Published in final edited form as:

Med Care. 2015 November ; 53(11): 989–995. doi:10.1097/MLR.0000000000000429.

Do Patients at Sites with High RCT Enrollment Propensity have Better Outcomes?

A. James O'Malley, Ph.D.^{1,*}, Katya Zelevinsky, B.A.², Yulei He, Ph.D.³, and Alisa B. Busch, MD, MS^{6,4,5}

¹Department of Biomedical Data Science and The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine at Dartmouth, 35 Centerra Parkway, Lebanon, NH 03766, USA.

²Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, USA. zelevins@hcp.med.harvard.edu; Phone: 617-432-0372; Fax: 617-432-0173

³Office of Research and Methodology, National Center of Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD 20782-2064, USA. wdq7@cdc.gov; Phone: 1-800-232-4636; Fax: 1-888-232-6348

⁶Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, USA. abusch@hcp.med.harvard.edu; Phone: (617) 855-2989; Fax: 617-432-0173

⁴McLean Hospital, Belmont, MA, 115 Mill St., Belmont, MA 02478

⁵Department of Psychiatry, Harvard Medical School

Abstract

Background—Concerns about randomized controlled trial (RCT) generalizability typically center on characteristics of RCT patient participants. Possibly there are RCT site characteristics that distinguish RCT outcomes from those that can be expected in non-RCT settings.

Objectives—To examine whether site **propensity** toward RCT enrollment is associated with **recovery** outcomes for patients and **if the association differs between patients who** participate in a RCT compared to those who remain in an observational treatment environment.

Data—Study participants with acute bipolar-depression from The Systematic Treatment Enhancing Program for Bipolar Disorder acute depression pharmacotherapy RCT (N=337) and observational treatment arm (N=1581).

Methods—A longitudinal observational study comparing the likelihood of recovery in the RCT to the observational arm, allowing effect modification by site **high** RCT enrollment **propensity (defined as > the median)** and other predictors over a six-month follow-up period.

Results—Non-RCT participants who received care in sites with high RCT enrollment propensity had a higher probability of recovering from their bipolar-depression episode compared to participants from low propensity sites (OR [95%CI] = 2.13 [1.28-3.55]). RCT enrollment propensity was not associated with recovery outcomes for RCT participants (1.03 [0.35-3.03]).

*Corresponding author James.OMalley@Dartmouth.edu; Phone: 603-653-0854; Fax: 603-653-0896..

Conclusions—Sites with high propensity to enroll patients in RCTs appear to have unobserved characteristics which play a significant role in outcomes for non-RCT patients. For RCT participants in low-enrollment sites, possibly RCT protocols, which proscribe care delivery and monitoring, attenuate this effect. These results have implications for future research to improve outcomes in non-research care settings.

Keywords

Bipolar disorder; Effect modification; Generalizability; **Mixed**-effects model; STEP-BD

INTRODUCTION

Concerns about the generalizability of randomized controlled trial (RCT) outcomes to usual care stem from the observation that RCT populations do not accurately reflect the diversity of patients or providers delivering treatment in routine care settings.¹⁻⁴ Possibly, within this heterogeneity are patient or provider/setting characteristics that affect outcomes, perhaps differently between RCT and usual care settings. The increasing focus on improving healthcare quality demands a better understanding of which treatments will improve outcomes for usual care patients. Thus, efforts to pinpoint the variables that compromise generalizability are important in understanding how best to translate RCT findings into usual care settings and populations.

The National Institute of Mental Health (NIMH) funded Systematic Treatment Enhancing Program for Bipolar Disorder (STEP-BD) offers a unique opportunity to examine concerns of generalizability in a large observational study in which RCTs that are embedded within the larger observational study structure recruit from the same patient and provider/clinic population.⁵ Previously, we investigated how representative STEP-BD RCT participants were compared to participants in its observational arm and found that study site (representing unmeasured site-level factors) was the largest predictor.⁶ Therefore, an important follow-up question is whether site is associated with important patient outcomes (i.e., is a confounder of the patient-level effect of RCT enrollment) and the extent to which any such association is accounted for in observed quantities such as site RCT enrollment propensity. In this paper, we evaluate whether sites most aggressive at enrolling their patients in an RCT are associated with better outcomes and if any such association differs between RCT and observational environments, where the latter is thought of as a surrogate for usual care settings. The importance of identifying RCT-usual care effect heterogeneity is that it provides a mechanism for more accurately generalizing results from RCTs to usual care or, alternatively, of gaining insight into why generalizability may be compromised. In this paper we specifically examine if site high RCT enrollment propensity is associated with better outcomes and if so whether the association is different for RCT participants compared to the observational care population. This investigation is motivated by a basic model in which site high RCT enrollment propensity is a surrogate for various unobserved site traits that influence care delivery to patients and ultimately patient outcomes. We hypothesize that sites with high RCT enrollment have characteristics or approaches to patient care that lead to better outcomes (e.g., perhaps more or better staffing resources, or an overall greater ability of the site to assign appropriate treatment and provide appropriate care). Because RCT care

is subjected to study protocols (e.g., involving the evaluation and monitoring of patients), we secondly hypothesize that site high RCT enrollment propensity will have a larger effect on patient recovery in the STEP-BD observational arm than its RCT.

METHODS

Data Source

The STEP-BD aimed to conduct research in a diverse bipolar disorder population. As a rough initial screening tool, sites were required to have bipolar specialty programs with at least 100 active patients. Then sites were selected based on geographic balance and demographic diversity.⁷ STEP-BD included an observational arm which served as a pool of bipolar disorder patients for whom clinical treatments and course were monitored but who could also choose to participate in several RCTs, if clinically appropriate. In the observational arm, providers received additional education about bipolar disorder treatment, but treatment options were not constrained.⁵ Twenty-two STEP-BD sites were recruited from across 12 states. To ensure that all patients were eligible for the outcome, we restricted our sample to patients who had observations through to at least week 5 of follow-up; N=17 sites with at least one RCT and one observational arm participant remained (Figure 1 and see Outcome section).

STEP-BD participants gave informed consent at each site to participate in the observational arm, and additional consent if participating in an RCT. Further IRB approval was obtained from McLean Hospital, Harvard Medical School, and Geisel School of Medicine at Dartmouth College for this analysis.

We focus on the comparison of participants enrolled in the STEP-BD pharmacotherapy RCT for the treatment of acute bipolar depression to STEP-BD participants with acute depression who never enrolled in the RCT.⁷ The goal of the RCT was to test whether antidepressant medication, adjunctive with a mood stabilizer or antipsychotic, improved bipolar-depression symptoms compared to mood stabilizers/antipsychotics plus a placebo. The findings were a null result, that is, the intervention and control arms yielded similar outcomes. We chose this RCT because it had the largest sample size among the STEP-BD RCTs, maximizing our ability to compare outcomes between the RCT and observational arm care.

RCT participants were required to be adults age 18 who met the American Psychiatric Association's Diagnostic Statistical Manual-IV (DSM-IV) criteria⁸ for bipolar-I or -II disorder. Bipolar diagnoses were determined by a modified Structured Clinical Interview for DSM Disorders (SCID)⁹, and confirmed by the Mini-International Neuropsychiatric Interview (MINI).¹⁰ Study participants were also required to meet the DSM-IV criteria for a major depressive episode within bipolar disorder. RCT exclusion criteria were individuals: requiring short term intensive treatment for a co-occurring substance use disorder; having a history of either lack of response or non-tolerance of the antidepressant study medications; and requiring either a dose change to a long-standing antipsychotic prescription, or addition of a new antipsychotic medication. Time-varying clinical characteristics (such as mood state [e.g., depressed, manic], symptom severity scores (e.g., the Clinical Global Impression-Severity [CGI-S] scales) were noted on the Clinical Monitoring Form (CMF), a standardized

progress note for STEP-BD observational and RCT study arms.¹¹ More complete details of the STEP-BD pharmacotherapy RCT study criteria and protocol can be found in Sachs et al.⁷

Because the goal **was** to compare outcomes between those diagnosed with bipolar-depression but never enrolled in the RCT versus those enrolled in the RCT, we sought to apply eligibility criteria to the never enrolled in the RCT cohort that mimicked criteria used for the RCT by STEP-BD investigators. We first defined an “index acute bipolar-depressed visit.” For the RCT, we defined the index visit as the CMF that was completed closest to the date of randomization. Preliminary analyses indicated that this occurred ± 7 days of randomization for 92.25% of the RCT sample. We then selected from the pool of observational study arm participants those newly diagnosed with an acute depressive episode not previously enrolled in the RCT. **In so doing, we used their** first observed “depression” determination in a CMF as the index acute bipolar-**depression** CMF visit for **each** observational arm **participant**.

The RCT protocol called for four follow-up visits over the first 6 weeks. Then, participants who responded to treatment received monthly visits for the next 20 weeks, while those who did not received biweekly visits (i.e., more intensive monitoring) over the next 10 weeks.⁷ Therefore, we include RCT study outcomes for up to 26 weeks following a participant's index CMF. In the observational arm there was no specific protocol under which treatment visits were planned but to retain parity with the RCT we followed participants for 26 weeks from their index CMF. We excluded participants without a CMF, or RCT participants for whom we were not able to identify the date for RCT consent.

We treated multiple observations in a week as separate observations (i.e., repeated measurements), allowing both to contribute to the precision of an individual's status that week.

Outcome

Our outcome of interest was whether a participant was recorded as *recovered* (yes or no) in a given week. In STEP-BD, *recovered* was defined as having no more than two symptoms meeting DSM-IV threshold criteria for a mood episode and no significant symptoms present for 4 weeks. This differs from the primary outcome used by the RCT investigators, which required an eight-week period of “recovered”.⁷ Fewer visits in the observational arm make our analysis amenable to a four rather than an eight-week recovered period. Because individuals could not meet our outcome “recovered” until week 5, our analysis studies patients beginning week 5. The CMF recorded mood status at each visit with the STEP-BD psychiatrist. We reduced the eight-category clinical status measure used in the STEP-BD observational and RCT arms to a three level measure reflecting the mood symptom severity/course: the first level includes the four *diagnostic* states (depression, mania, mixed/cycling, hypomania) and continued symptoms, the second level consists of roughening and recovering *process* states (transition points from and to recovered), and the third is *recovered*. This grouping was chosen to avoid levels that occurred rarely and maximize the ratio of within- to between-group transitions between CMFs.

Predictors

The predictor of primary interest is *high enrollment propensity*, the indicator of whether the proportion of STEP-BD participants at a site exceeded the median across the sites. The reduction of a site's RCT enrollment propensity to a binary variable was performed because the relationship to the outcome did not appear to be purely linear and because **binary predictors can be more easily interpreted**. (The Appendix explores the sensitivity of the results to instead representing **site RCT enrollment propensity** as a continuous predictor in the model.)

To estimate the association of site RCT enrollment propensity on patient outcomes, high enrollment propensity is included in the model as a single main effect predictor. To test for effect heterogeneity between the RCT and observational arms, an interaction term of high enrollment propensity with RCT participation is added to the model. Alternatively, to obtain stratified effects, separate high enrollment propensity variables for the RCT and observational arm participants **are included in the model**. **The difference of** the coefficient for high enrollment propensity between RCT participants and observational arm participants corresponds to the interaction effect.

Demographic predictors obtained upon entry to STEP-BD include age, gender, race/ethnicity (Black, Hispanic, or Other), education (less than high school, high school graduate, some college, or college graduate or more), income (divided by the median income of < **versus** \$40,000), and insurance type (private insurance, Medicare only, Medicaid only, both Medicare and Medicaid, or none). Additional baseline predictors measured co-occurring psychiatric and general medical conditions that could complicate or otherwise influence bipolar disorder medication choices (psychiatric comorbidities: substance use **disorders**, anxiety disorders, attention deficit hyperactivity, and eating disorders; general medical comorbidities: pregnancy or hepatic, renal, pancreatic, seizure, thyroid or inflammatory disorders). Psychiatric comorbidity, including past substance use disorder, was derived from the MINI;¹⁰ general medical comorbidities were based on participant self-report. Comorbidities (psychiatric and general medical) were represented as separate categorical variables (e.g., 0, 1, or 2+ co-occurring conditions). Other clinical predictors obtained at baseline were: the bipolar disorder symptom history from the Bipolarity Index [BPI] and the Clinical Global Impression-Severity [CGI-S]).¹² The CGI-S ranges from 1-7, with higher scores indicating greater illness severity. For the model, we categorized the CGI-S as: 1-3 = no or mild symptoms, 4 = moderate symptoms and 5-7 = severe symptoms. Unlike the demographic and health status predictors obtained from the ADE, the CGI-S is a time-varying clinician rated scale. However, we only use the baseline values (measured on or before week 1) as predictors. Notably, **in the RCT**, Sachs et al. found no difference in outcomes between the study medications (bupropion and paroxetine) or placebo.⁷ Therefore, we ignore the treatment a participant received within the RCT.

Although we do not include bupropion, paroxetine, or placebo treatment indicators in our statistical model, some of the RCT subjects also participated in an adjunctive psychotherapy study that found that the studied psychotherapies improved outcomes.¹³ To adjust for such effects we included psychotherapy study participation as a predictor in the present study.

Occasionally, multiple CMFs were available on the same day, possibly due to the information being entered by both the RCT and observational arm data managers. We selected one CMF per day to be included in the analysis and defined a hierarchy in which the **(pharmacotherapy) RCT CMF took precedence over the psychosocial RCT CMF and in-turn the psychosocial study CMF took precedence over the observational arm CMF.**

Statistical Analysis

Our primary analysis conditions on baseline covariates and **estimates** the probability of a participant being “Recovered” in a given week for weeks 5 - 26. We use a longitudinal design because preliminary analyses of the data found that out of **850** times when an individual was in the recovered state prior to their final observed CMF, on **233** (approximately **27%**) they were non-Recovered (i.e., in either the *diagnostic* or *process* sub-categories of the clinical status measure) at their next CMF (Appendix Table A1).

Therefore, subsequent observations contain additional information **that can be utilized in a longitudinal modeling approach.** We compressed the time-scale across the weeks of follow-up so that linear trends would fit the data (see Appendix Figure A and accompanying text for details), allowing straight-forward interpretation of estimates.

As described in the Predictors section, we characterized site RCT enrollment propensity based on a binary variable, *high enrollment propensity*, which indicated whether a site's enrollment propensity exceeded the median of the sites. (See Appendix Table A2 and accompanying text in Appendix A2 for sensitivity analyses that use a Winsorized transformation of site RCT enrollment propensity as an alternative characterization of this predictor). We first estimate a single overall effect of high enrollment propensity and then develop a model allowing different regression coefficients of high enrollment propensity between the RCT and observational arms.

To account for individual-level heterogeneity and the varied frequency and timing of follow-up observations between enrollees, **each** model **included** random effects for individual-specific intercepts and time-slopes. The random **effects** component of the model allows **data for all** individuals with at least two CMFs to be used in the analysis irrespective of which weeks they had CMFs. Thus, in a sense there **are** no missing outcomes as the time-slope for an individual imputes outcomes for them in weeks for which they did not have a CMF. Missing values of predictors **were confined** to two patients contributing four observations in the RCT arm who had unknown ages. Because the amount of missing data was minimal we excluded these four observations as opposed to performing multiple imputation.

We also estimated a model with site random effects in addition to patient random effects. We found that no unexplained variation existed at the site level (results not reported) implying that it was sufficient to account only for heterogeneity between patients within sites in our primary analyses.

RESULTS

Upon **applying the STEP-BD exclusion criteria and** restricting the **sample** to sites with at least one RCT and one observational (OBS) arm member with a follow-up observation in week 5-26, our final study sample consisted of **1,916** individuals (**335** RCT; **1,581** OBS) contributing **9,236** (**2,391** RCT; **6,845** OBS) **complete non-missing** observations **over weeks 5-26 (Figure 1)**. An average of 7.11 and 4.33 observations per participant were obtained from the RCT **arm** and the **OBS** arm, respectively, reflecting that RCT participants were seen more frequently. Some participants were followed for less than six months; the average weeks of follow-up (capped at 26) for participants in the RCT and OBS arms was 21.3 and 19.9, respectively (Table 1).

Table 1 shows unadjusted means of the predictors stratified by high enrollment propensity (High = Above Median, Low = Below Median) and participant study type (RCT arm, OBS arm). The sample is composed primarily of participants who are white, college graduates, and privately insured. They typically have moderate symptoms, a clear bipolar history, and no general medical conditions that commonly complicate bipolar pharmacotherapy choices. The sample is more balanced with respect to gender (60% female).

Site **high RCT** enrollment propensity has a significant overall effect on Recovered (odds-ratio [95% confidence interval] = (1.90 [1.19, 3.03]). In a further analysis that allowed for differential effects by study arm, we found that high RCT enrollment propensity has a strong positive association with Recovered for participants who do not participate in the RCT (**2.13 [1.28, 3.55]**) despite having no association for RCT patients (**1.03 [0.35, 3.03]**) (Table 2). The smaller number of RCT participants meant that there was limited power to find a statistically significant difference in the high RCT enrollment propensity coefficient between the RCT and observational arms ($p = 0.23$). The significantly larger coefficient of the time trend for the RCT compared to observational participants (**5.49 [3.83, 7.89]** vs 2.95 [2.46, 3.54]; p -value of the difference = 0.003), implies that the RCT patients recovered more quickly than those in the observational arm. However, the flatness of Figure 2 beyond week 15 suggests that the overall proportion of individuals who recover at some point is about the same across the arms – a result consistent with Sachs et al.⁷

DISCUSSION

Our primary finding is that unmeasured factors associated with **site RCT enrollment propensity** are associated with **greater** clinical improvement in patient populations who are not RCT participants at that site. However, **site high RCT enrollment propensity** was not associated with **differential outcomes among RCT patients**. We conjecture that **RCT enrollment propensity may** reflect the culture at a site that is conducive to improved patient outcomes. However, because RCTs have strict protocols, site culture might be neutralized in the RCT environment. Several reasons have been posited in the literature that would explain why observational arm participants at a site with a high RCT enrollment propensity may do better than their counterparts at a site with a low RCT enrollment propensity when there appears no such benefit for RCT participants. These include superior quality of patient monitoring **and** care giving.^{14, 15}

The above finding is all the more notable given that *all* sites were selected based on their prior track record of success in conducting RCTs. The requirement to have an active bipolar disorder specialty program and clinical research experience likely resulted in the sites used for this analysis being more advanced than in the US as a whole. If this resulted in greater uniformity across the sites (e.g., the restrictions to the eligibility criteria narrowed the diversity of the sites in the study compared to all US sites), the effect of RCT enrollment propensity on the outcome would not be as precisely estimated as it would be if the sample was more diverse. If sites with a broader range of recruitment success were included one would then anticipate finding more precise results as there would be more statistical information to estimate the relationship of site RCT enrollment propensity to the outcome.

Limitations of our study include the relatively small number of subjects in the RCT, which reduce our power to detect a significant difference in the associations of site high RCT enrollment propensity between the RCT and observational arms, despite the fact that we found a null effect in the RCT arm and a significant effect within the observational arm. However, the pharmacotherapy RCT is by far the largest RCT in STEP-BD and few other studies contain both RCT and observational arms (the Woman's Health Initiative³ and the Radiation Therapy Oncology Group combined RCT and usual care study¹⁶ are exceptions). To the extent that our findings are limited by insufficient power to detect associations, this paper is a catalyst for investigating more studies like the STEP-BD.

Ideally we would have known which patients were offered the chance to participate in the RCT (from which we could determine who accepted and who declined) as this would be a candidate instrumental variable for whether they participated in the RCT, potentially allowing the effects of unmeasured confounders to be negated. Unfortunately such information was not available.

The findings from this research provide important information that enhances our understanding of RCT generalizability to non-RCT settings. Important future work would be to isolate the site characteristics of high RCT enrolling sites that are associated with improved patient outcomes so that this information could be used to inform and improve usual care settings. Another avenue for further work would be to examine whether there are patient characteristics that have heterogeneous effects on outcomes across the RCT and observational settings. The identification of such effects may provide additional information on the generalizability of RCT results to usual care settings.

Our results highlight the importance of site effects, and possibly culture – even among sites that are experienced in conducting RCTs – that are associated with different outcomes among non-RCT participants. These findings also point to the need for rigorous analysis within RCTs to determine whether heterogeneous effects exist across sites, as opposed to simply estimating a model with homogeneous treatment effects, although such analyses are typically limited by sample size. Better understanding why different outcomes may be obtained under different settings or across different types of patients would improve our ability to understand how RCT outcomes can generalize to usual care settings, as well as help determine strategies for improving patient outcomes in usual care settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors gratefully acknowledge support of the NIMH: RC4 MH092717 (O'Malley and co-authors) and K01 MH071714 (Busch). We also thank Andrew Nierenberg, MD and Gary Sachs, MD for sharing their knowledge and expertise about the STEP-BD study.

REFERENCES

1. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine*. 2000; 342:1878–1886. [PubMed: 10861324]
2. Feinstein AR. Epidemiologic analyses of causation: the unlearned scientific lessons of randomized trials. *Journal of Clinical Epidemiology*. 1989; 42:481–489. [PubMed: 2661727]
3. Hartz A, He T, Wallace R, et al. Comparing hormone therapy effects in two RCTs and two large observational studies that used similar methods for comprehensive data collection and outcome assessment. *British Medical Journal Open*. *BMJ Open*. 2013; 3:e002556. doi:10.1136/bmjopen-2013-0025562013.
4. Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials: race-, sex-, and age-based disparities. *Journal of the American Medical Association*. 2004; 291:2720–2726. [PubMed: 15187053]
5. Sachs GS, Thase ME, Otto MW, et al. Rationale, design, and methods of the systematic treatment enhancement program for bipolar disorder (STEP-BD). *Biological Psychiatry*. 2003; 53:1028–1042. [PubMed: 12788248]
6. Busch AB, He Y, Zelevinsky K, et al. Which Patients Participate in Clinical Trials? *Psychiatric Services*. 2015 doi: 10.1176/appi.ps.201300557:1-7.
7. Sachs GS, Nierenberg AA, Calabrese JR, et al. Effectiveness of adjunctive antidepressant treatment for bipolar depression. *New England Journal of Medicine*. 2007; 356:1711–1722. [PubMed: 17392295]
8. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association; Washington, D.C.: 1994.
9. Spitzer RL, Williams JB, Gibbon M, et al. The Structured Clinical Interview for DSM-III-R (SCID). I. History, rationale, and description. *Archives of General Psychiatry*. 1992; 49:624–629. [PubMed: 1637252]
10. Sheehan DV, Lecrubier Y, Sheehan KH, et al. 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. *Journal of Clinical Psychiatry*. 1998; 59:22–57. [PubMed: 9881538]
11. Sachs GS, Guille C, McMurich SL. A clinical monitoring form for mood disorders. *Bipolar Disorder*. 2002; 4:323–327.
12. Guy, W. *ECDEU Assessment Manual for Psychopharmacology--Revised*. Rockville, MD: 1976.
13. Miklowitz DJ, Otto MW, Frank E, et al. Psychosocial treatments for bipolar depression: A 1-year randomized trial from the Systematic Treatment Enhancement Program. *Arch Gen Psychiatry*. 2007; 64:419–427. [PubMed: 17404119]
14. Als-Nielsen, B.; Chen, W.; Gluud, LL., et al. Are trial size and reported methodological quality associated with treatment effects?. *Observational study of 523 randomised trials*. 12th International Cochrane Colloquium.; Ottawa, Canada. 2004;
15. Sterne JAC. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Statistics in Medicine*. 2002; 21:1513–1524. [PubMed: 12111917]
16. Efsthathiou, J. Radiation Therapy Oncology Group (RTOG) proposed combined randomized control arm and usual care arm trials.. 2010. Personal communication

Figure 1.

Visual schema detailing patient and site inclusion and exclusion. The loss of patients at the final step is due to 2 patients with missing age who contributed a total of 4 observations to the data set on which our statistical models were estimated.

Figure 2.
Likelihood of Recovered by week.

Table 1

Descriptive characteristics of individuals at baseline, stratified by (i) whether the participant was enrolled in a site that had a low or high RCT enrollment propensity (low= below median; high=above median) and (ii) whether the participant was in the RCT or the observational (OBS) arm.

Characteristic\Level	Site RCT Enrollment Propensity		Individual Study Type	
	Low (N = 1025)	High (N = 893)	Obs (N = 1581)	RCT (N = 337)
Individual RCT participation (%)	9.9	26.4		
Site RCT Enrollment Above Median (%)			41.6	70.0
Follow-up weeks (Mean $\pm$ SD, Max 26)	20.1 $\pm$ 6.0	20.3 $\pm$ 5.8	19.9 $\pm$ 6.0	21.3 $\pm$ 5.5
Psychosocial study participation (%)	8.7	20.5	3.2	65.6
Substance use disorder	48.1	51.7	48.6	55.2
Income above 40K (%)	43.3	45.0	45.4	38.0
Age (Mean $\pm$ SD)	40.7 $\pm$ 12.8	40.0 $\pm$ 11.7	40.3 $\pm$ 12.5	40.4 $\pm$ 11.6
Male (%)	41.6	38.3	39.6	42.1
Ethnicity\ethnicity				
White	85.2	90.5	87.3	89.0
Black	4.2	4.6	4.2	5.3
Hispanic	6.8	2.6	5.2	3.3
Other	3.8	2.4	3.3	2.4
Education (% in category)				
Incomplete High School	8.4	5.4	7.5	4.7
High School graduate	15.6	14.9	14.7	17.8
Some College	26.1	25.1	24.5	31.2
College graduate +	49.9	54.6	53.3	46.3
Insurance (% in category)				
None	16.2	18.6	15.1	27.6
Medicare	14.2	9.4	12.5	9.8
Medicaid	4.9	5.2	4.7	6.2
Medicare and Medicaid	1.3	3.2	2.1	2.7
Private	63.4	63.6	65.6	53.7
General health (% in category)				
0 conditions	74.1	69.5	71.7	73.6
1 conditions	3.4	4.9	4.0	4.5
2+ conditions	22.4	25.5	24.3	22.0
Mental health (% in category)				
0 conditions	41.3	32	37.5	34.4
1 conditions	45.8	49.8	47.1	50.1
2+ conditions	13	18.1	15.4	15.4
Baseline Symptoms (% in category)				

Characteristic\Level	Site RCT Enrollment Propensity		Individual Study Type	
	Low (N = 1025)	High (N = 893)	Obs (N = 1581)	RCT (N = 337)
None or Mild	26.8	18.8	24.4	16.9
Moderate	51.6	55.7	52.6	57.9
Severe	21.6	25.5	23	25.2
Baseline Bipolar Disorder Course of Illness Index (% in category)				
No evidence of disorder	4.3	2.9	3.8	3
Non-specific	14.1	14	13.5	16.6
Known convincing	81.6	83.1	82.7	80.4
Baseline clinical status (% in category)				
Diagnosis phase	99.6	98.4	99.8	95.5
Process phase	0.4	1.6	0.2	4.5
Recovered	0.0	0.0	0	0

Note: Bipolarity Index (BPI) = measure of how pathognomonic history is for bipolar disorder.

General health conditions = pregnancy or hepatic, renal, pancreatic, seizure, thyroid or inflammatory disorders.

Table 2

Estimated Odds Ratios of Predictors on Probability of Recovered for STEP-BD Participants.

Explanatory Variable	Estimate	95% CI		P (effect)	
Site enrollment propensity					
Above Median (OBS study subjects)	2.13	1.28	3.55	0.00	**
Above Median (RCT study subjects)	1.03	0.35	3.03	0.96	
Follow-up time					
Linear trend(OBS study subjects)	2.95	2.46	3.54	0.00	***
Linear trend (RCT study subjects)	5.49	3.83	7.89	0.00	***
RCT participation	0.53	0.13	2.27	0.40	***
Psychosocial study participation	0.70	0.29	1.67	0.43	
Baseline clinical status = Process	17.6	2.69	114.8	0.00	**
Past substance abuse	1.41	0.88	2.26	0.16	
No insurance	0.63	0.33	1.20	0.16	
Income 40000 or greater	1.69	1.05	2.70	0.03	*
Age	0.98	0.96	1.00	0.03	*
Male	0.87	0.55	1.39	0.57	
Race/Ethnicity (base = white)					
Black	1.19	0.37	3.85	0.77	
Hispanic	1.83	0.63	5.26	0.26	
Other	0.61	0.15	2.44	0.49	
Education (base = Less than high school graduate)					
High School Graduate	0.94	0.28	3.11	0.91	
Some College	1.35	0.44	4.12	0.60	
Greater than college	2.20	0.76	6.41	0.15	
General Health Comorbidities					
1 condition	2.36	0.79	7.01	0.12	
2+ condition	1.18	0.68	2.06	0.55	
Mental Health Comorbidities					
1 condition	0.58	0.35	0.97	0.04	*
2+ condition	0.31	0.15	0.66	0.00	**
Baseline symptoms (Base = None or Mild)					
Moderate	0.73	0.42	1.26	0.26	
Severe	0.26	0.13	0.53	0.00	***
Baseline Bipolar Disorder Course of Illness Index (% in category)					

Explanatory Variable	Estimate	95% CI		P (effect)
Non-specific	1.20	0.28	5.12	0.81
Known convincingly	1.68	0.44	6.35	0.45

Note:

RCT indicates that the patient participated in the RCT while OBS indicates they remained in the observational arm. The p-values for the interaction of site enrollment propensity above median and linear time trend with RCT participation were $p = 0.230$ and $p = 0.003$, respectively.

p-value < 0.001

**

0.001 p-value < 0.01

*

0.01 p-value < 0.05; . = $0.05 < \text{p-value} < 0.10$.