

Generalizability in the Family-to-Family Education Program Randomized Waitlist-Control Trial

Sue M. Marcus, Ph.D.

Deborah Medoff, Ph.D.

Li Juan Fang, M.S.

James Weaver, M.P.H.

Naihua Duan, Ph.D.

Alicia Lucksted, Ph.D.

Lisa B. Dixon, M.D., M.P.H.

Objective: Randomized controlled trials (RCTs) may have limited generalizability for the community when a high proportion of individuals refuse randomization or otherwise do not participate—a not uncommon phenomenon. A randomized waitlist-control trial of the Family-to-Family (FTF) education program, a 12-week course offered by the National Alliance on Mental Illness for family members of adults with mental illness, was previously reported. This study assessed whether the RCT-derived estimates of effectiveness of FTF were generalizable to individuals who participated in FTF but declined participation in the RCT. **Methods:** Propensity score matching was used to create five quintiles, each containing scores for individuals in FTF or waitlist conditions and for decliners; scores were matched on multiple baseline characteristics (N=442) within each quintile. Effectiveness estimates, with standard errors, were derived for the decliner population on the basis of effectiveness estimates derived from participants in the RCT; estimates were weighted to the baseline distribution of quintiles for the decliners. **Results:** For each outcome, estimates of the effect sizes observed in the RCT were very similar to the effect sizes observed for the decliner population; confidence intervals also had a high degree of overlap. **Conclusions:** This study suggests that the benefits of FTF observed in the RCT are generalizable to the group of individuals who declined RCT participation, providing further evidence of FTF's effectiveness. Propensity score matching was a useful statistical tool for addressing selection bias resulting from high rates of nonconsent in randomized waitlist-control trials. (*Psychiatric Services* 64:754–763, 2013; doi:10.1176/appi.ps.002912012)

Randomized controlled trials (RCTs) are considered the most rigorous test of an intervention's effectiveness. The internal validity of RCTs gives confidence that study findings can be attributed to the differences between the experimental and control conditions. However, RCTs may have limited external validity (generalizability) for the community of potential users of the program being tested if a high proportion of individuals refuse randomization or otherwise do not participate—a not uncommon phenomenon (1).

Individuals may decline to consent to random assignment to a treatment if it differs greatly from those currently received or familiar (medication versus psychotherapy, for example) (1). A similar situation arises when an RCT control group is placed on a waitlist for an experimental intervention; some people may withhold consent for random assignment if they are unwilling to wait for the experimental treatment. In prior work, we proposed the parallel randomized and nonrandomized (PRN) clinical trial design (also known as the partially randomized preference design) as a solution for this problem (1,2). Most RCTs exclude individuals who do not consent to randomization. However, in the PRN trial design, those who consent to randomization are randomly assigned, and those who

Dr. Marcus and Dr. Duan are affiliated with the Departments of Psychiatry and Biostatistics in the Division of Biostatistics at New York State Psychiatric Institute (NYSPI), Columbia University, New York City. Dr. Dixon is affiliated with the Department of Psychiatry, Columbia University, and with NYSPI, New York City. Dr. Medoff, Ms. Fang, and Dr. Lucksted are with the Department of Psychiatry, University of Maryland School of Medicine, Baltimore. Mr. Weaver is with the Research Foundation for Mental Health, New York City. Send correspondence to Dr. Dixon, Department of Psychiatry, Columbia University/NYSPI, 1051 Riverside Dr., New York, NY 10032 (e-mail: dixonli@nyspi.columbia.edu).

do not are assigned to their treatment of choice and are followed in a manner similar to those in the RCT. This design can enhance generalizability by enabling the estimation of effectiveness for those who decline randomization (1).

We reported the results of an RCT that tested the effectiveness of the National Alliance on Mental Illness (NAMI) Family-to-Family (FTF) program, a 12-week course for family members of adults with mental illness (3). In this study, 318 consenting participants in five Maryland counties and Baltimore City were randomly assigned to participate in FTF immediately or to wait at least three months for the next available class and to freely use in the meantime any other NAMI, community, or professional supports. We found that FTF participants had significantly greater improvements in coping, family problem solving, knowledge and distress. However, less than one-third of the potential sample was willing to consider study participation. The most common reason for declining was unwillingness to undergo random assignment because of the potential delay in FTF participation (3).

The study's consent rate created a concern that the RCT participant sample was not representative of individuals who generally participate in FTF. To address this, we offered nonrandomized study participation to a cohort of 124 individuals who refused to enroll in the RCT and who were planning to take the class immediately. We evaluated these individuals (called the "decliner" sample) according to the same schedule as the participants in the RCT. The aims of the study were to apply innovative statistical methods to determine whether the findings of the RCT could generalize to the sample of decliners and therefore potentially to the population of individuals who enroll in FTF through usual NAMI programming.

Methods

Participants

Individuals were eligible to participate in the primary study if they were between ages 21 and 80, desired enrollment in the next FTF class

regarding a family member or significant other, and spoke English. A total of 1,532 individuals who expressed interest in FTF were screened for study participation; 1,168 were found to be eligible. From this group, 318 individuals consented to participate in the randomized portion (RCT) of the overall study; 160 were randomly assigned to FTF, and 158 were assigned to the waitlist. An additional sample of 124 individuals from the 850 who had declined enrollment in the RCT and who were planning to take the class enrolled in the nonrandomized portion of the overall study. This decliner group was recruited approximately midway through the RCT when the need to address the modest consent rate was recognized; of the persons deemed eligible for the RCT, those who consecutively refused to participate in the RCT were offered enrollment as decliners until we achieved our target enrollment. Participants in both the RCT and decliner (nonrandomized) portions of the study completed identical baseline and follow-up interviews. We refer to three groups: decliners, RCT FTF participants, and RCT waitlist participants.

The institutional review board (IRB) at the University of Maryland approved all study activities; because interviews were conducted over the telephone, the IRB permitted consent to be obtained over the telephone after complete description of the study to the participants. Participants were recruited between March 2006 and September 2009.

Variables

This study considered three sets of variables. The first set includes all of the variables that were obtained in the participant interview. [This set is available online as a data supplement to this article.] Each variable was tested for inclusion in the propensity score analyses.

The second set of variables, described below, included those that differed between the decliner and RCT samples and therefore were used to generate the propensity scores. These included consumer race-ethnicity, consumer gender, living siblings of the consumer, family

member income, family member marital status, consumer hospitalization in the past six months, and information about objective illness burden and required supervision obtained from the Family Experience Interview Survey (4).

Also in the second variable set, the Family Empowerment Scale provided measures of family, community, and service system empowerment (5); the Experience of Caregiving Inventory provided measures of positive aspects of the relationship, need of backup, problem with service system, stigma, and total positive and total negative subscales (6,7); and the NAMI Family Member Questionnaire provided measures of empowerment, coping with consumer's illness, subjective burden and worry, and understanding of the mental health system (8). We also used the physical composite score of the 12-Item Short Form Health Survey (9), the Global Severity Index of the Brief Symptom Inventory-18 (BSI-18) (10,11), the percentage correct on the FTF mental illness knowledge test, and whether the family member self-reported having ever attended any formal NAMI educational programs (8).

The third relevant set of variables consisted of outcomes that improved with FTF for participation in the RCT. Knowledge was measured with a 20-item true-false test of factual information covering material drawn from the FTF curriculum that tapped general knowledge about mental illnesses (3). The five-item anxiety subscale of the BSI-18 measured psychological distress. It is designed for use primarily in nonclinical, community populations and has well-established reliability and validity (10,11). Family functioning was measured with the five-item problem-solving subscale of the Family Assessment Device, which evaluates family functioning and family relations. It is widely used in studies of family response to general medical illness and has well-established reliability and validity (12). The four-item acceptance dimension of the COPE measures emotion-focused coping (13) and family, service system, and community empowerment as described above. The analyses used both

Table 1

Baseline covariates of participants in a randomized controlled trial (RCT) of the National Alliance on Mental Illness (NAMI) Family-to-Family psychoeducation, by propensity score quintiles

Covariate and group	Overall			Quintile 1			Quintile 2			Quintile 3			Quintile 4			Quintile 5		
	Total N	N	%	Total N	N	%	Total N	N	%	Total N	N	%	Total N	N	%	Total N	N	%
RCT waitlist for Family-to-Family	140	88	63	18	13	72	26	13	50	31	26	84	30	18	60	35	18	51
White consumers	140	83	59	18	13	72	26	17	65	31	17	55	30	20	67	35	16	46
Married or living as if married	140	95	68	18	13	72	26	21	81	31	22	71	30	21	70	35	18	51
Family income >\$50,000																		
Objective daily living assistance rating (M±SD) ^a	140	1.2±1.0	18	1.3±1.0	26	1.6±1.1	31	1.4±1.0	30	1.0±.9	35	.9±.9						
Attended any formal NAMI educational programs	140	33	24	18	8	44	26	4	15	31	6	19	30	8	27	35	7	20
Male consumers	140	72	51	18	11	61	26	15	58	31	18	58	30	14	47	35	14	40
Any psychiatric hospitalization in past 6 months	140	42	30	18	10	56	26	15	58	31	9	29	30	3	10	35	5	14
Knowledge (M±SD % correct)	140	57.3±17.3	18	43.9±20.6	26	55.3±14.9	31	60.4±17.7	30	60.5±12.1	35	60.1±18.0						
Problem solving (M±SD) ^b	136	13.2±2.9	18	13.8±4.2	26	13.0±2.5	30	12.8±2.3	30	13.1±2.4	32	13.3±3.3						
Anxiety (M±SD) ^c	140	53.2±10.1	18	52.7±8.5	26	49.8±9.2	31	52.3±9.5	30	54.1±10.7	35	56.1±11.0						
Global Severity Index (M±SD) ^d	140	53.2±9.8	18	51.8±6.2	26	49.6±8.5	31	51.9±10.0	30	52.6±9.1	35	58.3±11.0						
Empowerment in aspect of family (M±SD) ^e	140	3.4±.7	18	2.9±.7	26	3.4±.7	31	3.3±.6	30	3.3±.6	35	3.5±.7						
Empowerment in aspect of service (M±SD) ^e	140	3.1±.9	18	2.6±.8	26	3.3±1.1	31	3.0±.9	30	2.9±.8	35	3.5±.9						
Empowerment in aspect of community (M±SD) ^e	140	2.4±.8	18	1.9±.5	26	2.4±.8	31	2.2±.6	30	2.3±.5	35	2.9±.9						
Acceptance (M±SD) ^f	139	12.4±2.6	18	11.7±3.1	26	12.8±2.3	30	12.5±2.2	30	12.5±2.3	35	12.5±2.9						
Depression (M±SD) ^g	139	10.1±8.7	17	8.5±6.5	26	8.5±6.3	31	8.4±7.0	30	9.5±8.1	35	14.0±11.8						
Worry (M±SD) ^h	140	1.8±.5	18	1.6±.5	26	1.8±.5	31	1.7±.4	30	1.9±.5	35	1.8±.5						
Subjective burden (M±SD) ^h	140	2.6±.5	18	2.4±.5	26	2.5±.5	31	2.7±.5	30	2.7±.5	35	2.7±.5						
RCT Family-to-Family participants	152	90	59	18	14	78	21	16	76	34	23	68	41	21	51	38	16	42
White consumers	152	97	64	18	15	83	21	16	76	34	26	76	41	22	54	38	18	47
Married or living as if married	152	104	68	18	14	78	21	19	90	34	21	62	41	29	71	38	21	55
Family income >\$50,000																		
Objective daily living assistance rating (M±SD) ^a	152	1.1±.9	18	1.2±1.0	21	1.4±.9	34	1.0±.9	34	1.0±.9	41	.9±.9	38	1.0±1.0				
Attended any formal NAMI educational programs	152	18	12	18	3	17	21	1	5	34	6	18	41	5	12	38	3	8
Male consumers	152	84	55	18	14	78	21	15	71	34	19	56	41	21	51	38	15	39
Any psychiatric hospitalization in past 6 months	152	51	34	18	10	56	21	10	48	34	12	35	41	14	34	38	5	13
Knowledge (M±SD % correct)	152	58.3±17.8	18	44.4±18.2	21	61.0±12.4	34	55.9±18.5	41	61.7±18.6	38	61.9±15.7						
Problem solving (M±SD) ^b	151	12.8±2.9	18	14.0±3.3	21	13.2±3.0	34	12.4±2.8	40	12.3±2.7	38	12.8±2.9						
Anxiety (M±SD) ^c	152	52.6±9.2	18	51.6±10.1	21	48.7±9.9	34	54.4±10.2	41	51.7±8.1	38	54.5±8.0						
Global Severity Index (M±SD) ^d	152	51.9±9.3	18	50.3±11.6	21	48.6±8.9	34	53.6±9.1	41	50.7±7.9	38	54.4±9.3						

Continues on next page

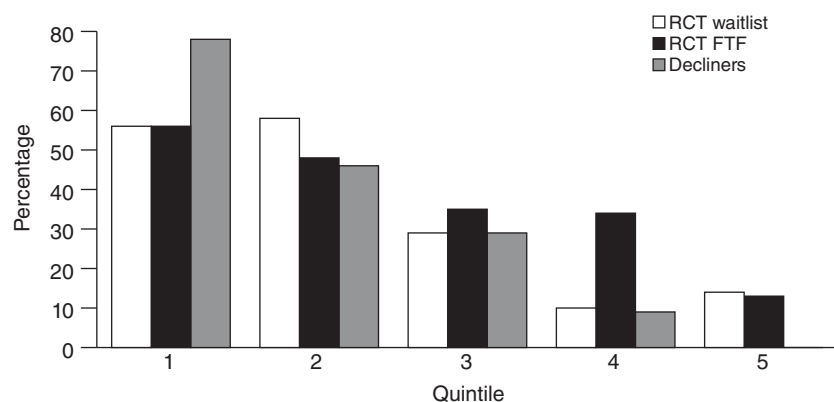
Table 1*Continued from previous page*

Covariate and group	Overall			Quintile 1			Quintile 2			Quintile 3			Quintile 4			Quintile 5		
	Total	N	%	Total	N	%	Total	N	%	Total	N	%	Total	N	%	Total	N	%
Empowerment in aspect of family (M±SD) ^e	152	3.5±.6	18	3.0±.7	21	3.4±.7	34	3.4±.6	41	3.6±.5	38	3.6±.5	41	3.6±.5	38	3.6±.5		
Empowerment in aspect of service (M±SD) ^e	152	3.2±.9	18	2.7±1.0	21	3.0±1.0	34	3.3±.9	41	3.3±.7	38	3.4±.8	41	3.3±.7	38	3.4±.8		
Empowerment in aspect of community (M±SD) ^e	152	2.6±.8	18	2.0±.9	21	2.5±.6	34	2.3±.7	41	2.8±.7	38	2.8±.8	41	2.8±.7	38	2.8±.8		
Acceptance (M±SD) ^f	151	12.9±2.3	18	12.5±2.5	21	13.8±1.4	33	12.4±2.7	41	12.9±2.3	38	13.0±2.2	41	12.9±2.3	38	13.0±2.2		
Depression (M±SD) ^g	149	8.6±7.2	18	8.7±6.2	21	6.1±6.9	33	9.2±6.0	39	8.5±8.2	38	9.3±7.7	39	8.5±8.2	38	9.3±7.7		
Worry (M±SD) ^h	152	1.8±.5	18	1.7±.4	21	1.8±.4	34	1.8±.5	41	1.9±.6	38	1.8±.5	41	1.9±.6	38	1.8±.5		
Subjective burden (M±SD) ^h	152	2.6±.5	18	2.5±.3	21	2.6±.5	34	2.7±.5	41	2.7±.4	38	2.6±.5	41	2.7±.4	38	2.6±.5		
Decliners (no RCT random assignment)																		
White consumers	117	82	70	46	34	74	35	29	83	17	12	71	11	5	45	8	2	25
Married or living as if married	117	82	70	46	35	76	35	26	74	17	9	53	11	6	55	8	6	75
Family income >\$50,000	117	91	78	46	42	91	35	27	77	17	13	76	11	7	64	8	2	25
Objective daily living assistance rating (M±SD) ^a	117	1.3±1.0	46	1.6±1.0	35	1.4±1.1	35	1.4±1.1	17	1.2±.9	11	.6±.7	11	.7±.6	8	.6±.7		
Attended any formal NAMI educational programs																		
Male consumers	117	30	26	46	18	39	35	6	17	17	4	24	11	2	18	8	0	—
Any psychiatric hospitalization in past 6 months	117	74	63	46	33	72	35	21	60	17	10	59	11	5	45	8	5	63
Knowledge (M±SD % correct)	117	58	50	46	36	78	35	16	46	17	5	29	11	1	9	8	0	—
Problem solving (M±SD) ^b	117	54.0±17.5	46	50.2±18.6	35	57.3±17.5	35	57.3±17.5	17	60.8±11.7	11	52.4±18.8	11	52.4±18.8	8	48.8±14.8		
Anxiety (M±SD) ^c	109	13.1±2.4	43	12.7±2.2	31	13.2±2.2	31	13.2±2.2	17	13.2±3.1	10	12.9±2.6	10	14.6±2.5	8	12.9±2.6		
Global Severity Index (M±SD) ^d	117	51.6±10.2	46	51.7±11.4	35	49.7±9.3	35	49.7±9.3	17	52.9±6.8	11	55.3±8.9	11	55.3±8.9	8	51.5±14.3		
Empowerment in aspect of family (M±SD) ^e	117	50.7±10.7	46	50.2±12.4	35	49.4±9.2	35	49.4±9.2	17	51.3±7.3	11	55.3±10.3	11	55.3±10.3	8	51.8±13.6		
Empowerment in aspect of service (M±SD) ^e	117	3.3±.6	46	3.3±.5	35	3.3±.6	35	3.3±.6	17	3.4±.7	11	3.2±.6	11	3.2±.6	8	3.6±.7		
Empowerment in aspect of community (M±SD) ^e	117	3.1±.8	46	3.2±.7	35	3.1±.8	35	3.1±.8	17	3.1±.7	11	2.9±.7	11	2.9±.7	8	3.7±1.0		
Acceptance (M±SD) ^f	117	2.3±.7	46	2.1±.6	35	2.2±.7	35	2.2±.7	17	2.4±.7	11	2.8±.8	11	2.8±.8	8	2.5±1.0		
Depression (M±SD) ^g	116	12.6±2.2	45	12.6±2.3	35	12.9±1.9	35	12.9±1.9	17	13.4±1.5	11	11.4±2.7	11	11.4±2.7	8	11.6±3.5		
Worry (M±SD) ^h	117	8.8±7.0	46	9.0±8.0	35	7.6±5.0	35	7.6±5.0	17	8.3±6.1	11	11.2±9.0	11	11.2±9.0	8	10.8±6.9		
Subjective burden (M±SD) ^h	117	1.7±.4	46	1.8±.4	35	1.8±.4	35	1.8±.4	17	1.7±.4	11	1.7±.5	11	1.7±.5	8	2.0±.5		
	117	2.6±.5	46	2.5±.5	35	2.7±.4	35	2.7±.4	17	2.6±.5	11	2.3±.6	11	2.3±.6	8	2.9±.5		

^a From the Family Experience Interview Schedule. Possible scores range from 0 to 4, with higher scores indicating more frequent assistance in daily life.^b From the Family Assessment Device. Possible scores range from 6 to 24, with higher scores indicating worse problem solving.^c T scores range from 38 to 81, with higher scores indicating more anxiety symptoms.^d T scores range from 33 to 81, with higher scores indicating more global symptoms.^e Family Empowerment Scale. Possible scores range from 1 to 5, with higher scores indicating more empowerment.^f From the COPE. Possible scores range from 4 to 16, with higher scores indicating better coping.^g As measured on the Center for Epidemiological Studies Depression Scale. Possible scores range from 0 to 42, with higher scores indicating more severe depression symptoms.^h Family Member Questionnaire. Possible scores range from 1 to 4, with higher scores indicating less worry and fewer burdens.

Figure 1

Percentage of participants reporting that a family member had a psychiatric hospitalization in the past 6 months, by propensity score quintile^a



^a Respondents received or were waitlisted to receive the 12-week Family-to-Family (FTF) psychoeducation program in a randomized controlled trial (RCT). Persons who declined random assignment (decliners) received FTF but did not participate in the RCT.

baseline and three-month measures of FTF outcomes.

Statistical approach

Although this FTF study began with a traditional randomization process, the addition of the decliners transformed the combined RCT plus the decliner nonrandomized portion into a PRN study, a hybrid of randomized and observational study that is used in effectiveness analyses (14). Observational studies often suffer from selection bias; that is, people who receive treatment may differ systematically from those who do not. One approach is to match those in the treatment group with those in the control group, so that treatment effects can be attributed to the treatment rather than to baseline differences between the treated and untreated participants.

Propensity score matching is a useful statistical tool for adjusting for many covariates simultaneously (15). Matching participants on the unidimensional propensity score between those who receive or do not receive treatment has been shown formally to be statistically comparable to matching separately on each of the multiple covariates used to create the propensity score, but the former is preferable because separate matching becomes infeasible when there are more than a few covariates.

The propensity score is defined as the probability that a participant received the treatment versus the

control condition, which is contingent on a set of potential measured confounders. Commonly, the propensity score is estimated with logistic regression to model the propensity of an individual to receive the treatment versus the control condition (16). However, to assess the generalizability of the treatment effect for the decliners, we chose to evaluate the propensity to be in the RCT versus in the decliner sample, in order to clarify characteristics that were associated with being a decliner. This approach has been shown to be mathematically equivalent to the more common method of estimating propensity to receive the treatment versus the control condition but provides more useful information for this study design (14).

First, propensity scores for each person were estimated with logistic regression. Variables were selected for the regression model by using bivariate statistics (chi square and *t* tests) to compare the RCT sample with the decliner sample for all baseline measures we collected in the RCT, including, for example, consumer and family demographic variables, family member-reported objective and subjective burden, coping, empowerment, family functioning, and other supports (3). Variables showing significant differences between the RCT and decliner samples (23 variables, $p < .2$) were then entered into a logistic regression model comparing the RCT sample and the

decliner sample. Missing data were handled by using missing-data indicators (17). Propensity scores were calculated from this model. Participants with higher propensity scores had profiles more closely resembling RCT enrollees. All participants were placed into quintiles according to their propensity score (18). We then examined the distribution of participants by propensity score quintile for the randomized FTF, randomized waitlist, and decliner samples.

A sample of covariates by quintile and group is listed in Table 1. Analyses of variance (ANOVAs) and chi square tests were used to assess heterogeneity across quintiles, with respect to each selected covariate.

Effect size estimates for the decliner sample

Our primary goal was to determine whether the estimate of the effect of FTF versus waitlist observed in the RCT generalized to the decliner sample. We planned to derive estimates of FTF's impact on the outcomes of knowledge, family problem solving, empowerment, acceptance aspects of coping, anxiety, and subjective burden (worry) for the decliner sample and compare these with the estimates of benefits of FTF observed in the published RCT (3). Our approach was to build on the internally valid estimates of benefit derived from the RCT and enhance external validity by weighting the estimates of effectiveness observed in the RCT to fit the distribution of the propensity score quintiles for the baseline decliner population.

Although similar to age-adjusted estimates that are commonly used in life tables, propensity score matching enabled us to adjust for many covariates simultaneously. Estimates of the effectiveness of RCT FTF versus RCT waitlist were calculated for decliners. [Calculations and corresponding standard errors are available online as a data supplement to this article.] This approach provided an estimate of how individuals similar to the decliners would do if they received FTF versus how they would do if they could also be observed after assignment to (hypothetically) the waitlist. We next used these estimates

Table 2

Three-month outcomes in a randomized controlled trial (RCT) of participants receiving or waiting to receive Family-to-Family (FTF) psychoeducation, by propensity score quintile

Outcome and group	Overall			Quintile 1			Quintile 2			Quintile 3			Quintile 4			Quintile 5		
	Total			Total			Total			Total			Total			Total		
	N	M	SD	N	M	SD	N	M	SD	N	M	SD	N	M	SD	N	M	SD
RCT FTF waitlist	114	58.8	17.4	15	58.7	13.8	22	60.6	17.5	25	59.6	16.8	28	56.6	16.5	24	58.7	21.6
Knowledge (% correct)	113	12.9	2.9	15	13.2	3.9	22	12.5	2.4	24	12.5	1.8	28	13.5	3.1	24	12.8	3.2
Problem solving ^a	114	52.4	9.4	15	52.1	7.4	22	49.3	6.8	25	51.3	10.4	28	54.6	10.1	24	53.8	10.6
Anxiety ^b	114	51.9	9.4	15	50.5	5.7	22	49.0	9.1	25	51.1	8.7	28	53.5	10.2	24	54.5	10.9
Global Severity Index ^c	114	3.5	.6	15	3.2	.7	22	3.5	.6	25	3.5	.5	28	3.4	.8	24	3.6	.6
Empowerment in aspect of family ^d	114	3.1	.9	15	2.8	.8	22	3.2	.8	25	3.2	.9	28	2.9	.9	24	3.2	.9
Empowerment in aspect of service ^d	114	2.5	.8	15	2.1	.6	22	2.3	.6	25	2.3	.6	28	2.5	.8	24	2.9	1.0
Empowerment in aspect of community ^d	114	12.7	2.5	15	12.0	2.4	22	13.4	2.2	25	12.6	2.6	28	12.4	2.7	24	12.9	2.3
Acceptance ^e	114	8.5	6.9	15	7.0	4.6	22	8.0	5.9	25	6.3	5.1	28	9.8	8.1	24	10.7	8.6
Depression ^f	114	1.9	.5	15	1.8	.5	22	1.8	.4	25	2.0	.4	28	2.1	.6	24	2.0	.5
Worry ^g	114	2.7	.5	15	2.5	.5	22	2.5	.4	25	2.9	.6	28	2.7	.5	24	2.8	.5
Subjective burden ^g																		
RCT FTF participants	129	65.2	16.8	16	58.0	18.7	21	70.5	13.2	26	67.6	16.9	36	64.3	18.5	30	64.1	15.1
Knowledge (% correct)	126	12.1	2.6	16	13.1	2.6	21	11.6	2.5	26	12.0	2.5	34	11.6	2.4	29	12.6	3.0
Problem solving ^a	129	50.6	8.1	16	48.1	7.6	21	48.0	7.4	26	53.6	7.2	36	50.6	7.8	30	51.2	9.4
Anxiety ^b	129	50.3	9.0	16	46.9	9.2	21	47.0	7.8	26	53.2	7.2	36	51.2	7.9	30	50.9	11.1
Global Severity Index ^c	129	3.7	.6	16	3.4	.9	21	3.7	.7	26	3.7	.6	36	3.8	.5	30	3.7	.6
Empowerment in aspect of family ^d	129	3.4	.8	16	3.2	1.1	21	3.5	.8	26	3.4	.8	36	3.5	.7	30	3.5	.8
Empowerment in aspect of service ^d	129	2.9	.8	16	2.4	.8	21	2.8	.7	26	2.7	.8	36	3.1	.7	30	3.2	.7
Empowerment in aspect of community ^d	128	13.6	2.0	16	13.0	2.5	21	14.0	1.8	26	14.1	2.0	35	13.7	1.8	30	13.1	2.1
Acceptance ^e	129	7.5	7.0	16	6.4	6.1	21	4.6	4.3	26	9.4	6.9	36	6.9	5.9	30	9.3	9.4
Depression ^f	129	2.0	.5	16	1.7	.7	21	1.9	.4	26	2.0	.4	36	2.1	.6	30	2.0	.5
Worry ^g	129	2.7	.5	16	2.6	.4	21	2.6	.4	26	2.8	.5	36	2.9	.4	30	2.7	.5
Subjective burden ^g																		
Decliners (no RCT random assignment)	91	63.8	18.9	38	65.2	19.8	27	65.8	20.1	14	61.9	16.3	6	61.9	10.0	6	53.2	20.7
Knowledge (% correct)	84	12.6	2.6	34	12.1	2.1	26	12.9	2.5	13	13.2	2.3	6	15.2	3.4	5	10.4	3.8
Problem solving ^a	89	50.1	8.1	37	49.3	7.7	26	52.2	7.7	14	50.0	6.4	6	48.0	7.8	6	48.7	14.9
Anxiety ^b	89	49.4	8.6	37	47.8	8.8	26	50.5	8.0	14	51.1	6.2	6	48.0	10.0	6	51.5	12.8
Global Severity Index ^c	90	3.5	.6	37	3.6	.5	27	3.5	.5	14	3.4	.7	6	3.8	.6	6	3.7	.7
Empowerment in aspect of family ^d	90	3.4	.7	37	3.4	.7	27	3.3	.7	14	3.3	.7	6	3.5	.9	6	3.5	1.1
Empowerment in aspect of service ^d	90	2.7	.7	37	2.6	.6	27	2.7	.8	14	2.9	.8	6	3.2	.9	6	2.8	.9
Empowerment in aspect of community ^d	91	13.3	2.2	38	13.4	2.0	27	13.7	2.3	14	13.0	2.5	6	11.8	2.8	6	12.7	2.0
Acceptance ^e																		

Continues on next page

Table 2*Continued from previous page*

Outcome and group	Overall			Quintile 1			Quintile 2			Quintile 3			Quintile 4			Quintile 5		
	Total		N	M	SD	N	M	SD	N	M	SD	N	M	SD	N	M	SD	
Depression ^f	89	7.8	6.6	37	5.7	6.0	26	9.3	5.9	8.4	6.7	6	10.7	6.4	6	10.2	10.0	
Worry ^g	90	2.0	.4	37	1.9	.4	27	2.0	.4	1.9	.5	6	2.0	.4	6	2.2	.8	
Subjective burden ^g	90	2.7	.5	37	2.7	.4	27	2.8	.5	2.7	.5	6	2.7	.5	6	3.0	.9	

^a From the Family Assessment Device. Possible scores range from 6 to 24, with higher scores indicating worse problem solving.^b From the Brief Symptom Inventory. Anxiety T scores range from 38 to 81, with higher scores indicating more anxiety symptoms.^c T scores range from 33 to 81, with higher scores indicating more global symptoms.^d From the Family Empowerment Scale. Possible scores range from 1 to 5, with higher scores indicating more empowerment.^e From the COPE. Possible scores range from 4 to 16, with higher scores indicating better coping.^f As measured on the Center for Epidemiological Studies Depression Scale. Possible scores range from 0 to 42, with higher scores indicating more severe depression symptoms.^g From the Family Member Questionnaire. Possible scores range from 1 to 4, with higher scores indicating less worry and fewer burdens.

and standard errors to calculate 95% confidence intervals and corresponding effect sizes regarding FTF versus waitlist effectiveness for the decliner sample. These confidence intervals were then compared with the RCT estimates.

Results

Baseline differences

Of the total sample of 409 consumers whose race was reported by family member participants, 260 (64%) were white, 11 (3%) were Asian, 105 (26%) were black, seven (2%) were Hispanic, and 26 (6%) were other. When compared with the decliners, participants in the RCT were significantly more likely to report having family income greater than \$50,000 per year ($\chi^2=4.39$, $df=1$, $p<.036$), to report that the consumer required more assistance in daily living ($t=2.13$, $df=435$, $p<.033$), and to report that the family member had a psychiatric hospitalization in the past six months ($\chi^2=11.35$, $df=1$, $p<.001$). With respect to study outcomes, decliners had less knowledge about mental illness ($t=2.43$, $df=435$, $p<.016$) and less community empowerment ($t=2.51$, $df=434$, $p=.012$) at baseline.

Figure 1 provides an example of how the propensity score approach allowed the decliner sample to be matched with the RCT sample when one of the variables contributed to the propensity score. An important aspect of matching groups via propensity score quintiles was to examine the percentage of participants within each quintile by sample. Specifically, the critical question was whether there was a comparable percentage of each sample (decliner, RCT waitlist, or RCT FTF) for any particular variable represented in each quintile.

The figure shows that the variable, percentage of participants whose family member experienced a psychiatric hospitalization in the past six months, decreased from quintile 1 (most like the decliners) to quintile 5 (least like the decliners); in other words, compared with the RCT participants, a higher percentage of the decliners tended to have a family member who had been hospitalized, consistent with the bivariate analysis presented above. However, within

each quintile, the percentage was generally similar across the three samples. It is important to note that matching does not require perfect balance. Other covariates were also effectively balanced by propensity score quintile matching. Table 1 gives baseline covariate and outcome data by group (RCT waitlist, RCT FTF, or decliner) and propensity score quintile. ANOVAs and chi squares across quintiles of all covariates demonstrated significant heterogeneity.

Generalizability estimates

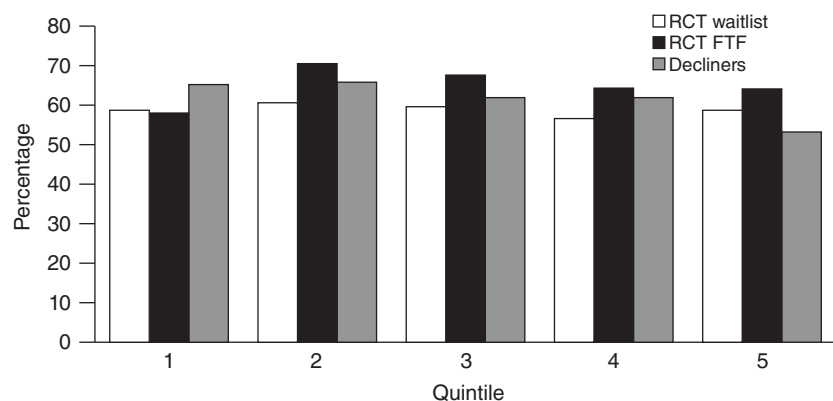
Table 2 provides mean outcome levels by group and propensity score quintile for three-month outcomes. The three groups appeared to be comparably distributed within each quintile, as shown in Figure 2 for the knowledge test. Table 3 presents the effectiveness estimates with confidence intervals and effect sizes derived from the RCT for comparison of individuals receiving FTF with individuals on the waitlist. It also provides estimates of the RCT FTF versus RCT waitlist effect for the decliner sample—estimates that reflect the capacity of the propensity scoring process and assignment of quintiles to predict what the effects of FTF versus waitlist would have been for the decliners. We note that the effect sizes were remarkably similar despite the selection differences for being in the decliner population. For example, with respect to knowledge, the effect size observed in the RCT was .31. The estimated effect size for the decliner population was .29. Also, the confidence intervals had a high degree of overlap.

Discussion

RCTs are vulnerable to selection bias that can reduce the external validity of study findings. Without external validity, the overall value of RCTs for informing care delivery and policy is critically limited. Programs that are widely available prior to effectiveness evaluation may face special challenges in avoiding significant selection bias when attempting to conduct RCTs. This creates difficulty in amassing high-quality practice-based evidence sufficient to merit the program's determination as an evidenced-based

Figure 2

Knowledge scores at 3 months, by propensity score quintile^a



^a Respondents received or were waitlisted to receive the 12-week Family-to-Family (FTF) psychoeducation program in a randomized controlled trial (RCT). Persons who declined random assignment (decliners) received FTF but did not participate in the RCT.

Table 3

Effectiveness of Family-to-Family program versus waitlist and estimated generalizability for persons declining random assignment in the randomized controlled trial (RCT)

Measure	Generalizability estimate ^a	95% CI	Effect size
Knowledge			
RCT	5.28	3.33 to 7.23	.31
Decliners	4.94	2.09 to 7.79	.29
Problem solving			
RCT	-.70	-.99 to -.41	-.23
Decliners	-.56	-1.09 to -.03	-.19
Anxiety			
RCT	-1.95	-2.89 to -1.01	-.21
Decliners	-2.00	-3.23 to -.77	-.22
Global Severity Index			
RCT	-1.62	-2.53 to -.71	-.18
Decliners	-2.17	-3.56 to -.78	-.24
Family empowerment			
RCT	.14	.08 to .20	.23
Decliners	.21	.08 to .34	.35
Service system empowerment			
RCT	.23	.15 to .31	.26
Decliners	.35	.19 to .51	.39
Community empowerment			
RCT	.26	.19 to .33	.37
Decliners	.40	.28 to .52	.58
Acceptance			
RCT	.74	.48 to 1.00	.32
Decliners	.93	.52 to 1.34	.40
Depression			
RCT	-1.23	-2.18 to -.28	-.16
Decliners	-1.17	-2.13 to -.21	-.15
Worry			
RCT	.04	-.04 to .12	.08
Decliners	-.01	-.11 to .09	-.02
Subjective burden			
RCT	-.10	-.09 to .07	-.20
Decliners	.07	-.01 to .15	.13

^a Comparisons were as follows: RCT, estimate from the RCT, excluding decliners (3). For decliners, the estimate is of Family-to-Family recipients versus the waitlist effect for the decliner group.

practice. This study's overall significance derives from its development and application of methods to meet that challenge.

Our RCT of NAMI's FTF education program appeared to be vulnerable to selection bias because individuals could access FTF without participating in our study. We were able to empirically evaluate this threat to our analysis by recruiting a sample of persons (decliners) who declined to participate in the randomization process. We showed that the estimated RCT FTF versus RCT waitlist effect sizes for the decliner sample were quite similar to the effect sizes observed in the RCT in which the individuals randomly assigned to FTF were compared with a waitlisted group. We thus conclude that FTF may indeed be effective for a target population that includes people similar to the decliners as well as those similar to the RCT participants.

This study therefore reinforced our previous findings that the NAMI FTF program is a valuable resource to family members of individuals with mental illness. FTF has been found to increase knowledge about mental illness, improve self-reported family problem-solving skills, and reduce distress. The RCT also demonstrated that FTF improves family members' coping skills and empowerment (3). The positive and generalizable impact of FTF observed in this study further reinforces the value of this program as an evidence-based practice and the imperative for mental health providers and clinicians to consider it a resource for struggling family members. These findings also underscore the unique contributions of peer-based support programs in the service array for persons with mental illnesses and their relatives (19,20).

The differences between RCT decliners and RCT participants may suggest some unique sampling vulnerabilities for RCTs with waitlisted control groups. As could be expected, individuals with higher indicators of need (greater objective burden and greater likelihood of a consumer's recent hospitalization) were less willing to take the chance of random assignment to the waitlist condition.

In addition, individuals with greater income were more likely to refuse RCT enrollment. Such patterns could have plausibly produced estimates suggesting that FTF would not have been effective with the decliners. The analyses presented therefore underscore the importance of adopting a systematic, empirical approach to evaluating external validity.

This use of propensity scores to evaluate such potential biases in non-randomized samples was limited by the fact that the estimates for the decliners were unbiased only when we could adjust for all confounders. Thus it is important for confounders to be considered during the design phase, so they can be measured. This involves collecting information about characteristics that might be related to being a decliner, as well as characteristics that are thought to influence the primary outcomes of a study. Collecting more covariates can add cost and complexity, but it is important not to overlook those necessary to determine whether results are convincing. Qualitative methods can be helpful in identifying additional confounders, particularly for areas in which there is not much existing research.

Our approach can be modified to other situations that commonly occur in psychiatric services research, for example, when people do not consent to randomization processes because they have strong preferences about treatment (1,14). Studies comparing medication to psychotherapy, two different medications, or two different psychotherapies exemplify this circumstance. In these situations, it is essential to document reasons for nonconsent and to subsequently collect outcome data when possible.

Conclusions

This study used innovative statistical methods to assess whether the benefits observed in a RCT of NAMI's FTF education program could be extended to a majority of eligible individuals who declined to participate in the RCT. By including a cohort of decliners and evaluating their status before and after participating in FTF, the analyses suggest that FTF versus waitlist benefits of improving knowl-

edge, reducing distress, and improving family problem solving generalize to the larger group. The significance of this study rests not only in the demonstration of benefits of FTF but also provides an important example of how RCTs of interventions available in the community can address the problem of external validity for the valid designation of programs as evidence-based practices.

Acknowledgments and disclosures

This project was supported by grant 1R01-MH72667-01A1 from the National Institute of Mental Health, by grant P20 MH085983 from the Center for Collaborative Inner-City Child Mental Health Services Research, and by grant P30 MH090322-01 A1 from the Advanced Center on Implementation-Dissemination Science in States for Children and Families.

The authors report no competing interests.

References

1. Marcus SM: Assessing non-consent bias with parallel randomized and nonrandomized clinical trials. *Journal of Clinical Epidemiology* 50:823-828, 1997
2. Paradise JL, Bluestone CD, Bachman RZ, et al: Efficacy of tonsillectomy for recurrent throat infection in severely affected children: results of parallel randomized and nonrandomized clinical trials. *New England Journal of Medicine* 310:674-683, 1984
3. Dixon LB, Lucksted A, Medoff DR, et al: Outcomes of a randomized study of a peer-taught Family-to-Family Education Program for mental illness. *Psychiatric Services* 62:591-597, 2011
4. Tessler R, Gamache G: Family Experiences Interview Schedule (FEIS); in the Toolkit on Evaluating Family Experiences With Severe Mental Illness. Cambridge, Mass, Human Services Research Institute, Evaluation Center, 1995. Available at www.hsri.org
5. Koren P, DeChillo N, Friesen B: Measuring empowerment in families whose children have emotional disorders: a brief questionnaire. *Rehabilitation Psychology* 37:305-321, 1992
6. Sznukler GI, Burgess P, Herrman H, et al: Caring for relatives with serious mental illness: the development of the Experience of Caregiving Inventory. *Social Psychiatry and Psychiatric Epidemiology* 31:137-148, 1996
7. Joyce JL, Leese M, Sznukler G: The Experience of Caregiving Inventory: further evidence. *Social Psychiatry and Psychiatric Epidemiology* 35:185-189, 2000
8. Dixon L, Lucksted A, Stewart B, et al: Outcomes of the peer-taught 12-week family-to-family education program for severe mental illness. *Acta Psychiatrica Scandinavica* 109:207-215, 2004
9. Ware J, Jr, Kosinski M, Keller SD: A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of

- reliability and validity. *Medical Care* 34: 220–233, 1996
10. Derogatis LR, Melisaratos N: The Brief Symptom Inventory: an introductory report. *Psychological Medicine* 13:595–605, 1983
 11. Derogatis LR: BSI-18: Administration, Scoring and Procedures Manual. New York, Pearson, 2001
 12. Sawin KJ, Harrigan MP: Measures of Family Functioning for Research and Practice. New York, Springer, 1995
 13. Carver CS, Scheier MF, Weintraub JK: Assessing coping strategies: a theoretically based approach. *Journal of Personality and Social Psychology* 56:267–283, 1989
 14. Marcus SM, Stuart EA, Wang P, et al: Estimating the causal effect of randomization versus treatment preference in a doubly randomized preference trial. *Psychological Methods* 17:244–254, 2012
 15. Rosenbaum PR, Rubin DB: The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55, 1983
 16. Stuart EA, Marcus SM, Horvitz-Lennon MV, et al: Using non-experimental data to estimate treatment effects. *Psychiatric Annals* 39:41451, 2009
 17. Haviland AM, Nagin DS, Rosenbaum PR: Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods* 12:247–267, 2007
 18. Rosenbaum PR, Rubin DB: Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79: 516–524, 1984
 19. Segal SP, Silverman CJ, Temkin TL: Self-help and community mental health agency outcomes: a recovery-focused randomized controlled trial. *Psychiatric Services* 61: 905–910, 2010
 20. Brown LD: Consumer-Run Mental Health: Framework for Recovery. New York, Springer, 2012

Submissions Invited for New *Psychiatric Services* Column on Integrated Care

The integration of primary care and behavioral health care is a growing research and policy focus. Many people with mental and substance use disorders die decades earlier than other Americans, mostly from preventable chronic medical illnesses. In addition, primary care settings are now the gateway to treatment for behavioral disorders, and primary care providers need to provide screening, treatment, and referral for patients with general medical and behavioral health needs.

To stimulate research and discussion in this critical area, *Psychiatric Services* is launching a new column on integrated care. The column will focus on service delivery and policy issues encountered on the general medical–psychiatric interface. Submissions are welcomed on topics related to the identification and treatment of (a) common mental disorders in primary care settings in the public and private sectors and (b) general medical problems in public mental health settings. Reviews of policy issues related to the care of comorbid general medical and psychiatric conditions are also welcomed, as are descriptions of current integration efforts at the local, state, or federal level. Submissions that address care integration in settings outside the United States are also encouraged.

Benjamin G. Druss, M.D., M.P.H., is the editor of the Integrated Care column. Prospective authors should contact Dr. Druss to discuss possible submissions (bdruss@emory.edu). Column submissions, including a 100-word abstract and references, should be no more than 2,500 words.