Randomized-controlled trial of virtual reality for chronic low back pain to improve patient-reported outcomes and physical activity

Statistical Analysis Plan

National Clinical Trial (NCT) Identified Number: NCT04409353

Principal Investigator: Brennan Spiegel, MD

Sponsor: NIAMS

Funded by: NIH

Revision: 09/14/21

Original: 02 April 2020

Table of Contents

6.2 Interim Analyses and Data Monitoring

6.3 Multiple Comparisons/Multiplicity

6.4 Missing Data

MODIFICATION HISTORY

Version	Version Date	Author	Changes
1.0	04/02/20	Garth Fuller	n/a
2.0	06/07/20	Garth Fuller	n/a
3.0	09/14/21		Necessary edits

1. STUDY RATIONALE AND BACKGROUND

1.1 Study

Chronic low back pain (cLBP) is a prevalent and costly condition that markedly impairs physical, emotional, and social function. The 2015 Global Burden of Disease Study estimated that the prevalence of cLBP increased by more than 17% between 2005 and 2015.[1] In 2010, low back pain ranked third in disability-adjusted life years in North America, and the prevalence is expected to increase further due to the aging population and rise in obesity rates.[1,2] The National Health Interview Survey found that more than a quarter of all workers reported low back pain in the prior 3 months.[3] More than half reported the pain caused disability related to self-care, work, or social activities.[4]

Diminished work productivity attributable to cLBP is conservatively estimated at more than $28 billion annually.[5] Escalating treatment expenses include an increase of 423% in the cost of opioid prescriptions for patients with spinal disorders from 1997-2004, increases of 307% in the volume of lumbar MRI and 231% in the number of spinal injections reimbursed by Medicare from 1994-2001, and a 220% increase in lumbar spinal fusion surgeries between 1990 and 2001.[6] The Back Pain Survey administered as part of the National Health and Nutrition Examination Survey (NHANES) found that opioids were the most commonly prescribed pain medication taken by patients with cLBP (18.8%), followed by antidepressants (17.8%).[7] More than three quarters of people using prescribed opioids took them long term. Yet, patients with cLBP often discover that opioids fall short in delivering meaningful pain reduction or improving health-related quality of life (HRQOL).[7] Additionally, opioids are associated with a host of adverse effects, including but not limited to fall risk, constipation, sedation, physical dependency, opioid use disorder, and drug related mortality.[8] Hence, there is a critical gap in pain management in cLBP; it is vital to address this evidence gap in a way that maximizes benefits for patients while minimizing harms from medical therapy.

2.2 Background

The dynamic nature of clinical medicine, coupled with limited time to spend with individual patients, pose challenges to offering holistic care for patients with pain. Treatment of pain is often focused on pharmacological management, which can yield inconsistent and sub-optimal pain control. However, extensive data reveal that adjunctive non-pharmacological techniques, such as cognitive behavioral therapy and relaxation techniques, can modify cognitions and behaviors that influence the perception of pain. Therapeutic Virtual Reality (VR) technology provides an immersive, multisensory, and three-dimensional (3D) environment that enables users to have modified experiences of reality by creating a sense of “presence.” To date, VR has been used in numerous clinical settings to help treat anxiety disorders, control pain, support physical rehabilitation, and distract patients during wound care.[9-15] For example, VR coupled with medication is effective in decreasing pain during bandage changes for severe burns.[10,16] Similarly, VR reduces pain and provides positive distraction during routine procedures, such as intravenous

line placements[14] and dental procedures[11]. Other studies reveal that VR helps manage chronic pain conditions such as complex regional pain syndrome[17], lower back pain[18,19], and chronic neck pain. Our own research shows that VR can reduce pain by an average of 24% among hospitalized patients with a wide range of somatic and visceral pain.[20] A more recent randomized comparative effectiveness study conducted by our group (N = 120) also demonstrated the analgesic benefits of VR across a wide range of pain syndromes, with greatest effectiveness in patients with the highest levels of pain, defined as >7 points on a 0-10 numeric rating scale.[21]

By stimulating the visual, auditory, and proprioception senses, VR acts as a distraction to limit the user's processing of nociceptive stimuli. However, despite the evidence and increasing media attention surrounding VR, there have been no controlled trials using VR at scale to manage outpatient chronic lower back pain. Addressing this lack of evidence is the purpose of our randomized-controlled trial.

2. PURPOSE OF THE STATISTICAL ANALYSIS PLAN

The purpose of the statistical analysis plan (SAP) is to detail technical specifications for final analyses of data collected from protocol 00000631, a randomized-controlled trial of virtual reality therapy for chronic low back pain to improve patient-reported outcomes and physical activity. The SAP will support completion of the clinical study report for submission to the Back Pain Consortium (BACPAC).

3. SUMMARY OF CLINICAL TRIAL FEATURES

3.1 General Description

Protocol (STUDY00000631) describes a randomized, double-blind, sham-controlled, 3-arm, Phase 2 study. Two immersive VR arms use a skills-based VR therapy program, EaseVRx, and a distraction-based VR therapy program, EaseVRx-Distraction, while the sham VR arm uses a VR headset to deliver two-dimensional (non-immersive) content. The primary hypothesis is that participants randomized to either skills-based VR therapy or distraction VR therapy will report meaningful improvements in patient-reported outcomes (PROs), improved biometric outcomes, and reduced opioid use compared to participants receiving a non-immersive sham VR control intervention. It is currently planned as a single-site study. However, the self-administered nature of the intervention and remote data collection would accommodate additional sites from the BACPAC consortium if deemed necessary in time.

Patients in the study will receive access to two devices: (1) PICO G2 4K headset with one of three therapeutic visualization software developed by AppliedVR; and (2) Fitbit Charge 3 activity monitor. Study staff will monitor patient progress remotely and provide technical support. Patients allocated to the Sham VR arm will be exposed to 2D nature footage with neutral music. Patients allocated to the Distraction VR arm will be exposed to a content library called RelieVR (AppliedVR; Los Angeles, CA), which offers 360-videos and interactive 3d games. Patients allocated to the Skills-based VR arm will be exposed to a multi-modal, skills-based, self-

management VR program, called EaseVRx (AppliedVR; Los Angeles, CA), that incorporates evidence-based principles of CBT, mindful meditation, and physiologic biofeedback therapy enabled by embedded biometric sensors.

3.2 Randomization and Blinding

We will allocate study participants using a random number generator to assign blocks of 3, 6, 9, or 12 to ensure there is an equal distribution in the EaseVRx skills-based group vs. the EaseVRx-Distraction group vs. the EaseVRx-Sham group. Participants, their clinical providers, and study statisticians will be blinded to the study arm. The groups will be labeled as A, B, or C at random. Datasets will be provided to the statistician using these group labels. We anticipate there will be no circumstances during the study that require unblinding of an individual participant or a whole group because we do not expect any related serious adverse event(SAEs) to occur with this low risk intervention. If a research coordinator unintentionally reveals a participant's assignment,

3.2 Study Objectives and Endpoints

OBJECTIVES	ENDPOINTS
Primary	
To assess the efficacy of immersive Skills-Based VR and Distraction VR in improving perceived pain from baseline to Day 30. The trial will be considered a success if there is statistical evidence of improvement in either VR group compared to sham VR.	The change from study baseline to Day 30 in pain interference as measured by the 8-item PROMIS Pain Interference (PI) scale is the primary endpoint. This scale measures the consequences of pain on relevant aspects of life, including the extent to which pain hinders engagement with social, cognitive, emotional, physical, and recreational activities. The study will be considered a success if there is a statistically significant difference of 5 points in the Pain Interference score between participants in either the Skills-Based or Distraction VR arm compared with the Sham VR arm.
Secondary	
To assess the efficacy of immersive Skills-Based VR and Distraction VR in improving perceived pain interference from baseline to Day 60 and Day 90. The objective will be considered achieved if there is statistical evidence of improvement in either VR group compared to control (sham) VR.	The change from study baseline to day 60 and day 90 in pain interference as measured by the 8-item PROMIS PI scale is a secondary endpoint. This scale measures the consequences of pain on relevant aspects of life, including the extent to which pain hinders engagement with social, cognitive, emotional, physical, and recreational activities. We will test for a statistically significant difference of 5 points in the PROMIS PI score from baseline, and compare differences between either VR group and control (sham) VR.

OBJECTIVES	ENDPOINTS
To assess the efficacy of Skills-Based and Distraction VR in improving self-reported perceptions of sleep quality, sleep depth, and restoration associated with sleep from baseline to Day 90. The objective will be considered achieved if there is statistical evidence of improvement in either VR group compared to control (sham) VR.	The change from study baseline to day 90 in sleep disturbance as measured by the 6-item PROMIS Sleep Disturbance scale is a secondary endpoint. This scale assesses self-reported perceptions of sleep quality, sleep depth, and restoration associated with sleep. This includes perceived difficulties and concerns with getting to sleep or staying asleep, as well as perceptions of the adequacy of - and satisfaction with - sleep. We will test for a statistically significant difference of 5 points in the PROMIS Sleep Disturbance score from baseline, and compare differences between either VR group and control (sham) VR.
To assess the efficacy of Skills-Based and Distraction VR in improving self-reported perceptions of anxiety from baseline to Day 90. The objective will be considered achieved if there is statistical evidence of improvement in either VR group compared to control (sham) VR.	The change from study baseline to Day 90 in anxiety as measured by the 6-item PROMIS Anxiety scale is a secondary endpoint. This scale assesses self-reported perceptions of fear, anxious misery (worry, dread), hyperarousal, and somatic symptoms related to arousal. We will test for a statistically significant difference of 5 points in the PROMIS anxiety score from baseline, and compare differences between either VR group and control (sham) VR.
To assess the efficacy of Skills-Based and Distraction VR in improving self-reported pain catastrophizing from baseline to Day 90. The objective will be considered achieved if there is statistical evidence of improvement in either VR group compared to control (sham) VR.	The change from study baseline to Day 90 in pain catastrophizing as measured by PCS SF-6 is a secondary endpoint. We will test for a difference in rates of high catastrophizing as defined by a score of ≥7 on the PCS-SF6, and compare these differences between either VR group and control (sham) VR.
To assess the efficacy of Skills-Based and Distraction VR in reducing use of opioids from baseline to Day 90. The objective will be considered achieved if there is statistical evidence of improvement in either VR group compared to control (sham) VR.	The change from study baseline to Day 90 in weekly average opioid dosage calculated as a 7-day average of daily maximum milligrams morphine equivalent (MME) is a secondary endpoint. We will test for a greater than .5 SD in change from baseline, and compare differences between either VR group and control (sham) VR.
Tertiary/Exploratory	
To assess the efficacy of Skills-Based and Distraction VR in improving self-reported physical function from baseline to Day 90. The objective will be considered achieved if there is statistical evidence of improvement in either VR group compared to control (sham) VR.	The change from study baseline to Day 90 in physical function as measured by the 6-item PROMIS Physical Function scale is an exploratory endpoint. This scale measures self-reported functioning of one's upper extremities (dexterity), lower extremities (walking or mobility), and central regions (neck, back), as well as instrumental activities of daily living. We will test for a statistically significant difference of 5 points in the PROMIS Physical Function score from baseline, and compare differences between either VR group and control (sham) VR.

OBJECTIVES	ENDPOINTS
To assess the efficacy of Skills-Based and Distraction VR in improving self-reported depression from baseline to Day 90. The objective will be considered achieved if there is statistical evidence of improvement in either VR group compared to control (sham) VR.	The change from study baseline to Day 90 in depression as measured by the 8-item PROMIS Depression scale is an exploratory endpoint. This scale measures self-reported negative mood (sadness, guilt), views of self (self-criticism, worthlessness), and social cognition (loneliness, interpersonal alienation), as well as decreased positive affect and engagement (loss of interest, meaning, and purpose). We will test for a statistically significant difference of 5 points in the PROMIS Depression score from baseline, and compare differences between either VR group and control (sham) VR.
To assess the efficacy of Skills-Based and Distraction VR via patients' global impression of change (PGIC). The objective will be considered achieved if there is statistical evidence of higher rates of self-reported improvement in either VR group compared to control (sham) VR.	The overall effect of treatment from study baseline to Day 90 as measured by the PGIC is an exploratory endpoint. This scale measures self-reported belief regarding efficacy of treatment. A favorable response of 5-7 on the PGIC indicates significant improvement occurred over the course of the study. We will test for a statistically significant difference in PGIC responses between either VR group and control (sham) VR.
To assess the efficacy of Skills-Based and Distraction VR in improving measures of physical activity from baseline to Day 90. The objective will be considered achieved if there is statistical evidence of measured improvement in either VR group compared to control (sham) VR.	The change from study baseline to Day 90 in weekly steps as measured by Fitbit is an exploratory endpoint. Change in steps will also be examined for interaction effects on the primary outcome.
To assess the effect of presence on the efficacy of Skills-Based and Distraction VR in improving measures of pain interference from baseline to Day 30.	The effect of presence - as measured by a custom, 6-item survey - on the efficacy of VR treatment for pain is an exploratory endpoint. Scores on the survey at Day 1 will be examined for interaction effects on the primary outcome.
To assess the effect of Immersive Tendencies on the efficacy of Skills-Based and Distraction VR in improving measures of pain interference from baseline to Day 30.	The effect of immersive tendencies - as measured by the ITQ - on the efficacy of VR treatment for pain is an exploratory endpoint. Scores on the ITQ at baseline will be examined for interaction effects on the primary outcome.
To assess the effect of dose on the efficacy of Skills-Based and Distraction VR in improving measures of pain interference from baseline to Day 30.	The effect of dose - as measured by minutes spent accessing therapeutic content on the VR device - on the efficacy of VR treatment for pain is an exploratory endpoint. Minutes spent accessing therapeutic content will be examined for interaction effects on the primary outcome.

3.3 Number of Subjects

360 patients (120 in each study-arm) are planned.

3.4 Treatment, Study Duration, and Assessment Schedule

Patients in all arms will be identified, screened, enrolled, on-boarded, randomized, and assessed according to the following schedule:

Procedures	Pre-screening	Screening Week Day -8 to -2	Enrollment Day -1 to 0	Day 1 (+ 6 days)	Day 7 (+ 6 days)	Day 15 (+ 6 days)	Day 21 (+ 6 days)	Day 30 (+6 days)	Day 45 (+ 6 days)	Day 60 (+ 6 days)	Day 75 (+ 6 days)	Day 90 (+ 6 days)
Prospective patient identified (DEEP 6 or provider registry)	X											
Confirm eligibility and Informed Consent via telephone and REDCap	X											
Online Pain Intensity Journal (7 days)		X										X
Screening Week Online Baseline Survey (see below)		X										
Randomization			X									
Participant receives study intervention kit			X									
Technical onboarding call				X								
Intervention: Participant uses VR therapy program and wears Fitbit Charge 4 watch				X	----------							X
Online survey: NIH HEAL Minimum Dataset		X										X
Immersive Tendency Questionnaire (ITQ)[1] and Motion Sickness Propensity Survey[2]		X										
Treatment expectation question		X										
Simulator Sickness Questionnaire (SSQ)[3] and Presence Survey				X								
Online survey: Primary outcome: PROMIS® Pain Interference		X			X	X	X	X	X	X	X	X
Discontinuation of Treatment Questionnaire					X	X	X	X	X	X	X	X
PROMIS® Physical Function[4], Anxiety[4], Depression[4], Sleep disturbance[4]; Pain intensity/interference with Enjoyment of life/interference with General activity (PEG)[5], PCS-6[6]		X				X		X	X	X	X	X
EMR data: Charlson Comorbidity Index[7], CURES and EMR: Prescription Data		X										X
Perceived study arm question, treatments in the last 90 days question												X
Online survey: PHQ-2, GAD-2, PGIC, TAPS [1/2]		X										X
Remote technical support				X	----------							X
End of study procedure question												x
Event Assessment: AE, SAE and UP review and reporting				X	X	X	X	X	X	X	X	X

Procedures	Pre-screening	Screening Week Day -8 to -2	Enrollment Day -1 to 0	Day 1 (+ 6 days)	Day 7 (+ 6 days)	Day 15 (+ 6 days)	Day 21 (+ 6 days)	Day 30 (+6 days)	Day 45 (+ 6 days)	Day 60 (+ 6 days)	Day 75 (+ 6 days)	Day 90 (+ 6 days)
ITQ: Immersive Tendency Questionnaire PEG: Pain, Enjoyment, General Activity PCS-SF: Pain Catastrophizing Scale-Short Form 6 PHQ-2: Patient Health Questionnaire-2 GAD-2: Generalized Anxiety Disorder 2-item PGIC: Patient Global Impression of Change TAPS: Tobacco, Alcohol, Prescription Medication, and Other Substance Use AE: Adverse event SAE: Serious adverse event UP: Unanticipated Problem												

3.5 Determination of Sample Size

The primary aim is to test the efficacy of immersive VR in improving perceived pain. The trial will be a success if there is statistical evidence that either immersive VR group is better than control (sham) VR.

Preliminary studies showed that the PROMIS pain interference scale has a SD of 10. Assuming that the SD at baseline (SD0) and at 30 days (SD1) are similar and equal to 10 for this population, the variance for the difference in PROMIS from baseline to 30 days after intervention is $Var_{diff}= SD0^2 + SD1^2 - 2\ \rho\ SD0\ SD1$, where ρ is the correlation coefficient between measurements at baseline and 30 days. A conservative estimate of this variance is achieved when ρ = 0. Therefore, the estimate $SD_{diff} = (10^2 + 10^2)^{0.5} = 14.14$. Let mc, mv1, and mv2 be the mean change in PROMIS score from baseline to 30 days for the control, VR1, and VR2 groups, respectively. We estimate power by simulating 10000 trial replicates and testing the null hypothesis that Ho: mc = mv1 = mv2 versus the alternative hypothesis that H1: mc ≠ mv1 or mc ≠ mv2. To maintain the familywise error rate at 0.05, a two-sample t-test is used to compare the control arm to each VR arm and the test is declared statistically significant if the p-value of the two-sided test is less than 0.025. Under the alternative hypothesis that | mc - mv1| = | mc – mv2| = 5, data from 120 patients in each of the three arms achieve 83% power to detect a clinically meaningful effect 5 units in the PROMIS score. The actual type I error rate is 0.049. For the secondary outcome of comparison between the two VR arms, we test the null hypothesis that Ho: mv1 = mv2 versus the alternative hypothesis that H1: mv1 ≠ mv2 at the two-sided 0.05 level of significance if there is statistical evidence that both VR arms are better than control. Using the same assumptions as above and simulating 10000 trial replicates, then if both VR arms are better than the control arm, we can achieve 71.4% power to detect 5 units in PROMIS score between the two VR arms. This power was derived under the alternative hypothesis | mc - mv1|= 5 and | mv1 – mv2| = 5. The actual type I error rate for this conditional test is 0.02.

3.5.1 Power Estimation

For the primary endpoint of the study, there is 83% power to detect a clinically meaningful effect size of 5 units in PROMIS score between control and either VR arms at the 0.05 level of significance. This requires 360 patients randomized on a 1:1:1 ratio.

For one of the secondary endpoints of comparison between the two VR arms, we test the null hypothesis that Ho: mv1 = mv2 versus the alternative hypothesis that H1: mv1 ≠ mv2 at the two-sided 0.05 level of significance if there is statistical evidence that both VR arms are better than control. Using the same assumptions as above and simulating 10000 trial replicates, then if both VR arms are better than the control arm, we can achieve 71.4% power to detect 5 units in PROMIS score between the two VR arms. This power was derived under the alternative hypothesis | mc - mv1|= 5 and | mv1 – mv2| = 5. The actual type I error rate for this conditional test is 0.02.

For assessing the effect of VR on treatment response, we estimate power using a logistic regression model, accounting for all possible confounding factors as described in the protocol.

The outcome variable is treatment response (binary) and the predictor of interest is VR type, Distraction versus Skills-based. Table 1 gives the minimum odds ratio that can be detected with 80% power with the two-sided 0.05 level of significance as a function of baseline probability of positive response when a patient is treated with traditional VR and R^2, the proportion of variability in the predictor of interest (VR type) that is explained by all relevant baseline covariates in the model using data from 120 patients in the Distraction VR arm and 120 in the Skill-based VR arm. For example, data from 240 patients achieve 80% power to detect an odds ratio of 2.19 if the probability of positive response when a patient is treated with Distraction VR is 0.3 and 10% of the variability in VR type is explained by all other baseline covariates in the model. These odds ratios vary between 2.01 and 2.8 and are clinically meaningful. Therefore, we have enough power to test statistical significance of predictors of interest in the multivariable logistic regression model.

Table 1. Minimum detectable odds ratio as a function of the proportion of variability in VR type variable that is explained by all other relevant covariates in the model and the baseline probability of positive treatment response.

Baseline probability	R^2			
	0.0	0.1	0.2	0.4
0.2	2.26	2.36	2.48	2.82
0.3	2.11	2.19	2.30	2.61
0.5	2.01	2.18	2.30	2.64

4. STATISTICAL ANALYSES

4.1 General Methodology

All statistical analyses will be performed jointly by the Cedars-Sinai Biostatistics Core and the Cedars-Sinai Center for Outcomes Research and Education (CS_CORE) using SAS® software version 9.3 or higher (SAS Institute, Cary, NC, USA), R pack 4.1.1 (R Foundation for Statistical Computing, Vienna, Austria) or Stata software version 14 or higher (StataCorp LLC, College Station, TX, USA).

Continuous variables will be summarized using descriptive statistics (N, mean, standard deviation, minimum, median, and maximum). Categorical variables will be summarized using the number and percentage of patients in each category. Data will be summarized with respect to patient demographic and baseline characteristics both across the study and by study-arm. The efficacy endpoints, safety assessments, and other outcome results for each treatment group will be summarized descriptively unless otherwise indicated. In addition, statistical model estimates of least squares means, treatment differences, p-values and 95% confidence intervals will also be provided where relevant. The fit of linear models will be assessed using residual plots and/or other diagnostic plots as appropriate. The fit of logistic models will be assessed using Hosmer-Lemeshow goodness-of-fit and/or receiver operating characteristic (ROC) curves as appropriate. All statistical tests will be 2-sided and performed at the 0.05 level of significance unless stated

otherwise. Baseline is defined as the value at the screening week for all parameters, unless specified otherwise.

4.2 Analysis Populations and Patient Disposition

All efficacy and safety data summaries and analyses will be performed by study arm using an Intent-to-Treat (ITT) population defined as all randomized patients.

The number of patients identified as candidates will be reported, as will the number screened (i.e. consented). The number and percentage of patients randomized, patient population (ITT), and treatment status (completed, discontinued/withdrew) will be summarized both by treatment group and overall. Reasons for discontinuation/withdrawal will be presented.

An exploratory, per protocol (PP) analysis will focus on patients who use the assigned intervention on at least 50% of days during the first 30-day period. Usage meta-data on the headsets will facilitate sample definition in this population.

4.3 Deviations from Protocol

Protocol deviations/violations impacting participant safety will be reported to National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) and the Data and Safety Monitoring Board (DSMB) Safety Officer through theExecutive Secretary (Navitas Life Sciences) within 48 hours of the investigator becoming aware of the event; all other deviations/violations that do not impact participant safety can be reported as part of the routine DSMB meeting report. The investigator will also report deviations within 2 working days of identification of the protocol deviation to the IRB. All deviations must be addressed in study source documents and reported to NIAMS. Protocol deviations must be sent to the reviewing Institutional Review Board (IRB) per their policies. The site investigator is responsible for knowing and adhering to the reviewing IRB requirements. Protocol deviations will be presented descriptively in results but will not be analyzed.

4.4 Demographic and Baseline Characteristics

We will summarize demographic and baseline characteristics both by treatment group and overall. Any time-to-event counts will be measured from the day of randomization. Body Mass Index (BMI) will be calculated according to: BMI=weight (kg)/(height (m))2. Age will be calculated according to: Age=(date of event-birth date+1)/365.25. Weekly average opioid dosage will be calculated as a 7-day average of daily maximum milligrams morphine equivalent (MME) as prescribed. Mean activity measures will be calculated as weekly averages (e.g. weekly steps, weekly sleep minutes). Activity classification (i.e. sedentary versus light versus moderate, etc.) will be provided by FitBit and reported as such. Zip code will be matched to median income using census data as an aggregate measure of socio-economic status.

4.5 Treatment Compliance and Dosage

The cumulative time patients access content and/or engage interactive elements will be summarized descriptively in total and by study-arm. Duration will be calculated by subtracting

the content start time from the content end time. We will summarize treatment compliance in terms of percent of scheduled engagement across all days within patient (i.e. [(Minutes of Content Received) ÷ (Expected Minutes of Content)] *100) at 30-days and 90-days. The participants will be considered to have completed a module, if they have completed 70% of the module on a given day.

4.6 Analysis of Efficacy

4.6.1 Primary Analysis

PROMIS-PI

The goal of this analysis is to test the efficacy of immersive VR in improving measures of pain interference. The trial will be a success if there is statistical evidence that either immersive VR group is better than sham VR.

The primary efficacy endpoint is the change from study baseline to Day 30 in pain interference as measured by PROMIS-PI t-score. The primary analysis will compare the treatment groups, separately, to the control group using a linear mixed model repeated measures (MMRM) analysis.[1] The repeated measures are the change from baseline PROMIS-PI score obtained at Days 7, 15, 21, and 30, respectively. The model will include fixed categorical effects for treatment, week, treatment by week interaction, and the baseline PROMIS-PI t-score as a continuous covariate. We will employ the Stata *xtmixed* command or the R lmer function from the lme4 package with restricted maximum likelihood estimation (REML) and an unstructured within-patient covariance structure for this model. We will evaluate the assumptions of the model, including normality, using residual and other diagnostic plots of model fit.

From this model, we will estimate least squares means, standard errors, treatment differences in least squares means, and 95% confidence intervals for each time period. Primary inference will be based on the treatment comparison of least squares means for Day 30, and a p-value will be presented for this time period only. The null hypothesis is that the mean difference in the primary endpoint between the treatment groups and the sham control group is zero, versus the alternative hypothesis that these differences are not zero. The hypotheses can be expressed as follows:

$$H0_a: \mu SB - \mu control = 0 \text{ versus } H1_a: \mu SB - \mu control \neq 0$$

$$H0_b: \mu D - \mu control = 0 \text{ versus } H1_b: \mu D - \mu control \neq 0$$

Where μSB refers to the mean change from baseline to Day 30 in Promis-PI t-score in the Skills-based VR treatment group, μcontrol refers to the mean change from baseline to Visit Day 30 in Promis-PI t-score in the sham treatment group, and μD refers to the mean change from baseline to Day 30 in Promis-PI t-score in the Distraction VR treatment group. The test will be performed using the final MMRM model with a two-sided significance level of 5%. Estimated least squares

means for change from baseline and the observed absolute values (± SE) by treatment group will be plotted over time.

4.6.2 Secondary Analyses

PROMIS-Pain Interference

The change from study baseline to Day 60 and 90 in Pain Interference as measured by PROMIS-PI is a secondary endpoint. This analysis will compare the treatment groups, separately, to the control group using a MMRM analysis. The repeated measures are the change from baseline PROMIS-PI score obtained for Day 60 and 90, respectively. The model will include fixed categorical effects for treatment, week, treatment by week interaction, and the baseline PROMIS-Pain Interference score as a continuous covariate. We will employ the Stata *xtmixed* command or the R lmer function from the lme4 package with REML and an unstructured within-patient covariance structure for this model. We will evaluate the assumptions of this model, including normality, using residual and other diagnostic plots of model fit.

From this model, we will estimate least squares means, standard errors, treatment differences in least squares means, and 95% confidence intervals for each time point. Primary inference will be based on the treatment comparison of least squares means for Day 90 from this model, and a p-value will be presented for this time period only. The null hypothesis is that the mean difference in the primary endpoint between the experimental arms and the sham arm is zero, versus the alternative hypothesis that this difference is not zero.

PROMIS-Sleep Disturbance

The change from study baseline to Day 90 in sleep disturbance as measured by PROMIS-Sleep Disturbance is a secondary endpoint. This analysis will compare the treatment groups, separately, to the control group using a MMRM analysis. The repeated measures are the change from baseline PROMIS-Sleep Disturbance score obtained at Days 15, 30, 60, and 90, respectively. The model will include fixed categorical effects for treatment, week, treatment by week interaction, and the baseline PROMIS-Sleep Disturbance score as a continuous covariate. We will employ the Stata *xtmixed* command or the R lmer function from the lme4 package with REML and an unstructured within-patient covariance structure for this model. We will evaluate the assumptions of this model, including normality, using residual and other diagnostic plots of model fit.

From this model, we will estimate least squares means, standard errors, treatment differences in least squares means, and 95% confidence intervals for each time point. Primary inference will be based on the treatment comparison of least squares means for Day 90 from this model, and a p-value will be presented for this time period only. The null hypothesis is that the mean difference in the primary endpoint between the experimental arms and the sham arm is zero, versus the alternative hypothesis that this difference is not zero.

PROMIS-Anxiety

The change from study baseline to Day 90 in anxiety as measured by *PROMIS–Anxiety* is a secondary endpoint. This analysis will compare the treatment groups, separately, to the control group using a MMRM analysis. The repeated measures are the change from baseline *PROMIS–Anxiety* score obtained at Day 15, 30, 60, and 90, respectively. The model will include fixed categorical effects for treatment, week, treatment by week interaction, and the baseline *PROMIS–Anxiety* score as a continuous covariate. We will employ the Stata *xtmixed* command or the R lmer function from the lme4 package with REML and an unstructured within-patient covariance structure for this model. We will evaluate the assumptions of this model, including normality, using residual and other diagnostic plots of model fit.

From this model, we will estimate least squares means, standard errors, treatment differences in least squares means, and 95% confidence intervals for each time point. Primary inference will be based on the treatment comparison of least squares means for Day 90 from this model, and a p-value will be presented for this time period only. The null hypothesis is that the mean difference in the primary endpoint between the experimental arms and the sham arm is zero, versus the alternative hypothesis that this difference is not zero.

PCS SF-6

The change from study baseline to Day 90 in pain catastrophizing as measured by PCS SF-6 is a secondary endpoint. This analysis will compare the treatment groups, separately, to the control group using a MMRM analysis. The repeated measures are the change from baseline PCS SF-6 score obtained for Day 15, 30, 60, and 90, respectively. The model will include fixed categorical effects for treatment, week, treatment by week interaction, and the baseline PCS SF-6 score as a continuous covariate. We will employ the Stata *xtmixed* command or the R lmer function from the lme4 package with REML and an unstructured within-patient covariance structure for this model. We will evaluate the assumptions of this model, including normality, using residual and other diagnostic plots of model fit.

From this model, we will estimate least squares means, standard errors, treatment differences in least squares means, and 95% confidence intervals for each time point. Primary inference will be based on the treatment comparison of least squares means for Day 90 from this model, and a p-value will be presented for this time period only. The null hypothesis is that the mean difference in the primary endpoint between the experimental arms and the sham arm is zero, versus the alternative hypothesis that this difference is not zero.

MME

A key secondary endpoint is the change from study baseline to Day 90 in weekly MME of prescribed medication. Descriptive summary statistics will be presented for weekly MME of prescribed medication at two timepoints: baseline and Day 90. Analysis of this endpoint will compare the difference in weekly MME of prescribed medication between the treatment groups

and control group at baseline and Day 90 using analysis of covariance (ANCOVA), adjusting for baseline MME. We will assess the assumptions underpinning the ANCOVA model graphically and will undertake appropriate transformation of the weekly average MME outcome as deemed appropriate. If no suitable transformations can be found, bootstrapped confidence intervals for the between-group differences in weekly MME will be produced. Both adjusted and unadjusted between-group comparisons will be presented with 95% confidence intervals.

4.6.3 Exploratory Endpoint Analyses

PROMIS–PF

The change from study baseline to Day 90 in physical function as measured by PROMIS–PF is an exploratory endpoint. This analysis will compare the treatment groups, separately, to the control group using a MMRM analysis. The repeated measures are the change from baseline PROMIS–PF score obtained for Day 15, 30, 60, and 90, respectively. The model will include fixed categorical effects for treatment, week, treatment by week interaction, and the baseline PROMIS–PF score as a continuous covariate. We will employ the Stata xtmixed command or the R lmer function from the lme4 package with REML and an unstructured within-patient covariance structure for this model. We will evaluate the assumptions of this model, including normality, using residual and other diagnostic plots of model fit.

From this model, we will estimate least squares means, standard errors, treatment differences in least squares means, and 95% confidence intervals for each time point. Primary inference will be based on the treatment comparison of least squares means for Day 90 from this model, and a p-value will be presented for this time point only. The null hypothesis is that the mean difference in the primary endpoint between the experimental arms and the sham arm is zero, versus the alternative hypothesis that this difference is not zero.

PROMIS-Depression

The change from study baseline to Day 90 in depression as measured by PROMIS–Depression is an exploratory endpoint. This analysis will compare the treatment groups, separately, to the control group using a MMRM analysis. The repeated measures are the change from baseline PROMIS–Depression score obtained for Day 15, 30, 60, and 90, respectively. The model will include fixed categorical effects for treatment, week, treatment by week interaction, and the baseline PROMIS–Depression score as a continuous covariate. We will employ the Stata *xtmixed* command or the R lmer function from the lme4 package with REML and an unstructured within-patient covariance structure for this model. We will evaluate the assumptions of this model, including normality, using residual and other diagnostic plots of model fit.

From this model, we will estimate least squares means, standard errors, treatment differences in least squares means, and 95% confidence intervals for each time point. Primary inference will be based on the treatment comparison of least squares means at Day 90 from this model, and a p-value will be presented for this time point only. The null hypothesis is that the mean difference

in the primary endpoint between the experimental arms and the sham arm is zero, versus the alternative hypothesis that this difference is not zero.

PGIC

An exploratory endpoint is the change in disease severity as measured by the PGIC at Day 90. This secondary analysis will compare the treatment groups, separately, to the control group using a logistic regression analysis. Based on the PGIC assessment, a dichotomous scale of "Yes" or "No" will be derived. A favorable response of 5-7 on the PGIC indicates "Yes", significant improvement occurred over the course of the study. An unfavorable response of 1-4 indicates "No", significant improvement did not occur over the course of the study.

The percentage of patients in each treatment arm responding in the "Yes" and "No" categories will be compared to the percentage of patients in the control arm responding in the "Yes" and "No" categories via the Cochran Mantel-Haenszel test. We will analyze responder status (i.e. "Yes" vs "No" as the dependent variable) using logistic regression, with terms for study-arm and baseline PROMIS-PI.

Biometric Data – Steps

We will assess the change from study baseline to Day 90 by comparing the treatment groups, separately, to the control group using a MMRM analysis. The repeated measures are the change from baseline weekly steps obtained for all weeks between Baseline and Day 90. The model will include fixed categorical effects for treatment, week, and the baseline weekly steps as a continuous covariate. We will employ the Stata xtmixed command or the R lmer function from the lme4 package with restricted maximum likelihood estimation (REML) and an unstructured within-patient covariance structure for this model. We will evaluate the assumptions of this model, including normality, using residual and other diagnostic plots of model fit.

From this model, we will estimate least squares means, standard errors, treatment differences in least squares means, and 95% confidence intervals for each time point. Primary inference will be based on the treatment comparison of least squares means at Day 90 from this model, and a p-value will be presented for this time point only. The null hypothesis is that the mean difference in the primary endpoint between the experimental arms and the sham arm is zero, versus the alternative hypothesis that this difference is not zero.

Biometric Data – Sleep

We will assess the change from study baseline to Day 90 by comparing the treatment groups, separately, to the control group using a MMRM analysis. The repeated measures are the change from baseline weekly minutes classified as "sleep" obtained at weekly intervals between baseline and Day 90. The model will include fixed categorical effects for treatment, week, and baseline weekly steps as a continuous covariate. We will employ the Stata xtmixed command or the R lmer function from the lme4 package with restricted maximum likelihood estimation (REML) and an unstructured within-patient covariance structure for this model. We will evaluate the

assumptions of this model, including normality, using residual and other diagnostic plots of model fit.

From this model, we will estimate least squares means, standard errors, treatment differences in least squares means, and 95% confidence intervals for each time point. Primary inference will be based on the treatment comparison of least squares means at Day 90 from this model, and a p-value will be presented for this time point only. The null hypothesis is that the mean difference in the primary endpoint between the experimental arms and the sham arm is zero, versus the alternative hypothesis that this difference is not zero.

4.6.3 Subgroup Analyses

Subgroup analyses are planned for the change from baseline in PROMIS-PI. An MMRM model will be used to test for treatment by subgroup interactions. Interactions with a significance level of less than 10% will be considered potentially important and flagged for further assessment. In general, the models will include fixed categorical effects for treatment, week, treatment by week interaction, subgroup, and treatment by subgroup interaction. The p-values of interaction terms will be presented, as well as the least squares means and 95% confidence intervals by treatment and subgroup classification factor. Descriptive statistics of the observed and change from baseline PROMIS-PI t-score will also be presented by treatment and week within each subgroup.

Subgroup analyses will be performed for the following subgroups:

- Dosage of VR (minutes per week)
- Previous experience with VR
- ITQ cutoff score
- Presence cutoff score
- Patient comorbidities
- History of spinal surgery
- Pain severity and duration cutoffs
- Socio-demographics (i.e. age, sex, race, ethnicity, marital status, zip code)
- Other Medications
- TAPS-1, TAPS-2

The primary MMRM models will be refit to include terms describing subgroups. The change from baseline PROMIS-PI score obtained for Day 30 will be compared between the treatment groups using MMRM analysis. The repeated measures are the change from baseline PROMIS-PI score obtained for Day 7, 15, 21, and 30, respectively. The model will include fixed categorical effects for treatment, week, treatment by week interaction, relevant subgroup, treatment by subgroup interaction, and the baseline PROMIS-PI score as a continuous covariate. The estimated least squares means and 95% confidence intervals on the treatment comparison for Day 7, 15, 21, and 30 will be presented in forest plots for each subgroup of relevance.

4.7 Analysis of Safety

All safety analyses will be descriptive. No statistical testing will be performed.

5. CHANGES IN THE CONDUCT OF THE STUDY OR PLANNED ANALYSES

5.1 Detailed Changes in the Conduct of the Study

For a broad overview, please see the amendment summaries for Protocol 00000631.

CHANGES WILL BE DETAILED HERE

5.2 Detailed Changes in the Analyses Planned in the Protocol

For a broad overview, please see the amendment summaries for Protocol 00000631.

CHANGES WILL BE DETAILED HERE

6. GENERAL STATISTICAL/ANALYTIC ISSUES AND APPROACHES

6.1 Handling of Dropouts or Missing Data

For randomized patients who discontinue early from the study, measurements will not forward to the date of discontinuation. For purposes of analysis, their last day in the study will be the last day on which data was contributes. Unless stated otherwise on a case-by-case basis, no further imputation will be used for descriptive analyses, or for primary and secondary efficacy analyses utilizing MMRM methodology.

6.2 Interim Analyses and Data Monitoring

No formal interim analysis or interim statistical testing for treatment comparisons is planned.

Response rates and patient safety data will be monitored by an independent DMSB until the last patient completes his last scheduled assessment. The DMSB will be an external group overseeing the safety of the study treatment through an executive secretary (Navitas Life Sciences). The principal investigator(s) will meet with the DSMB on a biannual basis, once via teleconference (approximately 2 hours) and once in-person (approximately 4 hours). Additional meetings may be held ad hoc or routinely, if determined necessary. The research team submits a monthly enrollment report to NIAMS viathe Executive Secretary to review enrollment progress and safety monitoring data.

6.3 Multiple Comparisons/Multiplicity

The family-wise type I error rate (FWER) for the statistical tests of the primary and secondary endpoints will be controlled at 0.05. To strongly control the FWER at this level, a gate-keeping approach will be utilized in which each family of statistical tests will be conducted in a sequential manner. A closed testing procedure will be employed to control the FWER. Hypothesis tests conducted for exploratory purposes will be conducted outside of any gatekeeping.

6.4 Missing Data and Sensitivity Analyses

If the primary or key secondary endpoint is missing in >15% of patients in either treatment group, then the pattern of baseline covariates with missing values will be examined using the method of Little[2] and in case the data is not missing completely at random, missing values will be imputed using fully conditional specification with the multivariate imputation by chained equations (MICE) algorithm under the missing at random (MAR) assumption.[3,4] Fifty or so data sets will be generated and analyzed separately, and the results will be combined using the formula according to Van Buuren.[5] Similar considerations will be applied to missing data on the dependent variable in a repeated measure mixed model by investigating patterns of missingness and presence of nonignorable missing response data.[6]

References

1. Malinckrodt et al. 2008; Siddiqui et al. 2009).
2. Little RJ. A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association. 1988;83(404):1198-202.
3. Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software. 2011;45(3).
4. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical methods in medical research. 2007;16(3):219-42.
5. Rubin DB. Multiple imputation for nonresponse in surveys: John Wiley & Sons; 2004.
6. Ibrahim J. G, Molenberghs G. Missing data methods in longitudinal studies: a review. 2009; PMC, 18(1): 1-43.