQ-9 Self-Esteem/Guilt	Any anxiety disorder (yes/no)
PHQ-9 Concentration	GAD-7 Total (α = .91)
PHQ-9 Psychomotor symptoms	Prior inpatient treatment (yes/no)
PHQ-9 Suicidality	Sex (male/female)
BASIS-24 Self-Harm (α = .83)	Age
BASIS-24 Emotional Lability (α = .58)	Education level

Note. PHQ-9 = 9-item Patient Health Questionnaire (Kroenke & Spitzer, 2002); BASIS-24 = 24-item Behavior and Symptom Identification Scale (Cameron et al., 2007); GAD-7 = 7-item Generalized Anxiety Disorder scale (Spitzer et al., 2006); MSI-BPD = McLean Screening Instrument for Borderline Personality Disorder (Zanarini et al., 2003); CSSRS = Columbia Suicide Severity Rating Scale (Posner et al., 2011); The PHQ-9, BASIS-24 and GAD-7 were assessed at discharge, the remaining variables (e.g., demographics, treatment history) were assessed at intake.

Table 4. *Variables retained in elastic net models*

Predictors of BT skills-PA relation	***B***
(Intercept)	0.12
BASIS-24 Emotional Lability	0.04
Any anxiety disorder (yes/no)	0.06
PHQ-9 Self-Esteem/Guilt	-0.02
PHQ-9 Psychomotor symptoms	-0.03
Predictors of DBT skills-PA relation	***B***
(Intercept)	0.05
Non-suicidal self-injury (past month)	0.06
PHQ-9 Insomnia/Hypersomnia	0.03

Note. BASIS-24 = 24-item Behavior and Symptom Identification Scale; PHQ-9 = 9-item Patient Health Questionnaire. Given that 10,000 elastic net regularization (ENR) models were run (1,000 repetitions of 10-folds cross-validation), each model may have retained a different set of predictors. In this table, we present those predictors that were retained in at least 90% of the 10,000 models. Beta values represent the median across the 10,000 models.

Figure Captions

Figure 1. Schematic of the data analysis pipeline.

Figure 2. Distribution of observed BT skills-PA correlations (red) and DBT skills-PA (green) correlations.

Figure 3. ROC curve (Area Under the Curve = 0.80) representing the ability to discriminate between individuals with a stronger relation between BT skills and PA vs. DBT skills and PA.

Figure 1

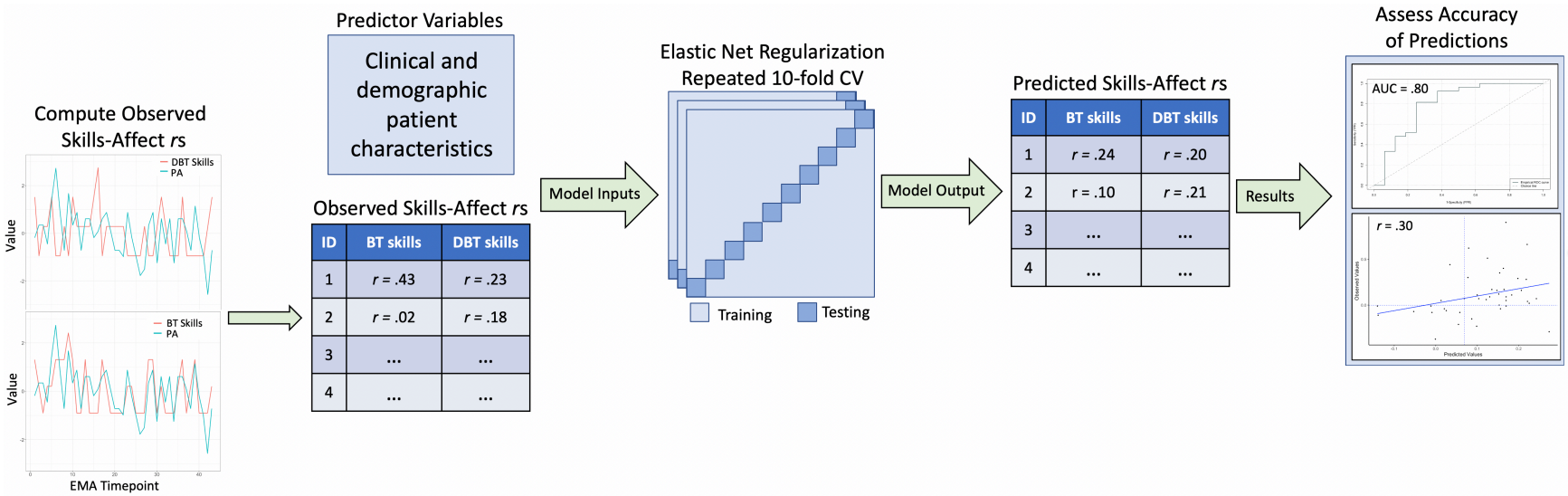

Compute Observed Skills-Affect rs
DBT Skills
PA
BT Skills
PA
Value
Value
EMA Timepoint
Predictor Variables
Clinical and demographic patient characteristics
Observed Skills-Affect rs
ID | BT skills | DBT skills
1 | r = .43 | r = .23
2 | r = .02 | r = .18
3 | ... | ...
4 | ... | ...
Model Inputs
Elastic Net Regularization
Repeated 10-fold CV
Training
Testing
Model Output
Predicted Skills-Affect rs
ID | BT skills | DBT skills
1 | r = .24 | r = .20
2 | r = .10 | r = .21
3 | ... | ...
4 | ... | ...
Results
Assess Accuracy of Predictions
AUC = .80
r = .30

Figure 2

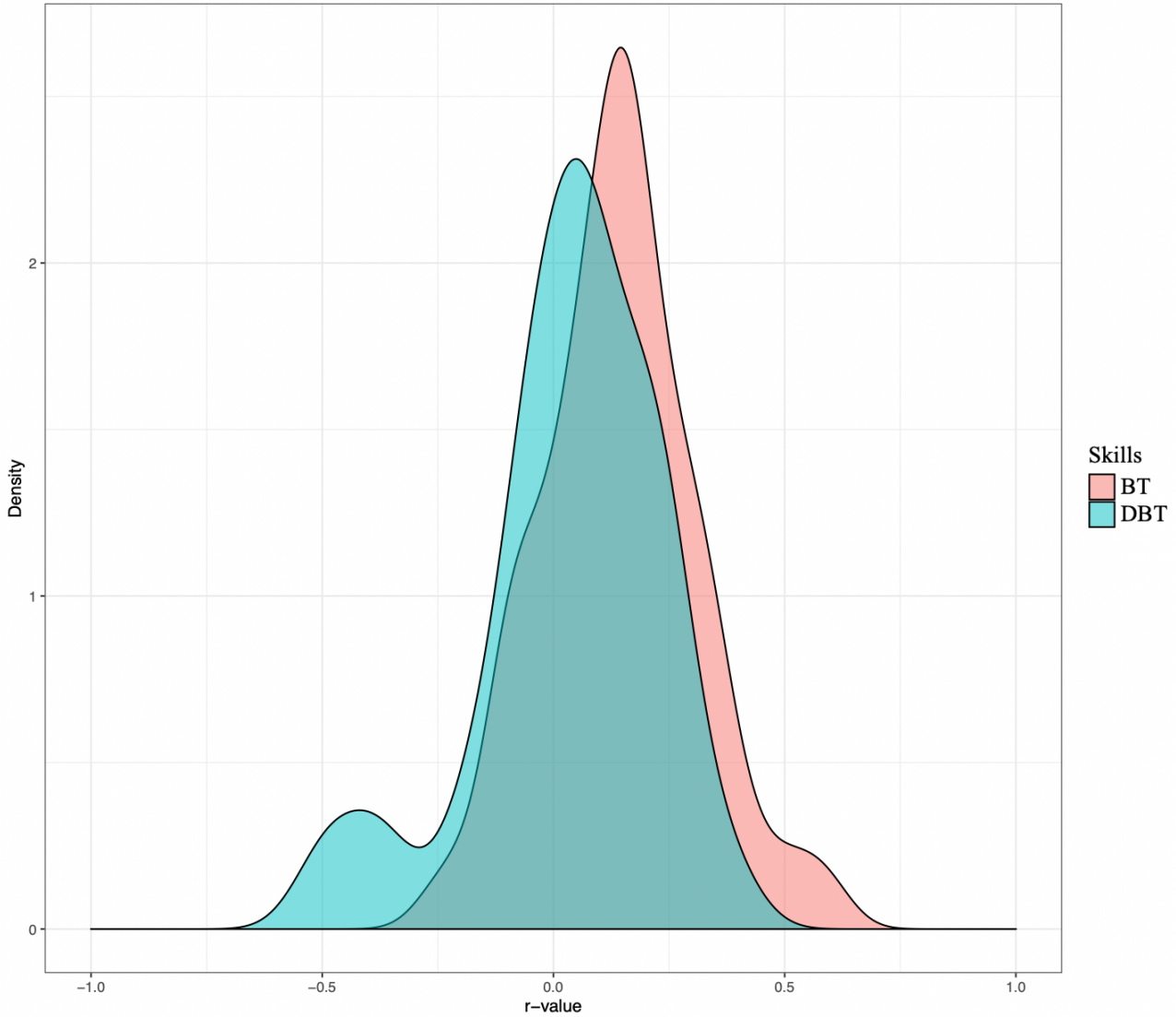
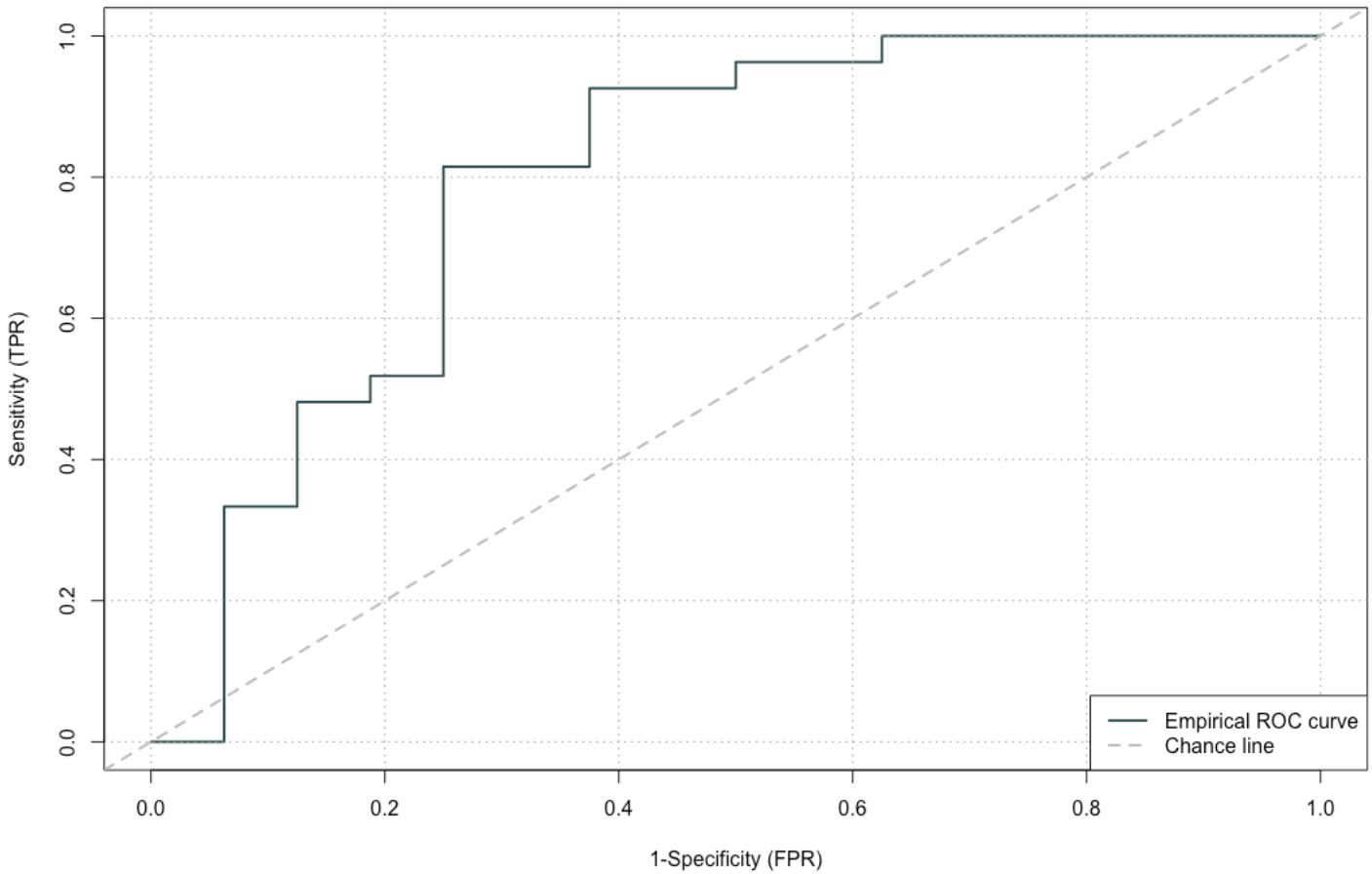


Figure 3


Sensitivity (TPR)
0.0
0.2
0.4
0.6
0.8
1.0
1-Specificity (FPR)
0.0
0.2
0.4
0.6
0.8
1.0
Empirical ROC curve
Chance line

Appendix

Data Transparency Statement: The data reported in this manuscript were collected as part of a larger ecological momentary assessment (EMA) project. One manuscript has been published from this dataset, which focused on a different set of research questions (i.e., the relationship between stress, social contact, skills use and affect following hospital discharge) in a partially overlapping sample (44% of participants were shared between both studies). In contrast, the current manuscript used information from 24 clinical and demographic characteristics (see Table 3) to develop personalized predictions of the relationship between use of skills (DBT and BT skills) and positive vs. negative affect.