Social Psychiatry and Psychiatric Epidemiology (2019) 54:291–301
https://doi.org/10.1007/s00127-018-1631-x

ORIGINAL PAPER

A comparison of neurocognition and functioning in first episode psychosis populations: do research samples reflect the real world?

Emily Kline[1,2] · Victoria Hendel[2] · Michelle Friedman-Yakoobian[1,2] · Raquelle I. Mesholam-Gately[1,2] · Ann Findeisen[2] · Suzanna Zimmet[1,2] · Joanne D. Wojcik[1,2] · Tracey L. Petryshen[1,3] · Tsung-Ung W. Woo[1,2,4] · Jill M. Goldstein[1,5] · Martha E. Shenton[1,6,7] · Matcheri S. Keshavan[1,2] · Robert W. McCarley[1,7] · Larry J. Seidman[1,2,3]

Received: 22 February 2018 / Accepted: 13 November 2018 / Published online: 28 November 2018

Abstract

Purpose The current study evaluates the demographic, clinical, and neurocognitive characteristics of a recruited FEP research sample, a research control group, and a FEP clinic sample that were assessed and treated within the same center and time period.

Methods This study utilized data collected through an observational study and a retrospective chart review. Samples were ascertained in the Longitudinal Assessment and Monitoring of Clinical Status and Brain Function in Adolescents and Adults study and the Prevention and Recovery in Early Psychosis clinic. FEP clinic patients ($n=77$), FEP research participants ($n=44$), and age-matched controls ($n=38$) were assessed using the MATRICS consensus cognitive battery and global functioning social and role scales. Between-group differences were assessed via one-way ANOVA and Chi-square analyses.

Results No significant differences were observed between groups with regard to age and gender. The FEP research sample had a higher proportion of white participants, better social and role functioning, and better neurocognitive performance when compared with the FEP clinical population. The clinic sample also had more diagnostic variability and higher prevalence of substance use disorders relative to the FEP research sample.

Conclusions Researchers should be aware of how study design and recruitment practices may impact the representativeness of samples, with particular concern for equal representation of racial minorities and patients with more severe illness. Studies should be designed to minimize burden to promote a wider range of participation.

Keywords Early psychosis · Schizophrenia · Neuroscience · Selection bias · Generalization

Introduction

The external validity or "generalizability" of clinical research hinges on the extent to which participant samples faithfully represent real-world populations of interest [1]. Despite researchers' best efforts, however, sample recruitment is an inexact science. Non-representative sampling may stem from systemic factors (for example, location of research hospitals), study design (stringent eligibility criteria, aggressive recruitment practices, high or low reimbursement for participants' time), or from participants themselves (mistrust of researchers, or conversely, eagerness to try anything that might help). Regardless of the source, it is typical for researchers to briefly acknowledge limitations to the generalizability of their findings, but more serious scrutiny of this potential problem is rare [2].

Evidence of sampling skew has been found in schizophrenia research studies, resulting in possible under-representation of women and racial minority participants. Although epidemiological studies have found that males represent about 58% of the population with schizophrenia, non-epidemiological studies tend to oversample men at a ratio of about

Some analyses included in this manuscript were first presented in poster format at the 2017 Department of Mental Health Research Centers of Excellence Conference in Boston, MA and at the 2017 Harvard Psychiatry Mysell Research Conference in Boston, MA.

Drs. Robert W. McCarley and Larry J. Seidman are deceased.

✉ Emily Kline
ekline@bidmc.harvard.edu

Extended author information available on the last page of the article

two to one [3]. For instance, in clinical trials of new atypical antipsychotic medications, the average number of women in the total samples was around 33.3% and the minimum was 6.7% [4]. The relative paucity of women participating in schizophrenia research limits understanding of gender differences in the presentation and course of the disease, and may result in women receiving sub-optimal treatment.

Under-representation of racial minorities in clinical research may also compromise the generalizability of findings. A review of 637 schizophrenia studies found that of the studies analyzed, only 27% of the studies reported the racial composition of the study participants, thus precluding serious consideration of the issue [5]. The impact of race on patients' interest in research participation is also contested. Although a legacy of racist research practices (e.g., the Tuskegee syphilis experiment) is thought to impede African Americans' willingness to participate in medical research in the United States, there is some evidence that minorities are equally interested in research participation relative to nonminority patient groups [6–8]. In the United Kingdom, where ethical guidelines emphasize internal (but not external) validity of research designs [9], clinicians' lack of confidence or possible misconceptions about research participation, as well as patients' concerns that research might harm their mental health, were noted as barriers to research participation in Woodall and colleagues' review on this topic [10]. A study of early psychosis patients in Norway found that those with a longer duration of illness were more likely to refuse participation in research [11]. Since duration of untreated psychosis (DUP) has been associated with poor prognosis but also with barriers to obtaining treatment [12, 13], the lower rate of participation among those with long DUP may have the effect of underestimating illness severity and barriers to treatment within this literature.

Stringent eligibility criteria may represent the greatest threat to external validity in the context of schizophrenia research. Despite widely accepted evidence that comorbidity is the norm in the early course of psychotic disorders—40% enter treatment with more than one psychiatric diagnosis, with substance use disorders accounting for the majority of these secondary diagnoses [14]—patients with comorbid conditions are often excluded by design. Researchers examining the standard eligibility criteria for antipsychotic trials found that only 14 of 50 (28%) patients with schizophrenia screened for clinical trial participation met the study criteria [15]; a second, larger study found that 1320 out of 6012 (22%) schizophrenia patients were eligible for trial participation [2]. Ineligible participants were more likely to be female, African American, and older [2]. Finally, participant-related factors such as attitudes toward research, lack of transportation to attend multiple appointments, mental illness stigma, cognitive impairment, and severe negative symptoms may also present barriers to enrollment in research protocols [10].

Despite the burgeoning research and clinical interest in establishing specialized care for individuals experiencing a first episode of psychosis (FEP), the question of whether research samples involving this group accurately reflect the clinical population of individuals seeking treatment for a first episode has not received substantial attention. The FEP category is somewhat more inclusive than schizophrenia; FEP samples include patients with schizophrenia as well as schizoaffective, schizophreniform, and psychosis "not otherwise specified" diagnoses. Further, different studies variously define FEP as the first months, year, or several years of psychotic illness. Given these vague diagnostic parameters, as well as the high hopes for the impact of intensive FEP intervention, ensuring that FEP research studies generalize to real world samples remains critical. A few researchers have examined this issue. In one study comparing early psychosis patients participating in a large drug trial to a large epidemiological cohort, the authors concluded that the stricter eligibility criteria of the drug trial did not appear to yield notable differences between the two samples with regard to age of onset, gender and premorbid functioning [16]. Conversely, "epidemiological" FEP samples (i.e., groups of patients tracked from clinical centers rather than selected for research) have tended to demonstrate worse prognosis (more hospital re-admissions, lower employment rate, and lower GAF scores) relative to selected research samples [17]. Conus and colleagues note that the process of informed consent is itself a powerful selection factor, since studies report refusal rates of 24–46% among FEP patients approached for participation in prospective research [17]. These observations are limited, however, by the inherent problems of comparing samples recruited at different sites, which are likely to differ in many meaningful ways. To date, there have been no published studies comparing FEP research participants to a naturalistic sample of early psychosis patients treated at the same center.

The aim of this study is to examine the comparability of a number of key variables in FEP research derived from a recruited FEP research sample ($n = 44$), a naturalistic FEP clinical sample ($n = 77$), and an age-matched control group ($n = 38$) obtained through a cross-sectional study and chart review. Although the primary comparisons of interest are between the two FEP samples, the control group was included to both serve as a check on site differences and to provide additional context for interpreting clinical and neurocognitive tests. Specifically, we compared demographic and clinical features including primary diagnosis, social and role functioning, and neurocognitive task performance in these three samples. In line with prior findings, we hypothesized that the clinical sample would contain higher

proportions of women, racial/ethnic minorities, and affective and substance use disorders relative to the research samples.

Methods

Settings and procedures

The Boston Center for Intervention Development and Applied Research (CIDAR). The Boston CIDAR study [18] aimed to identify markers of progression in schizophrenia (2007–2012). CIDAR recruited throughout the metropolitan Boston area via outreach to clinicians, local hospitals and clinics, advertisements, and word of mouth. Eligible CIDAR first episode psychosis (hereafter, "CIDAR-FEP"; $n=44$) participants met criteria for a DSM-IV-TR diagnosis of schizophrenia, schizophreniform, or schizoaffective disorder, were ages 13–45, and had psychosis onset within the past 5 years. CIDAR healthy volunteers ("CIDAR-HV"; $n=38$) were generally recruited from the same geographic areas as the case groups, with comparable age, gender, race and ethnicity, handedness, and parental socioeconomic status. No controls met criteria for any current or past major Axis I disorders. Controls were also excluded for any history of psychiatric hospitalizations, prodromal symptoms, schizotypal or other Cluster A personality disorders, first degree relatives with psychosis, or any current or past antipsychotic use (past psychotropic medication use before study enrollment was permitted). Exclusion criteria for all participants were: sensory-motor handicaps, neurological disorders, medical illnesses impacting neurocognitive function, diagnosis of mental retardation, lack of English fluency, DSM-IV-TR substance abuse in the past month or dependence in the past 3 months, current suicidality, a history of electroconvulsive therapy (ECT) within the past 5 years for cases and any history of ECT for controls.

CIDAR assessments took place at offices within 2 miles from the PREP® clinic. All CIDAR participants provided informed consent and demographic information and were assessed using measures described below. Within the CIDAR-FEP group, 23 (52%) were receiving psychotherapy at the time that they participated in the study and 35 (80%) were currently taking psychiatric medications. Of these, 15 were receiving care at PREP®. Clinical diagnoses were obtained using a Structured Clinical Interview for DSM-IV (SCID) administered by a masters- or doctoral-level clinician. All CIDAR procedures were approved by the Beth Israel Deaconess Medical Center IRB. Participants were paid for their time.

Prevention and Recovery in Early Psychosis (*PREP®*). PREP® [19] is an intensive outpatient program treating FEP patients ages 16–30 who have psychosis onset within the past 3 years. PREP® has few concrete exclusion criteria, but tends to refer elsewhere patients who have substance-induced psychosis, history of intellectual or developmental disability, lack of English fluency, or are perceived to require acute or inpatient level of care. PREP® services are funded by the Massachusetts Department of Mental Health. There are no insurance requirements or costs to PREP® patients, ensuring that ability to pay does not serve as a barrier to accessing care for this program. Most patients come from within the metro Boston area, but some travel from well outside the city to receive this specialized care. Thus PREP® does not represent a particular catchment, but instead illustrates a treatment-seeking clinical population.

As part of routine clinical practice, patients entering the program received a baseline assessment consisting of a diagnostic interview with a psychiatrist and one or more clinicians or clinical trainees as well as a psychological assessment including the Wechsler Test of Adult Reading (WTAR) oral word reading test, MATRICS consensus cognitive battery [20], and Global Functioning (GF) social and role scales [21]. De-identified clinical data were then entered into an outcome database. The vast majority (nearly all) of incoming patients does take advantage of this testing service and thus the database is an accurate reflection of the patients engaged in care at this program. Only PREP® patients assessed within the time frame coinciding with the CIDAR study (2007–2012) were included in the current analysis database ($n=77$). Because CIDAR recruited FEP participants at PREP®, the groups were cross-checked by members of the PREP® (EK) and CIDAR (RMG) teams who had access to identifying information. This check identified 15 CIDAR FEP participants who were also PREP® patients. These 15 participants were included in the CIDAR-FEP group only (i.e., deleted from the PREP® group), with the aim of defining these individuals within the research-participant population. The Massachusetts Department of Mental Health has approved use of this de-identified data for research.

Individuals administering assessments for both PREP® and CIDAR were supervised by licensed psychologists (authors MFY and RMG) who worked closely together and coordinated training procedures between the two projects. At PREP®, clinicians administering the GF scales and MCCB were trained annually by a licensed psychologist. They observed two administrations of the battery, and then were observed administering the battery at least once, before implementing assessments on their own. Further, the PREP® assessment team met for weekly supervision to discuss scoring and assessment issues. Within CIDAR, the same protocol was used to ensure reliability. Further, the psychologists supervising the clinical team and research team worked closely together to coordinate training protocols.

Measures

MATRICS Consensus Cognitive Battery (*MCCB*). The MCCB was designed to create a standard for measurement of treatment effects on cognition in schizophrenia, and has demonstrated strong reliability and minimal practice effects [20, 22]. The current investigation used nine MCCB subtests: Trail Making Test-A, BACS Symbol Coding, Hopkins Verbal Learning Test-Revised, Wechsler Memory Scale III: Spatial Span, Letter Number Span, Neuropsychological Assessment Battery: Mazes, Brief Visual Memory Test-Revised, Category Fluency (animal naming), and Continuous Performance Test-Identical Pairs (the Mayer–Salovey–Caruso Emotional Intelligence Test was not used due to problems with item comprehension among some PREP® clients). Because the samples were well-matched with regard to age and gender and MCCB T-scores are not reliable for individuals below age 20, raw scores were used in analyses.

Wechsler Test of Adult Reading (*WTAR*) *and Wide Range Achievement Test* (*WRAT*). PREP® patients were assessed using the Wechsler Test of Adult Reading assesses premorbid intellectual functioning [23]. CIDAR research participants (FEP and control) were assessed via the word reading list of the Wide Range Achievement Test, 4th Edition [24]. Both consist of brief word reading lists and provide estimates of premorbid IQ. The tests both provide a normed IQ estimate with mean of 100. Although they are different tests, previous research suggests that IQ estimates from the WRAT-4 and WTAR can be used interchangeably [25].

Global Functioning Scales. The Global Functioning Social and Role (GFS and GFR) Scales are 10-point scales that assess a person's interpersonal relationships and primary role performance, respectively. Both scales have displayed high construct validity and interrater reliability [21, 26].

Clinical diagnoses were based on interviews with the Structured Clinical Interview for DSM-IV-TR (SCID), Research Version [27–29] or the child version [30] for subjects < 18.

Data management and analyses

Data were screened for normality and outliers. One outlier was identified for the CPT task; this score was deleted but the individual was retained in the dataset. There was a substantial amount of missing data from the clinical (PREP®) dataset, which was not surprising. For example, only 55 of the 77 individuals represented in the PREP® database had a recorded diagnosis. We examined potential demographic (age, gender, educational attainment) difference between participants with and without complete data; no significant differences emerged. Individuals with missing data were excluded per analysis but retained within the overall sample. Sample sizes are specified per analysis within Tables 1, 2, 3 and 4.

Chi-square tests were performed to examine differences between groups with regard to overall diagnostic patterns,

Table 1 Demographic differences between clinical and research FEP and control groups

	PREP® (clinical)				CIDAR-FEP (research)				CIDAR-HV (research)				Test statistic	*df*	*p*
	N	Mean	SD	%	*N*	Mean	SD	%	*N*	Mean	SD	%			
Age	69	21.45[a]	3.13	–	44	21.92[a]	4.54	–	38	21.72[a]	3.20	–	$F=0.24$	2	0.79
Years education	47	12.05[a]	2.18	–	44	13.11[ab]	2.56	–	38	14.13[b]	2.33	–	$F=8.21$	2	<0.01
Est. premorbid IQ	44	96.86[a]	18.10	–	43	109.77[b]	16.43	–	38	110.79[b]	14.21	–	$F=9.51$	2	<0.01
Gender	77	–	–	–	44	–	–	–	38	–	–	–	$\chi^2=0.19$	2	0.91
Male	52	–	–	68	28	–	–	64	25	–	–	66			
Female	25	–	–	32	16	–	–	36	13	–	–	34			
Ethnicity	66	–	–		43	–	–		38	–	–		$\chi^2=1.04$	2	0.60
Hispanic or Latino	10	–	–	15	8	–	–	19	4	–	–	11			
Not Hispanic or Latino	56	–	–	85	35	–	–	81	34	–	–	89			
Race	66	–	–	–	44	–	–	–	38	–	–	–	$\chi^2=25.84$	12	0.01
White	24	–	–	36	27	–	–	61	23	–	–	61			
Black or Af. Am	32	–	–	48	6	–	–	14	8	–	–	21			
Amer. Indian/Alaskan native	0	–	–	0	1	–	–	2	0	–	–	0			
Asian	2	–	–	3	4	–	–	9	4	–	–	11			
Multi-racial	7	–	–	11	4	–	–	9	3	–	–	8			
Other	1	–	–	2	2	–	–	5	0	–	–	0			

[abc]Means that share a superscript did not significantly differ in Bonferroni-corrected post-hoc comparisons

Table 2 Primary axis one diagnoses in PREP® and CIDAR-FEP

DSM-IV-TR diagnoses	PREP® (clinical)		CIDAR-FEP (research)		χ^2 ($df = 1$)	p
	N	%	*N*	%		
Primary diagnosis						
Schizophrenia	25	46	27	61	2.48	0.16
Schizoaffective disorder	6	11	13	30	5.47	0.02
Psychosis "not otherwise specified"	15	27	0	0	14.14	0.00
Schizophreniform disorder	0	0	4	9	5.21	0.04
Bipolar disorder with psychotic features	6	11	0	0	5.11	0.03
Major depressive disorder with psychotic features	3	5	0	0	2.48	0.25
Total	55	100	44	100		

gender ratios, and racial and ethnic composition. One-way ANOVAs were employed to test for age, educational attainment, and premorbid IQ. Given identified group differences in racial composition and educational attainment, two subsequent ANCOVAs controlling for these variables was used to compare (1) MCCB subtest differences and (2) social and role functioning between the three groups. Post-hoc Bonferroni correction was used for estimates of paired differences between each group. An additional ANOVA compared MCCB and social and role functioning among the subgroup of CIDAR FEP participants who were recruited from PREP® vs. PREP® patients who did not participate in research. Raw MCCB subtest scores were converted to within-sample z scores (using the full sample means and standard deviations) for ease of interpretation in Fig. 1. All analyses were performed using SPSS.

Results

Demographic characteristics of each group and relevant test statistics are displayed in Table 1. No significant differences were observed between groups with regard to age and gender. The groups differed with regard to overall racial composition, but had similar proportions of Hispanic/Latino participants. The groups differed significantly with regard to educational attainment, in that control participants had on average a year more education than FEP research participants, who in turn had on average a year more education than FEP clinic patients. There were also significant group differences with regard to estimated premorbid IQ, with both research FEP and control samples scoring substantially higher than the FEP clinic patients.

Primary diagnoses for the PREP® and CIDAR-FEP groups are displayed in Table 2. Psychosis "NOS" diagnoses were more common in the clinical group (27%) compared to the research group (0%), and schizoaffective disorder diagnoses were more prevalent among the research participants relative to the clinic patients (30% vs. 11%). Consistent with entry criteria for the study, no FEP research participants had current substance use disorder diagnoses, but 5% of clinic patients did have a current alcohol use disorder and 14% had another substance use disorder at the time of their admission to the program. Diagnostic differences between groups were significant (see Table 2).

An ANOVA revealed group differences with regard to groups' racial composition and educational attainment (Table 1). CIDAR control participants had significantly more years of education than CIDAR FEP participants, who in turn had more education than PREP® patients. White participants were overrepresented in the two research groups (61%) relative to the clinic group (36%). Thus, we included years of education and race (dichotomized as white vs. non-white) in an ANCOVA examining MCCB and social/role functioning scores in the three groups. This ANCOVA was significant for all MCCB subtests, with an overall pattern of PREP® patients scoring lower than CIDAR-FEP, who in turn did worse relative to CIDAR-HV participants (Table 3; Fig. 1). The same pattern was found with regard to social and role functioning among the groups (Table 3; Fig. 2).

The subgroup of CIDAR FEP participants who were recruited from PREP® was directly compared to PREP® patients who did not participate in research with regard to neuropsychological, social, and role functioning. CIDAR participants from this subgroup performed significantly better than PREP® patients on three MCCB subtests: symbol coding, category fluency, and letter-number span (see Table 4).

Discussion

The aim of the current study was to examine the issue of research-to-clinic sample generalization by comparing research participants with a naturalistic (i.e., obtained via typical referral pathways to FEP care and meeting clinic-defined criteria for inclusion in the program) clinical sample assessed using the same measures with many of the same

Table 3 ANCOVAs testing group differences in neurocognitive, role, and social functioning

Matrics consensus cognitive battery	PREP® (clinical) $N=36$		CIDAR-FEP (research) $N=35$		CIDAR-HV (research) $N=34$		F	p	Effect size (partial η^2)
	Mean	SD	Mean	SD	Mean	SD			
TMT-A	37.36[a]	18.75	29.31[ab]	10.38	21.41[b]	9.68	7.34	<0.01	0.211
BACS-SC	43.47[a]	10.99	54.23[b]	12.70	67.71[c]	14.77	15.83	<0.01	0.327
HVLT-R	20.89[a]	5.93	26.06[b]	4.48	29.41[c]	3.45	17.39	<0.01	0.385
WMS-SS	14.61[a]	3.87	17.29 [ab]	3.81	19.00[b]	3.57	10.08	<0.01	0.349
LNS	12.08[a]	3.57	14.57[b]	3.74	17.71[c]	2.86	13.10	<0.01	0.344
NAB Mazes	15.81[a]	7.72	18.80 [ab]	5.98	21.18[b]	6.31	5.56	<0.01	0.182
BVMT-R	20.61[a]	7.32	26.29[b]	6.79	28.47[b]	5.48	8.04	<0.01	0.243
Cat. Fluency-A	18.25[a]	6.11	21.80[a]	7.08	26.15[b]	7.27	6.19	<0.01	0.198
CPT-IP	2.13 [a]	0.63	2.43[a]	0.94	3.08[b]	0.58	11.52	<0.01	0.315
Global functioning	PREP® $N=46$		CIDAR-FEP $N=43$		CIDAR-HV $N=38$		F	p	Effect size (partial η^2)
Social	4.67[a]	1.73	5.84[b]	1.48	8.66[c]	1.02	46.53	<0.01	0.604
Role	3.20[a]	2.04	5.91[b]	1.87	8.68[c]	1.04	62.73	<0.01	0.673

Race (dichotomized as white/non-white) and years of education included as covariates

TMT-A Trail Making Test- A; possible scores range from 0 to 300 with higher scores indicating poorer performance, *BACS-SC* Brief Assessment of Cognition in Schizophrenia-Symbol Coding; possible scores range from 0 to 110 with higher scores indicating better performance, *HVLT-R* Hopkins Verbal Learning Test-Revised; possible scores range from 0 to 36 with higher scores indicating with higher scores indicating better performance, *WMS-SS* Wechsler Memory Scale III: Spatial Span; possible scores range from 0 to 36 with higher scores indicating better performance, *LNS* Letter Number Span; possible scores range from 0 to 24 with higher scores indicating better performance, *NAB Mazes* Neuropsychological Assessment Battery: Mazes; possible scores range from 0 to 26 with higher scores indicating better performance, *BVMT-R* Brief Visual Memory Test-Revised; possible scores range from 0 to 36 with higher scores indicating better performance, *Cat. Fluency-A* Category Fluency (animal naming); possible scores range from 0 or greater with higher scores indicating better performance, *CPT-IP* Continuous Performance Test-Identical Pairs; possible scores range from 0 to 4 with higher scores indicate better performance

[abc]Means that share a superscript did not significantly differ in Bonferroni-corrected post-hoc comparisons

Table 4 ANOVA testing subgroup differences in neurocognitive, role, and social functioning

Matrics consensus cognitive battery	PREP® (clinical) $n = 37–52$		CIDAR-FEP recruited from PREP® (research) $n = 8–11$		F	p	Effect size (η^2)
	Mean	SD	Mean	SD			
TMT-A	38.09	17.89	36.78	13.31	0.04	0.84	0.001
BACS-SC	20.79	6.51	24.22	4.32	2.27	0.14	0.065
HVLT-R	43.19	10.73	51.09	15.06	4.25	0.04	0.043
WMS-SS	14.63	3.94	14.89	5.40	0.03	0.87	0.001
LNS	11.58	3.67	14.44	3.28	4.67	0.04	0.085
NAB Mazes	15.72	7.65	17.33	7.68	0.33	0.57	0.007
BVMT-R	19.80	7.67	22.22	7.81	0.73	0.40	0.015
Cat. Fluency-A	2.08	0.68	2.21	0.74	0.21	0.65	0.103
CPT-IP	17.90	6.04	23.00	5.15	5.50	0.02	0.005
Global functioning							
Social	4.66	1.64	5.00	1.25	0.39	0.54	0.006
Role	3.07	1.97	3.80	2.10	1.12	0.29	0.018

TMT-A Trail Making Test- A; possible scores range from 0 to 300 with higher scores indicating poorer performance, *BACS-SC* Brief Assessment of Cognition in Schizophrenia-Symbol Coding; possible scores range from 0 to 110 with higher scores indicating better performance, *HVLT-R* Hopkins Verbal Learning Test-Revised; possible scores range from 0 to 36 with higher scores indicating with higher scores indicating better performance, *WMS-SS* Wechsler Memory Scale III: Spatial Span; possible scores range from 0 to 36 with higher scores indicating better performance, *LNS* Letter Number Span; possible scores range from 0 to 24 with higher scores indicating better performance, *NAB Mazes* Neuropsychological Assessment Battery: Mazes; possible scores range from 0 to 26 with higher scores indicating better performance, *BVMT-R* Brief Visual Memory Test-Revised; possible scores range from 0 to 36 with higher scores indicating better performance, *Cat. Fluency-A* Category Fluency (animal naming); possible scores range from 0 or greater with higher scores indicating better performance, *CPT-IP* Continuous Performance Test-Identical Pairs; possible scores range from 0 to 4 with higher scores indicate better performance

clinical researchers, recruited from the same geographic area, and seen during the same period of time. The groups were well-matched with regard to age distribution and gender ratios. Although there were differences in the frequency of diagnoses between the FEP groups, this is likely due to the research group receiving a more structured and careful diagnostic interviews. The clinical sample had a higher proportion of "psychosis NOS" diagnoses; these may represent individuals who could technically meet criteria for a "schizoaffective" diagnosis if assessed with a structured diagnostic interview. The higher frequency of "NOS" diagnoses within the clinic group likely reflects the greater tolerance of diagnostic uncertainty, and the use of less structured assessment methods, within clinical (as compared to research) settings.

In other ways, however, the variety and magnitude of the differences we identified between groups was large and clinically meaningful. Consistent with our expectations, there was a higher prevalence of current substance abuse disorders within the clinic FEP group relative to the research FEP sample. This is clearly not surprising; individuals with active substance use disorders were excluded from participating in CIDAR. However, the presence of substance use disorders in the PREP® sample is still important, as it highlights one potential source of differences between the samples.

With regard to race, white participants were over-represented (and African Americans under-represented) in both the research FEP and control groups relative to the clinic population. In the United States, mistreatment of ethnic minorities in research contexts may lead to skepticism about the benefits (or even the safety of) research participation, thus leading to lower rates of representation in research samples [6–8]. Additionally, participating in research is less possible for individuals who have more stringent caregiver or financial responsibilities or more precarious employment or economic circumstances; these factors unfortunately often do align with race in the United States, which is another potential explanation for the observed underrepresentation of minorities in the CIDAR samples.

FEP research participants had more education and higher estimated premorbid IQ relative to clinic patients. There were also statistically significant and clinically meaningful differences between clinical and research groups with regard to neurocognitive and clinical characteristics. The CIDAR-FEP research sample performed around 0.6 standard deviations below the age-matched control group on most MCCB subtests, which is somewhat attenuated relative to prior findings in the field [31]. More notably, PREP® patients performed an additional 0.6 SD below the FEP research group

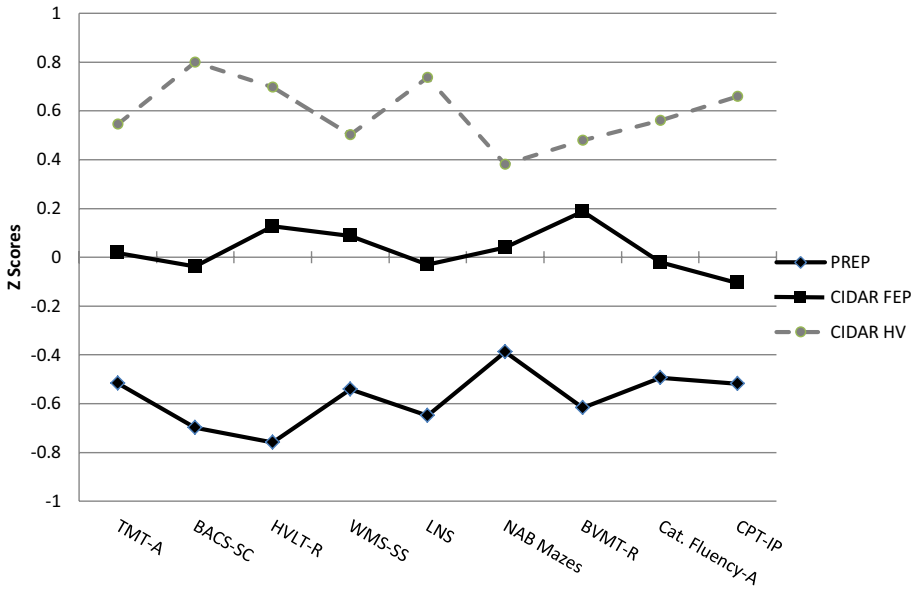
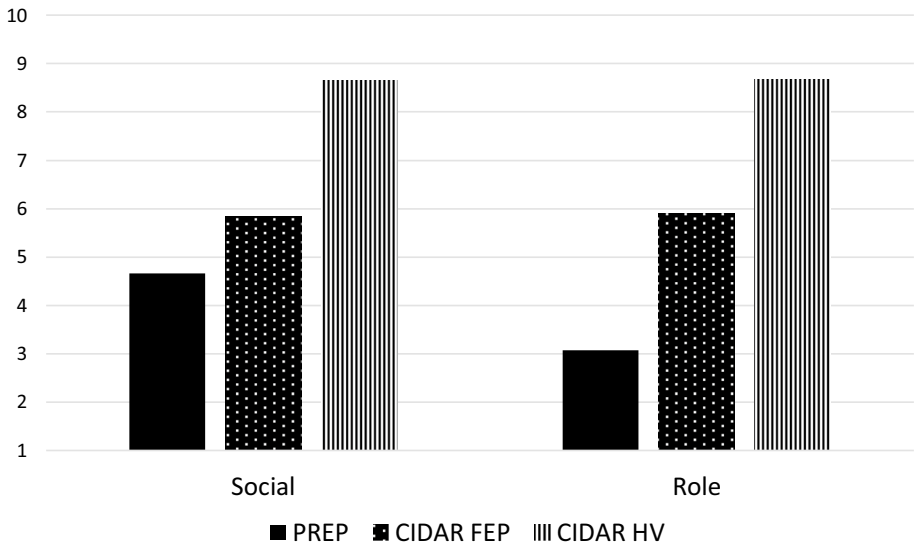


Fig. 1 MCCB Task Performance. *TMT-A* Trail Making Test-A; higher scores indicate poorer performance, *BACS-SC* Brief Assessment of Cognition in Schizophrenia-Symbol Coding; higher scores indicate better performance, *HVLT-R* Hopkins Verbal Learning Test-Revised; higher scores indicate better performance, *WMS-SS* Wechsler Memory Scale III: Spatial Span; higher scores indicate better performance, *LNS* Letter Number Span; higher scores indicate better performance, *NAB Mazes* Neuropsychological Assessment Battery: Mazes; higher scores indicate better performance, *BVMT-R* Brief Visual Memory Test-Revised; higher scores indicate better performance, *Cat. Fluency-A* Category Fluency (animal naming); higher scores indicate better performance, *CPT-IP* Continuous Performance Test-Identical Pairs; higher scores indicate better performance. Raw scores converted to z scores based on full sample means and standard deviations; TMT z scores multiplied by −1 for ease of interpretation

Fig. 2 Global functioning: social and role scales



(i.e., > 1 SD below controls) on most neurocognitive tasks, which is more consistent with the extant literature. This is particularly surprising given the research sample had a longer allowable duration of illness (less than 5 years) compared to the clinical sample (less than 3 years).

The same pattern was seen for social and role functioning: although nothing would look amiss comparing the relatively lower scores of the CIDAR-FEP group to the healthy volunteers, the functioning of the clinic patients was significantly worse relative to their research-participating peers. This pattern of results was confirmed when we narrowed focus to examine only the subgroup of CIDAR participants recruited from PREP® relative to other PREP® patients who either were not referred to the study, not eligible to participate, or declined to participate. In other words, while it is clear that the FEP research sample had much better

functioning than their clinic counterparts, it appears that even within the clinic, those who opted (or were encouraged) to participate in research had better social, role, and neuropsychological performance than non-research-participating PREP® patients. This is unsettling, in that it introduces problems with the generalization of research findings and may result in underestimating the true extent of the impairments associated with first episode psychosis.

One possible explanation for the differences found between the FEP research vs. clinical groups is the differential proportion of racial minorities. The FEP-clinical (PREP®) group had higher proportions of non-white patients relative to the FEP research sample. Some measures may have differential validity across ethnic and cultural groups. Further, assessment measures that are culturally inappropriate increase the probability of making incorrect assumptions about their study sample [1]. Estimates of IQ, in particular, may be vulnerable to social factors (e.g., language barriers or lack of educational opportunities) rather than reflecting true premorbid IQ differences between white vs. non-white participants, artificially deflating scores in the FEP-clinical group.

Another possible explanation concerns the high demands of research participation. Research participation requires a degree of organization, motivation, and skill: participants must be able to read and understand a long consent form, manage multiple appointments, and have the stamina and motivation to complete assessments. Studies may inadvertently exclude participants with more severe negative symptoms, disorganization, or cognitive deficits by simply expecting too much of them. Additionally, patients with pronounced negative symptoms may be unmotivated by research benefits such as gift cards, money, or clinical consultations. Clinicians are also sensitive to these factors, and may prioritize symptom stabilization over referrals to research for patients struggling with organization and adherence. Additionally, patients with a conceptualization of their experience that does not align with a medical model of illness may less likely to volunteer for studies that identify participants as belonging in defined disease categories. For instance, a qualitative study conducted in the U.K. found that older people, especially Black African-Caribbean seniors, do not necessarily view depression as a mental illness, thus rendering their participation in depression-focused medical illness unlikely [32].

Stringent eligibility criteria represent the most likely factor contributing to the differences between research and clinical samples in the current study. Although research focusing on biological/brain-based variables strives to eliminate potential confounds such as substance use disorders, neurodevelopmental disorders, and traumatic brain injuries, this must be balanced against the competing scientific need to faithfully represent populations of interest. CIDAR eligibility criteria were typical of the field in this respect: potential participants with substance use disorders, neurological disorders, premorbid IQ less than 70, or educational attainment of less than ninth grade were excluded from the first episode sample. For instance, at least 14% of PREP® clients had a substance use disorder, and three had estimated IQ scores below 70. Additionally, the most severely ill patients may more often be deemed not competent or otherwise ready for research participation. However, these factors are part of the "true" picture of FEP presentation in the real world.

This study has several limitations. First, history of head injury and neurologic disorders, which could account for cognitive impairment in some cases, was not assessed systematically in the clinical sample and thus could not be included as a covariate in our analyses. Additionally, the PREP® group was not assessed using the SCID, which likely accounts for the differences in diagnostic distribution between the FEP groups. Second, some of the clinical data was missing, thus findings should be interpreted with that caution in mind. Third, this study compares only identified clinic patients to research participants; we cannot speak to the similarities or differences of either of these groups to the larger population of true FEP in the community, including many individuals who do not seek treatment for their psychosis. We also do not know whether clinic patients may have participated in studies other than CIDAR. Additionally, this study lacks any patient self-report data that could corroborate clinician impressions of illness severity and functioning. Another limitation is that reliable data reflecting participants' socioeconomic status and experiences of childhood adversity both of which have been implicated as potential risk factors for psychosis development and chronicity [33, 34] was not collected. Finally, the current study is a secondary analysis of two datasets originally collected for different purposes, and does not reflect the primary aims of either project.

It is also worth pointing out that a small sample size is a limitation. While generalization is important for studies involving application of best practice principles to the wider community, studies of neurobiology will still need homogeneous samples for appropriate interpretation of the findings. Larger samples will allow potential stratification of the populations and thereby enable investigations of a broader range of questions. However, it is possible that even large samples of treatment-seeking individuals with psychosis would fail to reflect the broader United States FEP population, many of whom do not receive treatment or routinely intersect with health care settings [35].

Despite these limitations, the findings represent a call to the field to examine study design and recruitment practices in psychiatric research. We conduct research to facilitate awareness, to advance knowledge, and most importantly

to support ideas with scientific facts. If our research samples reflect only a subset of the populations of interest, the general application of our findings carries less weight. This becomes especially important for the dissemination of the so-called "best practices": clinicians tasked with implementation may be rightly skeptical if they believe the research does not accurately represent their clientele. Researchers, on the other hand, must balance a challenging set of competing priorities: the elimination of confounds and a desire for hefty datasets serving many research goals, vs. faithful representation of the clinical population and generalizability of results. Research studies, if available, should be offered to patients in a standard way to ensure that certain groups are not overlooked or excluded. Study measures should therefore be carefully selected for cultural as well as psychometric validity to maximize the generalizability of research findings, and protocols should be designed to minimize burden and promote a wide range of participation.

Acknowledgements This work was supported in part by the National Institutes of Health via grants R01 MH103831, P50MH080272, UO1 MH081928, R01MH103831, R01MH102377, 1S10RR023401,1S10RR019307, and 1S10RR023043; by the Massachusetts Department of Mental Health (SCDMH82101008006); by a VA Merit Award (MES); by the National Alliance for Research in Schizophrenia and Depression via the Distinguished Investigator Award (MES), and by a Clinical Translational Science Award UL1RR025758 and General Clinical Research Center Grant M01RR01032 to Harvard University and Beth Israel Deaconess Medical Center from the National Center for Research Resources. We also wish to acknowledge the patients and their families for working with us in PREP® as well as the many individuals who contributed to the PREP® program, including Cynthia Berkowitz, Brina Caplan, Margaret Guyer, Jude Leung, Thomas Monteleone, and Ginger Smith. We also thank the clinical, research assistant, and data management staff from the Boston CIDAR study, including Caitlin Bryant, Ann Cousins, Grace Francis, Molly Franz, Lauren Gibson, Anthony Giuliano, Andréa Gnong-Granato, Maria Hiraldo, Sarah Hornbach, Kristy Klein, Grace Min, Corin Pilo-Comtois, Janine Rodenhiser-Hill, Julia Schutt, Shannon Sorensen, Reka Szent-Imry, Alison Thomas, Lynda Tucker, Chelsea Wakeham, and Kristen Woodberry.

Compliance with ethical standards

Conflict of interest The authors have no conflicts of interest to declare.

Ethical standards All CIDAR participants provided informed consent prior to participating in the study. The use of de-identified data from the PREP® clinic for research has been approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center and the Massachusetts Department of Mental Health.

References

1. Sue S (1999) Science, ethnicity, and bias: where have we gone wrong? Am Psychol 54:1070
2. Robinson D, Woerner MG, Pollack S et al (1996) Subject selection biases in clinical trials: data from a multicenter schizophrenia treatment study. J Clin Psychopharmacol 16:170–176
3. Longenecker J, Genderson J,et al (2010) Where have all the women gone? participant gender in epidemiological and non epidemiological research of schizophrenia. Schizophr Res 119:240–245
4. Chaves AC, Seeman MV (2006) Sex selection bias in schizophrenia antipsychotic trials. J Clin Psychopharmacol 26:489–494
5. Chakraborty BH, Steinhauer SR (2010) Reporting of minority participation rates and racial differences in schizophrenia and psychophysiological research: improving but still not adequate. Schizophr Res 123:90–91
6. Wendler D, Kington R, Madans J et al (2009) Are racial and ethnic minorities less willing to participate in health research? PLoS Med 3:e19
7. Hamilton LA, Aliyu MH, Lyons PD et al (2006) African-American community attitudes and perceptions toward schizophrenia and medical research: an exploratory study. J Natl Med Assoc 98:18
8. Thompson EE, Neighbors HW, Munday C et al (1996) Recruitment and retention of African American patients for clinical research: an exploration of response rates in an urban psychiatric hospital. J Consult Clin Psychol 64:861
9. Rothwell P (2005) External validity of randomized controlled trials: "to whom do the results of this trial apply?". Lancet 365:82–93
10. Woodall A, Morgan C, Sloan C et al (2010) Barriers to participation in mental health research: are there specific gender, ethnicity and age related barriers? BMC Psychiatry 10:103
11. Friis S, Melle I, Larsen TK et al (2004) Does duration of untreated psychosis bias study samples of first-episode psychosis? Acta Psychiatr Scand 110:286–291
12. Marshall M, Lewis S, Lockwood A et al (2005) Association between duration of untreated psychosis and outcome in cohorts of first-episode patients: a systematic review. Arch Gen Psychiatry 62:975–983
13. Compton MT, Ramsay CE, Shim RS et al (2009) Health services determinants of the duration of untreated psychosis among African-American first-episode patients. Psychiatr Serv 60:1489–1494
14. Pope MA, Joober R, Malla AK (2013) Diagnostic stability of first-episode psychotic disorders and persistence of comorbid psychiatric disorders over 1 year. Can J Psychiatry 58:588–594
15. Khan AY, Preskorn SH, Baker B (2005) Effect of study criteria on recruitment and generalizability of the results. J Clin Psychopharmacol 25:271–275
16. Rabinowitz J, Bromet EJ, Davidson M (2003) Are patients enrolled in first episode psychosis drug trials representative of patients treated in routine clinical practice? Schizophr Res 61:149–155
17. Conus P, Cotton S, Schimmelmann BG et al (2017) Rates and predictors of 18-months remission in an epidemiological cohort of 661 patients with first-episode psychosis. Soc Psychiatry Psychiatr Epidemiol 1–1
18. Woodberry KA, Serur RA, Hallinan SB et al (2014) Frequency and pattern of childhood symptom onset reported by first episode schizophrenia and clinical high risk youth. Schizophr Res 158:45–51
19. Caplan B, Zimmet SV, Meyer EC et al (2013) Prevention and recovery in early psychosis (PREP1): building a public-academic partnership program in Massachusetts, united states. Asian J Psychiatr 6:171–177
20. Nuechterlein KH, Green MF, Kern RS et al (2008) The MATRICS Consensus Cognitive Battery, part 1: test selection, reliability, and validity. Am J Psychiatry 165:203–213

21. Cornblatt BA, Auther AM, Niendam T et al (2007) Preliminary findings for two new measures of social and role functioning in the prodromal phase of schizophrenia. Schizophr Bull 33:688–702
22. Kern RS, Nuechterlein KH, Green MF et al (2008) The MATRICS Consensus Cognitive Battery, part 2: co-norming and standardization. Am J Psychiatry 165:214–220
23. Wechsler D (2001) Wechsler test of adult reading: WTAR. Psychological Corporation, San Antonio
24. Wilkinson GS, Robertson GJ (2006) Wide range achievement test. Psychological Assessment Resources, Lutz, FL
25. Mullen CM, Fouty HE (2014) Comparison of the WRAT4 reading subtest and the WTAR for estimating premorbid ability level. Appl Neuropsychol Adult 21:69–72
26. Piskulic D, Addington J, Auther A et al (2011) Using the global functioning social and role scales in a first-episode sample. Early Interv Psychiatry 5:219–223
27. First MB, Spitzer RL, Gibbon M (2002) Structured clinical interview for DSM-IV-TR axis I disorders. Research version, patient edition. (SCID-I/P) New York. Psychiatric Institute, New York
28. Lobbestael J, Leurgans M, Arntz A (2011) Inter-rater reliability of the structured clinical interview for DSM-IV axis I disorders (SCID I) and axis II disorders (SCID II). Clin Psychol Psychother 18:75–79
29. Zanarini MC, Skodol AE, Bender D et al (2000) The collaborative longitudinal personality disorders study: reliability of axis I and II diagnoses. J Pers Disord 14:291–299, 2000
30. Hien D, Matzner FJ, First MB et al (1994) Structured clinical interview for DSM–IV–child edition (version 1.0). Columbia University, New York
31. Mesholam-Gately RI, Giuliano AJ, Goff KP et al (2009) Neurocognition in first-episode schizophrenia: a meta-analytic review. Neuropsychology 23:315–336
32. Marwaha S, Livingston G (2002) Stigma, racism or choice. why do depressed ethnic elders avoid psychiatrists? J Affect Disord 72:257–265
33. Varese F, Smeets F, Drukker M et al (2012) Childhood adversities increase the risk of psychosis: a meta-analysis of patient control, prospective-and cross-sectional cohort studies. Schizophr Bull 38:661–671
34. Longden E, Sampson M, Read J (2016) Childhood adversity and psychosis: generalised or specific effects? Epidemiol Psychiatr Sci 25:349–359
35. Schoenbaum M, Sutherland JM, Chappel A et al (2017) Twelve-month health care use and mortality in commercially insured young people with incident psychosis in the United States. Schizophr Bull 43:1262–1272

Affiliations

Emily Kline[1,2] · Victoria Hendel[2] · Michelle Friedman-Yakoobian[1,2] · Raquelle I. Mesholam-Gately[1,2] · Ann Findeisen[2] · Suzanna Zimmet[1,2] · Joanne D. Wojcik[1,2] · Tracey L. Petryshen[1,3] · Tsung-Ung W. Woo[1,2,4] · Jill M. Goldstein[1,5] · Martha E. Shenton[1,6,7] · Matcheri S. Keshavan[1,2] · Robert W. McCarley[1,7] · Larry J. Seidman[1,2,3]

1 Department of Psychiatry, Harvard Medical School, Boston, MA 02115, USA

2 Massachusetts Mental Health Center, Public Psychiatry Division of the Beth Israel Deaconess Medical Center, 75 Fenwood Road, Boston 02115, MA, USA

3 Department of Psychiatry, Center for Human Genetic Research, Massachusetts General Hospital, 185 Cambridge St., Boston 02114, MA, USA

4 Laboratory of Cellular Neuropathology, McLean Hospital, 115 Mill St., Belmont, MA 02478, USA

5 Department of Medicine, Department of Psychiatry, Connors Center for Women's Health and Gender Biology, Brigham and Women's Hospital, 1620 Tremont St., Boston, MA 02120, USA

6 Department of Psychiatry and Radiology, Brigham and Women's Hospital, 1249 Boylston Street, 02215 Boston, MA, USA

7 VA Boston Healthcare System, 940 Belmont St., Brockton, MA 02301, USA