Distribution Agreement

Signature:

Shiyu Chen Date

Assess Improvement of Suicide Ideation by the Empower Veteran Program using PHQ-9

By

Shiyu Chen

Master of Public Health

Biostatistics and Bioinformatics Department

Xiangqin Cui, PhD

Thesis Advisor

Yi-An Ko, PhD

Reader

Assess Improvement of Suicide Ideation by the Empower Veteran Program using PHQ-9

By

Shiyu Chen

B.S.

China Pharmaceutical University

2018

Thesis Committee Chair: Xiangqin Cui, PhD

Reader: Yi-An Ko, PhD

An abstract of

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of
Master of Public Health in Public Health,

in Biostatistics and Bioinformatics

2020

Abstract

Assess Improvement of Suicide Ideation by the Empower Veteran Program using PHQ-9

By Shiyu Chen

Background: For veterans undergoing major depression, the Empower Veteran Program (EVP) is a 10-week program which provides better access and self-care training than traditional therapy. However, the effect of this intervention program has not been evaluated. In addition, it is unknown who would benefit most from this program. The objective of the study is to focus on the PHQ-9 data to assess the impact of the EVP program based on before and after survey results.

Methods: A total of 639 Veteran patients enrolled in Week 1 or 2 of the EVP's 10-week program through the Atlanta VA Health Care System (AVAHCS). Demographic and baseline descriptive statistics from the baseline Quality Improvement (QI) clinical assessment were summarized stratified by subjects who completed at least 8 weeks of the 10-week program (Completers) and those who did not complete at least 8 weeks of the program (non-completers). The nine-item Patient Health Questionnaire (PHQ-9) is shown to be a reliable and validated measure of depression severity. A random intercept mixed linear model and fixed effect multiple linear regression model were used for analyzing the impact of EVP program on PHQ-9 measurement.

Results: Veterans enrolling in the EVP program had more serious mental health illnesses compared to a normal population, as indicated by a larger mean score of Quality Improvement (QI) assessments. Completers and non-completers were significantly different for some assessments (p-value<0.05). Comparing the PHQ-9 data pre and post completion, the EVP significantly reduced the PHQ-9 scores. Women veterans benefited more than men veterans from the EVP. In addition, EVP showed different impact on patients with different symptoms of depression at week 1. Comparing the baseline PHQ-9 total scores to the endline PHQ-9 scores, greater change in the total score was observed among participants with higher level depression. This suggests that veterans who have more severe depression likely benefit the most from EVP.

Conclusion: EVP generated meaningful improvements for veterans who are suffering from chronic pain based on the PHQ-9 survey. Female veterans benefitted more in terms of depression symptoms than their male counterparts. Moreover, veterans who had more severe depression at baseline, benefited the most from EVP.

Assess Improvement of Suicide Ideation by the Empower Veteran Program using PHQ-9

By

Shiyu Chen

B.S.

China Pharmaceutical University

2018

Thesis Committee Chair: Xiangqin Cui, PhD

Reader: Yi-An Ko, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of
Master of Public Health in Public Health,

in Biostatistics and Bioinformatics

2020

CONTENTS

1. Introduction

Chronic pain is highly prevalent among Veterans, especially when it is combined with major depression, which results worse outcomes including death by elevated suicide rates among veteran population compared with the general adult population.[1] Thus, suicide prevention in United States is a national priority of Veterans Health Administration (VHA). High impact chronic pain touches multiple domains, bio-psycho-social-spiritual, usually as a whole life dysfunction. More severely affected persons, e.g. with increasing levels of depression and physical deconditioning, are less likely to engage in, complete, and benefit from standard physical therapy.

To treat complex chronic pain conditions, interdisciplinary pain rehabilitation program (IPRP) takes a functional restoration approach. With a greater appreciation for a biopsychosocial approach to more effectively manage patients with chronic pain, a more comprehensive treatment program-IPRP with less of a biomedical emphasis and more of a biopsychosocial one has been developed such as interventional therapy, unimodal physical therapy, and passive modalities. IPRP involves the use of multiple disciplines such as physical and occupational therapy, pain psychology, medical pain management, vocational rehabilitation, relaxation training, and nursing educations. To better assess its impact on outcomes, multiple psychometric tools are used in the assessment process.[2] In general, Multi- or Inter-disciplinary Pain Rehabilitation Program (IPRP) bring multiple professionals' individual skills to bear for one client. However, traditional application of IPRP may be costly and is often difficult to access anywhere, whether in or outside of VHA.[3]

To overcome the challenges of IPRP, Empower Veteran Program (EVP) was developed, which is a 10-week program allowing better access to clients and promoting self-care training. VHA providers may not be able to help the most challenged veterans with most severe pain and bio-psycho-social-spiritual needs to make successful recovery. However, EVP matches best choices of effective self-care training in spite of high impact chronic pain. Standard of care from a range of prior IPRPs guided EVP's initiation of coaching self-care skills and iterative quality improvements for addressing high impact chronic pain. Following the Plan-Research-Research-Act (PDSA) cycle of the American Healthcare Improvement Association's improvement model, multimodal therapies can be gradually modified to achieve functional rehabilitation of chronic pain.[4] Evidence Based Practice (EBP) and the biopsychosocial model laid the foundation for EVP theoretical frameworks.[5] EVP versions integrate three core components: Behavioral Therapy (BT); Whole Health coaching with Motivational Interviewing (MI) style and in-session Mindful Awareness practices; and

Physical Therapy with therapeutic neuroscience education and guidance in body awareness in whole body movement.

Many Quality Improvement (QI) clinical assessments measurements have been conducted in EVP. The nine-item Patient Health Questionnaire (PHQ-9) is one of the widely used measurements shown to be a reliable and validated measure of depression severity.[6,7] As a quick screening tool, PHQ-9 provides valid information to identify patients at greater risk of suicide and direct them toward appropriate treatment. Each question assesses presence and frequency of the nine DSM-IV diagnostic criteria for depression over the last two weeks with response of 0 (not at all), 1 (several days), 2(more than half of days), and 3(nearly every day).[8] 9 questions of PHQ-9 indicate different symptoms of depression. The total score range for PHQ-9 is 0-27, scores of 10-12 have been proposed as clinically significant cut points for determining if respondents have depression. A score of 10 was shown to have acceptable diagnostic sensitivity in a recent meta-analysis.[9] Thus, a score of 0-9 indicates mild depression, 10-14 indicates moderate depression, 15-19 indicates moderately severe depression, and 20-27 indicates severe depression.

To our knowledge, only two studies have evaluated PHQ-9 in suicidal risk among veterans. The study by Samantha A. Louzon *et al.* used multivariable proportional hazards regressions to evaluate associations between responses to item 9 of PHQ-9 and suicide mortality[1]. Rossom RC *et al.* furthered the understanding of whether item 9 of the PHQ-9 could predict risk for suicide ideation.[10]

Compare to the previous studies, our study assesses the improvement of major depression of EVP program by using PHQ-9 total score. We focus on the before and after PHQ-9 total score using 639 sequential Veteran patient data collected from Atlanta VA Health Care System (AVAHCS) to evaluate the impact of EVP.

2. Methods

2.1 Patient population and data collection

From Atlanta VA Health Care System (AVAHCS), from May, 2015 through early December, 2017, a total of 639 Veteran patients enrolled at Week 1 or 2 of EVP's 10 - week program. This period included EVP version 3.0-4.8. Patients were stratified by whether complete at least 8 weeks of the 10 three hours weekly EVP training program (three hours per session per week). Overall, there were 444 completers and 195 non-completers.

Demographics (age, gender, race and service connected) and Quality Improvement (QI) assessments were recorded for to allow for control of covariates and quantification of the impact of QI assessments in the analysis of outcome. Quality Improvement (QI) assessments measurements included are average Numerical Rating Scale (NRS) score, PCS (Pain Catastrophizing Scale) total score, FFMQ (mindfulness) total score, AAQ II (Acceptance and Action Questionnaire II) total score, PHQ-9 (Patient Health Questionnaire for Depression) total score, CPAQ (Chronic Pain Acceptance Questionnaire) Willingness for Activity Engagement scale score, CPAQ Willingness for Painful Sensation scale score, CPAQ total score, PROMIS (Patient-Reported Outcomes Measurement Information System) Physical Function T Score, PROMIS Fatigue T Score, PROMIS Sleep Disturbance T Score, PROMIS Satisfaction with Social Roles T Score, PROMIS Anxiety score, PROMIS Depression score, PROMIS Pain Interference T Score, WHOQOL (The World Health Organization Quality of Life) Environment T Score, WHOQOL Physical Health T Score, WHOQOL Social Relationships T Score.

2.2 Statistical Analysis

2.2.1 Descriptive Analysis

Prior to the analysis, the range of all items from the QI clinical Assessment Scores were checked, and negatively phrased items were reverse coded. Then, domain scores were calculated and check. For PROMIS and WHOQOL scales, raw scores were computed and transformed to T-scores using "short form conversion table". The baseline characteristics and QI clinical Assessment Scores of the first 639 sequential participants enrolled at Week 1 or 2 of EVP was were summarized. For continuous variables, the mean and standard deviation (SD) were summarized. For binary and categorical variables, the frequencies and percentage were presented. Each baseline variable was compared between completers and non-completers using two sample t-test for continuous variables and chi-square test for binary and categorical variables.

2.2.2 Correlation among the Baseline Quality Improvement (QI) Assessments

To examine the relationship among the QI assessment measures, Pearson Correlation Coefficient (PCC) was calculated. In our study, we have total 19 QI assessments, correlation matrix was created using R function cor(). To better reorder the correlation matrix according to the correlation coefficient, R function hclust() was used for hierarchical clustering order. Finally, correlation heatmap was created with R ggplot2 package to visualize the correlation.

2.2.3 PHQ-9 measurement Assessment over the 10 - week EVP

The main objective of our study is to evaluate the impact of EVP by comparing PHQ-9 total scores before and after the intervention. Boxplot was generated to visualize the trend of PHQ-9 total score changes from week 1 to week 10 across all ranges of baseline PHQ-9 total score.

Then PHQ-9 total score percent change were visualized by barplot. Percent change of PHQ-9 total score was defined as (week 10 PHQ-9 total score - baseline week 1 PHQ-9 total score)/ baseline week 1 PHQ-9 total score. PHQ-9 percent change was categorized into 4 scales with 30% cut-off, which are >30% worsen, 0-30% worsen, 0-30% improvement, >30% improvement.

To assess the relationship between PHQ-9 total score differences from week 1 to week 10 and the baseline week 1 PHQ-9 total score for women and men veterans, dot plot was generated using R ggplot2 package. Shading with progressively darker grays indicate traditional cut off points for [0,9] “mild”, [10, 14] “moderate”, [15,19] “moderately severe” and [20,27] “severe” Major Depression.

The PHQ-9 survey has a substantial amount of missing data. We used a mixed effect model, which has the advantage of fully utilizing the data which included a total of 517 among all 639 participants. A random intercept model can account for the the correlation between the week 1 and week 10 measurements within the same participant by treating the participant as a randome effect. The fixed-effects included in the model were, intervention (binary), different levels of depression based on different ranges of baseline week 1 PHQ-9 total score (categorical), gender, age, race, intervention×baseline (categorical), intervention×gender, intervention×race and intervention×age.

The random intercept mixed linear model includes 517 participants among all 639 participants. The R pakage lme4Test were used to fit the model. To investigate what factors predict the change in PHQ-9 score after the program, backward variable selection method with an α=0.02 removal criteria was implemented for model selection. The final model includes intervention, baseline (categorical), gender, and the interaction between intervention and gender, and the interaction between intervention and baseline (categorical).

To examine how much information we gained using the mixed model versus a fixed effect model using only the complete data (281 among 639 participants), we tested a fixed effect multiple linear regression model on the change of score. The difference PHQ-9 total score between week 1 and week 10 was used as the outcome for model building. The variables included baseline PHQ-9 total score, age, gender, race. The baseline PHQ-9 total scores were categorized as "mild" [0,9], "moderate" [10-14], "moderately severe" [15-19] and "severe" Major Depression [20-27]. A total of 348 observations among all 639 participants were excluded due to missing either week 1 or week 10 data. R function lm() was used to fit this model. Similar as random intercept mixed linear model, backward variable selection method with an $\alpha = 0.02$ removal criteria was implemented for model selection. The final model includes categorized baseline PHQ-9 total score and gender.

Model diagnostics were conducted for both the random intercept linear mixed model and fixed effect multiple linear regression model. The random variables of a mixed model add the assumption that observations within a level, the random variable group, are correlated. Mixed model is designed to address this correlation and do not cause a violation of the independence of observations assumptions. The assumption is relaxed to observations are independent of the other observations except where there is correlation specified by the random variable groups. Other assumptions for the random intercept mixed model are the same as the assumptions of the underlying model and need to be tested. The relationship of residuals and the predictor was intended to test the assumption of linearity. Q-Q plot could test the normality of residuals, significant deviations from linearity of the observations or non-symmetric scales indicate a deviation from normality of the residuals. In addition, the model should not be influenced by one or a small set of observations. Leverage was used to check this assumption. No major problems were shown for both of the models after the diagnostics.

The data was analyzed using R Studio version 1.1.463. P-values less than 0.05 were considered statistically significant.

3. Results

3.1 Study population

The demographic characteristics of the first 639 participants enrolled at the week 1 or 2 of the 10-week program are summarized in **Table 1**. About two thirds (444 out of 639) of these participants attended at least 8 weeks of the EVP program (completers). There was no significant difference between completers and non-completers in gender, service connected, and week 1 data availability. However, there was a significant difference in age, race, and week 10 data availability between completers and non-completer. Completers tended to be about 4 years older than non-completers (57.26 vs. 53.61, p<0.001). Approximately 79% of completers were African American, however, only about 67% of non-completers were African American (p=0.002). This suggests that African Americans were more willing to participate in this program compared to participants of other races and ethnic backgrounds. Week 10 data were more likely to be available for completers than non-completers (83.1% vs. 10.3%, p<0.001).

Table 1: Baseline Characteristics for EVP Participants

	Overall (N=639)	Missing	Non-completers (N=195)	Completers (N=444)	P Value
Age -year	56.15 ±10.75	6	53.61 (11.37)	57.26 (10.29)	<0.001
Female sex - n(%)	186 (29.1)	0	66 (33.8)	120 (27.0)	0.098
Race - n(%)					
African American	477 (75.4)	6	128 (67.0)	349 (79.0)	0.002
Service Connected - n(%)					
Yes	518 (81.7)	5	158 (82.7)	360 (81.3)	0.74
Week 1 data availability - n(%)	572 (89.5)	67	170(87.2)	402 (90.5)	0.25
Week 10 data availability - n(%)	389 (60.9)	250	20(10.3)	369 (83.1)	<0.001

Service Connected: Veterans whose medical condition is a result of their military service get discounted or free medical care. Week 1 data availability: Attend and/or submit Quality Improvement (QI) self-assessment at week 1; Week 10 data availability: Attend and/or submit Quality Improvement (QI) self-assessment at week 10; Race was dichotomized as either African - American or non-African – American

Baseline characteristics of Quality Improvement (QI) assessments for EVP participants are summarized in **Table 2**. Our QI assessments had four principle measurements, which included pain measurements (Numerical Rating Scale (NRS)), mental health measurements (PCS (Pain Catastrophizing Scale) total score, FFMQ (mindfulness) total score, AAQ II (Acceptance and Action Questionnaire II) total score, PHQ-9 (Patient Health Questionnaire for Depression) total score, CPAQ Willingness for Activity Engagement scale, CPAQ Willingness for Painful Sensation scale, CPAQ total score), quality of life measurements (PROMIS Physical Function T Score, PROMIS Fatigue T Score, PROMIS Sleep Disturbance T Score, PROMIS Satisfaction with Social Roles T Score, PROMIS Anxiety, PROMIS Depression, PROMIS Pain Interference T Score, WHOQOL Environment T Score, WHOQOL Physical Health T Score, WHOQOL Social Relationships T Score), and suicidal ideation measurement (PHQ-9 Question 9).

Veterans enrolling in EVP had a higher burden of chronic pain, more serious mental health illnesses, and other comorbidities compared to the normal population. For example, the average pain scale of Numeric Rating Scale (NRS) (ranging from 0 to 10) was 7.2 (sd 1.7), following into the severe range (7-10). The Pain Catastrophizing Scale (PCS) (total score ranging from 0 to 52) had a mean of 32.6 (sd 13.0), which was classified as high (>29 shows high Pain Catastrophizing). The average total score of Patient Health Questionnaire for Major Depression (PHQ-9) (range from 0 to 27) was 16.2 (sd 6.7), which was consistent with moderately severe major depression. More than 25% of Veterans at week 1 had some degree of recent suicidal ideation (SI) based on the larger than 0 response to PHQ-9 Q9. Other evidences of high degree of Mental Illness included the T Scores for the PROMIS subscales for Depression with a mean of 61.6 (sd 10.3), those for Anxiety with a mean 63.2 (sd 9.8), and those for WHOQOL Psychological with a mean 46.1 (sd 20.1). Participants had relative "lack of acceptance" i.e. rigid approach to painful sensations as shown by both the Chronic Pain Acceptance Questionnaire (CPAQ) with a total score mean of 41.3 (sd 15.4) and the Acceptance and Action Questionnaire II (AAQ II) scores with mean 31.3 (sd 10.9). These CPAQ and AAQ II scores highlight the initial high degree of unwillingness to engage in activity or to experience sensation perceived as painful. For most of the Quality Improvement (QI) assessment measurements, completers and non-completers were not statistically different except for CPAQ Willingness for Activity Engagement, PROMIS Physical Function, and PROMIS Satisfaction with Social Roles. However, completers were more likely to participate in their normal activities while in pain (CPAQ willingness for Activity Engagement) than non-completers (26.2 vs. 23.7, p=0.027). They also had significantly better physical functioning (39.4 vs. 40.5, p=0.031) and satisfaction with social roles (39.2 vs. 36.03. p=0.038). Of those self-reporting via PHQ-9 Q9 on week 1, significantly more veterans with suicidal ideation completed EVP compared to those who dropped out (35.1% vs. 25.1%, p=0.016).

Table2. Baseline Quality Improvement (QI) clinical Assessment Scores of EVP Participants.

	Overall	**Non-Completers**	**Completers**	**P Value**
	(N=639)	(N=195)	(N=444)	
"Pain" Measure				
NRS Average Numerical Rating Scale	7.2 (1.7)	7.4 (1.7)	7.1 (1.7)	0.111
"Mental Health" Measures				
PCS Total Pain Catastrophizing Scale	32.6 (13.0)	33.6 (12.9)	32.1 (13.0)	0.260
PHQ-9 Total Patient Health Questionnaire for Depression	16.2 (6.7)	16.8 (6.5)	15.9 (6.7)	0.178
PROMIS Depression *	61.6 (10.3)	61.9 (9.8)	61.4 (10.5)	0.656
PROMIS Anxiety *	63.2 (9.8)	63.9 (9.5)	62.9 (9.9)	0.284
WHOQOL Psychological *	46.1 (20.1)	44.8 (20.1)	46.7 (20.1)	0.346
CPAQ Willingness for Activity Engagement	25.5 (11.0)	23.7 (11.7)	26.2 (10.7)	0.027
CPAQ Willingness for Painful Sensation	15.8 (8.4)	16.1 (8.8)	15.7 (8.2)	0.604
CPAQ Total	41.3 (15.4)	39.7 (17.3)	42.0 (14.5)	0.144
AAQ II Acceptance and Action Questionnaire II	31.3 (10.9)	31.3 (11.0)	31.4 (10.9)	0.930
FFMQ Total (Mindfulness)	121.0 (20.4)	117.9 (21.4)	122.2 (19.9)	0.055
Other "Physical" and/or Quality of Life Measures				
PROMIS Physical Function *	39.7 (5.2)	40.5 (5.1)	39.4 (5.2)	0.031
PROMIS Fatigue *	63.6 (8.5)	63.9 (8.2)	63.4 (8.6)	0.537
PROMIS Sleep Disturbance *	61.4 (8.3)	62.0 (8.0)	61.2 (8.4)	0.334
PROMIS Satisfaction with Social Roles *	51.1 (7.0)	52.2 (7.1)	50.7 (7.0)	0.038
PROMIS Pain Interference *	67.7 (5.8)	67. 9 (6.2)	67.6 (5.6)	0.616
WHOQOL Environment *	59.2 (18.1)	58.2 (18.4)	59.6 (18.0)	0.428
WHOQOL Physical Health *	31.8 (16.4)	30.2 (15.7)	32.4 (16.6)	0.151
WHOQOL Social Relationships *	38.3 (23.5)	36.03 (24.8)	39.2 (22.9)	0.165
Suicidal Ideation Drill-Down				
PHQ-9 Question 9				
Not Answered or No Available Data		35 (17.9%)	55 (12.4%)	0.082
0 ("Not at all")		111 (56.9%)	233 (52.5%)	
>0 (i.e. 1 or 2 or 3)		49 (25.1%)	156 (35.1%)	0.016
1 ("Several Days")		23	79	

2 (“More than half of the days”)	11	43
3 (“Nearly every day”)	15	34

*T-score transformation

To assess the correlation among the baseline variables, pairwise Pearson correlation was calculated (**Figure 1**.) Different measurements from the same domain show relatively high correlation ranging from 0.6 to 1. For example, the correlation between WHOQOL Environment and WHOQOL Physical Health was 0.73, and the correlation between WHOQOL Social Relationships and WHOQOL Environment was 0.82. Interestingly, PROMIS Fatigue, PROMIS Anxiety, PROMIS Depression, PCS (Pain Catastrophizing Scale), AAQ II (Acceptance and Action Questionnaire II) and PHQ-9 (Patient Health Questionnaire for Depression) were even more correlated to each other with correlation coefficients ranging from 0.5 to 1. This high level of correlation is likely because these variables are all mental health measurements. Another interesting observation was that most of the PROMIS measurements were negatively correlated with those of WHOQOL and CPAQ measurements.

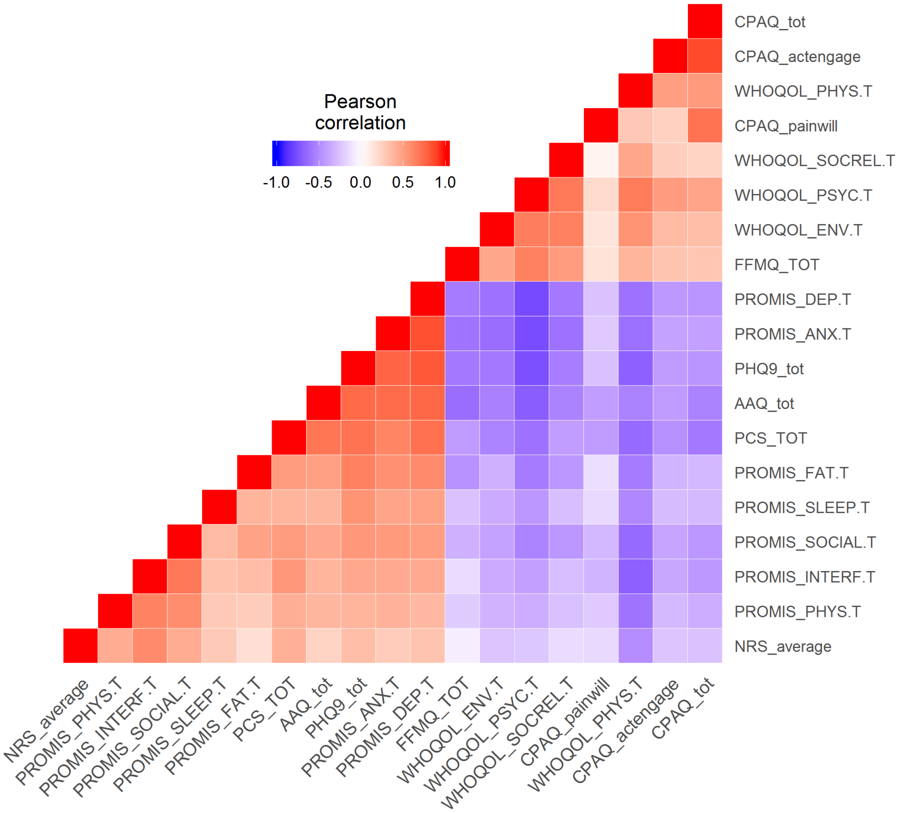

Figure 1. Pearson Correlation among Clinical Assessment Measurements. Correlation values range from -1 to 1 represented by color and intensity. Red indicates 1 and purple indicates -1.

3.2 Improvement in PHQ-9 after the 10-week EVP intervention

A total of 291 participants submitted responses for all 9 questions of PHQ-9 screening for major depression at week 1 and week 10. Comparing these two time points showed that PHQ-9 total score decreased from week 1 to week 10 across all ranges of baseline PHQ-9 total scores (**Figure 2**). In addition, higher levels of depression at baseline appears to be associated with a larger change in total PHQ-9 score assessed at endline. For example, among the participants with the highest depressive scores at baseline assessment, these individuals tended to experience the largest decrease in total score as assessed by endline. For participants with very low PHQ-9 score at baseline (0 to 9), the median score change was small, which is because these low scores fall into the mild depression range. For participants who had baseline scores above 9, significant improvements were seen based on the large median score difference from week 1 to week 10, attributable to the EVP program.

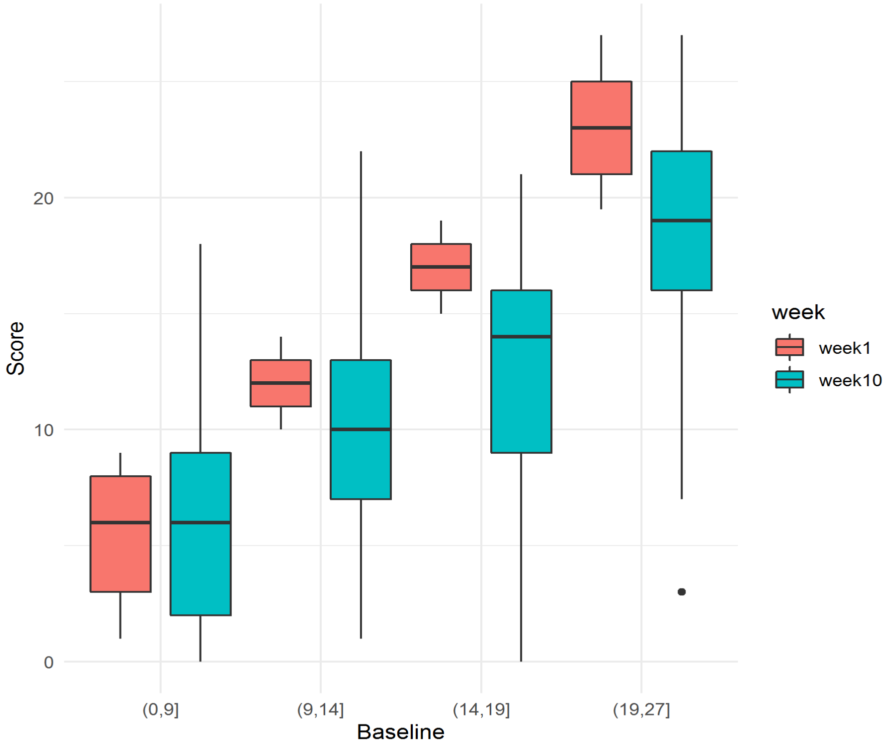

Figure 2. Relationship between baseline PHQ-9 total score and week 1 week 10 PHQ-9 total scores. Baseline PHQ-9 total score was divided into four categories according to different levels

of depression, which are [0,9] mild depression, [10,14] moderate depression, [15,19] moderately severe depression and [20,27] severe depression.

The change and percent change of PHQ-9 total scores were calculated for each participant and compared across different depression group levels. A larger proportion of participants with moderately severe depression [15, 19] and severe depression [20,27] at week 1 showed improvement due to the relatively high baseline PHQ-9 total score. However, more participants (44.9%) with mild depression at week 1 tended to experience a percent change of 30% or more for worse depression by week 10 (**Figure 3**). The reason is that they have lower baseline score, thus the space for improvement is limited.

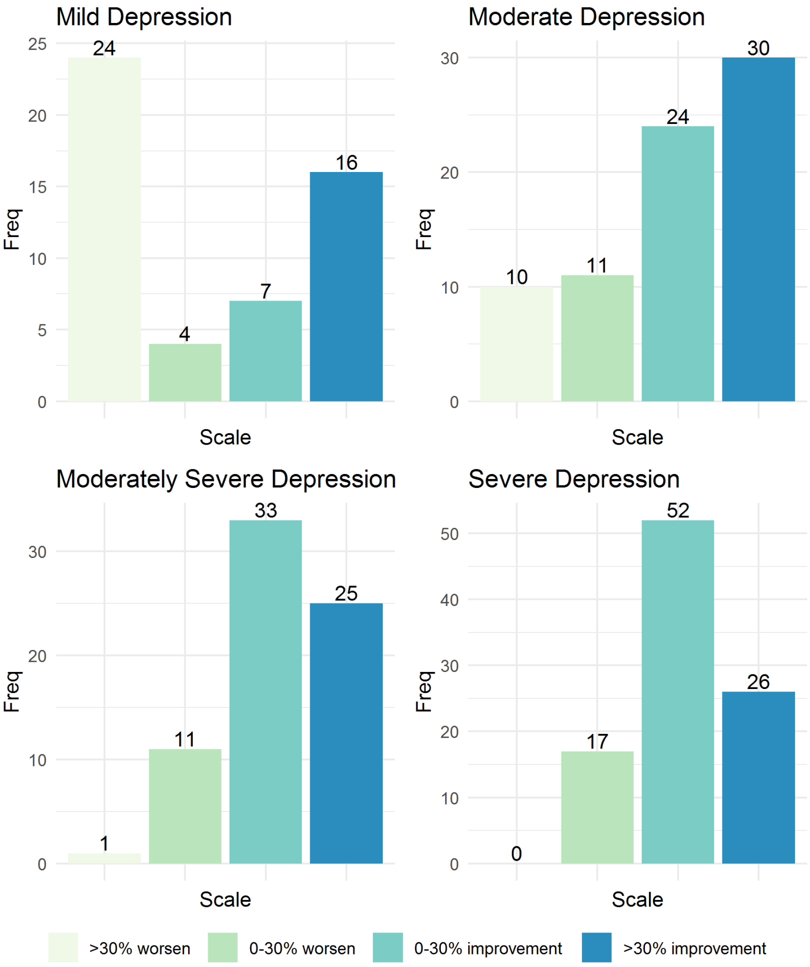

Figure 3. PHQ-9 Percent Change with 30% Cut-off. Baseline PHQ-9 total score was divided into four groups according to different levels of depression, which are [0, 9] mild depression, [10, 14] moderate depression, [15,19] moderately severe depression and [20, 27] severe depression.

PHQ-9 percent change was defined as score difference (week 10 PHQ-9 total score – baseline week 1 PHQ-9 total score) divided by baseline week 1 PHQ-9 total score. PHQ-9 percent change was categorized into 4 scales with 30% cut-off, which are >30% worsen, 0-30% worsen, 0-30% improvement, >30% improvement.

More specifically, 23.6% had scores consistent with "moderate" Major Depression, 24.7% "moderate-severe", and 35.2% "severe". Of the 270 Veterans initially scoring with some degree of depression (5 or more points out of 27 at week 1), 39.3% met criteria of at least a partial response in their depression in week 10. Another 33.3% of Veterans had 5 more points decrease in the total score. A smaller proportion of Veterans (4.8%) demonstrated 25% improvement, and 1.1% had additional change from above to below score of 10.

We also considered the baseline PHQ-9 total score as continuous. As shown in **Figure 4**, most of these Veterans entered the program with "moderately severe" or "severe" Major Depression, as suggested by PHQ-9 total score ranges of 15-19 or 20-27, respectively. The downward trend of score change along the baseline score indicates that larger change is observed in subjects with more severe depression. We also stratified the participants based on gender. A similar trend was observed for both genders, but female veterans benefitted more in terms of depressive symptoms.

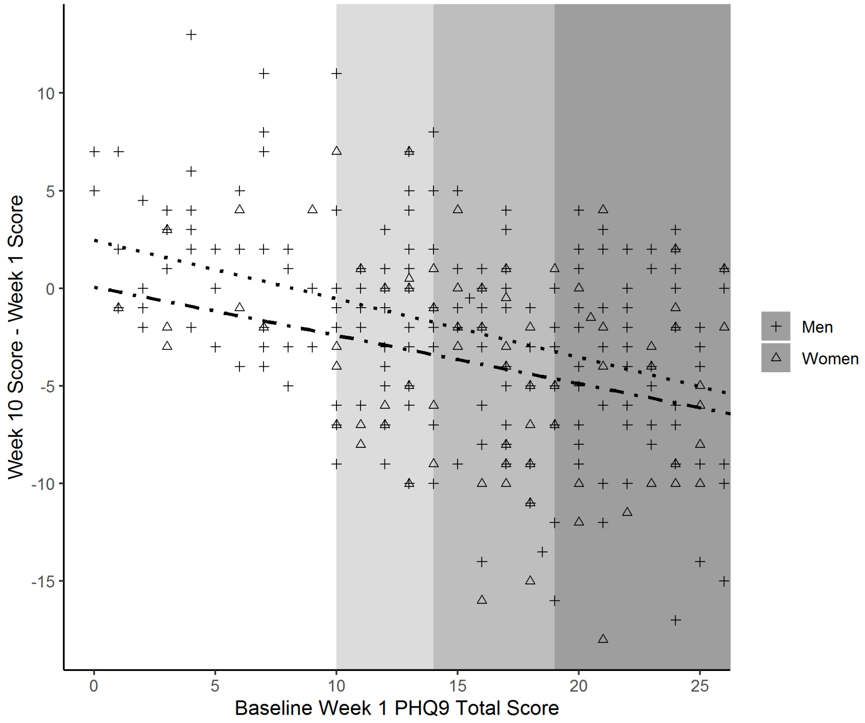

Figure 4. **Relationship between PHQ-9 total Score Differences from Week 1 to Week 10 and the baseline Week 1 PHQ-9 Total Score, for Female and Male Veterans**. Traditional cut points for baseline week 1 PHQ-9 total score with "moderate" [10-14], "moderately severe" [15-19] and "severe" Major Depression [20-27] are shown as progressively darker grays shade. The hollow triangle and cross shaped dots represent for women and men veterans separately. Dotted and dotted-dash linear regression lines refer to men and women veterans, respectively.

To comprehensively analyze the PHQ-9 total score, we used a random-intercept mixed model by treating the individual subject as a random effect. Race, age and gender covariates were included. Race and age effects were not significant, and therefore, were removed from the model. Other terms remained significant after rerunning the model. The interaction term between gender and intervention was significant. The difference of the EVP impact between man and women was 1.24 in PHQ-9 total score. Female veterans achieved an extra 1.24 reduction in PHQ-9 total score than male veterans on average, while controlling for baseline PHQ-9 total score. This result is consitent with the gap between the two regression lines in **Figure 4**. Meanwhile, the interaction effect of baseline and intervention was also significant. The EVP also shows a different impact on patients with different degrees of depression at week 1. Patients with moderate depression at week 1 achieved an extra 2.39 reduction in PHQ-9 total score compared with patients with mild depression at week 1 (p=0.00084) after controlling for gender. Patients with moderately severe depression and severe depression at week 1 had similar extra reductions of PHQ-9 total score (4.53 and 4.61 respectively) than patients with mild depression at week 1, after controlling for gender (p<0.0001). Overall, veterans who had severe depression at week 1 benefit more from the EVP program.

After testing a fixed effect multiple linear regression model on the change of score, similar conclusions were made as compared to the random-intercept mixed model. Female veteran had significantly 1.24 lower PHQ-9 total scores from week 1 to week 10 compared to male veteran, while controlling for baseline PHQ-9 total score. Although the estimate of coefficients remained similar for the two models, the fixed effect multiple linear regression model had slightly larger absolute estimates of coefficients. The random-intercept mixed model incorporated some amount of shrinkage for individual-specific effects. More power was lost for the mutiple regression model after excluding the missing values. A more narrow 95% confidence interval (CI) and around 20% smaller standard estimate were seen with the random-intercept mixed model.

Table 2. Fixed Effect Multiple Linear Regression Model and Random Intercept Linear Mixed Model for PHQ-9 Total Score

	Effect	Estimate	95% CI	SE	P-value
Fixed Effect Multiple Linear Regression Model	Intercept	1.46	(0.14, 2.79)	0.67	0.030
	Baseline (9,14]	-3.00	(-4.69,1.32)	0.86	0.0005
	Baseline (14,19]	-5.17	(-6.89, -3.45)	0.87	<0.0001
	Baseline (19,27]	-5.22	(-6.84, -3.61)	0.82	<0.0001
	Gender	-1.39	(-2.61, -0.18)	0.62	0.024
Random Intercept Mixed Linear Model	Intercept	5.59	(4.89, 6.29)	0.35	<0.0001
	Intervention	0.84	(-0.25, 1.92)	0.55	0.13
	Baseline (9,14]	6.50	(5.59, 7.42)	0.47	<0.0001
	Baseline (14,19]	11.33	(10.44, 12.23)	0.46	<0.0001
	Baseline (19,27]	17.39	(16.57,18.23)	0.42	<0.0001
	Gender	-0.021	(-0.65, 0.61)	0.32	0.95
	Intervention* Baseline (9,14]	-2.39	(-3.79, -1.00)	0.71	0.00084
	Intervention* Baseline (14,19]	-4.53	(-5.94, -3.00)	0.71	<0.0001
	Intervention* Baseline (19,27]	-4.61	(-5.93, -3.13)	0.67	<0.0001
	Intervention* Gender	-1.24	(-2.24, -0.26)	0.51	0.014

Reference Group: Intervention: Week 1; Race: non-African America; Gender: Male; Baseline (0, 9]. For fixed effect multiple linear regression model, 348 observations excluded due to missing, while random intercept mixed model includes 517 participants. The outcome for fixed effect multiple linear regression model is the difference of PHQ-9 total score between week 10 and week 1, while the outcome for random intercept mixed linear model is week 1 and week 10 PHQ-9 total score.

4. Discussion

In this study we evaluated the impact of the Empower Veterans Program (EVP) program on veterans with major depression. The EVP program was found to meaningfully improve depression for veterans who suffer from chronic pain. Completers benefited more from this 10-week program compared to non-completers in most Quality Improvement (QI) assessments. Using PHQ-9 total scores as criteria, we could see that EVP has a statistically significant impact on patients. Compared to male veterans, female veterans had statistically more improvement. Moreover, veterans who had severe depression benefited more from the EVP program.

The EVP program details an operational quality improvement to deliver intensive, integrated self-care training for veterans with high impact chronic pain, who failed earlier biomedical strategies and who failed and/or declined Behavioral Therapy with Mental Health professionals. The improvement of Major Depression is impressive by assessing the PHQ-9 measurement in light of this challenging population which was presumedly referred to EVP after a lack of response to other mental health treatments and therapies. Such improvements in PHQ-9, even if modest, should be of great societal interest as findings of recent Suicidal Ideation or of increasing severity of major depression, are associated with increasing rates among veterans of death by suicide.

In our study, we only focused on the PHQ-9 data to assess the impact of the EVP program based on before and after survey results. For future analysis, Principal Components Analysis (PCA) could be used to reduce the dimensionality from a large number of interrelated variables, while retaining as much of the information as possible. In our study, for example, PCA could be conducted for mental health assessments. After calculating Principle components (PC), we could use PC1 for model building to assess the impact of EVP since it retains most of the variation present in all of the original variables. This would be a novel application particularly for these data and this population.

Previous studies focused on the item 9 of the Patient Health Questionnaire (PHQ-9) in predicting risk for suicide attempts and deaths across age groups among veteran populations. However, in this study we focused on the intervention effect of EVP for depression by assessing the PHQ-9 total score. Our random intercept mixed linear model allows us to use more information than the fixed effect multiple linear regression model and results in increased power and accuracy. However, in this analysis we did not add complete status (completers and non-completers) as one potential covariate since there were fewer non-completers than completers (195 vs. 444). Thus, future work might want to add complete status as a covariate to the models to further assess the impact of EVP.

One limitation in our study are related to the handling of missing values. For instance, with the fixed effect multiple linear regression model, we only used complete cases, which resulted in approximately 20% of missing data. While this model is easy to conduct, it will lose efficiency compared to other models. Furthermore, for the random-intercept mixed linear model, while it could handle missing values, Missing Completely at Random (MCAR), Missing at Random (MAR) or Missing not at Random (NMAR) tests should be considered for further research. For MACAR and MAR, the random intercept model gives unbiased estimates, while NMAR could lead to bias. In this situation, regression-based imputation technologies could be used for NMAR. This would involve fitting the random-intercept model with all available data, and then replacing missing outcome with fitted values. Finally, the random-intercept model could be used to re-estimate. After performing the regression-based imputation, smaller standard errors may result.

Another limitation in our study is that our study is a single arm pre and post study, of which patients measured before the intervention and again after the intervention. This could cause potential bias. The interpretation of the pre-post study might be difficult. Because the change of the outcome could theoretically be due to the implement of the intervention. However, it might also due to placebo effect (neither patients or providers are blinded) or to a natural history improvement. For a subject that has self-report improvement, it could be argued that the subject would have improvement even without treatment or because they thought that they were receiving efficacious intervention. For future study, randomized control trial could be considered. In this study, at least two separate groups are evaluated, one of which receive the intervention and another treated as control group.

Our work has demonstrated that the EVP program had meaningful improvements for veterans who suffer from chronic pain and severe depression as assessed by the PHQ-9 survey. These findings suggest the benefit of expanding the EVP program to other VA hospitals and networks. Similarly, the EVP can serve to specifically target veterans with the highest levels of depression severity as these populations were shown to have the greatest improvement over the 10-week program. Finally, future work in this area can focus on understanding the differential impact between male and female veterans of the EVP program.

Reference

1. Louzon SA, Bossarte R, McCarthy JF, Katz IR. Does suicidal ideation as measured by the PHQ-9 predict suicide among VA patients? *Psychiatric Services.* 2016;67(5):517-522.
2. Stanos S. Focused review of interdisciplinary pain rehabilitation programs for chronic pain management. *Curr Pain Headache Rep.* 2012;16(2):147-152.
3. Guzman J, Esmail R, Karjalainen K, Malmivaara A, Irvin E, Bombardier C. Multidisciplinary rehabilitation for chronic low back pain: systematic review. *BMJ.* 2001;322(7301):1511-1516.
4. Berwick DM. A primer on leading the improvement of systems. *BMJ.* 1996;312(7031):619-622.
5. Sperry L. Should spiritually oriented psychotherapy be evidence based? *Spirituality in Clinical Practice.* 2015;2(3):165.
6. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med.* 2001;16(9):606-613.
7. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. *JAMA.* 1999;282(18):1737-1744.
8. Wells TS, Horton JL, LeardMann CA, Jacobson IG, Boyko EJ. A comparison of the PRIME-MD PHQ-9 and PHQ-8 in a large military prospective study, the Millennium Cohort Study. *J Affect Disord.* 2013;148(1):77-83.
9. Manea L, Gilbody S, McMillan D. A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. *Gen Hosp Psychiatry.* 2015;37(1):67-75.
10. Rossom RC, Coleman KJ, Ahmedani BK, et al. Suicidal ideation reported on the PHQ9 and risk of suicidal behavior across age groups. *J Affect Disord.* 2017;215:77-84.