AIDS and Behavior
https://doi.org/10.1007/s10461-020-02883-5

ORIGINAL PAPER

Psychometric Data Linking Across HIV and Substance Use Cohorts

Benjamin D. Schalet[1] · Patrick Janulis[1,3] · Michele D. Kipke[2] · Brian Mustanski[1,3] · Steven Shoptaw[4] · Richard Moore[5] · Marianna Baum[6] · Soyeon Kim[7] · Suzanne Siminski[7] · Amy Ragsdale[8] · Pamina M. Gorbach[8,9]

Abstract

Psychometric data linking of psychological and behavioral questionnaires can facilitate the harmonization of data across HIV and substance use cohorts. Using data from the Collaborating Consortium of Cohorts Producing NIDA Opportunities (C3PNO), we demonstrate how to capitalize on previous linking work with a common linked depression metric across multiple questionnaires. Cohorts were young men who have sex with men (MSM), substance-using MSM, HIV/HCV cocaine users, and HIV-positive patients. We tested for differential item functioning (DIF) by comparing C3PNO cohort data with general population data. We also fit a mixed-effects model for depression, entering HIV-status and recent opioid/heroin use as fixed effects and cohort as a random intercept. Our results suggest a minimal level of DIF between the C3PNO cohorts and general population samples. After linking, descriptive statistics show a wide range of depression score means across cohorts. Our model confirmed an expected positive relationship between substance use and depression, though contrary to expectations, no significant association with HIV status. The study reveals the likely role of cohort differences, associated patient characteristics, study designs, and administration settings.

Keywords Linking · Harmonization · Depression · HIV · Substance use · MSM

Introduction

Substantial progress has been made in reducing new HIV infections among people who use drugs [1], largely driven by reductions in incidence among those who inject drugs (PWID) [2]. Nevertheless, substance use remains a risk factor for continuing HIV transmission, fed by the opioid use disorder epidemics across the United States and Canada in the past decade [3]. The vast scope of this problem demands data and analytic solutions that can assess unique mechanisms within and across different subpopulations of people who use drugs. Integrative data analysis using multiple datasets provides one such methodological solution [4] and may be particularly valuable for studies of drug use and HIV [5]. For example, effective and valid pooling of data permits investigators to increase power (and reduce Type 2 errors) to study risk factors, test novel hypotheses, and better understand potential sources of variation across subpopulations [6]. The Collaborating Consortium of Cohorts Producing NIDA Opportunities (C3PNO) was recently established to capitalize on the similarities among longitudinal observational cohort studies in the area of HIV-risk and substance use. C3PNO is the coordinating center for nine National

✉ Benjamin D. Schalet
b-schalet@northwestern.edu

1 Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, 633 N. St. Clair, 19th Floor, Chicago, IL 60611, USA

2 Department of Pediatrics, Children's Hospital Los Angeles, Los Angeles, USA

3 Institute for Sexual and Gender Minority Health and Wellbeing, Northwestern University, Chicago, USA

4 Department of Family Medicine, David Geffen School of Medicine, University of California, Los Angeles, USA

5 Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, USA

6 Department of Dietetics and Nutrition, Robert Stempel College of Public Health, Florida International University, Miami, USA

7 Frontier Science Foundation, Brookline, MA, USA

8 Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, USA

9 Division of Infectious Diseases, David Geffen School of Medicine, University of California, Los Angeles, USA

Published online: 19 May 2020

Institute on Drug Abuse (NIDA) cohorts that represent key populations and has a combined sample size of 12,000 active and 20,000 historical participants. Together, the cohorts allow for understanding the spectrum of responses across populations at risk of HIV who use substances.

Pooled data research, like all research, relies on reliable and valid measurement; for substance use research, key health outcome measures include pain, substance use behaviors, anxiety, and depression. The myriad of choices available in terms of instruments to measure the same or similar psychological and behavioral constructs presents challenges for appropriately combining data and drawing valid inferences. For example, a review into anxiety measures alone identified 92 empirically-based questionnaires that measure this construct [7]. Data coordinating centers for independently established cohorts are likely to encounter this problem, because different cohorts may have historical data on similar constructs collected with different instruments, and will be understandably constrained in their willingness and/or ability to change instruments in order to standardize across cohorts. Cohorts must balance the need to maintain standardization within their own cohort—having followed individuals longitudinally and accumulated extensive data on specific measures—with the need to make such measures comparable to those collected in other studies. The incommensurability of scores across instruments may also represent a problem for the downstream synthesis of research findings: are differences in findings real or do they reflect methodological artifacts?

A robust solution to the multiple measures problem is to conduct cross-sectional studies in which multiple instruments are administered to large samples to bridge across instruments. This should be followed with the application of Item Response Theory (IRT) and other linking methods in order to establish the equivalence of items across measures [8–10]. IRT is a set of mathematical models that allow researchers to assign unique "properties" (i.e., parameters) to each item on a questionnaire based on how likely people with different levels of the measured construct are to endorse each response option [11, 12]. With sufficient data (and meeting of assumptions), analysts can jointly estimate parameters for items on multiple questionnaires. This co-calibration allows total scores from one questionnaire to be reliably translated into the scores of another. In the field of patient-reported outcomes (PRO), we have applied these IRT linking methods in the PROsetta Stone project (Cella, D: 1RC4CA157236) in the area of depression [13], anxiety [14], and pain [15], among others. Similarly, these approaches have been used in studies of drug dependence [16, 17], alcohol use symptoms [18], and HIV knowledge [19]. This approach has also been used previously to link depression measures for a longitudinal cohort study that examines influences on HIV and substance use [20].

The current study is a necessary step to provide a rigorous basis to justify a comparison of scores from different depression scales across C3PNO cohorts. Our study leverages depression data from five C3PNO cohorts, as well existing general population data, to develop a common metric across these cohorts. We validate the measurement characteristics of depression measures of each C3PNO cohort against population level data and use previously established item parameters to provide a common metric for these measures. In addition, we demonstrate the value of such a pooled data set by examining the relationship between depression, HIV, and opioid use. Given the complex association between HIV, substance use, and depression [21, 22], we examine the main effects of HIV and opioid use on depression as well as the potential interaction effect indicating that the association between opioid use and depression may vary according to HIV status.

Methods

Study Populations

The nine C3PNO cohorts, spanning the US and Canada, follow a diverse group of high-risk HIV-negative and HIV-positive persons including substance-using youth, PWID, stimulant users, MSM, racial/ethnic minorities, transgender individuals, and HIV-positive individuals with transmissible viral loads. Some cohorts exclusively enroll PWID or HIV-positive persons. Most cohorts enrolled HIV-positive and negative individuals from community samples—only two of the nine enroll at clinical care sites. For these analyses, the five participating C3PNO cohorts were young MSM (HYM, RADAR and mSTUDY), HIV/HCV cocaine users (MASH), and HIV-positive patients in care (JHHCC).

Detailed descriptions of the clinical and demographic characteristics of the nine C3PNO cohorts have been published elsewhere [23]. Briefly, 39% of persons across cohorts are HIV-positive. Most cohort members who ever injected illicit substances are over 30 years of age and HIV-positive, reflecting the cohort characteristics. Recent substance use is as follows: 30% heroin or illicit (non-medical) prescription drugs; 30% heroin injection; 15% non-medical prescription drugs (opioids etc.); 44% illicit stimulants including methamphetamine and cocaine; and 24% injected any of these drugs. Viral suppression at last visit among HIV + heroin, cocaine/crack, prescription drugs, and methamphetamine users was 36, 50, 50, and 56%, respectively, and significantly lower than non-users of these substances [23].

For comparison purposes, data on depression measures were also obtained from two PROsetta Stone linking studies [13]. Participants were recruited from the US general population by internet panel companies. Polimetrix (now

YouGov) collected the self-report data on the CESD and PROMIS from 747 individuals who were part of the original "full bank" PROMIS calibration sample [24]. This first sample was 51.9% female, 9.5% Hispanic, 80.5% White, and 10.1% Black; the mean age was 51.3 (SD = 18.8). Similarly, PHQ-9 and PROMIS data were collected for 748 respondents during the calibration phase of the NIH Toolbox study [25] by Greenfield Online (now Toluna; www.tolunagroup.com). (Because the PHQ-8 was administered in a C3PNO cohort, NIH Toolbox data on the ninth item of the PHQ-9 was not used.) The second sample was 56.1% female, 15.2 Hispanic, 80.1 White, and 9.1 Black; the mean age was 47.2 (15.2) [13].

Measures

PROMIS Depression Bank

The PROMIS Depression bank v1.0 for adults consists of 28 items with a 7-day time frame and a 5-point scale, with response options ranging from "Never" to "Always" [24, 26]. Item content covers emotional, cognitive, and behavioral symptoms of depression rather than somatic symptoms. The items were calibrated with IRT such that different subsets of items from the instrument bank can be reported on the same metric [24]. The T-score metric of PROMIS Depression is computed from the IRT-based theta of person scores (T-score = [θ × 10] + 50). The HYM study administered 23 items from the PROMIS Depression bank, while the RADAR cohort completed the PROMIS Depression 8a short form [27].

PHQ-8

The PHQ-8 is an eight-item instrument designed to assess depression in primary care [28]. It was originally developed as a nine-item version, which also assesses suicidal ideation [29]. The PHQ-9 was developed to directly reflect the criteria for major depressive disorder in Diagnostic and Statistical Manual of Mental Disorders (4th ed.; DSM–IV; American Psychiatric Association, 1994). Respondents rate their symptoms over the last 2 weeks, using a 4-point scale ranging from 0 (Not at all) to 3 (Nearly every day).

CES-D

The CES-D is a 20-item measure designed to assess depressive symptoms in the general population [30]. The CES-D has been used in a variety of contexts, including community samples and clinical samples with both medical and psychiatric conditions [31–33]. Respondents rate their symptoms based on the past week using a 4-point scale that ranges from 0 (Rarely or none of the time) to 3 (Most or all of the time). Given the variability in the strength of association between the four reverse-coded items and the remainder of the scale across samples [34], we decided to limit our linking analysis to the 16 positively-coded items.

In addition, we obtained demographics, current HIV status (confirmed with rapid antibody and/or HIV viral load test), and self-reported recent substance use based on use during the past 3 to 6 months.

Data Collection

Questionnaires were administered by each C3PNO cohort according to their schedule of evaluations and provided by each cohort's data center. Both mSTUDY and RADAR provided baseline data, while JHHCC, JYM, and MASH provided data on the last full visit.

Statistical Methods

Our aim was to take advantage of previous cross-walk tables or item parameters that placed depression measures on a common metric [13]. Doing so directly, however, rests on the assumption of the absence of differential item functioning (DIF, or measurement invariance) between the general population used in linking and the HIV and substance use populations of the C3PNO cohort. Therefore, we tested for DIF by comparing C3PNO cohort data with general population data from linking studies, with the aim to delete any items displaying non-trivial DIF. Our DIF analysis is further enhanced by IRT-based scoring, which in turn assumes relative unidimensionality of the item set. To do this, we estimated the proportion of total variance attributable to a general factor known as omega-hierarchical (omega-h) using the **psych** package in **R** [35]. This method estimates omega-h from the general factor loadings in an exploratory bi-factor model [36–38]. Values of 0.70 or higher for omega-h suggest that the item set is sufficiently unidimensional for many purposes [39].

Next, we applied logistic ordinal regression using the **lordif** package in **R** [40]. The logistic regression approach is one of several DIF methods; one of its advantages is that it provides a measure of magnitude (i.e., an effect size for each item) [41]. The lordif program identifies DIF by matching individuals on an estimate of the underlying trait being measured, then examines differences in IRT parameters across matched individuals. The matching variable will be based on IRT scale scores, based on the graded response model [42] which is preferable to summed scores (which would assume that items have equal discrimination power). The application of IRT in this DIF context permits the use of an iterative purification procedure, whereby items initially flagged with DIF are assigned sample-specific parameters. In subsequent regressions, new IRT scaled scores (based on

both DIF and DIF-free items) are used as the matching criterion to re-identify DIF items. The process is repeated until the same set of items is found to have DIF over successive iterations. For DIF items, lordif generates item characteristic curves illustrating how the expected item score relates to the underlying trait (e.g., depression), and how group membership affects this relationship [43].

Comparisons between different regression models will allow us to determine whether (a) there is any DIF at all, (b) uniform DIF only, or (c) non-uniform DIF. We used the chi-squared likelihood-ratio statistic as the DIF detection criteria (alpha < 0.01). We will use a cut-off of McFadden pseudo $R^2\Delta \geq 0.016$ in model comparisons to further identify non-trivial DIF, a relatively conservative threshold in the field of self-reported health outcomes [41, 44]. DIF-free items were scored with linked (general population) item response theory (IRT) parameters and placed on the PROMIS T-score metric using parameters previously published [13].

Once cohort depression scores were on the T-score metric, we computed descriptive statistics. To ease interpretation of the differences among cohorts, we also calculated the percentage of patients meeting the T-score threshold of 60 (one standard deviation above the general population estimate) as a threshold of "moderate" depression. This is a reasonable threshold given the thresholds of other measures [13, 45], and work with patients diagnosed with depression [46].

Finally, we fit a mixed-effects model for depression, entering HIV-status and recent opioid/heroin use as fixed effects and cohort as a random intercept, controlling for participant age at the time of interview. The **lme4** package of **R** was used to estimate the mixed effect model [47].

Results

C3PNO Cohort Characteristics

Table 1 shows the demographic and clinical characteristics for C3PNO cohorts, as well as the particular depression instrument administered in each of the cohorts. The mean sample size from each cohort was 757 (range 448–1047). The mean age ranged from 21 (RADAR) to 55 (JHHCC and MASH). The proportion of participants who are HIV-positive varied among non-clinical cohorts, ranging from 11 to 51%, and is 100% for those in HIV care.

General Population Samples

Demographic characteristics of the two general population internet panel samples are described in detail elsewhere [13]. Briefly, the majority of the participants were white (80% for both samples), mostly female (CESD sample: 52%; PHQ-8 sample: 56%), and the average age was 51 in the CESD sample and 47 in the PHQ-8 sample.

Internal Consistency and Unidimensionality

Table 2 shows the internal consistency and unidimensionality estimates for the measures in their C3PNO cohort and the comparison general population sample. Estimates across all samples were high for Cronbach's alpha (range 0.88 to 0.98), as well as for omega-h (0.79 to 0.92), suggesting a sufficient level of reliability and unidimensionality to proceed with DIF analyses. Values for the general population samples were generally higher than for the C3PNO cohorts.

Table 1 Demographic and clinical characteristics of study population across C3PNO cohorts

	HYM	RADAR	mSTUDY	MASH	JHHCC
N	448	1040	521	862	918
Age, mean [IQR]	22.3 [3.1]	21.3 [4.4]	31.3 [11.0]	54.9 [10.0]	54.7 [11.2]
Race, % (n)					
Black	21.0 (94)	34.0 (354)	40.7 (209)	59.5 (513)	85.7 (787)
Hispanic/Latino	58.9 (264)	29.8 (310)	34.6 (183)	23.8 (205)	1.4 (13)
White	0.0 (0)	25.1 (261)	14.0 (72)	8.5 (73)	12.5 (115)
Other/multiple	20.1 (90)	11.1 (115)	9.7 (50)	61 (53)	0.3 (3)
HIV+, % (n)	11.4 (51)	16.4 (171)	50.4 (257)	50.1 (429)	100.0 (918)
Recent opioid use, % (n)	16.3 (66)	6.2 (64)	13.1 (67)	28.9 (249)	6.5 (56)
Depression instrument used	PROMIS Bank (23 items)	PROMIS SF 8a	CESD	CESD	PHQ-8

IQR interquartile range

Table 2 Estimates of internal consistency and unidimensionality by individual C3PNO cohort vs general population data

	PHQ-8		PROMIS 23 items		PROMIS 8a SF		CESD 16 items		
Cohort/sample	JHHCC	Gen Pop	HYM	Gen Pop	RADAR	Gen Pop	MASH	mSTUDY	Gen Pop
Cronbach's alpha	0.88	0.91	0.91	0.98	0.95	0.95	0.90	0.92	0.94
Omega hierarchical	0.79	0.81	0.85	0.90	0.88	0.92	0.81	0.80	0.89

Differential Item Functioning and IRT Scoring on a Common Metric

Tables 3, 4 and 5 show details on the uniform and non-uniform pseudo-R^2 effect sizes for DIF between C3PNO cohort and the general population sample. Among the 71 item comparisons, we found evidence for non-trivial DIF in 13 items across cohorts, using the $R^2 > 0.016$ criterion. For the CESD, 7 items showed DIF for the general population sample and MASH, but only 1 for the general population and mSTUDY. For PROMIS, 5 out of 23 items showed DIF between HYM and the general population. Figure 1 shows item characteristic curves for 4 items with non-trivial DIF. No DIF items were found for RADAR (PROMIS 8a) and JHHCC (PHQ-8) and the general population sample. Using only non-DIF items, we scored the C3PNO depression item data on a common PROMIS T-score metric using standard expected a priori scoring [48].

Distribution of Depression Across Cohorts

Figure 2 shows the depression distribution by cohort; T-scores ranged from a mean of 47.6 (JHHCC) to 55.8 (mSTUDY), representing a range of 8 T-score points (or 0.8 standard deviation). A T-score of 60 is equivalent to a PHQ-9 raw score of 10 [13], which is commonly used in medical settings to identify likely clinically significant depression [29, 49, 50]. Using T-score of 60 as a cut-point, percentages of probable positive depression cases in the cohorts were 11% (JHHC) 21% (HYM), 24% (RADAR), and 38% (mSTUDY). Depression was substantially lower for JHHCC relative to the other cohorts (> 4.4 T-score units).

Mixed Effects Model for Depression

Excluding JHHCC in the mixed-effects model, we found no significant interaction between HIV-status and opioid/heroin use (b = 1.71, 95% Confidence Interval (CI)

Table 3 Depression differential item functioning effect for two C3PNO cohorts vs general population data McFadden pseudo R^2: CESD items

Item order	CESD item stem text	mSTUDY and general population		MASH and general population	
		McFadden pseudo R^2		McFadden pseudo R^2	
		Uniform	Non-uniform	Uniform	Non-uniform
1	I was bothered by things that usually don't bother me	0.001	0.000	**0.013**	**0.006**
2	I did not feel like eating; my appetite was poor	0.000	0.000	**0.012**	**0.004**
3	I felt that I could not shake off the blues even with help from my family or friends	0.000	0.000	**0.004**	**0.018**
5	I had trouble keeping my mind on what I was doing	0.005	0.000	0.004	0.000
6	I felt depressed	0.000	0.000	**0.009**	**0.008**
7	I felt that everything I did was an effort	0.002	0.001	**0.012**	**0.006**
9	I thought my life had been a failure	0.001	0.000	0.000	0.000
10	I felt fearful	0.011	0.000	0.013	0.000
11	My sleep was restless	0.004	0.000	0.004	0.001
13	I talked less than usual	0.002	0.001	0.001	0.000
14	I felt lonely	0.000	0.000	0.001	0.001
15	People were unfriendly	0.004	0.003	0.012	0.000
17	I had crying spells	0.001	0.000	**0.023**	**0.000**
18	I felt sad	0.006	0.002	0.000	0.000
19	I felt that people dislike me	0.013	0.000	0.000	0.001
20	I could not get ''going''	**0.013**	**0.004**	**0.042**	**0.011**

Items with total R^2 effects (uniform + non-uniform) greater than .016 are in bold

Table 4 Depression differential item functioning for two C3PNO cohorts vs general population data McFadden pseudo R^2: PROMIS depression items

Item ID	PROMIS item stem text	HYM and general population		RADAR and general population	
		McFadden pseudo R^2		McFadden pseudo R^2	
		Uniform	Non-uniform	Uniform	Non-uniform
EDDEP04	I felt worthless	0.005	0.002	0.000	0.001
EDDEP05	I felt that I had nothing to look forward to	0.000	0.000	0.001	0.000
EDDEP06	I felt helpless	0.001	0.000	0.001	0.000
EDDEP07	I withdrew from other people	**0.019**	**0.001**	NA	NA
EDDEP09	I felt that nothing could cheer me up	0.003	0.000	0.001	0.000
EDDEP14	I felt that I was not as good as other people	0.010	0.001	NA	NA
EDDEP17	I felt sad	0.012	0.000	NA	NA
EDDEP19	I felt that I wanted to give up on everything	0.005	0.000	NA	NA
EDDEP21	I felt that I was to blame for things	0.000	0.001	NA	NA
EDDEP22	I felt like a failure	0.001	0.001	0.001	0.000
EDDEP23	I had trouble feeling close to people	0.003	0.004	NA	NA
EDDEP26	I felt disappointed in myself	0.000	0.000	NA	NA
EDDEP27	I felt that I was not needed	0.004	0.001	NA	NA
EDDEP28	I felt lonely	**0.018**	**0.000**	NA	NA
EDDEP29	I felt depressed	0.001	0.001	0.000	0.000
EDDEP31	I felt discouraged about the future	0.000	0.000	NA	NA
EDDEP35	I found that things in my life were overwhelming	**0.032**	**0.003**	NA	NA
EDDEP36	I felt unhappy	0.001	0.000	0.001	0.000
EDDEP39	I felt I had no reason for living	0.002	0.000	NA	NA
EDDEP41	I felt hopeless	0.005	0.002	0.004	0.000
EDDEP46	I felt pessimistic	0.011	0.001	NA	NA
EDDEP48	I felt that my life was empty	**0.016**	**0.000**	NA	NA
EDDEP54	I felt emotionally exhausted	**0.016**	**0.001**	NA	NA

Items with total R^2 effects (uniform + non-uniform) greater than 0.016 are in bold.

NA item was not administered in the RADAR cohort

Table 5 Depression differential item functioning for the JHHCC cohort vs general population data McFadden pseudo R^2: PHQ-8 items

Item order	PHQ-8 stem text	JHHCC and general population	
		McFadden pseudo R^2	
		Uniform	Non-uniform
1	Little interest or pleasure in doing things	0.000	0.003
2	Feeling down, depressed, or hopeless	0.000	0.000
3	Trouble falling or staying asleep, or sleeping too much	0.001	0.000
4	Feeling tired or having little energy	0.005	0.002
5	Poor appetite or overeating	0.001	0.000
6	Feeling bad about yourself—or that you are a failure or have let yourself or your family down	0.012	0.000
7	Trouble concentrating on things, such as reading the newspaper or watching television	0.004	0.000
8	Moving or speaking so slowly that other people could have noticed? Or the opposite—being so fidgety or restless that you have been moving around a lot more than usual	0.002	0.001

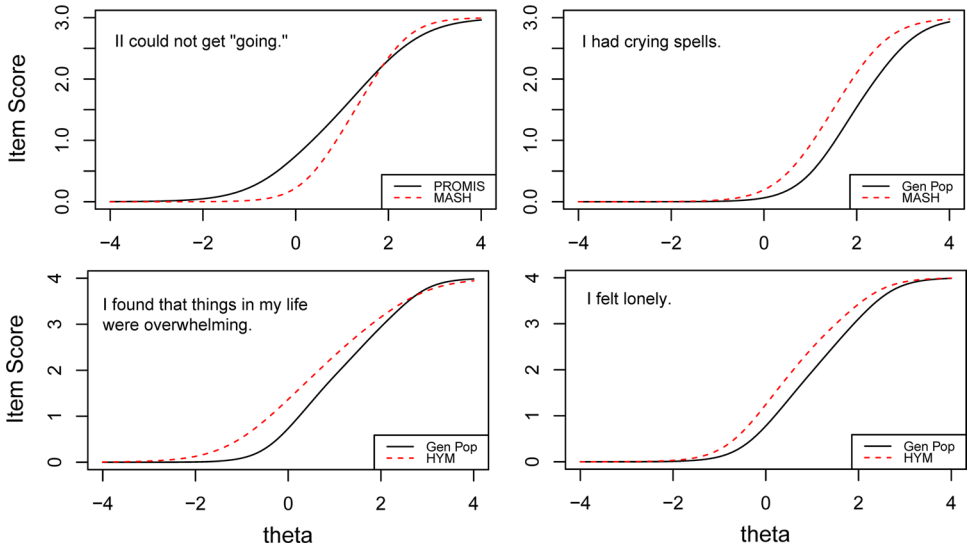

Fig. 1 Item characteristic curves for four items with non-trivial DIF compared to a general population sample (total pseudo $R^2 > 0.016$). The top two panels represent two CESD items used in the MASH cohort; the bottom two panels show two PROMIS items administered in the HYM cohort. Except for "I could not get 'going,'' these curves show that the C3PNO cohort participants were more likely to endorse these items relative to the general population samples, given participants' overall level of depression. For "I could not get 'going,'" this was reversed for low and mild levels of depression

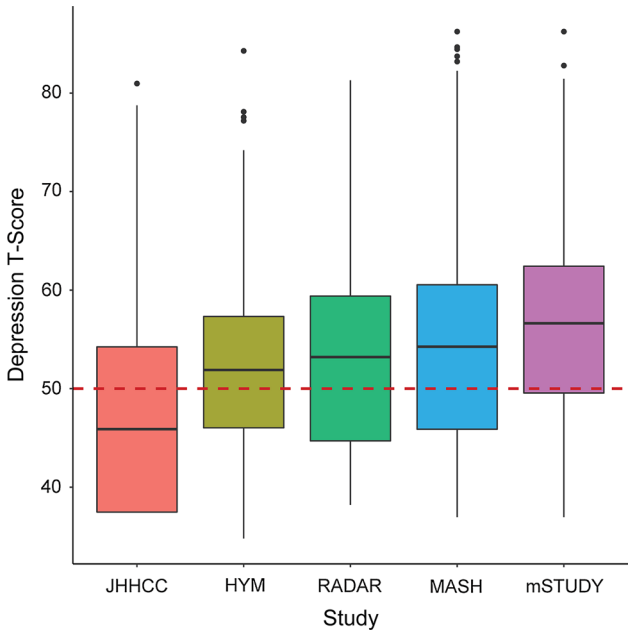

Fig. 2 Box and whisker plots showing PROMIS Depression T-score for the five C3PNO cohorts. For T-scores, a value of 50 represents the mean of the US general population (SD = 10). Higher scores indicate a greater degree depression

[−0.34, 3.79], p = 0.104). Examining the main effects, we found that recent opioid/heroin use was associated with depression with a mean increase of 2.13 T-score points (95% CI [0.81, 3.45], (p < 0.001). While HIV-positive persons showed a lower mean by 0.51 T-score points (95% CI [−1.43, 0.44]), it was not significant (p = 0.284). Age was also not significantly associated with a difference in T-score (b = −0.3 per 10 year increase, 95% CI [−0.9, 0.3], p = 0.260).

Conclusion

The use of multiple psychological and behavioral measures represents a possible impediment to the goal of data harmonization or integrative analysis of cohort studies of HIV risk populations [4]. Because the interpretation of scores and clinical-cut-off values differ for each questionnaire [13, 14], raw data cannot simply be pooled. This problem can be solved with new cross-sectional studies focused on linking or aligning multiple measures of the same health construct. In addition, the results and tools of previous linking studies [13, 51] can potentially be leveraged for use in new populations.

Our study with the C3PNO cohorts demonstrates how data capturing similar depression constructs in HIV-risk populations can be linked to a common depression metric. Taking advantage of previous linking studies with multiple measures of depression measures [13], our study illustrates how to take these results—IRT item parameters—and apply them in the service of data harmonization across cohorts in HIV/substance use populations. We found minimal levels of DIF between the C3PNO cohorts and general population internet panels (e.g., DIF was flagged for 13 out 71 item comparisons using a conservative R^2 effect size criterion). This suggests that measurement variance across populations and within substance using populations was not an impediment to the application of previous linking studies for C3PNO harmonization goals, a finding consistent with previous DIF and IRT studies of depression in multiple samples [52, 53] and linking studies of pain measures in a general population sample [15] and multiple sclerosis [54].

Our analytic approach allowed us to combine community-sample cohorts with sample sizes ranging from 448 to 1041 to achieve a combined sample size of 2871 for the mixed effects model. Whereas a study such as MASH had

83% power to detect a standardized effect size of 0.2 for depression when comparing HIV-infected and HIV–uninfected groups, the power to detect that difference increases to > 99% using the combined cohort. The utility of being able to combine cohorts is noteworthy when assessing a factor with low prevalence, such as opioid use. Whereas in the MASH study we could detect a standardized effect size difference of 0.2 in depression by opioid use with 54% power, the combined cohorts have > 90% power.

After placing scores on a common depression metric, descriptive statistics show a wide range of depression score means across cohorts (0.8 standard deviations on the linked general population PROMIS T-score metric). We note that this 8 point range would represent a meaningful difference in an individual clinical context. Studies establishing minimally important differences (MIDs) suggest that changes of 2 to 5 T-score points on the PROMIS metric are meaningful to patients [55–57]. Figure 1 visually identifies that one of the cohorts (JHHCC) scored more than 4 T-score points lower on depression relative to the other cohorts; this result is consistent with the fact that JHHCC is the only C3PNO clinical cohort in which all patients are undergoing comprehensive clinical care. The remaining cohorts are observational community cohorts, which offer referrals to mental health treatment, but the assessments do not occur in a care setting where research personnel and providers are integrated. The lower depression scores may be explained by the fact that participants are screened and aggressively treated for depression with antidepressant medication. While we acknowledge that many other factors could influence depression among C3PNO cohorts, the salient feature of integrated mental health treatment in the JHHCC cohort suggests the potential health value of making mental health treatment more widely available to persons at risk for HIV. This inference is consistent with findings on the positive effects of mental health treatments (particularly those of longer duration) for people living with HIV [58].

Our model confirmed an expected positive relationship between substance use and depression, though contrary to expectations, no significant association with HIV status. The main effect for substance use corresponds to a significant, but small effect size (0.2 standard deviation units on the PROMIS T-score metric). While unexpected, the lack of a significant association observed between HIV and depression further illuminates the important role of cohort differences, associated patient characteristics, inclusion criteria, study designs, and administration settings (e.g., clinical vs observational) that may confound cross-cohort analysis. Future analyses in the C3PNO cohorts will model a fuller range of variables associated with depression in an effort to more explicitly account for these known differences. This cross-cohort methodology could be extended to other key outcomes, such as pain, physical function, sexual risk behavior, and substance use scales.

Acknowledgements We thank all C3PNO Cohort Principal Investigators for participation in the consortium and making this study possible: Kora DeBeck, Kanna Hayashi, Thomas Kerr, Gregory Kirk, Shenghan Lai, Shruti Mehta, M-J Milloy, and Jeanne Keruly. This project is supported by the National Institute of Drug Abuse (NIDA) of the National Institutes of Health under Award Numbers U24DA044554, U01DA036935, U01DA040381 U01DA036267, U01DA036939, U01DA036926, and P30 MH058107.

References

1. Marshall BD, Friedman SR, Monteiro JF, Paczkowski M, Tempalski B, Pouget ER, et al. Prevention and treatment produced large decreases in HIV incidence in a model of people who inject drugs. Health Aff. (Millwood). 2014;33(3):401–9.
2. Crepaz N, Hess KL, Purcell DW, Hall HI. Estimating national rates of HIV infection among MSM, persons who inject drugs, and heterosexuals in the United States. AIDS. 2019;33(4):701–8.
3. Fauci AS, Redfield RR, Sigounas G, Weahkee MD, Giroir BP. Ending the HIV epidemic: a plan for the United States. JAMA. 2019;321:844–5.
4. Curran PJ, Hussong AM. Integrative data analysis: the simultaneous analysis of multiple data sets. Psychol Methods. 2009;14(2):81–100.
5. Chandler RK, Kahana SY, Fletcher B, Jones D, Finger MS, Aklin WM, et al. Data collection and harmonization in HIV research: the seek, test, treat, and retain initiative at the national institute on drug Abuse. Am J Public Health. 2015;105(12):2416–22.
6. Hussong AM, Curran PJ, Bauer DJ. Integrative data analysis in clinical psychology research. Annu Rev Clin Psychol. 2013;9:61–89.
7. McHugh RK, Rasmussen JL, Otto MW. Comprehension of self-report evidence-based measures of anxiety. Depress Anxiety. 2011;28(7):607–14.
8. Dorans NJ. Linking scores from multiple health outcome instruments. Qual Life Res. 2007;16(1):85–94.
9. Kolen MJ, Brennan RL. Test equating, scaling, and linking. New York: Springer; 2004.
10. Kolen MJ, Brennan RL. Observed score equating using the random groups design. Test equating, scaling, and linking. New York: Springer; 2014. p. 29–63.
11. De Ayala RJ. The theory and practice of item response theory. New York: Guilford Publications; 2013.
12. Embretson SE, Reise SP. Item response theory. London: Psychology Press; 2013.
13. Choi SW, Schalet B, Cook KF, Cella D. Establishing a common metric for depressive symptoms: linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. Psychol Assess. 2014;26(2):513.
14. Schalet BD, Cook KF, Choi SW, Cella D. Establishing a common metric for self-reported anxiety: linking the MASQ, PANAS, and GAD-7 to PROMIS anxiety. J Anxiety Disord. 2014;28(1):88–96.
15. Cook KF, Schalet BD, Kallen MA, Rutsohn JP, Cella D. Establishing a common metric for self-reported pain: linking BPI pain interference and SF-36 bodily pain subscale scores to the PROMIS pain interference metric. Qual Life Res. 2015;24(10):2305–18.
16. Rose JS, Dierker LC, Hedeker D, Mermelstein R. An integrated data analysis approach to investigating measurement equivalence of DSM nicotine dependence symptoms. Drug Alcohol Depend. 2013;129(1–2):25–322.

17. Greenbaum PE, Wang W, Henderson CE, Kan L, Hall K, Dakof GA, et al. Gender and ethnicity as moderators: integrative data analysis of multidimensional family therapy randomized clinical trials. J Fam Psychol. 2015;29(6):919–30.
18. Hussong AM, Gottfredson NC, Bauer DJ, Curran PJ, Haroon M, Chandler R, et al. Approaches for creating comparable measures of alcohol use symptoms: harmonization with eight studies of criminal justice populations. Drug Alcohol Depend. 2019;194:59–68.
19. Janulis P, Newcomb ME, Sullivan P, Mustanski B. Evaluating HIV knowledge questionnaires among men who have sex with men: a Multi-Study item response theory analysis. Arch Sex Behav. 2018;47(1):107–19.
20. Kaat AJ, Newcomb ME, Ryan DT, Mustanski B. Expanding a common metric for depression reporting: linking two scales to PROMIS® depression. Qual Life Res. 2016;26:1119–28.
21. Scherrer JF, Svrakic DM, Freedland KE, Chrusciel T, Balasubramanian S, Bucholz KK, et al. Prescription opioid analgesics increase the risk of depression. J Gen Intern Med. 2014;29(3):491–9.
22. Gonzalez JS, Batchelder AW, Psaros C, Safren SA. Depression and HIV/AIDS treatment nonadherence: a review and meta-analysis. J Acquir Immune Defic Syndr. 2011. https://doi.org/10.1097/QAI.0b013e31822d490a.
23. Gorbach PM, Siminski S, Ragsdale A, the C3PNO Investigators. Cohort profile: the collaborating consortium of cohorts producing NIDA opportunities (C3PNO). Int J Epidemiol. In press.
24. Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. Assessment. 2011;18(3):263–83.
25. Pilkonis PA, Choi SW, Salsman JM, Butt Z, Moore TL, Lawrence SM, et al. Assessment of self-reported negative affect in the NIH Toolbox. Psychiatry Res. 2013;206(1):88–97.
26. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. J Clin Epidemiol. 2010;63(11):1179–94.
27. Cella D, Choi SW, Condon DM, Schalet B, Hays RD, Rothrock NE, et al. PROMIS® adult health profiles: efficient short-form measures of seven health domains. Value Health. 2019;22(5):537–44.
28. Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord. 2009;114(1–3):163–73.
29. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16(9):606–13.
30. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Meas. 1977;1(3):385–401.
31. Myers JK, Weissman MM. Use of a self-report symptom scale to detect depression in a community sample. Am J Psychiatry. 1980;137:1081–4.
32. Naughton MJ, Wiklund I. A critical review of dimension-specific measures of health-related quality of life in cross-cultural research. Qual Life Res. 1993;2(6):397–43232.
33. Zimmerman M, Coryell W. Screening for major depressive disorder in the community: a comparison of measures. Psychol Assess. 1994;6(1):71.
34. Carleton RN, Thibodeau MA, Teale MJ, Welch PG, Abrams MP, Robinson T, et al. The center for epidemiologic studies depression scale: a review with a theoretical and empirical examination of item content and factor structure. PLoS ONE. 2013;8(3):e58067.
35. Revelle W. Psych: procedures for psychological, psychometric, and personality research. 1.3.2 ed. Evanston, IL: The comprehensive R archive network; 2013.
36. Schmid J, Leiman JM. The development of hierarchical factor solutions. Psychometrika. 1957;22(1):53–61.
37. McDonald RP. TEst theory: a unified treatment. Mahwah: L. Erlbaum Associates; 1999.
38. Zinbarg RE, Revelle W, Yovel I, Li W. Cronbach's α, Revelle's β, and McDonald's ω H: their relations with each other and two alternative conceptualizations of reliability. Psychometrika. 2005;70(1):123–33.
39. Reise SP, Scheines R, Widaman KF, Haviland MG. Multidimensionality and structural coefficient bias in structural equation modeling: a bifactor perspective. Educ Psychol Meas. 2013;73(1):5–26.
40. Choi SW, Gibbons LE, Crane PK. Lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. J Stat Softw. 2011;39(8):1.
41. Teresi JA. Different approaches to differential item functioning in health applications: advantages, disadvantages and some neglected topics. Med Care. 2006;44:S152–S170170.
42. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monogr Suppl. 1969. https://doi.org/10.1002/j.2333-8504.1968.tb00153.x.
43. Kleinman M, Teresi JA. Differential item functioning magnitude and impact measures from item response theory models. Psychol Test Assess Model. 2016;58(1):79.
44. Hays RD, Calderón JL, Spritzer KL, Reise SP, Paz SH. Differential item functioning by language on the PROMIS® physical functioning items for children and adolescents. Qual Life Res. 2018;27(1):235–47.
45. Kroenke K, Yu Z, Wu J, Kean J, Monahan PO. Operating characteristics of PROMIS four-item depression and anxiety scales in primary care patients with chronic pain. Pain Med. 2014;15(11):1892–901.
46. Pilkonis PA, Yu L, Dodds NE, Johnston KL, Maihoefer CC, Lawrence SM. Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS®) in a three-month observational study. J Psychiatr Res. 2014;56:112–9.
47. Bates D, Mächler M, Bolker BM, Walker SC. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67(1).
48. Bock RD, Mislevy RJ. Adaptive EAP estimation of ability in a microcomputer environment. Appl Psychol Meas. 1982;6(4):431–44.
49. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. Can Med Assoc J. 2012;184(3):E191–E196196.
50. Kroenke K, Spitzer RL, Williams JB, Löwe B. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen Hosp Psychiatry. 2010;32(4):345–59.
51. Cook KF, Jensen SE, Schalet BD, Beaumont JL, Amtmann D, Czajkowski S, et al. PROMIS measures of pain, fatigue, negative affect, physical function, and social function demonstrated clinical validity across a range of chronic conditions. J Clin Epidemiol. 2016;73:89–102.
52. Liegl G, Wahl I, Berghöfer A, Nolte S, Pieh C, Rose M, et al. Using Patient Health Questionnaire-9 item parameters of a common metric resulted in similar depression scores compared to independent item response theory model reestimation. J Clin Epidemiol. 2016;71:25–34.
53. Cook KF, Kallen MA, Bombardier C, Bamer AM, Choi SW, Kim J, et al. Do measures of depressive symptoms function differently in people with spinal cord injury versus primary care

patients: the CES-D, PHQ-9, and PROMIS®-D. Qual Life Res. 2017;26(1):139–48.
54. Askew RL, Kim J, Chung H, Cook KF, Johnson KL, Amtmann D. Development of a crosswalk for pain interference measured by the BPI and PROMIS pain interference short form. Qual Life Res. 2013;22(10):2769–76.
55. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008;61(2):102–9.
56. Purvis TE, Neuman BJ, Riley LH III, Skolasky RL. Discriminant ability, concurrent validity, and responsiveness of PROMIS health domains among patients with lumbar degenerative disease undergoing decompression with or without arthrodesis. Spine. 2018;43(21):1512–20.
57. Jensen RE, Moinpour CM, Potosky AL, Lobo T, Hahn EA, Hays RD, et al. Responsiveness of 8 Patient-Reported Outcomes Measurement Information System (PROMIS) measures in a large, community-based cancer study cohort. Cancer. 2016;123:327–35.
58. Remien RH, Stirratt MJ, Nguyen N, Robbins RN, Pala AN, Mellins CA. Mental health and HIV/AIDS: the need for an integrated response. AIDS (London, England). 2019;33(9):1411.