American Pain Society
RESEARCH
EDUCATION
TREATMENT
ADVOCACY
PUBLISHED BY ELSEVIER

The Journal of Pain, Vol 15, No 12 (December), 2014: pp 1360-1365
Available online at www.jpain.org and www.sciencedirect.com

Using Multiple Daily Pain Ratings to Improve Reliability and Assay Sensitivity: How Many Is Enough?

Alicia Heapy,[*,†] James Dziura,[†] Eugenia Buta,[†] Joseph Goulet,[*,†] Joseph F. Kulas,[*,†] and Robert D. Kerns[*,†]

[*]*Veterans Affairs Connecticut Healthcare System, West Haven, Connecticut.*
[†]*Yale School of Medicine, New Haven, Connecticut.*

Abstract: **The Initiative for Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) has reported diminished assay sensitivity in pain treatment trials and recommended investigation of the causes. Specific recommendations included examination of outcome measure reliability and lengthening the baseline measurement period to allow more measurements to be collected. This secondary data analysis evaluated the minimum number of daily pain intensity ratings required to obtain a reliability of at least .90 and whether a composite of this smaller number of ratings was interchangeable with the composite of all ratings. Veterans Affairs medical center patients made 14 daily calls to an automated telephone system to report their average daily pain intensity rating. A single daily rating produced less than adequate reliability (intraclass correlation coefficient = .65), but a composite of the average of 5 ratings resulted in reliability above .90. A Bland-Altman analysis revealed that the differences between a 5-day composite and the composite of all ratings were small (mean .09 points, standard deviation = .45; 95% confidence interval = −.05 to .23) and below the threshold for a clinically meaningful difference, indicating that the 2 measurements are interchangeable. Our results support the IMMPACT recommendations for improving assay sensitivity by collecting a multiple-day baseline of pain intensity ratings.**

Perspective: ***This study examined the minimum number of pain ratings required to achieve reliability of .90 and examined whether this smaller subset of ratings could be used interchangeably with a composite of all available ratings. Attention to measure reliability could enhance the assay sensitivity, power, and statistical precision of pain treatment trials.***

Published by Elsevier Inc. on behalf of the American Pain Society

Key words: *Interactive voice response, assay sensitivity, chronic pain, self-report assessment, reliability.*

The measurement of pain intensity is complicated by its very nature as a "personal, subjective experience influenced by cultural learning, the meaning of the situation, attention, and other psychological variables."[11] Because pain is subjective and susceptible to the effects of placebo,[20] its reliable and valid assessment in clinical trials is particularly important to ensure that any observed changes are attributable to specific rather than nonspecific effects (eg, participant expectancies) and/or study design. An extensive literature attests to attempts to reliably measure pain intensity (see review by Jensen and Karoly[9]). However, a consensus statement by the Initiative for Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) noted several recent trials where previously efficacious analgesic medications failed to demonstrate superiority to placebo.[3] This raised questions about the *assay sensitivity* of those trials or "the ability of a trial to distinguish an effective from an ineffective treatment."[7] Inadequate assay sensitivity undermines the interpretation of negative findings, particularly against the effects of placebo. In response, IMMPACT identified potential ways to improve assay sensitivity including examining outcome measure reliability and lengthening the baseline assessment period to increase reliability.

Received May 8, 2014; Revised September 9, 2014; Accepted September 19, 2014.
This material is based on work supported by Jansen Pharmaceutical. A.H., J.G., and R.D.K. were supported by the Veterans Health Administration Health Services Research and Development Service Center of Innovation (CIN 13-407).
The authors have no conflicts of interest to declare.
The project was not registered with clinicaltrials.gov since it began in 2003 prior to the requirement to register.
The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the U.S. Government.
Address reprint requests to Alicia Heapy, PhD, VA Connecticut Healthcare System (11ACLGS), 950 Campbell Avenue, West Haven, CT 06516. E-mail: alicia.heapy@va.gov
1526-5900/$36.00
Published by Elsevier Inc. on behalf of the American Pain Society
http://dx.doi.org/10.1016/j.jpain.2014.09.012

Improved reliability reduces the occurrence of statistical regression to the mean and reduces the possibility that it will be interpreted as a treatment effect.[14] Reliability also sets an upper bound for the power, effect size, and statistical precision of a trial.[13] In a simulation, Perkins and colleagues demonstrated that improving outcome reliability from .70 to .90 would result in a 22% decrease in required sample size and an increase in study power from .64 to .72,[16] thereby reducing study costs and increasing the likelihood of detecting treatment-related improvement. Classical test theory dictates that reliability will be improved by increasing the total number of measurements obtained or the number of items in a measure.[15]

Despite the benefits of improved reliability, there is little empirical guidance for determining how many pain intensity measurements are necessary to reach a reliable baseline and at what point additional measurements do not yield improvements in reliability but simply add to participant burden and trial costs. IMMPACT recommendations of 1 week of daily baseline measurements (and 2 weeks for longer studies) were based on expert opinion, and further empirical examinations were encouraged.

In 1 early study, Kerns and colleagues examined the stability of hourly pain intensity ratings over a 2-week period in 98 individuals with chronic pain.[12] Participants reported high levels of stability between pain intensity ratings collected during week 1 versus week 2 (r = .93) and variability in pain intensity (r = .92). Jensen and McFarland[10] found that single ratings were less reliable than composite scores in a sample of 200 participants who completed hourly pain ratings for 7 days. Composite scores containing ratings across multiple days were required to achieve validity and stability above r = .90. Prior work has been criticized, however, for examining reliability using bivariate correlations between a subset of pain ratings and the grand mean of all available ratings.[2] An association between 2 measures is not equivalent to demonstrating that the measures agree or can be used interchangeably, which is the presumed end goal of using a composite measure.

Given concerns about assay sensitivity in pain treatment trials, our aims were to 1) evaluate the reliability of a single average daily pain rating among patients with long-term chronic pain, 2) determine the minimum number of daily average pain ratings that were required to obtain a composite measure with reliability of .90 or higher, and 3) assess whether this composite of fewer than all of the daily pain ratings was interchangeable with the mean of all available ratings.

Method

Procedure Overview

This study was a secondary analysis of data collected from a 13-week randomized controlled open-label clinical trial examining the relative efficacy of transdermal fentanyl (TDF) compared to oral short-acting opioids for the treatment of chronic noncancer pain. The parent study[6] was designed to test the hypothesis that a long-acting medication, TDF, because it is associated with more stable and predictable levels of pain relief and improved sleep, would be associated with increased activity, reduced perceived disability, and improved sleep, functioning, and overall quality of life relative to short-acting opioids. In the parent study, participants were asked to make daily reports of pain intensity and other pain-related outcomes via automated telephone calls enabled by an interactive voice response (IVR) system. IVR is a computerized interface that allows participants to provide responses to prerecorded questions using their telephone's numeric keypad. During 3 prespecified 2-week data collection periods, trial participants called the IVR system daily and answered 18 automated questions about their pain, its effect on their physical and emotional functioning, and their adherence to prescribed pain medications.[5] The analysis presented here focuses solely on data obtained from the pain intensity question ("On a scale from 0 to 10, with 0 representing no pain and 10 representing the worst pain imaginable, rate your average pain today.") obtained during the first 2-week period and prior to randomization and any associated medication changes. The project was approved by the Veterans Affairs Connecticut Healthcare System's Human Studies Subcommittee and the Yale School of Medicine Human Investigation Committee. Participants were paid $5 for each completed daily call.

Participants

Participants were veterans receiving care at Veterans Affairs Connecticut Healthcare System who were recruited through advertisements, referrals from primary care providers, and direct invitation to patients seen in a multidisciplinary pain management clinic. Eligibility criteria included 1) presence of pain with average pain intensity of 4 or greater on a 0 (no pain) to 10 (worst pain imaginable) numeric rating scale, 2) daily use of an oral short-acting opioid equivalent of at least 60 mg oral morphine per day for at least 6 consecutive months prior to study enrollment, 3) age 18 or older, 4) no medical contraindications to TDF therapy, and 5) consent to urine drug screen to evaluate potential substance misuse. Persons with evidence of active alcohol or substance abuse or dependence, active psychotic disorder, active suicidal or homicidal risk, back surgery within the past 6 months, pregnancy, or lactation were excluded.

Data Analysis

We analyzed 14 daily baseline pain ratings of study participants. We excluded subjects with less than 7 ratings during the 14-day period. Daily baseline pain ratings were collected prior to randomization and any randomization-related medication change from oral opioid to TDF. Because our aim was to examine the number of consecutive daily ratings needed to achieve .90 reliability, we used only those ratings that were made