STATISTICAL ANALYSIS PLAN (SAP)

A Cluster-Randomized Trial of the Northwestern Embedded Emergency Department Physical Therapy (NEED-PT) Protocol for Acute Low Back Pain

February 20th, 2025

Version 2.0

NU IRB: STU00213134
Registration: NCT04921449

Statistical Analysis Plan (SAP) Revision History

Version Number: 1.0
Version Date: 10/04/2021
Summary of Revisions Made: Original Version

Version Number: 2.0
Version Date: 2/20/25
Summary of Revisions Made:

- Context for changes: The final NEED-PT trial participant was enrolled on 7/11/24 and we anticipate final data collection on 7/25/25 (two-week follow-up window for the 12-month follow-up timepoint). We are modifying the SAP in advance of our anticipated database lock on 7/25/25.
- Expanded the methodology of Exploratory Analysis #4. As primary treatment classification is available only among those participants in the treatment arm (NEED-PT), we must specify our process for matching participants in the control arm (usual care) to comparable treatment classifications.
- Additional detail has been added to Sensitivity Analyses #1-#3. We anticipate that the total number of participants meeting criteria for Sensitivity Analysis 1 and 2 may be low and have therefore specified combining these analyses should n≤10 in either case. We also clarify that Sensitivity Analysis 3 (per-protocol) is the same analysis previously described as a “complier analysis.”
- We now specify our approach to missing data, which we piloted in a similarly structured trial of an ED physical therapy intervention for dizziness (PMID 39951266).

STATISTICAL ANALYSIS PLAN (SAP)

A Cluster-Randomized Trial of the Northwestern Embedded Emergency Department Physical Therapy (NEED-PT) Protocol for Acute Low Back Pain

Principal Investigator: Howard S. Kim, MD MS
Statistical Team: Jody D. Ciolino, PhD; Jacob M. Schauer, PhD, Ann Kan, MS

1. INTRODUCTION

This document outlines the proposed analyses for the NEED-PT study. Briefly, NEED-PT is a physician-randomized (i.e., physicians serve as clustering or randomization units) trial evaluating efficacy of an embedded emergency department (ED) physical therapist in comparison to usual care. Individual ED patients will be consented and enrolled during their index ED visit and will follow the randomization assignment of their treating physicians. The primary outcome data will be analyzed at three months after the index ED visit, with additional data collection up to 12 months for evaluation of longer-term effects and exploratory endpoints.

Study Aims

The overarching study aims are as follows. **This SAP will focus on the details for Aim 2 analyses, which will guide the reporting of the primary study findings**. We reserve details of analyses surrounding additional aims for separate document(s):

Aim 1: Develop and field-test the Northwestern "Embedded" ED Physical Therapy (NEED-PT) intervention protocol for the routine co-evaluation of all ED patients with acute low back pain.

We will co-locate the ED physical therapist with the ED physician as part of the primary treatment team in order to remove biases in treatment selection and allow for earlier integration of ED-PT into patient care. A formalized protocol will enhance intervention fidelity in Aim 2 and facilitate dissemination of our care model.

Aim 2: Conduct a single-center, physician-randomized trial (n=40) comparing NEED-PT to usual care among ED patients (n=360) with acute low back pain to evaluate a primary outcome of pain-related functioning at three months and a secondary outcome of opioid use at three months.

H1: Patients receiving NEED-PT will report greater improvement in pain-related functioning compared to patients receiving usual care, as measured by average PROMIS Pain-Interference score
H2: Patients receiving NEED-PT will use fewer daily opioids on average.

Aim 3: Compare rates of diagnostic imaging utilization for ED visits with low back pain among ED physicians randomized to NEED-PT versus usual care.

H1: ED physicians randomized to NEED-PT will have a lower rate of diagnostic imaging utilization for low back pain compared to ED physicians randomized to usual care.

For patients enrolled in the study, study time points include baseline assessment (completed at the index ED visit), Week 1, Month 1, Month 2, Month 3 (primary endpoint), Month 6, and Month 12.

2. STUDY OUTCOMES

Primary Outcome

The primary efficacy outcome is **PROMIS-Pain Interference Score (PROMIS-PI)** three months after the index ED visit. PROMIS-PI measures the self-reported consequences of pain on relevant aspects of a person's life, including social, cognitive, emotional, physical, and recreational activities. We will use the computer-adaptive format to minimize respondent burden. Scores are standardized to the general U.S. population, with a score of 50 representing the population mean and a standard deviation of 10 points. The time frame of interest for the PROMIS-PI is "in the past 7 days," meaning that participants provide responses based on their symptoms over the last week. The minimum clinically important difference for low back pain is in the range of 3.5-5.5 points. We will treat this variable as continuous in analyses.

Secondary Outcomes

Secondary efficacy outcomes include:

1) **Oswestry Disability Index (ODI)** at three months. *ODI* is a disease-specific instrument that contains 10 questions relating to low back pain intensity, personal care, lifting, walking, sitting, standing, sleeping, social life, traveling, and employment/homemaking. The ODI score ranges from zero (no disability) to 100 (maximum disability), with an estimated minimum clinically important difference of six points for acute low back pain. The time frame of interest for the ODI is "today," meaning that participants provide responses based on their current symptoms on the day of survey response. The modified ODI replaces an item from the original ODI pertaining to sex life with a new item pertaining to employment/homemaking. We expect this outcome to be largely correlated with PROMIS-Pain Interference. We will treat this variable as continuous in analyses.
2) **Patient-Reported Opioid Use** at three months. This will be collected using a customized instrument assessing whether participants have taken any opioid medication in the last 24 hours. The 24-hour timeframe was selected to maximize accuracy in patient recall and has been used previously. In brief, opioid medications are listed by brand and generic names; a "yes" response to any medication triggers an additional query asking the participant to specify the medication dose (e.g., oxycodone 10mg) and quantity (e.g., four pills), allowing for standardization by morphine milligram equivalents (MME). We anticipate treating this variable as either count or a binary (any dose vs. none), or continuous (MME) for analyses.

Exploratory Outcomes

We expect the following outcomes to be related to the primary and the major secondary outcomes of interest. We deem the more exploratory in nature, and they thus carry less weight in analyses and overall inferences regarding efficacy of intervention.

1) **Opioid Prescription Filling** will be queried in the state prescription monitoring database. We anticipate treating this variable as count, binary, or continuous (MME).
2) **Patient-Reported Prescription Analgesic Use** in the last 24 hours will be collected using the same customized instrument described above for opioid use. Prescription analgesics include: opioids, benzodiazepines, skeletal muscle relaxants, and gabapentinoids. We anticipate treating this variable as either count or binary.
3) **Prescription Analgesic Filling** will be queried in the state prescription monitoring database. Prescription analgesics include: opioids, benzodiazepines, skeletal muscle relaxants, and gabapentinoids. We anticipate treating this variable as either count or binary.
4) **Numeric Pain Rating Scale (NPRS)** measures pain intensity from 0 to 10 and is easily understood by laypersons, clinicians, and researchers. We will assess a single item relating to average pain intensity over the last 24 hours. We plan to treat this as a continuous outcome, but we anticipate requiring transformation or nonparametric analyses, as this variable will likely be skewed and exhibit flooring / ceiling effects.
5) **Global Rating of Change (GROC)** is a single-item survey widely used by clinicians and researchers to quantify functional disability in low back pain and evaluate the overall effectiveness of therapy. This item ranges from zero (a very great deal worse) to 14 (a very great deal better). We plan to initially treat this measure as continuous, but we anticipate exploring this outcome as a count variable, requiring transformation, or using nonparametric analyses.
6) **Pain Catastrophizing Scale (PCS-4)**. The original PCS is a 13-item survey measuring the degree to which an individual catastrophizes in response to pain. PCS scores correlate closely with pain intensity and disability over time; higher PCS scores are associated with progression from acute to chronic pain. We will utilize the brief 4-item PCS measure containing original items 3, 6, 8, and 11 to reduce respondent burden. We will treat this variable as continuous in analyses.
7) **Pain Self-Efficacy Questionnaire (PSEQ-4).** The original PSEQ is a 10-item survey measuring the confidence with which individuals can do things despite pain. We will utilize the brief 4-item PSEQ measure containing original items 4, 6, 8, and 9 to reduce respondent burden. We will treat this variable as continuous in analyses.
8) **Advanced Healthcare Resource Utilization.** We will assess the proportion of participants who utilized advanced healthcare resources for low back pain after their index ED visit, defined as advanced imaging (e.g., magnetic resonance imaging) or procedures/surgery (e.g., epidural steroid injection, lumbar discectomy).

9) **ED Diagnostic Imaging Utilization.** We will assess the proportion of ED visits in which diagnostic imaging of the lower back was performed, including plain radiography, computed tomography, and magnetic resonance imaging.
10) **Additional outcomes (not discussed in detail in this SAP) that are a part of the third study aim include:** ED length of stay, ED disposition (admit, observation, discharge), total costs/charges.

3. DEMOGRAPHICS AND BASELINE ASSESSMENTS

The following are specific demographic / baseline assessments of interest for analyses. **Primary analyses will adjust for these covariates as we anticipate they will influence outcome**. We plan to report both model-adjusted and simple unadjusted intervention effect estimates:

1) Sex
2) Age
3) Keele STarT Back Screening Tool: a nine-item survey which assesses risk for progression to chronic base pain

Additional demographics and clinical characteristics we plan to collect and summarize (i.e., we **do not plan to include as covariates in analyses**) include:

1) Race / ethnicity
2) Education level
3) Marital status
4) Employment status
5) Activity level at work for those that are working at baseline
6) Income level
7) Physical activity level according to self-report
8) Nature of injury
9) Length of pain at baseline
10) Primary diagnosis
11) Medications administered / prescribed during initial ED visit

Note that some additional exploratory analyses may examine these additional variables as covariates and/or effect modifiers as well. We will label any exploratory analyses involving additional potential covariates as post hoc in any dissemination materials.

4. DATA STORAGE

Data will be collected and managed using Research Electronic Data Capture (REDCap) housed at Northwestern University's Clinical and Translational Sciences Institute (CTSA), NUCATS (1, 2) . REDCap is a secure, web-based application designed for research studies that provides an intuitive interface for validated data entry, audit trails for tracking data manipulation and export procedures, and automated export procedures for seamless data downloads to common statistical packages, and procedures for importing data from external sources. Individualized REDCap survey links will be sent to participants using Mosio, a secure text messaging research platform that is 21 CFR Part 11 compliant and integrates with REDCap.

5. RANDOMIZATION METHODS

We plan for equal allocation (1:1) of physicians across study arms; thus, there will be inevitable imbalance in patient numbers across study arms. Physicians will be randomized to either the intervention (NEED-PT) or "control" (usual care). Physicians randomized to the NEED-PT intervention will have a physical therapist assigned to their treatment team who will automatically evaluate all patients with low back pain. Physicians randomized to "usual care" will not have a physical therapist assigned to their treatment team, and their patients with low back pain will not be automatically evaluated by the physical therapist. Due to the inherent risk of cluster-level (i.e., physician-level) covariate imbalance between study arms in cluster-randomized trials, we will employ covariate-constrained randomization techniques to control for possible imbalance in key physician-level characteristics. Covariate-constrained randomization methods tend to ensure the most efficient control over covariate imbalance between study arms at randomization (3, 4). With 40 total physicians, there are over 137 billion ways (40 choose 20) in which we can achieve equal allocation of physicians across study arms. The constrained randomization procedure will involve:

1) Enumerating a 10 thousand possible allocation schemes at the 1:1 physician allocation ratio.
2) Calculating imbalance in the following baseline physician-level variables across study arms for each of the schemes simulated in step 1:
    a. Physician gender
    b. Physician years' experience (since first year of residency)
    c. Physician race
    d. Physician opioid prescription rate
    e. Number of "fast track" zone shifts for a physician per month, on average – fast track shifts are those with the highest likelihood of receiving low-back pain patients
        i. This variable is highly correlated with the number of day shifts a physician tends to have per month
        ii. It is also correlated with the mean number of patients the physician sees per hour
        iii. While we will control imbalance in the randomization algorithm for this "fast track" zone variable, we anticipate reporting summary statistics on day shifts and patients per hour
3) Constraining the randomization space to a subset of allocation schemes that do not surpass some threshold of "allowable" imbalance for each of the variables (a-e in step 2) above. The thresholds will be guided by the following restrictions; however, the distribution of these physician-level variables may require modification(s) to these thresholds. Any updates will be documented in a later version of this SAP:
    a. Physician gender counts may not differ by more than two for any one category across study arms.
    b. Mean number of years' experience may not differ more than one year.
    c. Physician race will likely require dichotomization into White vs. Minority for randomization. We will not allow physician racial category counts to differ by more than two for any one category across arms.
    d. Mean physician opioid prescription rate may not differ by more than 0.5 standard deviation units across study arms.
    e. Mean number of orange or red zone shifts may not differ by more than 0.25 across study arms.
4) Of the possible allocation schemes meeting the criteria outlined in Step 3, randomly select one for implementation in the study.

6. STATISTICAL METHODS

We plan to use descriptive statistics to summarize baseline patient and physician-level variables both overall and by arm. We will use mean±standard deviation (or median and interquartile range [IQR] as appropriate) for continuous variables and frequency / percentage for categorical variables. Specifically, we will summarize age, sex, Keele STarT score, baseline patient-reported outcome scores (PROMIS-PI and ODI), analgesic medication prescription at ED discharge, and the variables listed above. Analyses will involve normal theory methods in general, and in cases of violations of assumptions, we will consider transformation and / or nonparametric / exact methods as appropriate.

Analyses will assume a two-sided 5% significance level. All primary efficacy and safety analyses will be pre-specified as outlined in this SAP, and deviations from planned analyses or post hoc analyses will be labeled as such in any reports or dissemination materials. We do not plan to control for multiple hypothesis tests.

In analyses for each outcome, we plan to control for the respective outcome value at baseline (i.e., in an analysis of covariance [ANCOVA] approach). Analyses for the primary outcome (Y) will involve a linear mixed model (LMM) with repeated measures with fixed effects for: study arm, baseline outcome score (Y0), timepoint, timepoint-by-arm interaction, and known influential predictor effects (age, sex, Keele STarT score). Inference will focus on treatment impacts for the outcome at three months. We will include a random physician effect to account for both within and between physician variability and also to allow for estimation of the intra-cluster correlation coefficient (ICC). The repeated measures on the same participant over time will also introduce a correlation structure across time points, providing the justification for modeling the correlation structure at the participant level over time. We will use an unstructured correlation matrix to account for the repeated measures within a participant as this has the least assumptions. If the model does not converge or parameters cannot be estimated under this unstructured covariance pattern, we will explore simpler covariance patterns using residual estimated maximum likelihood (REML) comparisons. Including repeated measures per participant will allow us to make most use of all participant data after baseline. We will use assume an unstructured covariance across time.

To evaluate efficacy, the Wald model type III test for fixed arm effect will be evaluated assuming a two-sided 5% type I error rate. The primary contrast of interest to address the primary research aims involves the comparison of the model-estimated mean outcome score at three months (T4) across study arms. This modeling strategy is robust to unbalanced (i.e., incomplete) data across study time points. We will also provide results for unadjusted analyses (i.e., without accounting for the pre-specified covariates). Analyses of additional outcomes will follow the same general analytic strategy: LMM with fixed arm, baseline outcome value, influential baseline covariate effects, and a random physician effect and covariance patterns to account for repeated measures within participants. We chose to incorporate baseline outcome as a covariate in the model, rather than as a time point, based on clinical reasoning. As these baseline values (e.g., PROMIS-PI score at the index ED visit) are assessed pre-intervention and primary analyses aim to assess outcome(s) as follow-up accounting for pre-intervention state. Incorporating this baseline value in the analytic model as a fixed effect will increase precision and reduce bias on the intervention effect estimate for primary outcome at the time point of interest as the baseline value will likely be highly correlated with outcome at follow-up (previous data: $p<0.001$ for both PROMIS and ODI).

Residual diagnostics will assess model fit and assumptions, and in the case of violation, we will explore transformations / nonparametric methods as indicated above. In the event of poor model fit, we may explore different distributional assumptions as appropriate (e.g., Poisson for count or rate data) with the corresponding canonical link (e.g., log) function. As above, we will assess model fit via residual diagnostics and may consider transforming or nonparametric methods as needed.

Analyses for outcomes that are either binary or count will follow the same general approach as above; however, they will involve generalized linear mixed effects (GLMMs) models with the appropriate distributional (e.g., binomial or Poisson) and link (e.g., logit or log) assumptions. Modeling the covariance structure for these outcomes may result in unstable model estimates. If this occurs, we anticipate removing the random physician effect and including a random participant effect instead to account for correlation.

Exploratory Analyses

In addition to repeating the above analyses with exploratory outcomes, we will conduct exploratory analyses to study effects among subgroups of patients (moderator analyses) and examine the potential impact of PT use among patients in the control arm.

Planned moderator analyses will include the following moderators:

1. Opioid naivete as measured by whether patients report taking opioids within the last 24 hours at their index ED visit or have a history of opioid prescription filling in the Illinois prescription monitoring program within the last 3 months.
2. Initial symptom burden measured as “moderate/severe” if their baseline measures of PROMIS pain scores are ≥60 or their STarT score registers as “high risk,” defined as a subscore ≥4 (questions 5-9).
3. Age ≥ 65 years old
4. Primary treatment classification, as per the clinical care protocol (directional preference, traction, stabilization, manipulation, nociplastic presentation). As primary treatment classification is assigned only in those participants in the treatment arm (NEED-PT), we will use two approaches to study the effects of treatment classification. First, we will modify GLMM to include treatment classification effects (as opposed to just a study arm effect). Second, we will use matching methods (e.g., propensity score matching or augmented synthetic controls) to match individuals in the NEED-PT arm who have a given primary treatment classification to control arm participants on baseline covariates. We would then use GLMM to estimate the impact of each treatment classification separately.

Analyses will focus on PROMIS-PI scores measured three months after patients’ index visits, as well as ODI scores and opioid use (proportion using an opioid within the last 24 hours) at the same time point. Analyses will involve generalized linear models with appropriate link functions (identity for PROMIS-PI and logit for opioid use) that include fixed effects for baseline measures of the outcome of interest, treatment assignment, a moderator variable, and a treatment-moderator interaction. As above, PROMIS-PI will be modeled with standard normality assumptions, which will be evaluated via residual diagnostics and appropriate transformations will be used as necessary. Separate models will be fit for each outcome and moderator. Tests for the treatment-moderator interaction will be two-sided with a 5% type I error rate, and we will report point estimates and 95% confidence intervals. For the logistic regression involving opioid use, we will use Wald confidence intervals and Wald tests. We will not make multiple comparison adjustments.

Mediation analyses will focus on PCS and PSEQ as possible mediators. Our hypotheses are that embedding a PT in the ED can impact patients downstream pain catastrophizing and self-efficacy which will in turn lead to lower reported pain and less frequent opioid use. Our key dependent variables will be PROMIS-PI and opioid use at three months after the index visit. PCS and PSEQ measured at one month will be the mediators of interest. We will use a nonparametric approach to analyses, running separate models for each outcome and mediator (5). In addition, we will examine the possible correlation between mechanisms by using a joint nonparametric estimation framework (6).

In addition, we will conduct a complier analysis. Based on pilot data, we expect some patients in the control arm will receive a discretionary PT consultation as part of usual care. These consultations will be operationally different from those in the treatment group, as the PT will not be embedded with the care teams in the control arm. Conversely, it is possible some treatment arm patients may not receive an embedded PT consult, though we expect this will be rarer. Since we hypothesize that PT consultation will play a large role in this intervention's effectiveness, we propose to examine the impact of these differential PT consultations (discretionary, embedded) in two ways. First, we will re-create the proposed confirmatory analyses excluding control patients receiving a PT consultation and intervention arm patients who do not. Second, we will use a generalized mediation analysis that includes all patients that treats receipt of a PT consult as a mediator to estimate the direct and indirect effects of treatment assignment and PT consultation. This mediation analysis will focus on PROMIS-PI at three months post-index visit as the outcome of interest, and use a generalized nonparametric estimation approach (5).

7. ANALYTIC DATASET

Primary and secondary outcomes will be evaluated across arms under a modified intention-to-treat (mITT) principle, (1) whereby all participants will be included in analyses, regardless of their or their physicians' adherence to their assigned study arm, and (2) only participants contributing at least one follow-up data point will be included. That is, we will exclude patients who are lost to follow-up before Week 1. Sensitivity analyses will be detailed after data collection; however, we plan to conduct sensitivity analyses that would involve:

1) Excluding patients who are ultimately admitted to the hospital after their ED visit. We anticipate that the number of participants meeting the criteria for this sensitivity analysis will be low. If n≤10 participants, we will combine with sensitivity analysis #2 to create a single sensitivity analysis.
2) Excluding patients with an alternative diagnosis after enrollment that would have deemed them otherwise ineligible (e.g., discovery of kidney stones or shingles after enrollment). We anticipate that the number of participants meeting the criteria for this sensitivity analysis will be low. If n≤10 participants, we will combine with sensitivity analysis #1 to create a single sensitivity analysis.
3) Excluding patients who cross over to the study arm to which they were not assigned (i.e., per-protocol analysis). If this occurs frequently, we may explore instrumental variables or propensity score methods as sensitivity analyses. This per-protocol analysis is also described in the previous section as a "complier analysis." We clarify that this these two descriptions refer to the same single analysis.

Power and sample size considerations allowed for some missing data (20%). Any analysis involving missing data inherently makes assumptions about why data are missing and such assumptions are often not testable, nor are the missing data mechanisms to which they pertain necessarily known prior to the completion of data collection. Thus, we will evaluate potential missingness mechanisms and patterns that will ultimately inform multiple analyses, but we highlight here our anticipated steps and delineate possible primary and sensitivity analyses.

We will examine rates of missing data for all variables and determine whether the rates vary by participant characteristics and study arm. These summarizations will inform potential biases resulting from missing data.

Given that the mechanism of missing data may not be explicitly known in many cases, we plan to present multiple analyses conducted under various assumptions of missingness. Mixed effects models planned for longitudinal analysis are generally robust for unbalanced data across study time points, and so we will present analyses that ignore missing observations, which assumes that they are missing completely at random (MCAR). Additional analyses may be explored to evaluate overall robustness of inferences should greater than 10% of data wind up missing. These analyses will again serve as sensitivity analyses and provide inference under different assumptions regarding missingness. The details of these analyses will be documented at the time of analyses (if needed). We now specify the following additional analyses:

4) Multiple Imputation (MI) for mITT analyses will involve the imputation of missing follow-up data among the modified intention to treat population. This approach assumes data are missing at random (MAR, a less stringent assumption), as opposed to MCAR (more stringent). We will use multilevel imputation models to generate m=40 imputations using observed data in our mITT population (i.e., those participants providing follow-up data at least one timepoint). This will allow us to confirm the robustness of our primary mITT analysis to the assumption of MAR vs. MCAR data.
5) ITT analysis: Because the mITT population excludes participants lost to follow-up (LTFU) (i.e., excludes those participants who provided follow-up data at zero of the possible timepoints), we will conduct a sensitivity analysis using data on all eligible patient participants (i.e., including those participants who provided zero follow-up data) and combining MI with inverse probability weighting (IPW).(7) The probability of *not* being LTFU was estimated as a function of baseline characteristics and baseline measures of outcomes using a logistic regression model. Then, among our mITT imputations, GLMMs will be estimated using weights inversely proportional to that probability. This method adjusts analyses for differences between the LTFU and mITT populations and provides a measures of how sensitive mITT analyses are to excluding LTFU patient participants.

8. POWER AND SAMPLE SIZE CONSIDERATIONS

Power calculations focus on the primary endpoint of PROMIS-PI at three months, and we desire adequate (at least 80%) power to detect the minimum clinically important PROMIS-PI score difference of 3.5-5.5 points as previous literature suggests (8). If we assume a standard deviation of 10 points, which is the defined standard deviation of PROMIS-PI, this corresponds to a desired minimal detectable effect size of d=0.35 standard deviation unit difference across arms. Power considerations also account for a 20% drop-out rate for physician clusters (e.g., physician leaves the practice or refuses participation after randomization) and a 20% lost to follow-up rate among recruited participants. We used "The Shiny CRT Calculator" to explore varying assumptions on cluster size (i.e., average number of participants per physician), number of clusters/physicians, and ICC. Under the parallel-arm, "cohort" design, with baseline measurement of primary outcome (PROMIS), the calculator also allows for an assumption on correlation between baseline and follow-up. The table below illustrates power to detect at least a 3.5 mean difference across study arms if we assume just two time points (baseline and three months, which we deem conservative as we will have up to seven time points of observation, including baseline) per participant with a correlation between the two of approximately 0.50. We conservatively estimate that we will need to enroll up to 360 total participants to account for worst-case (20%) scenario dropout for both physicians and participants. **Thus, after accounting for physician and participant dropout, a final sample size of 16 physicians per arm and 7 participants per physician (n=224 total or 112 per arm) achieves 84% power** to detect a mean between-arm difference of 3.5 PROMIS-PI points assuming standard deviation of 10 points, ICC of 0.10, and a two-sided 5% level of significance. In our pilot work, we found a small ICC (0.01-0.04), indicating minimal within-physician effects that were not significant; however, we utilize a more conservative estimate of the ICC at 0.10 in the event that greater than anticipated within-physician effects are encountered. In the event that ICC is lower than expected or dropout rate is lower than 20%, we anticipate often over 90% to detect a meaningful difference across arms. Similar effect size in secondary outcomes (e.g., 0.35 standard deviation units difference in ODI across arms) are also detectable with at least 80% power under similar assumptions. Additionally, we plan to conduct secondary longitudinal analyses involving multiple time points per participant (i.e., more data observations) using likelihood-based methods that are robust to missing data. Therefore, we anticipate adequate power to evaluate differences across arms in outcome trajectories. **Since our target final analytic sample size is 224 total participants, if we can reach our target with fewer participants enrolled than 360, we will consider stopping enrollment.** We will plan to monitor dropout rates, ICC, standard deviation, and within-participant correlation throughout the course of the trial, and we will seek advice from the External Advisory Board and DSMB as we make any interim decisions on stopping enrollment prior to the planned 360 participants.

ICC	Physicians (Total)	Physician % Dropout	Average N participants per Physician	Participant % Dropout	Power: Mean 3.5-point Difference
0.01	40	0	9	0	97%
	40	0	8	5 to 10	95%
	40	0	7	15 to 20	92%
	38	5	9	0	96%
	38	5	8	5 to 10	94%
	38	5	7	15 to 20	90%
	36	10	9	0	95%
	36	10	8	5 to 10	92%
	36	10	7	15 to 20	89%
	34	15	9	0	94%
	34	15	8	5 to 10	91%
	34	15	7	15 to 20	87%
	32	20	9	0	92%
	32	20	8	5 to 10	89%
	32	20	7	15 to 20	85%
0.05	40	0	9	0	96%
	40	0	8	5 to 10	94%
	40	0	7	15 to 20	91%
	38	5	9	0	95%
	38	5	8	5 to 10	93%
	38	5	7	15 to 20	90%
	36	10	9	0	94%
	36	10	8	5 to 10	91%
	36	10	7	15 to 20	88%
	34	15	9	0	93%
	34	15	8	5 to 10	90%
	34	15	7	15 to 20	86%
	32	20	9	0	91%
	32	20	8	5 to 10	88%
	32	20	7	15 to 20	84%
0.10	40	0	9	0	96%
	40	0	8	5 to 10	94%
	40	0	7	15 to 20	91%
	38	5	9	0	95%
	38	5	8	5 to 10	93%
	38	5	7	15 to 20	90%
	36	10	9	0	94%
	36	10	8	5 to 10	92%
	36	10	7	15 to 20	88%
	34	15	9	0	93%
	34	15	8	5 to 10	90%
	34	15	7	15 to 20	86%
	32	20	9	0	91%

	32	20	8	5 to 10	88%
	32	20	7	15 to 20	84%

We will not need to adjust sample size calculations for the covariate-constrained randomization approach, as this merely controls imbalances across arms on physician-level (i.e. cluster) covariates, such as physician productivity (e.g., patients seen per hour) while preserving the 1:1 study arm allocation ratio. Therefore, controlling imbalance on these physician-level covariates is intended to translate to both equal allocation of physician participant numbers and comparable participant-level covariate distributions across arms. As mentioned above, we anticipate that this increased control over imbalance coupled with the analytic strategies will increased precision and reduce bias in estimating intervention effects. Since the amount of increased precision is unknown, we deem the sample size and power calculations conservative.

9. TECHNICAL DETAILS

The SAP is subject to version control, and we anticipate modifications to analytic plans be documented herein. As in any study, the analytic plan may change due to assumption violations, logistical issues, unexpected empirical distributions of study outcomes, or a combination thereof. In these cases, the SAP will be updated accordingly. All analyses will be performed via SAS version 9.4 or higher (The SAS Institute; Cary, NC) or R version 4.0.4 or higher (The R Foundation for Statistical Computing platform). Table and figure formatting and style may be dictated by mode of dissemination or specific target journal(s) for results dissemination.

10. TIMELINE FOR ANALYSES

The analysis plan does not include any formal interim statistical analyses involving hypothesis testing or any pre-specified stopping criteria for efficacy or futility on primary or secondary outcomes. Interim reports to the study team, external advisory board, or Data and Safety Monitoring Board (DSMB) will consist of process measures such as protocol adherence, missing values, missing forms, etc. We also plan to use simple descriptive statistics on primary and safety outcomes of interest in aggregate (not stratified by arm). Regular bi-weekly meetings with the study team will utilize central statistical monitoring techniques as a method of quality control and quality assurance for trial data on an ongoing basis. We foresee the DSMB requiring specific data listings or summarizations, but these will be specified at the time of the relevant DSMB meeting(s); at this time, however, we do not plan for formal statistical analyses involving hypothesis testing for DSMB interim review.

To preserve the integrity of the study, no formal statistical analyses will occur until the REDCap database has been locked and all known queries/discrepancies resolved; the date of database lock will be documented.

References:

1. Harris PA, editor Research Electronic Data Capture (REDCap)-planning, collecting and managing data for clinical and translational research. BMC bioinformatics; 2012: BioMed Central.
2. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: Building an international community of software platform partners. Journal of biomedical informatics. 2019;95:103208.
3. Ivers NM, Halperin IJ, Barnsley J, Grimshaw JM, Shah BR, Tu K, et al. Allocation techniques for balance at baseline in cluster randomized trials: a methodological review. Trials. 2012;13(1):120.
4. Raab GM, Butcher I. Balance in cluster randomized trials. Statistics in medicine. 2001;20(3):351-65.
5. Imai K, Keele L, Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. Statistical science. 2010;25(1):51-71.
6. Imai K, Yamamoto T. Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. Political Analysis. 2013;21(2):141-71.
7. Seaman SR, White IR, Copas AJ, Li L. Combining multiple imputation and inverse-probability weighting. Biometrics. 2012;68(1):129-37.
8. Amtmann D, Kim J, Chung H, Askew RL, Park R, Cook KF. Minimally important differences for Patient Reported Outcomes Measurement Information System pain interference for individuals with back pain. Journal of pain research. 2016;9:251.