ANALYSIS PLAN

Study Title: Strategies for Prescribing Analgesics Comparative Effectiveness (SPACE) Trial

NCT01583985

This protocol supplement contains the original statistical analysis plan excerpted from the funded research proposal (date of document: December 1, 2011).

Details of the final protocol and recruitment outcomes are in the following publication: Krebs EE, Jensen AC, Nugent S, DeRonne B, Rutks I, Leverty D, Gravely A, Noorbaloochi S, Bair MJ, Kroenke K. Design, recruitment outcomes, and sample characteristics of the Strategies for Prescribing Analgesics Comparative Effectiveness (SPACE) trial. Contemporary Clinical Trials. 2017 Nov;62:130-139. PMID: 28893675 DOI 10.1016/j.cct.2017.09.003

C7. Analysis Plan

C7a. Sample size estimates

Sample size is calculated based on estimated effects on the primary outcome measure, the BPI Interference score. A between-group treatment difference of 1 point in the BPI Interference score represents a minimum meaningful intervention effect.[108] For our calculations, we assume the standard deviation (SD) of the BPI Interference score in both arms will be 2.7, based on the intervention arm of the SCAMP trial.[52] A sample size of 115 in each arm is estimated to provide 80% power to detect a 1-point difference in the mean BPI Interference score between groups, assuming 2-sided alpha of 0.05. Based on our prior studies, we anticipate a dropout rate of 8-12%; allowing for a conservative estimate of 20% attrition, we will aim for 138 participants in each arm. This sample size will also provide 86% power to detect a 1-point difference in the BPI Severity score and 94% power to detect a 3-point difference in the Roland disability score. Contamination effects related to primary care assignment are expected to be minimal because primary care providers will not be directly involved in the study interventions.

We will examine response rate as a secondary measure of effectiveness. The standard definition of response to chronic pain treatment is a 30% improvement.[108] For this study, we define pain response as a 30% difference from baseline in the BPI Severity score and functional response as a 30% difference in the BPI Interference score. Our power to detect a 20% difference in response rate (0.25 vs. 0.45) is 0.86. This analysis will also allow us to calculate numbers-needed-to-treat (NNT) and harm (NNH).

C7b. Primary analyses

We will use an intent-to-treat analysis approach, including all participants in the arm to which they were originally assigned. Preliminary analyses will compare baseline characteristics and potential confounding variables between the two treatment arms. Any imbalance in a measure will lead to additional analyses, *as* described in C7c. Medication use at the end of the trial in each arm will be presented in a descriptive table, including numbers using each drug and mean daily dose. For opioids, doses will be described as morphine-equivalent mg per day; for non-opioids, doses will be described both as the actual daily dose and, as a standardized value to facilitate comparisons between drugs, as percent of the maximum daily dose.[109]

Aim 1: To compare 12-month effects of opioid-intensive and opioid-avoidant prescribing strategies on pain-related function and pain intensity.

The primary outcome is BPI Interference score measured at 12 months. Consistent with recommendations for pain clinical trials, we will assess group differences on additional core pain-related domains as described in Section C6b. Preliminary analyses will use intent-to-treat analyses to compare mean scores on primary and secondary pain measures between groups at 12 months (with last observation carried forward for missing data) and at each time point. These will be based on analysis of variance (ANCOVA), which controls for the baseline score as a covariate. For analyses of the primary outcome, all repeated measurements of BPI Interference score will be fitted in a mixed model for repeated measures (MMRM)[111] as a function of the group assignment, while controlling for time points and baseline values of the outcome as fixed effects, with patients as random effects. Between-group differences at month 12 will be estimated and tested using an appropriate contrast as the primary test of intervention effect. BPI Severity score and other pain-related outcomes will be similarly analyzed, using appropriate simple two-group comparisons at month 12 as preliminary analysis. Where the secondary outcome variable has a non-normal distribution, an appropriate link function will be chosen for the

outcome in the mixed model, for a generalized MMRM. In particular, we will compare response rates for pain-related function and pain intensity between arms at 12 months using chi-square tests as preliminary analyses, followed by a generalized MMRM with a logit link for the binomial outcomes. If our findings are robust, we expect findings on the each of the secondary pain measures to be consistent with findings on the primary outcomes and the preliminary comparisons to agree with results from the mixed models.

C7c. Handling missing data and potential confounding

Although we expect a low rate of missing data, especially in the primary outcomes, reasons for missing data (dropouts and missing observations) will be documented and reported, and the observed outcomes at time points prior to dropout will be compared between groups. As long as the missingness depends only on the observed data but not the unobserved missing data, the missing at random assumption in the MMRM is met. Imputation of missing primary outcomes based on last-observation carried forward is only used for the preliminary analyses. To evaluate the robustness of our primary findings based on the mixed models, we will perform sensitivity analyses using different imputation strategies for missing data, including last observation carried forward for all outcomes and predicted values for the primary outcomes from multiple regression on observed data. We will not have enough missing data to fit complex models that assume certain missing data models, especially when the assumptions cannot be checked.

If potential confounding variables (e.g., prior failed analgesic trials, prior pain treatments, treatment expectations, co-interventions) are not balanced between groups, we will conduct additional analyses to evaluate whether findings may be due at least in part to these imbalances. We will adjust for potential confounders by adding them as covariates in models. Second, we will evaluate potential effects of interactions between confounders and treatment group by adding relevant interaction terms to the models.

Treatment groups may differ in the mean number of care manager contacts during the course of the trial, but we expect the number of contacts will overlap substantially between groups. We will conduct sensitivity analyses to examine effects of care manager contacts; specifically, the number of contacts will be added to models as a covariate. If between-group differences in outcomes are eliminated or attenuated after this adjustment, it would suggest that intervention effects are, at least in part, explained by non-specific care manager effects.

C7d. Secondary Analyses

Aim 2: To compare harms of opioid-intensive and opioid-avoidant prescribing strategies over 12 months.

Analyses of harms will be conducted to better understand the risk of harms between treatment groups and, secondarily, as associated with the specific drug or opioid dose received. We will assess potential medication-related harms in three domains: 1) patient-reported adverse symptoms; 2) adverse events; and 3) adverse effects on physical and cognitive performance. Given the limited available evidence for long-term opioid safety, we aim to assess the harms domain as comprehensively as possible; therefore, our assessment of potential adverse effects is broad and exploratory at the risk of finding spurious associations. No adjustment is made for multiple testing because we want our analyses to be sensitive to any potential harm signals. Findings from these analyses will be reported cautiously as needing prospective evaluation in future research.

Patient-reported adverse symptoms: The primary patient-reported symptom outcome is the number of symptoms (range 0-15) reported on the Symptom Checklist. Using repeated measures with four follow-up time points and assuming SD=4 and r=0.5, our power is 87% to detect a between-group difference of 1.0 symptom (from 2.5 to 3.5).[87] As a secondary outcome, we will examine the number of symptoms causing "a lot" of bother (range 0-15). Between-group differences will be tested using MMRM. If the distribution is non-normal, an appropriate link function will be used. We will secondarily explore the risk of events as functions of drug class and, in the subset of those receiving opioids, opioid dose (in daily morphine equivalents); these factors will be added to the models as time varying covariates within 60-day exposure windows preceding each follow-up time point. For this analysis, medications received by the participant will be included as exposures regardless of their source. The same approach (generalized MMRM to test between-group differences followed by the addition of time varying drug class and opioid dose to test their effects on the outcome) will be used to analyze the other harm outcomes using appropriate link functions for their respective distributions.

Power estimates for secondary patient-reported adverse symptom measures are presented in Table 4. In general, 115 subjects per group will yield 80% power to detect between-group differences >0.375 SD. If the 12-month outcome is correlated with its baseline measure with r=0.5, the detectable difference drops to 0.28 SD.

Table 4: Power Estimates for Secondary Patient-Reported Harm Measures

Measure	**Estimated power***	**Clinical context**
Multidimensional Fatigue Inventory	81% to detect 6 points (60 vs. 66, SD=16)	MCID in rheumatoid arthritis = 16.6 points.[112]
Athens Insomnia Scale	87% to detect 2 points (5 vs. 7, SD=4.9)	Difference between persons with and without insomnia = 8.8 points[113]
Headache Impact Test	88% to detect 2.5 points (50 vs. 50.5, SD=6)	Difference between mild and moderate headache severity = 7 points[90]
Arizona Sexual Experience	88% to detect 2 points (11 vs. 13, SD=4.8)	Difference between male patients and controls = 6.3[91]

*Estimated power to detect a between-group difference, assuming n=115 per group and 2-sided alpha=0.05.

Adverse events: The adverse event outcomes are 1) falls and 2) analgesic-related hospitalization or emergency department (ED) visit. Because the risk of misclassification exists with all methods of adverse event causality assessment,[114] we will examine all-cause events in sensitivity analyses. Adverse events will be evaluated in 30-day intervals during the study period. We will test the between-group differences using generalized MMRM as described above, with a logit link for the binary outcomes and a log link for the number of events. The number of expected events is difficult to estimate with precision. A longitudinal study estimated annual fall rates in a relatively healthy and affluent community population to be 21% for middle-aged (46-65 years) and 35% for older (>65 years) adults;[115] whereas a study of predominantly middle-aged fibromyalgia patients found a fall rate of 41%.[98] Assuming the event occurs in 20% of patients in the group with the lower event rate, we have 80% power to detect a 15% increase in proportion of patients with events in the other group using two-sided tests at 5% significance. A recent study found 12-month ED visit rates of 24-28% among patients receiving

opioids[116] and another reported adverse effect-related hospitalization rates of 100-105 per 1000 for patients on non-opioids and 155 for those on opioids.[117] Assuming 15% of patients in the opioid-avoidant group have a hospitalization or ED visit, we have >80% power to detect a 15% increase in proportion of patients with events in the opioid-intensive group. Using data from multiple intervals and basing analyses on event counts (instead of presence/absence) will allow us to detect smaller effect sizes.

Physical and cognitive performance: The primary outcome in this domain is the Fullerton advanced balance scale total score. A prior study of fibromyalgia patients found that the Fullerton scale differentiated between fallers (mean=29.8, SD=7.1) and non-fallers (mean=33.1, SD=5.5). Using two follow-up time points and assuming SD=7 and r=0.5, we have 86% power to detect a between-group difference of 2.0 (from 30 to 32).[98]

Aim 3: To compare effects of opioid-intensive and opioid-avoidant prescribing strategies on secondary outcomes, including health-related quality of life, pain sensitivity, and aberrant drug-related behaviors.

Health related quality of life: The approach described above (generalized MMRM to test between-group differences followed by the addition of time varying drug class and opioid dose to test their effects on the outcome) will be used to analyze SF-12 data. We estimate 81% power to detect a 3-point difference in the Physical Component Score (PCS-12).[118]

Pain sensitivity: Pain threshold and tolerance at 12 months will be compared between study arms using linear models, with outcomes transformed to normality if necessary. Covariates will then be added to determine whether pain sensitivity after 12 months of treatment is affected by the drug class and opioid dose, controlling for baseline pain sensitivity. We will also use generalized linear models to test whether pain sensitivity at baseline predicts the number and distribution of patient-reported pain symptoms or headache severity.

Aberrant drug-related behavior: Data from patient, clinician, and chart-review sources will be categorized according to nature and severity into the following 3 groups: 1) serious ADRB, meaning documentation of prescription drug diversion, buying prescription drug from illicit sources, or simultaneously obtaining controlled substances for the same condition from multiple prescribers; 2) minor ADRB, including behaviors other than those included under serious ADRB; and 3) substance-related ADRB, including documented alcohol disorders or illicit drug use, health or legal consequences of alcohol or illicit drug use, or any urine drug screen positive for an illicit drug or non-prescribed controlled medication (e.g., opioid or benzodiazepine). Using ordinal logistic regression, we will compare rates of misuse between arms; secondarily, we will examine predictors of misuse.