Measurement properties of quality-adjusted life year (QALY) measures among older adults with chronic neck pain

A Thesis
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Brent David Leininger, DC

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Adviser: Gert Brønfort, DC, PhD

April 2016

Acknowledgements

I would like to acknowledge and thank a number of individuals who made this work possible. First and foremost, I wish to express my gratitude to Gert Bronfort, my primary advisor, for his thoughtful and considerate counsel and mentorship throughout my young research career. Roni Evans, who in addition to being a great mentor, is always there to make sure the nitty gritty work required to perform high quality research isn't ignored. I am forever grateful to Gert and Roni, for their support of my work and commitment to my professional development. In addition, I'm grateful to John Nyman and Todd Rockwood for serving on my thesis committee and providing thoughtful feedback which has substantially improved the thesis. Linda Hanson and Corrie Vihstadt, my research fellow compadres, who've been there to share the every-day ups and downs associated with Graduate School. I'd also like to thank the research team responsible for performing the clinical trial which this work is based on. Your attention to detail and commitment to excellence has not gone unnoticed. I'd also like to acknowledge NIH's National Center for Complementary and Integrative Health (#R25AT003582 & F32AT007507) and the NCMIC foundation for providing financial support towards my degree. Finally, I'd like to thank my family, who has graciously supported me as I worked a number of late nights and weekends to complete this degree.

Dedication

This thesis is dedicated to my caring wife Bettina and our three wonderful children, Carter, Preston, and Bailey.

Abstract

Background

Quality-adjusted life year (QALY) measures are an important outcome for assessing the cost-effectiveness of healthcare interventions. Ideally, the choice of QALY measures will be informed by the measurement properties within the population of interest. Currently, the EQ-5D and SF-6D are the most commonly used QALY measures within cost-effectiveness analyses for spine pain. A number of studies have assessed the measurement properties of QALY measures for individuals with spine pain, but primarily within surgical populations. The psychometric properties of QALY measures may vary substantially within non-surgical populations. The primary aim of this thesis is to assess the psychometric properties (reliability, validity, and responsiveness) of commonly used QALY measures (SF-6D, EQ-5D, EQ Visual Analog Scale) among older U.S. adults with chronic mechanical neck pain managed non-surgically. The secondary aim of the thesis is to assess differences in the psychometric properties of QALY measures derived from the same instrument (SF-6D), but using different valuation methods.

Methods

Data for the study was collected within a randomized clinical trial comparing different combinations of non-invasive interventions (home exercise and advice, supervised exercise therapy, spinal manipulation) for the management of chronic neck pain in older adults. Quality-adjusted life years (QALYs) were measured with the 1) SF-6D, 2) EQ-5D, and 3) Euroqol visual analogue scale (EQ VAS) using U.S. population values for the primary aim. Test-retest reliability was determined using intraclass correlation coefficients (ICCs). The Bland-Altman method for limits of agreement and the smallest detectable change (SDC) were used to assess agreement. The longitudinal known-group validity and responsiveness of QALY measures was estimated using four external criteria: 1) global perceived change in health; 2) global improvement in neck symptoms; 3) neck pain; and 4) neck disability. Known-group validity was assessed by calculating mean QALY changes for each category of global perceived change in health and neck symptoms in addition to quintiles of neck pain and disability improvement. The relative responsiveness of QALY changes was estimated using correlation and area under the receiver operating characteristic (ROC) curve analyses.

Results

The SF-6D demonstrated better test-retest reliability (ICC = 0.81; 95% CI 0.77 to 0.85) relative to the EQ-5D (ICC = 0.44; 95% CI 0.33 to 0.53) and EQ VAS (ICC = 0.68; 95% CI 0.61 to 0.75). In addition, the smallest detectable change was lowest for the SF-6D (0.16; 95% CI 0.14 to 0.17), followed by the EQ-5D (0.18; 95% CI 0.16 to 0.20), and EQ VAS (0.22; 95% CI 0.20 to 0.25). Differences in QALYs during the one-week baseline period were evenly spread over the range of mean QALYs for the SF-6D, but not the EQ-5D or EQ VAS. The SF-6D and EQ VAS demonstrated better longitudinal known-group validity relative to the EQ-5D. Mean SF-6D and EQ VAS QALY changes were monotonically decreasing across levels of improvement for three of the four external criteria. All three QALY measures demonstrated similar responsiveness to change.

Correlations between QALY measures and three of the external criteria were similar and very low to low in strength (-0.233 to -0.391). Correlations with neck disability were low to moderate in strength with the SF-6D demonstrating the strongest association (-0.596; p-values for differences with EQ-5D and EQ VAS = 0.01). There were no significant differences among the QALY measures when measuring responsiveness with area under the ROC curve. SF-6D based QALY measures had similar reliability, agreement, validity, and responsiveness.

Conclusions

There were minor differences between U.S. QALY measures in terms of responsiveness; however, the SF-6D was more reliable and demonstrated less measurement error relative to the EQ-5D and EQ VAS, in addition to better known-group validity relative to the EQ-5D. The different methods for obtaining QALY values from the same instrument (SF-6D) had little to no impact on the psychometric properties.

Table of Contents

List of Tables

List of Figures

Background

Cost-effectiveness analyses (CEAs) value healthcare interventions not only by their clinical effectiveness, but also by their costs. The standard outcome for CEAs is the incremental cost-effectiveness ratio (ICER), which describes the additional costs required to improve health outcomes by one unit (e.g. costs per additional year of life, cost per heart attack avoided) for interventions that are both more effective and more costly. A growing number of organizations around the world are using CEAs to inform where to most efficiently use healthcare resources. One of the best known examples comes from England, where the National Institute for Health and Care Excellence (NICE) uses evidence from CEAs to inform coverage decisions within England's National Health Service.[1] Although U.S. healthcare agencies such as Medicare avoid the systematic use of CEAs for policy, there are examples of coverage decisions that have been informed by CEAs. [2] Furthermore, the recent public healthcare debate has sparked increased interest in CEAs within the broader U.S. healthcare community.[3]

The **quality adjusted life year (QALY)** is the most popular health outcome for assessing effectiveness within CEAs. Both the U.S. Panel on Cost Effectiveness in Health and Medicine[4] and England's National Institute for Health and Care Excellence[5] recommend using QALYs as the primary outcome within CEAs. QALYs combine morbidity and mortality into a single measure by multiplying the amount of time spent in a particular health state by the health state's value or impact on quality of life. An important property of QALYs is their ability to capture changes in morbidity and mortality across a number of diseases, which allows decision makers to assess the relative value of healthcare interventions across a variety of conditions.

While there is little debate about using QALYs as the primary effectiveness outcome for CEAs, multiple approaches for deriving QALYs exist. A number of different health state classification instrument have been created to measure QALYs. The EQ-5D [6], Health Utilities Index 2 or 3(HUI)[7], SF-6D (derived from the SF-36)[8], 15D [9], Assessment of Quality of Life (AQoL) [10], and the Quality of Wellbeing (QWB) instrument [11] are all

examples of measures developed for QALY estimation. The most commonly used instruments are the EQ-5D, SF-6D, and the HUI.[12] A 2006 systematic review assessing the use of QALY measures within CEAs found the EQ-5D was used in almost half of all studies.[13] Although the SF-6D is a relatively newer measure, it has grown in popularity since it was first developed in 1998, likely because it is derived from a commonly used quality of life measure (i.e. SF-36).[8] The EQ-5D, SF-6D and HUI vary considerably in length and the subsequent number of possible health states. There are 243 unique health states possible within the EQ-5D, compared to 18,000 within the SF-6D, and 972,000 within the HUI3. In theory, QALYs derived from one instrument should measure the same construct as QALYs derived from other instruments. However, a number of studies comparing QALY measures among common patient populations have found important differences between instruments.[14-16]

Another source of variability in QALY estimation is the method used to value possible health states within each instrument. [17] QALY valuation requires the elicitation of "preferences" for each health state on a 0 to 1 scale where 0 represents death and 1 represents perfect health. A number of preference elicitation techniques exist including time trade-off (TTO), standard gamble (SG), ordinal ranking (ORD), discrete choice experiments (DCE) and visual analog rating scales (VAS). A 2010 meta-analysis of studies assessing QALY values when using different elicitation techniques found no difference between the TTO and VAS methods, but noted SG methods result in QALY values that are approximately 0.2 standard deviations higher than either TTO or VAS. [18]

After selecting an instrument and elicitation technique for QALY valuation, the next question is what population should be used to value the health states. Patients, physicians, and the general population are all possible choices. Although patients and physicians have the most direct knowledge of health conditions and their impact, the general consensus among health economists is that the general public should be used to value health states. [12,19] There are a number of arguments for this approach. First, the general public is the most representative of all possible conditions and health states and would

not be biased towards a particular health state. There is also evidence that patients adapt to their condition and value their health state higher than non-patients.[20]

Researchers conducting cost-effectiveness analyses must choose the most appropriate QALY measure for their specific population. Ideally, this decision will be informed by the measurement properties of the QALY measures within the population of interest. Spinal pain, including neck and low back pain, is the most common form of chronic pain, the leading cause of disability, and accounts for 9% of total healthcare expenditures in the U.S.[21-23] The management of spinal pain in the U.S. has been heavily criticized due to an overutilization of expensive interventions with questionable effectiveness. [24] Accordingly, there is a growing need to assess the value of healthcare interventions for managing spinal pain. Given the importance of QALYs within CEAs, a better understanding of their psychometric properties among individuals with spinal pain is much needed. Currently, the EQ-5D and SF-6D are the most commonly used QALY measures within CEAs for spine pain.[25] A number of studies have assessed the reliability, validity, and responsiveness of QALY measures within adults with spinal pain, but most of the existing studies have been conducted within surgical populations. [14-16,26-35] The psychometric properties of QALY measures may be substantially different within non-surgical populations.

Purpose of thesis

The purpose of this thesis is two-fold. The primary aim is to estimate the psychometric properties of commonly used health state classification systems valued by U.S. populations when applied to older adults with chronic neck pain. The secondary aim is to estimate the psychometric properties of QALY measures derived from the same health state classification instrument (i.e. SF-6D) using different preference elicitation methods, populations, and modeling techniques. The following methods will be used to estimate the psychometric properties of QALY measures:

- Test-retest reliability and agreement using intraclass correlation coefficients (ICC) and Bland-Altman's level of agreement, respectively
- Construct validity using longitudinal known-group validity
- Responsiveness using correlation and area under the receiver operating characteristic (ROC) curve analyses

An overview of the methods is provided in figure 1. The results will be compared to previous studies assessing the psychometric properties of QALY measures within spinal pain populations.

Methods

Data for the study was collected within a randomized clinical trial conducted at a University-based outpatient research clinic in the Minneapolis, MN metropolitan region from 2004 to 2008. The clinical trial compared different combinations of non-invasive interventions (home exercise and advice, supervised exercise therapy, spinal manipulation) for the management of chronic neck pain in older adults. Details on the methodology and primary results from the clinical trial have previously been published.[36,37]

Population

Participants were recruited from the general public within the greater Minneapolis, MN metropolitan area using postcard mailings, newspaper advertisements, and community outreach. Participants were community dwelling men and women 65 years or older with a primary complaint of chronic (>3 months duration) mechanical neck pain. Exclusion criteria included a pain rating less than 3 (on a 0-10 scale), contraindications to study interventions, severe disabling health problems, substance abuse, litigation for neck pain, or ongoing non-pharmacological healthcare for neck pain.

Outcome assessments

Self-reported clinical outcomes were collected twice at baseline (one week apart) and at 4, 12, 26 and 52 weeks post-randomization using self-administered paper questionnaires. Participants completed the baseline, 4, and 12 week self-administered questionnaires at the University-based research clinic without the influence of study investigators or research staff. The 26 and 52 week questionnaires were completed by mail.

Quality-adjusted life year measures

Primary Aim

SF-6D

Health related quality of life was measured using the Medical Outcomes Study Short Form 36-item Health Survey (SF-36).[38] Quality-adjusted life years (QALYs) were estimated using the SF-6D which is derived from 11 of the items recorded within the SF-36.[8,39] The SF-6D includes six dimensions (each with 4-6 levels) to describe health: physical functioning, role limitations, social functioning, pain, mental health, and vitality. Higher scores within each domain indicate higher morbidity. SF-6D health states range from no problem in any of the six dimensions (perfect health = 111111) to the worst case for each dimension (health state = 645655, also known as "the pits"). Preferences for SF-6D health states were taken from a study which used discrete choice experiments within a U.S. population.[40] The nationwide sample included 666 participants. The sample was under representative of the US population in terms of 18 to 44 year olds, males, Hispanics, and African-Americans. Participants were asked to choose between a pair of health states that only varied in terms of two health domains (e.g. greater physical function and less social function vs. less physical function and greater social function). By limiting the difference in health states to a trade-off between two domains, discrete choice experiments reduce the complexity of valuing health states. A stacked probit model using maximum likelihood estimation to account for repeated comparisons from each participant was used to estimate US preferences for SF-6D health states. Final health state values ranged from 0.0129 "for the pits" to 1 "perfect health".

EQ-5D

Quality adjusted life years were additionally valued using the Euroqol EQ-5D. The EQ-5D measures health on five dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) over 3 levels (no problem, moderate problem, extreme problem) which results in 243 potential health states.[6,41] EQ-5D-derived QALYs were calculated using preferences for health states elicited from a sample of the US population using time trade-off methods.[42] Hispanics and non-Hispanic blacks were oversampled for the valuation survey which used US census demographic data to construct an otherwise representative sample frame. Participants were asked to choose between 10 years of living in one of the EQ-5D health states or living in perfect health for a variable amount of time (0-10 years) followed by immediate death. The amount of time in perfect health was varied until the participant was indifferent between this and 10 years of life in the presented health state. The participant's value for the health state was equal to the amount of time in perfect health they were willing to trade divided by 10 years. The final model included interaction terms for health states with more than one moderate or extreme problem. Health states values ranged from 1 (i.e. perfect health) to -0.109 (i.e. extreme problems in all 5 health domains).

EQ visual analogue scale (EQ VAS)

Participants also evaluated their current health state using a visual analogue scale included within the EQ-5D (EQ VAS).[41] The EQ VAS scale ranges from 0 "worst imaginable health state" to 100 "best imaginable health state". The EQ VAS measure was used to determine the study participant's direct estimate of QALYs. Other QALY measures use preferences for health states from the general population to value the participant's QALYs.

Secondary Aim

In addition, to the three QALY measures valued by U.S. populations, we also compared the psychometric properties of SF-6D-derived QALYs using different preference elicitation methods, populations, and modeling techniques. Preferences for 18,000

possible SF-6D health states have been collected from U.S. and U.K. populations using the following methods:

SF-6D U.K. standard gamble (SG)

The original description of the SF-6D and subsequent preferences for health states were obtained from a U.K. population using standard gamble methods.[39] The final model used data from 611 representative adults from the U.K. population in terms of age, education, and social class. Each participant valued 6 of the 249 the SF-6D health states chosen for valuation. The authors designed the study so each of the 249 possible health states was valued by a similar number of participants. During the standard gamble procedure, the participant was asked to choose between living in one of the SF-6D health states or taking a gamble between living in perfect health (the best possible SF-6D health state) and "the pits" (the worst possible SF-6D health state). The probability of the gamble between perfect health and the worst SF-6D health state was varied until the participant was indifferent between taking the gamble and living in the presented health state. The final model for health state preferences included an interaction for individuals valuing more than one of the six dimensions at the worst possible level. Health state values ranged from 0.301 for "the pits" to 1 for perfect health.

SF-6D U.K. SG bayesian model (BYS)

An alternative value set from this population has been published using more flexible nonparametric Bayesian modelling methods.[43] The parametric model used for the initial SF-6D valuation assumed an additive model with one interaction term for individuals with the worst possible response in one of the health dimensions. The non-parametric Bayesian model does not impose any assumptions on possible interactions between health domains. The authors used an additive model (i.e. no interactions between health states) as a prior distribution, but Bayesian methods allowed the data to modify and replace prior distributions, if indicated. The nonparametric model exhibited less non-monotonicity (worse health states receiving higher utility scores) than the parametric

model (20% vs 10% of randomly selected health states). Health state values ranged from 0.203 for "the pits" to 1 for perfect health.

SF-6D U.K. ordinal ranking (ORD)

In addition to the standard gamble method used to determine UK preferences for SF-6D health states, a value set using ordinal ranking of health states has been published.[44] The 611 participants from the original UK SF-6D study also completed a series of ranking exercises prior to completing the standard gamble valuation. Each participant ranked 8 health states, including the best possible state, worst possible state "the pits", and death. Conditional logistic regression was used to model preferences for health states using the rank data. Similar to the Bayesian model, there were fewer issues of non-monotonicity within each health domain when using the rank model compared to the standard gamble model. Health state values ranged from 0.179 for "the pits" to 1 for perfect health.

SF-6D U.S. discrete choice experiments (DCE)

Preferences for SF-6D health states have also been elicited from a U.S. population using discrete choice experiments.[40] A description of this value set was provided under the methods for SF-6D valuation.

QALY measurement

QALY measures are designed to incorporate both morbidity and mortality by accounting for the amount of time spent in different health states. A QALY represents one full year in perfect health. QALY values of 0.5 could be attained by living for ½ a year in full health (0.5 years x 1(value of health state)) or by living for a full year in a health state valued at 0.5 (1 year x 0.5 (value of health state)) (see Figure 2). When assessing construct validity and responsiveness, we used an area under the curve approach to measure change in QALYs over the one year trial duration. The area under the curve approach uses the trapezoid method for calculating change in QALYs from baseline and is recommended by a number of health economic texts ($\sum_{i=1}^{k}((QALY_i + QALY_{i+1})/$

$2)(t_{i+1} - t_i)$ where $QALY_i$ is the cross-sectional QALY estimate at each time point and $t_i$is time period for each of the measurements valued in years).[12,45] For the reliability and agreement analyses, we used the cross-sectional estimates for each baseline QALY assessment.

Analysis

Statistical analyses were performed using Stata version 13.0 unless noted otherwise.

Test-retest reliability

The test-retest reliability of the QALY measures was determined using the two baseline administrations of the QALY measures which were performed one week apart. An intraclass correlation coefficient (ICC) using a two-way random effects model for absolute agreement was used to determine the test-retest reliability (STATA procedure *ICC*).The two-way random effect model ICC is the recommended statistic for measuring test-retest reliability as it accounts for systematic error.[46] The ICC for agreement is a ratio between 0 and 1 which divides the between person variability (σ^2_{bp}) by the total variability due to: 1) between-person variability (σ^2_{bp}), 2) within-person variability during the test-retest (σ^2_{wp}), and 3) the residual variability (σ^2_r) = [(σ^2_{bp})/ (σ^2_{bp} + σ^2_{wp} + σ^2_r)].[47] An ICC near 1 indicates the amount of measurement error due to test-retest and residual variability is low compared to the variability between persons. An ICC of 0.70 has been recommended as a minimum threshold for acceptable reliability.[46] Differences between ICC's $\geq$ 0.10 were considered noteworthy and tested for statistical significance using methods described by Ramasundarahettige.[48]

Agreement

In addition, Bland-Altman methods for limits of agreement using the STATA procedure *batplot* were performed to determine the absolute mean difference and variability between the two baseline administrations of the QALY measures.[49] Agreement analyses report the magnitude of differences between repeated assessments on their original scale, which is often more meaningful to clinicians. The 95% limits of agreement were defined

as the mean difference ± smallest detectable change (i.e. 1.96*the standard deviation of the difference between baseline QALY administrations).[47,49] Changes within the limits of agreement or less than the smallest detectable change may be due to measurement error. Differences between measures were noted when there was no overlap of 95% confidence intervals.

Sensitivity analysis: reliability and agreement

Patients completing the two baseline administrations of QALY measures should be "stable" with regards to the construct of QALYs, for a valid assessment of reliability. The time between baseline QALY survey administrations was limited (one week) and quality-adjusted life years are a relatively stable construct which decreases the likelihood of significant health changes within the test-retest period. To assess the potential impact of "unstable patients" on test-retest reliability, a sensitivity analysis dropping participants with ≥50% change in neck pain or disability was performed and compared to the main analyses for reliability and limits of agreement.

Known-group validity

Known-group validity is a form of construct validity which assesses how the measured instrument differs between groups of individuals with different health outcomes.[50] We assessed the longitudinal known-group validity of QALY measures against four external criteria.

External criteria

1) Global perceived change in health was determined using a question from the SF-36 which asks participants to rate their current health in general compared to one year ago on a 5-point scale [i.e. much better, somewhat better, about the same, somewhat worse, much worse]. [38]

2) Global improvement in neck symptoms was determined by asking participant's "Overall, how much has your neck pain changed since starting treatment in this study?" using a 9-point scale [i.e. no symptoms (100% improvement), much better (75%

improvement), somewhat better (50% improvement), a little better (25% improvement), no change (0% improvement), a little worse (25% worse), somewhat worse (50% worse), much worse (75% worse), twice as bad (100% worse)].[51] The response categories for a worsening in neck pain (i.e. a little worse, somewhat worse, much worse, and twice as bad) were combined as a limited number of participants (n=12) reported a worsening in neck pain.

3) The participant's typical level of neck pain during the past week was measured using an 11-box numerical rating scale (NRS) (0=no pain; 10 = worst possible pain). The NRS has been shown to be a reliable and valid outcome measure for older adults with persistent pain. [52]

4) Neck disability was measured using the Neck Disability Index (NDI), an instrument previously shown to be reliable and valid for adults with neck pain.[53,54] The NDI measures disability using 10 domains (pain intensity, personal care, lifting, reading, headaches, concentration, work, driving, sleeping, and recreation) over 6 levels. A total NDI score was obtained by summing the responses from individual domains and dividing by the maximum possible score, resulting in a 0-100 scale, where 0 equals no disability.[55]

Mean QALY changes for each possible response on global perceived change in health and global perceived change in neck symptoms were calculated along with QALY changes for each quintile of improvement in neck pain and neck disability. We hypothesized that mean QALY changes would monotonically decrease with each decreasing level of improvement indicated by the external criteria.

Responsiveness

The relative responsiveness of one year time-weighted QALY changes was determined using multiple methods. First, the percentage of participants at or near the ceiling and floor of the respective QALY scales (i.e. within 5% or 10%) was determined as this can potentially impact responsiveness. The correlation between changes in QALYs and each of the four external criteria was determined using pearson's correlation coefficient. The

absolute strength of association was characterized as very low (0.0 to 0.3), low (0.3 to 0.5), moderate (0.5 to 0.7), high (0.7 to 0.9), and very high (0.9 to 1).[56] Relative differences in the strength of correlation between QALY measures and external criteria greater than 0.10 were considered noteworthy and tested for statistical significance.[57-59] We hypothesized that QALY changes would be negatively correlated with global perceived change in health, global perceived change in neck symptoms, and reductions in both neck pain and disability.

In addition, area under the receiver operating characteristic (ROC) curves were calculated to determine the ability of QALY changes to identify improved and non-improved participants. Participants were categorized as improved or not improved at the final outcome assessment (week 52) based on 1) global perceived change in health (much better or somewhat better), 2) global perceived change in neck symptoms (no symptoms, much better, or somewhat better), 3) ≥30% improvement in neck pain, and 4) ≥30% improvement in neck disability. A 30% improvement in neck pain and disability has been suggested as a threshold where patients report moderate improvement.[60] Area under the ROC curve indicates the probability that an improved patient will report greater QALY gains than a non-improved patient. An area under the ROC curve of 0.50 would indicate that QALY changes were not able to identify improved and non-improved patients (based on the external criteria) better than random chance. An area of 0.70 has been proposed as satisfactory.[46] Differences in area under the ROC curves were determined using a nonparametric approach specified within the *roccomp* procedure in Stata.[61]

Sensitivity analysis: responsiveness

In addition to measuring change in QALYs using an area under the curve approach, we also assessed the impact of using 1) the cross-sectional QALY estimate at week 52 (i.e. $QALY_{W52}$) and 2) change in QALYs from baseline to week 52 (i.e. $QALY_{W52} - QALY_{W0}$) on responsiveness as a sensitivity analysis. Global change scales have been

found to be more reflective of current health status than change in health, as patients have difficulty recalling prior health states when assessing change.[62]

Results

Baseline characteristics

Baseline characteristics for the 241 older adults with chronic mechanical neck pain who were enrolled in the randomized clinical trial are provided in table 1. Participant age ranged from 65 to 88 with a mean of 72.3. A limited proportion of participants experienced associated arm pain (17%), were currently employed (18.7%), or used tobacco (5.4%). Among the U.S. QALY measures, baseline values were lower for the SF-6D (0.766) and EQ VAS (0.766) compared to EQ-5D (0.795). SF-6D-based QALYs were highest for the US DCE measure (0.766), followed by the UK ORD (0.725), UK SG (0.714) and UK BYS (0.665) measures.

Attrition and completion rates of QALY measures

The number of completed QALY measures at each assessment period is outlined in table 2. Data collection rates for the trial's primary outcome, neck pain, are also presented for reference. The ability to derive QALYs using the SF-6D, EQ-5D, and EQ VAS was similar at baseline and over the duration of the trial. SF-6D QALY scores could not be derived for a small number of participants (n=3) due to missing responses within the SF36 survey. Data collection rates for the SF-6D, EQ-5D, and EQ VAS-derived QALYs were similar to the clinical trial's primary outcome (i.e. pain) when questionnaires were completed within the research clinic, but were slightly lower when returned by mail. Data collection rates at week 52 were above 93% for all measures.

Reliability and agreement

Results for test-retest reliability and agreement are provided in table 3. The Intraclass correlation coefficient (ICC) for the SF-6D was larger (0.81; 95% CI 0.77 to 0.85) than both the EQ-5D (0.44; 95% CI 0.33 to 0.53) and EQ VAS (0.68; 95% CI 0.61 to 0.75)

indicating better test-retest reliability. Significant differences between ICC's favoring the SF-6D over both the EQ-5D (difference = 0.38; 95% CI 0.29 to 0.47) and EQ VAS (difference = 0.13; 95% CI 0.07 to 0.20) were noted. Both the EQ-5D and EQ VAS were below the recommended ICC threshold of 0.70. There was little difference in reliability between the SF-6D based measures (0.76 to 0.81) with all methods resulting in ICCs above the recommended threshold. The baseline distributions of QALYs are displayed in Figures 3-4. Over one-quarter of the participants reported the same health state at baseline (moderate pain or discomfort, but no other problems) when using the EQ-5D, limiting the spread of the distribution relative to the other QALY measures.

The smallest detectable change among U.S. based QALY measures was the lowest for the SF-6D (0.16; 95% CI 0.14 to 0.17), followed by the EQ-5D (0.18; 95% CI 0.16 to 0.20), and EQ VAS (0.22; 95% CI 0.20 to 0.25). The SF-6D SDC confidence intervals did not overlap with the EQ VAS confidence intervals. Among the SF-6D measures, the UK SG (0.12; 95% CI 0.11 to 0.14) and UK BYS (0.12; 95% CI 0.10 to 0.13) measures demonstrated the lowest SDC which were 0.04 QALYs lower than the U.S. based SF-6D measure (DCE) with confidence intervals that did not overlap. Bland-Altman plots showing the differences in test-retest administrations of QALYs on the vertical axis and the mean QALYs on the horizontal axis are displayed in Figures 5-6. Test-retest differences for EQ-5D and EQ VAS-derived QALYs were larger among individuals reporting lower mean QALYs. Differences in test-retest administrations of QALYs were evenly spread over the range of mean QALYs for SF-6D based measures. 47 participants reported changes in neck pain or disability greater than 50% between baseline assessments. Sensitivity analyses removing these 47 participants to ensure the sample population was "stable" resulted in minor changes in the ICC and limits of agreement results (Table 4). The one notable change was the difference in ICC's between the SF-6D (0.81; 95% CI 0.75 to 0.85) and EQ VAS (0.72; 95% CI 0.65 to 0.78) fell below the threshold of 0.10.

Known-group validity

Mean QALY gains according to changes in external criteria are provided in tables 5-8 and figures 7-10 as an assessment of construct validity. QALY gains were monotonically decreasing across levels of global perceived change in health for the SF-6D, but not for the other U.S. based measures. The SF-6D resulted in greater mean QALY improvement relative to EQ-5D or EQ VAS measures among individuals reporting the most improvement in health (i.e. much better). The EQ VAS was the only QALY measure showing QALY decreases for participants reporting somewhat worse health status. QALY changes among SF-6D measures were similar across the different levels of global change in health. The only U.S. based QALY measure with monotonically decreasing QALY changes among individuals reporting changes in global neck symptoms was the EQ VAS. Mean QALY changes among individuals reporting a little improvement to no change in neck symptoms was not easily distinguishable for the U.S. based SF-6D. Additionally, the EQ-5D was not able to distinguish between individuals reporting their neck symptoms were "somewhat better" (mean QALY change = -0.001) to "worse" (mean QALY change = 0.005). In terms of improvement in neck pain and neck disability, the U.S. based SF-6D and EQ VAS measures were monotonically decreasing across quintiles of improvement. The EQ-5D was not able to distinguish between the second and third, or the fourth and fifth quintiles for either measure. Among SF-6D measures, the standard gamble and ordinal methods were not monotonically decreasing for neck pain improvement, but all SF-6D derived measures performed similarly across levels of neck disability improvement.

Responsiveness

Floor/Ceiling effects at baseline

The proportion of participants with baseline QALY values near the maximum and minimum possible score are presented in table 9. Floor effects were not present at baseline as none of the participants reported QALYs within 10% of the lowest possible value for each scale. Ceiling effects were minimal for most QALY measures (<2%),

except for the US SF-6D DCE and EQ VAS measures where 17% and 14% of participants reported QALYs within 10% of the highest possible score, respectively.

Correlation

The correlation between QALY measures and the four external criteria are provided in table 10. Correlations between the U.S. based QALY measures and three of the external criteria (global perceived change in health, global perceived change in neck symptoms, neck pain) were very low to low (-0.233 to -0.391). Among these three external criteria, the only difference ≥ 0.10 was between the SF-6D (-0.36) and EQ VAS (-0.26) for global perceived change in neck symptoms which was not statistically significant (p-value = 0.14). When using neck disability as an external criteria, correlations were low to moderate (-0.442 to -0.596). The SF-6D (-0.596) was more strongly correlated with neck disability than either the EQ-5D (-0.458; p-value for difference = 0.01) or EQ VAS (-0.442; p-value for difference = 0.01) QALY measures. Strengths of correlation among the SF-6D based measures were similar to the U.S. based measures. The ordinal ranking method (-0.426) demonstrated stronger correlation than the Bayesian model (-0.315; p-value for difference < 0.01) when using global perceived change in neck symptoms as an external criteria. In addition, the discrete choice experiment method (-0.596) was more strongly correlated with neck disability compared to the Bayesian model (-0.491; p-value for difference < 0.01).

The results of the sensitivity analyses for correlation are provided in table 11. Sensitivity analyses using change in QALYs from baseline to week 52 (i.e. $QALY_{W52} - QALY_{W0}$) resulted in stronger correlations with global perceived change in health (low correlation; -0.332 to -0.463). Sensitivity analyses using week 52 QALY estimates resulted in stronger correlations with neck disability (low to high correlation; -0.473 to -0.717). Differences in correlation between the SF-6D and EQ VAS with global perceived change in neck symptoms were still present when using change from baseline to week 52, but were below the 0.10 threshold when using week 52 as the outcome. Differences in correlation among the U.S. based measures with neck disability were similar to the primary analysis,

except for the correlations among the SF-6D (-0.569) and EQ-5D (-0.483) when using change from baseline to week 52, which fell below the 0.10 threshold for notable differences. Differences in strength of correlation among the SF-6D measures were similar when using change from baseline to week 52, but no notable differences were present when using week 52 as the outcome.

ROC curves

The predictive ability of QALY measures to distinguish between improved and non-improved participants using area under the ROC curve are provided in table 12 and Figures 11-12. There were no significant differences among the U.S. based QALY measures when measuring responsiveness with area under the ROC curve. Areas under the ROC curve were highest when detecting minimal improvement (≥ 30%) in neck disability (range 0.67 to 0.73) or neck improvement (range 0.67 to 0.72), and were the lowest for detecting minimal improvement (≥ 30%) in neck pain (range 0.60 to 0.64). In terms of meeting a satisfactory discrimination area of 0.70, the EQ-5D reached this threshold for three of the four external criteria, the SF-6D was above this threshold for two criteria, and the EQ VAS was below the threshold for all measured external criteria.

Among the SF-6D based measures, there was a significant difference in areas under the ROC curve between measures when using global perceived change in neck symptoms as the external criteria. The two rank based preference elicitation methods (DCE and ORD) resulted in areas above the discriminatory threshold of 0.70 for this threshold. All of the SF-6D measures were at or near the discriminatory threshold when neck disability was used as the external criteria. Sensitivity analyses using cross-sectional QALY estimates from week 52 and change from baseline to week 52 were similar to the primary analysis using area under the curve to measure QALY changes (Table 13).

Discussion

Researchers conducting cost-effectiveness analyses have an important decision to make when choosing the most appropriate QALY measure. Ideally, this decision will be

informed by the psychometric properties of the QALY instruments within the population of interest. The test-retest reliability, validity, and responsiveness of commonly used QALY measures was estimated within older adults with chronic neck pain to inform future economic evaluations within this population. An overview of the findings are presented in tables 14-15. The SF-6D demonstrated better reliability and agreement compared to the EQ-5D and EQ VAS, and better construct validity relative to the EQ-5D. There were few differences between U.S. based QALY measures when assessing responsiveness. The SF-6D derived QALY measures demonstrated similar reliability, agreement, construct validity, and responsiveness.

A Pubmed search was conducted to identify published studies assessing the psychometric properties of QALY measures within spinal pain populations for comparison. The search strategy placed no limitations on language or date of publication and used the following key terms: 1) Spinal pain = low back pain OR back pain OR neck pain OR lumbago OR cervicalgia 2) QALYs = quality adjusted life year OR QALY OR EQ-5D OR SF-6D OR HUI 3) Psychometric properties = psychometric OR reliability OR validity OR responsiveness OR clinimetric. In addition, the bibliographies of relevant studies were searched to identify additional studies of interest. The search strategy identified twelve studies assessing the psychometric properties of QALY measures among individuals with spinal pain. The important findings from studies using similar methods are highlighted throughout the discussion.

In terms of practicality or item completion, the different measures performed similarly. All participants completing the EQ VAS measure (the shortest measure) also completed the EQ-5D, and only 2 to 3 fewer participants completed the SF-6D (the longest measure) at each time point. Completion rates were similar to other studies reporting the practicality of QALY measures within spinal pain populations.[14,15,29,34] Prior studies have found that completion rates for the SF-6D were slightly lower than the EQ-5D, which is not surprising given the brevity of the EQ-5D relative to the SF36 from which the SF-6D is derived.

An important finding from this study is the difference in test-retest reliability between U.S. based QALY measures. The intraclass correlation coefficient for the SF-6D (0.81) was larger than the EQ VAS (0.68) and much larger than the EQ-5D (0.44). For older adults with chronic neck pain, the SF-6D is the most reliable measure. The reliability of the EQ-5D was likely limited by a small range of baseline QALY values. Over three-quarter of the population reported a mean baseline EQ-5D value between 0.77 and 0.83. The ICC relates the amount of variability from repeated measurements (i.e. test-retest) to the total amount of variability which includes variability between individuals. Measures with a small range of values will have a difficult time distinguishing between individuals which will invariably lead to more measurement error.

The reliability of QALY measures within spinal pain populations has not received much attention to date. Three prior studies assessing the test-retest reliability of QALY measures within spinal pain populations were identified for comparison. Ideally, test-retest reliability should be conducted over a long enough period where the participant cannot recall their prior rating but short enough that their health state does not change. Additionally, the time period between administrations should not include treatment interventions as this will likely have an impact on their health state. McDonough et al.[35] and Suarez et al.[32] assessed reliability among individuals reporting no change in spinal symptoms over a period of 3-6 months (Suarez) to one year (McDonough) during which they received spinal treatment. McDonough et al. reported lower reliability for the EQ-5D (0.62) compared to the SF-6D (0.76) among a surgical population. Suarez et al. found variable results for EQ-5D and EQ VAS reliability depending on the time frame used (3 or 6 months) for assessment among individuals reporting no change in back symptoms. ICC's ranged from 0.76 (3 months) to 0.48 (6 months) for the EQ-5D, and from 0.39 (3 months) to 0.54 (6 months) for the EQ VAS. Finally, a study by Solberg et al.[31] assessed the reliability of the EQ-5D among sub-samples of surgical patients and found higher reliability (ICC's from 0.82-0.87) than other studies.

The higher reliability of the EQ-5D within the two studies using individuals with spinal pain undergoing surgery may reflect the more heterogeneous mix of health states among this population. McDonough et al.[35] reported a bi-modal distribution of QALYs within their population of adults suffering from more complicated causes of spinal pain (i.e. intervertebral disc herniation, spinal stenosis, or degenerative spondylolisthesis) who enrolled within a clinical trial or cohort including surgery as a treatment option. The 25th percentile for QALYs was 0.08 with a significant proportion of adults reporting QALYs at or below 0 (the equivalent of death). In contrast, the 25th percentile for EQ-5D QALYs in this study was 0.79, which highlights the limited spread of QALY values and may explain the difference in reliability. These findings highlight the fact that the reliability of measurement scales is a property of the population in which it is measured and not the instrument itself.

The smallest detectable change (SDC) is often used to describe the amount of measurement error when conducting repeated measurements of a self-report survey instrument. The SDC corresponds to half the range of the 95% limits of agreement and changes smaller than the SDC may be due to measurement error.[46] Overall, the SDC among QALY measures was high, with no measures able to ensure changes below 0.116 were not due to measurement error. The SF-6D had the lowest SDC among U.S. based QALY measures, but was the highest among all the SF-6D-derived measures at 0.156. While a few studies have examined the reliability of QALY measures in spinal pain populations using ICCs, the search strategy was only able to locate one prior study which assessed the agreement of repeated QALY assessments.[27] Johnsen et al.[27] found a similar SDC for the SF-6D (0.157), but the SDC for the EQ-5D was much higher than our estimate (0.429). Johnsen et al. assessed agreement among individuals reporting no change in symptoms over a three month period, which may explain the difference in SDC for the EQ-5D.

The amount of measurement error relative to the expected changes in QALYs for spine pain is an issue when assessing the relative cost-effectiveness of healthcare interventions

for older adults with spinal pain. Mean QALY improvements after the application of common treatments (e.g. physical therapy, manual therapy, usual medical care) for spinal pain within outpatient settings have ranged from 0.02 to 0.11 and are below the SDC for U.S. QALY measures.[63-72] QALY improvements within surgical and hospital settings have been higher (~0.08 to 0.25) and are near or above the SDC for U.S. based QALYs.[69,73,74] Small changes in QALYs may be extremely important if they are associated with low additional costs; however, our current QALY measures are unable to detect these changes without a considerable amount of measurement error. Consequently, studies primarily designed to determine the cost-effectiveness of interventions not expected to result in QALY changes greater than 0.15 will require larger sample sizes to compensate for the measurement error.

The SF-6D and EQ VAS demonstrated better longitudinal known-group validity relative to the EQ-5D. Mean SF-6D and EQ VAS QALY changes were monotonically decreasing across levels of improvement for three of the four external criteria. The EQ-5D was the poorest U.S. based measure when distinguishing between individuals reporting different levels of improvement. The search strategy identified five studies assessing the longitudinal known-group validity of QALY measures. Previous studies have found the EQ-5D and SF-6D performed similarly when discriminating between levels of overall health change, spinal disability, treatment tolerability, and satisfaction.[15,28,31,33,35] McDonough et al. noted that both the EQ-5D and SF-6D had difficulty distinguishing between minor improvement and no change in overall health among adults with a lumbar disc herniation participating in a clinical trial comparing surgical to conservative care.[28] Whitehurst et al. assessed the longitudinal known-group validity of the SF-6D and EQ-5D among 346 adults enrolled in a clinical trial comparing advice and exercise alone with the addition of either spinal manipulation or diathermy.[15] They noted both measures performed well when discriminating between levels of overall health change. Whitehurst used UK preferences for both the SF-6D and EQ-5D which may explain the better performance of the EQ-5D when discriminating between changes in health states.

Compared to the U.S., U.K.-based EQ-5D values are lower with greater variability and result in larger changes following treatment for spinal pain.[28,35]

Another important finding of this study is the similar responsiveness among U.S. based QALY measures. ROC curve analyses showed each of the measures was near the recommended discriminatory threshold of 0.70 for most of the external criteria (range 0.65 to 0.73). There were no significant differences between U.S. based QALYs when assessing responsiveness with ROC curves. The overall association between QALY changes and the pre-specified external criteria was not strong. The impact of neck pain is theoretically captured within the pain, physical functioning, role limitations, social functioning, and vitality dimensions of the SF-6D, and the pain/discomfort, self-care, and usual activities dimensions of the EQ-5D. No one QALY measure stood out as having a consistently stronger association with external criteria relative to the other QALY measures. One would expect QALY measures would correlate more strongly with general health improvement than measures for neck symptoms; however, this was not the case. The only external criteria at least moderately correlated with QALY measures was neck disability. The SF-6D, EQ-5D, and neck disability (NDI) are multi-dimensional measures which share a number of common health domains. This overlap may explain the larger association with neck disability relative to other external criteria.
There are a couple of possible explanations for the low correlation between QALY measures and global perceived changes in health, global perceived changes in neck symptoms, and pain. The most likely explanation is that the magnitude of QALY changes relative to the amount of measurement error limited the responsiveness of these measures. Poor validity of our external criteria could also explain the low correlation. Assessing the responsiveness of survey instruments requires a comparison with a "gold standard" measure, which best captures the true status of the measured construct. Global change scales, although commonly used as an external criterion when assessing responsiveness, are not without their limitations.[75] These scales have been criticized for being strongly influenced by the current health state, rather than measuring change in health.[62] To assess the potential influence of current health state, a sensitivity analyses

was conducted to assess correlation with both change in health from baseline to week 52, and week 52 alone which generally resulted in higher correlations, but the overall associations were still low and few notable increases (≥ 0.10) were present. Global improvement scales are also individualized, in that they allow respondents to judge their change using the domains most important to them. Poor correlation with current QALY measures could indicate there are important health domains not currently captured within the health status instruments used to derive QALYs. However, the EQ VAS QALY measure also allows participants to value their current health state using the domains most important to them. The lack of correlation between the EQ VAS QALY measure and global improvement scales suggests that exclusion of important health domains within current QALY measures does not likely explain the low correlation.

A number of other studies have used similar methods to assess the responsiveness of the QALY measures within individuals with spinal pain. Suarez et al. found similar strengths of correlation to our findings between QALY measures (EQ-5D, EQ VAS) and both pain and disability among 46 patients treated at a Canadian outpatient clinic by a rheumatologist or chronic pain specialist.[32] They noted moderate correlation between the EQ-5D and global perceived change in health; however, the reported correlations varied dramatically with repeated assessment (e.g. 0.10 at three months and 0.53 at six months). Soer et al. noted moderate correlation between the EQ-5D and both pain and disability among 151 patients seen at a secondary or tertiary pain clinic in the Netherlands.[30] Johnsen et al. reported moderate and high correlation from the EQ-5D and SF-6D, respectively, with both disability and global change from treatment among 113 chronic low back pain patients participating in a randomized clinical trial comparing surgery to multidisciplinary rehab.[27] Previous studies reporting stronger correlations between QALY measures and external criteria were performed within specialty clinics or alongside surgical trials where the expected change in QALYs is less likely to be masked by measurement error. For example, participants included in the RCT by Johnsen et al. reported much lower baseline QALY values relative to our population (i.e. SF-6D = 0.56; EQ-5D = 0.29). The poorer health of the population at baseline allowed for greater

potential QALY gains following treatment. QALY gains were 0.28 and 0.38 for individuals randomized to rehabilitation and surgery, respectively.[76]

Four existing studies assessed the responsiveness of QALY measures using area under the ROC curve methods with global improvement as the external criteria. Soer et al. found the discriminatory ability of the EQ-5D and EQ VAS was satisfactory (0.70 to 0.71), but these measures did not perform as well when identifying improvements in pain and disability (0.59 to 0.65).[30] Chotai et al. reported area under the ROC curves of 0.62 and 0.65 for the EQ-5D and SF-6D following surgery for neck and arm pain.[26] Solberg et al. noted the EQ-5D resulted in a discriminatory area of 0.77 following surgery for lumbar degenerative disc disease.[31] Finally, Johnsen et al. found areas under the ROC curves of 0.83 for the EQ-5D and 0.90 for the SF-6D among patients participating in a RCT comparing surgery to multidisciplinary rehabilitation for chronic LBP.[27]

This study has a number of strengths that are worth mentioning. The psychometric properties of QALY measures were assessed within an older non-surgical population with chronic neck pain. Older adults make up a substantial proportion of the U.S. population with pain.[77] While prior studies have assessed the psychometric properties of QALY measures within older adults undergoing surgical interventions, this is a small percentage of the population with more substantial decrements in health. QALYs are used to assess the cost-effectiveness of interventions across the full spectrum of health states and it's important to assess their psychometric properties accordingly. Another strength of this study was the methodology for assessing test-retest reliability. The timing of repeat administrations of QALYs was short (1-2 weeks), and no interventions were applied within this time frame, making it unlikely for meaningful changes in health status to occur. A number of previous studies have used a sub-population reporting no improvement three months to one year after receiving treatment to assess test-retest reliability.[27,32,35] This study also assessed the psychometric properties of a number of U.S. based QALY measures (i.e. SF-6D, EQ-5D, EQ VAS) and SF-6D-derived QALY measures (SG, DCE, ORD, BYS) to better understand differences due to the instruments

themselves in addition to method used to elicit preferences for health states within the instruments.

This study also has a number of limitations. As mentioned previously, assessing responsiveness using anchor based methods requires the use of a "gold standard" measure for comparison. The validity of the external criteria are a potential limitation. Distribution based methods are also commonly used to assessed responsiveness and do not require a "gold standard" for comparison. However, these methods have their own limitations. Researchers frequently use distribution based methods to estimate which measure results in the largest effect size (i.e. mean change divided by the standard deviation). However, the responsiveness of an instrument depends on how accurately it measures change.[57] If the true expected change after the application of a healthcare intervention is small, measures demonstrating large effect sizes will not be responsive. Distribution based methods require a priori hypotheses regarding the expected changes following the application of healthcare interventions which can be used to judge responsiveness. Well-informed hypotheses of expected changes following treatment are difficult to construct which is why the anchor based approach was used. Another limitation of this study is the difference in scaling between the EQ-VAS and other QALY measures. Death was represented by 0 on the EQ-5D and SF-6D measures, but 0 represented the worst health state possible on the EQ-VAS measure. Rescaling the EQ-VAS so 0 represents death would result in a slight decrease in QALY values, but is unlikely to influence the reliability, validity, or responsiveness of the measure. A sensitivity analysis rescaling the EQ-VAS using the average rating for death from European countries [78] had little to no impact on the reliability, agreement, construct validity, or responsiveness of the measure. Finally, the reliability and responsiveness of QALY measures is a function of the population in which they are measured. Accordingly, findings from this study are only generalizable to older adults with chronic neck pain seeking non-surgical interventions.

Conclusions

There were minor differences between U.S. QALY measures in terms of responsiveness; however, the SF-6D was more reliable and demonstrated less measurement error relative to the EQ-5D and EQ VAS, in addition to better known-group validity relative to the EQ-5D. Consequently, the SF-6D is likely the preferred QALY measure when assessing the impact of non-surgical interventions for older adults with chronic neck pain, where the expected change in QALYs will be small and may be masked by measurement error. QALYs derived from the same instrument (SF-6D), but using different methods had similar psychometric properties.

Table 1. Baseline demographic and clinical characteristics

Characteristic	Mean or %
n	241
Women	46.9%
Age	72.3 (5.4)
- Range	65 - 88
Pain radiates to upper extremity (%)	17.0 %
Employed	18.7%
Tobacco use	5.4%
Exercise weekly	75.9%
Neck pain (0-10)	5.1 (1.4)
Neck disability (0-100)	23.3 (9.5)
QALYs	
U.S. Measures	
- EQ VAS	0.766 (0.130)
- EQ-5D	0.795 (0.074)
- SF-6D	0.766 (0.124)
SF-6D Measures	
- SF-6D (US) DCE	0.766 (0.124)
- SF-6D (UK) SG	0.714 (0.091)
- SF-6D (UK) BYS	0.665 (0.079)
- SF-6D (UK) ORD	0.725 (0.096)

DCE = discrete choice experiment; SG = standard gamble; BYS = Bayesian model; ORD = ordinal ranking

Table 2. Completed surveys by time point

QALY Measure	Pain	EQ VAS	EQ-5D	SF-6D
Baseline	241	241	240	238
Week 4	239	239	239	237
Week 12	236	236	236	235
Week 26	236	230	230	227
Week 52	229	226	226	224

Table 3. Test-retest reliability and agreement

	(n)	ICC (95% CI)	Mean difference (SD)	SDC (95% CI)	95% LOA
U.S. QALY Measures					
SF-6D	238	0.81 (0.77 to 0.85)	-0.005 (0.08)	0.16 (0.14 to 0.17)	-0.16 to 0.15
EQ-5D	240	0.44 (0.33 to 0.53)	0.006 (0.09)	0.18 (0.16 to 0.20)	-0.18 to 0.19
EQ VAS	241	0.68 (0.61 to 0.75)	0.008 (0.11)	0.22 (0.20 to 0.25)	-0.21 to 0.23
SF-6D-Derived QALY Measures					
DCE (US)	238	0.81 (0.77 to 0.85)	-0.005 (0.08)	0.16 (0.14 to 0.17)	-0.16 to 0.15
SG (UK)	238	0.79 (0.74 to 0.83)	-0.007 (0.06)	0.12 (0.11 to 0.14)	-0.13 to 0.12
SG BYS (UK)	238	0.76 (0.70 to 0.81)	-0.0006 (0.06)	0.12 (0.10 to 0.13)	-0.12 to 0.12
ORD (UK)	238	0.79 (0.74 to 0.83)	-0.007 (0.07)	0.13 (0.11 to 0.14)	-0.14 to 0.12

ICC = intraclass correlation coefficient; SDC = smallest detectable change; LOA = limits of agreement; DCE = discrete choice experiment; SG = standard gamble; BYS = Bayesian model; ORD = ordinal ranking

Table 4. Sensitivity analysis for test-retest reliability and agreement: participants with ≥ 50% change in neck pain or disability excluded

	(n)	ICC (95% CI)	Mean Difference (SD)	SDC (95% CI)	95% LOA
U.S. QALY Measures					
SF-6D	191	0.81 (0.75 to 0.85)	-0.008 (0.08)	0.15 (0.13 to 0.17)	-0.16 to 0.15
EQ-5D	193	0.43 (0.31 to 0.54)	0.003 (0.09)	0.17 (0.15 to 0.20)	-0.17 to 0.18
EQ VAS	194	0.72 (0.65 to 0.78)	0.007 (0.11)	0.21 (0.18 to 0.23)	-0.20 to 0.21
SF-6D-Derived QALY Measures					
DCE (US)	191	0.81 (0.75 to 0.85)	-0.008 (0.08)	0.15 (0.13 to 0.17)	-0.16 to 0.15
SG (UK)	191	0.79 (0.73 to 0.84)	-0.01 (0.06)	0.12 (0.10 to 0.13)	-0.13 to 0.11
SG BYS (UK)	191	0.75 (0.68 to 0.81)	-0.002 (0.06)	0.11 (0.10 to 0.12)	-0.11 to 0.11
ORD (UK)	191	0.78 (0.72 to 0.83)	-0.01 (0.06)	0.13 (0.11 to 0.14)	-0.14 to 0.12

ICC = intraclass correlation coefficient; SDC = smallest detectable change; LOA = limits of agreement; DCE = discrete choice experiment; SG = standard gamble; BYS = Bayesian model; ORD = ordinal ranking

Table 5. Mean QALY changes relative to global perceived change in health

	Much better (n=22)	Somewhat better (n=49)	About the same (n=118)	Somewhat worse (n=26)	Much worse (n=0)
U.S. QALY Measures					
SF-6D	0.084 (0.10)	0.054 (0.08)	0.020 (0.07)	0.011 (0.07)	--
EQ-5D	0.061 (0.09)	0.048 (0.07)	0.005 (0.06)	0.009 (0.08)	--
EQ VAS	0.045 (0.10)	0.049 (0.08)	0.012 (0.09)	-0.021 (0.10)	--
SF-6D QALY measures					
DCE (US)	0.084 (0.10)	0.054 (0.08)	0.020 (0.07)	0.011 (0.07)	--
SG (UK)	0.076 (0.09)	0.042 (0.07)	0.016 (0.06)	0.005 (0.05)	--
SG BYS (UK)	0.079 (0.09)	0.044 (0.06)	0.018 (0.07)	0.011 (0.05)	--
ORD (UK)	0.078 (0.09)	0.048 (0.06)	0.018 (0.05)	-0.000 (0.06)	--

DCE = discrete choice experiment; SG = standard gamble; BYS = Bayesian model; ORD = ordinal ranking

Table 6. Mean QALY changes relative to global perceived change in neck symptoms

	No symptoms (n=22)	Much better (n=74)	Some-what better (n=35)	A little better (n=31)	No change (n=41)	Worse (n=12)
U.S. QALY Measures						
SF-6D	0.075 (0.11)	0.053 (0.08)	0.048 (0.07)	0.003 (0.07)	0.007 (0.06)	-0.036 (0.06)
EQ-5D	0.065 (0.09)	0.045 (0.07)	-0.001 (0.07)	-0.006 (0.05)	-0.005 (0.05)	0.005 (0.09)
EQ VAS	0.050 (0.09)	0.034 (0.08)	0.032 (0.09)	0.004 (0.08)	-0.000 (0.10)	-0.045 (0.09)
SF-6D QALY measures						
DCE (US)	0.075 (0.11)	0.053 (0.08)	0.048 (0.07)	0.003 (0.07)	0.007 (0.06)	-0.036 (0.06)
SG (UK)	0.073 (0.09)	0.041 (0.07)	0.033 (0.05)	0.004 (0.05)	0.008 (0.05)	-0.034 (0.05)
SG BYS (UK)	0.071 (0.10)	0.042 (0.07)	0.038 (0.06)	0.000 (0.05)	0.01 (0.05)	-0.021 (0.09)
ORD (UK)	0.078 (0.10)	0.045 (0.06)	0.034 (0.05)	0.013 (0.05)	0.000 (0.04)	-0.040 (0.06)

DCE = discrete choice experiment; SG = standard gamble; BYS = Bayesian model; ORD = ordinal ranking

Table 7. Mean QALY changes relative to quintiles of neck pain improvement

	Quintile #1 mean = 4.3	Quintile #2 mean = 2.8	Quintile #3 mean = 1.8	Quintile #4 mean = 1.0	Quintile #5 mean = -0.2
U.S. QALY Measures					
SF-6D	0.065 (0.09)	0.049 (0.07)	0.042 (0.08)	0.006 (0.07)	-0.001 (0.06)
EQ-5D	0.072 (0.09)	0.016 (0.07)	0.017 (0.06)	-0.004 (0.06)	-0.002 (0.05)
EQ VAS	0.063 (0.10)	0.026 (0.08)	0.022 (0.08)	0.001 (0.07)	-0.014 (0.09)
SF-6D QALY measures					
DCE (US)	0.065 (0.09)	0.049 (0.07)	0.042 (0.08)	0.006 (0.07)	-0.001 (0.06)
SG (UK)	0.057 (0.08)	0.036 (0.06)	0.039 (0.06)	0.003 (0.05)	-0.003 (0.05)
SG BYS (UK)	0.065 (0.08)	0.036 (0.05)	0.030 (0.07)	0.010 (0.05)	0.003 (0.06)
ORD (UK)	0.063 (0.08)	0.034 (0.05)	0.039 (0.06)	0.006 (0.05)	-0.004 (0.05)

DCE = discrete choice experiment; SG = standard gamble; BYS = Bayesian model; ORD = ordinal ranking

Table 8. Mean QALY changes relative to quintiles of neck disability improvement

	Quintile #1 mean = 16.8	Quintile #2 mean = 10.2	Quintile #3 mean = 6.0	Quintile #4 mean = 2.4	Quintile #5 mean = -3.4
U.S. QALY Measures					
SF-6D	0.095 (0.09)	0.049 (0.06)	0.035 (0.07)	-0.003 (0.05)	-0.014 (0.07)
EQ-5D	0.064 (0.08)	0.027 (0.07)	0.028 (0.06)	-0.013 (0.06)	-0.005 (0.06)
EQ VAS	0.067 (0.08)	0.043 (0.08)	0.026 (0.08)	-0.005 (0.07)	-0.034 (0.09)
SF-6D QALY measures					
DCE (US)	0.095 (0.09)	0.049 (0.06)	0.035 (0.07)	-0.003 (0.05)	-0.014 (0.07)
SG (UK)	0.072 (0.08)	0.041 (0.05)	0.027 (0.05)	0.002 (0.05)	-0.009 (0.05)
SG BYS (UK)	0.074 (0.07)	0.044 (0.07)	0.023 (0.06)	0.009 (0.06)	-0.006 (0.07)
ORD (UK)	0.078 (0.08)	0.042 (0.05)	0.035 (0.05)	-0.004 (0.05)	-0.011 (0.05)

DCE = discrete choice experiment; SG = standard gamble; BYS = Bayesian model; ORD = ordinal ranking

Table 9. Proportion of individuals with potential floor or ceiling effects at baseline

QALY Measure	EQ VAS	EQ-5D US (TTO)	SF-6D US (DCE)	SF-6D UK (SG)	SF-6D UK (BYS)	SF-6D UK (ORD)
(n)	241	240	238	238	238	238
≥95%	5.4%	0.4%	8.3%	0.4%	0.4%	0.4%
≥90%	13.7%	1.2%	17.0%	0.4%	0.4%	1.3%
≤5%	-	-	-	-	-	-
≤10%	-	-	-	-	-	-

DCE = discrete choice experiment; SG = standard gamble; BYS = Bayesian model; ORD = ordinal ranking

Table 10. Correlation of QALY measures with external criteria

	Global Improvement	**Neck Improvement**	**Pain**	**NDI**
U.S. QALY Measures				
SF-6D	-0.277	-0.360	-0.340	-0.596
EQ-5D	-0.269	-0.312	-0.391	-0.458
EQ VAS	-0.233	-0.260	-0.316	-0.442
SF-6D-derived QALY Measures				
DCE (US)	-0.277	-0.360	-0.340	-0.596
SG (UK)	-0.303	-0.376	-0.364	-0.513
SG BYS (UK)	-0.280	-0.315	-0.323	-0.491
ORD (UK)	-0.338	-0.426	-0.403	-0.587

NDI = neck disability index; DCE = discrete choice experiment; SG = standard gamble; BYS = Bayesian model; ORD = ordinal ranking

Table 11. Sensitivity analysis for correlation analyses: Δ week 52 and week 52 outcomes

	Global Improvement		Neck Improvement		Pain		NDI	
	Δ W52	W52	Δ W52	W52	Δ W52	W52	Δ W52	W52
U.S. Measures								
SF-6D	-0.402	-0.337	-0.413	-0.338	-0.387	-0.421	-0.569	-0.717
EQ-5D	-0.372	-0.408	-0.338	-0.379	-0.397	-0.416	-0.483	-0.596
EQ VAS	-0.332	-0.386	-0.274	-0.290	-0.253	-0.314	-0.342	-0.473
SF-6D-derived Measures								
DCE (US)	-0.402	-0.337	-0.413	-0.338	-0.387	-0.421	-0.569	-0.717
SG (UK)	-0.424	-0.352	-0.431	-0.335	-0.389	-0.399	-0.515	-0.666
SG BYS (UK)	-0.391	-0.362	-0.344	-0.308	-0.363	-0.386	-0.459	-0.630
ORD (UK)	-0.463	-0.381	-0.476	-0.368	-0.398	-0.418	-0.521	-0.698

Δ W52 = W52 minus Baseline; NDI = neck disability index; DCE = discrete choice experiment; SG = standard gamble; BYS = Bayesian model; ORD = ordinal ranking

Table 12. Area under the ROC curve analyses

	Global Improvement (≥ somewhat better)	Neck Improvement (≥ somewhat better)	Pain (≥ 30% reduction)	NDI (≥ 30% reduction)
U.S. QALY Measures				
SF-6D	0.65 (0.57 to 0.73)	0.71 (0.64 to 0.78)	0.60 (0.52 to 0.67)	0.73 (0.66 to 0.80)
EQ-5D	0.70 (0.62 to 0.77)	0.72 (0.65 to 0.79)	0.64 (0.57 to 0.72)	0.71 (0.64 to 0.78)
EQ VAS	0.67 (0.59 to 0.75)	0.67 (0.60 to 0.75)	0.61 (0.53 to 0.69)	0.67 (0.59 to 0.74)
p-value	0.58	0.61	0.49	0.34
SF-6D-derived QALY Measures				
DCE (US)	0.65 (0.57 to 0.73)	0.71 (0.64 to 0.78)	0.60 (0.52 to 0.68)	0.73 (0.66 to 0.80)
SG (UK)	0.65 (0.57 to 0.73)	0.69 (0.62 to 0.76)	0.62 (0.54 to 0.69)	0.71 (0.64 to 0.78)
SG BYS (UK)	0.64 (0.56 to 0.72)	0.66 (0.59 to 0.73)	0.58 (0.51 to 0.66)	0.69 (0.62 to 0.76)
ORD (UK)	0.68 (0.60 to 0.75)	0.72 (0.66 to 0.79)	0.61 (0.54 to 0.69)	0.74 (0.67 to 0.80)
p-value	0.40	0.04	0.60	0.19

NDI = neck disability index; DCE = discrete choice experiment; SG = standard gamble; BYS = Bayesian model; ORD = ordinal ranking

Table 13. Sensitivity analysis for ROC curve analyses: Δ week 52 and week 52 outcomes

	Global Improvement (≥ somewhat better)		Neck Improvement (≥ somewhat better)		Pain (≥ 30% reduction)		NDI (≥ 30% reduction)	
	Δ W52	W52	Δ W52	W52	Δ W52	W52	Δ W52	W52
U.S. Measures – Area under ROC curve (95% confidence interval)								
SF-6D	0.69 (0.62 to 0.77)	0.63 (0.55 to 0.71)	0.71 (0.64 to 0.78)	0.64 (0.56 to 0.71)	0.65 (0.57 to 0.72)	0.62 (0.55 to 0.70)	0.75 (0.69 to 0.82)	0.71 (0.65 to 0.78)
EQ-5D	0.68 (0.61 to 0.76)	0.65 (0.57 to 0.73)	0.65 (0.58 to 0.72)	0.62 (0.55 to 0.69)	0.66 (0.58 to 0.73)	0.63 (0.56 to 0.70)	0.69 (0.62 to 0.76)	0.68 (0.61 to 0.74)
EQ VAS	0.72 (0.65 to 0.79)	0.71 (0.64 to 0.78)	0.65 (0.58 to 0.73)	0.63 (0.56 to 0.71)	0.63 (0.56 to 0.70)	0.63 (0.55 to 0.70)	0.67 (0.60 to 0.74)	0.65 (0.58 to 0.73)
p-value	0.72	0.07	0.21	0.84	0.86	0.93	0.06	0.13
SF-6D-derived Measures – Area under ROC curve (95% confidence interval)								
DCE (US)	0.69 (0.62 to 0.77)	0.63 (0.55 to 0.71)	0.71 (0.64 to 0.78)	0.64 (0.56 to 0.71)	0.65 (0.57 to 0.72)	0.62 (0.55 to 0.70)	0.75 (0.69 to 0.82)	0.71 (0.65 to 0.78)
SG (UK)	0.70 (0.63 to 0.78)	0.62 (0.54 to 0.70)	0.70 (0.63 to 0.77)	0.63 (0.55 to 0.71)	0.65 (0.58 to 0.72)	0.61 (0.54 to 0.69)	0.73 (0.67 to 0.80)	0.69 (0.62 to 0.76)
SG BYS (UK)	0.70 (0.63 to 0.77)	0.63 (0.55 to 0.71)	0.66 (0.58 to 0.73)	0.61 (0.53 to 0.68)	0.63 (0.56 to 0.71)	0.61 (0.54 to 0.69)	0.73 (0.66 to 0.79)	0.70 (0.63 to 0.77)
ORD (UK)	0.72 (0.65 to 0.80)	0.64 (0.56 to 0.71)	0.73 (0.66 to 0.80)	0.65 (0.57 to 0.72)	0.67 (0.60 to 0.74)	0.62 (0.55 to 0.70)	0.74 (0.67 to 0.80)	0.70 (0.63 to 0.77)
p-value	0.17	0.21	0.03	0.03	0.34	0.39	0.63	0.05

Δ W52 = W52 minus Baseline; NDI = neck disability index; DCE = discrete choice experiment; SG = standard gamble; BYS = Bayesian model; ORD = ordinal ranking

Table 14. Overview of findings for the primary aim

U.S. QALY Measures	SF-6D	EQ-5D	EQ VAS
Reliability			
Intraclass Correlation Coefficient	+	-	-
Smallest Detectable Change	**0.16**	**0.18**	**0.22**
Construct Validity			
Global Change in Health	+	-	-
Global Change in Neck Symptoms	-	-	+
Neck Pain	+	-	+
Neck Disability	+	-	+
Responsiveness			
Correlation			
Global Change in Health	**Very Low**	**Very Low**	**Very Low**
Global Change in Neck Symptoms	**Low**	**Low**	**Very Low**
Neck Pain	**Low**	**Low**	**Low**
Neck Disability	**Mod**	**Low**	**Low**
ROC Curve			
Global Change in Health	-	+	-
Global Change in Neck Symptoms	+	+	-
Neck Pain	-	-	-
Neck Disability	+	+	-

Criteria: ICC: "+" ≥0.70, "-" <0.70;
Limits of agreement: "green" = uniform variability, "red" = non-uniform variability;
Construct validity: "+" = Monotonically decreasing, "-" = non-monotonically decreasing;
Correlation "Very Low" = very low (0.0 to 0.29), "Low" = low (0.30 to 0.49), "Mod" = moderate (0.50 to 0.69);
ROC Curve: "+" ≥0.70, "-" <0.70

Table 15. Overview of findings for the secondary aim

SF-6D Derived QALY Measures	Discrete Choice Experiment (US)	Standard Gamble (UK)	Standard Gamble - Bayesian (UK)	Ordinal Ranking (UK)
Reliability				
Intraclass Correlation Coefficient	+	+	+	+
Smallest Detectable Change	**0.16**	**0.12**	**0.12**	**0.13**
Construct Validity				
Global Change in Health	+	+	+	+
Global Change in Neck Symptoms	-	-	-	+
Neck Pain	+	-	+	-
Neck Disability	+	+	+	+
Responsiveness				
Correlation				
Global Change in Health	**Very Low**	**Low**	**Very Low**	**Low**
Global Change in Neck Symptoms	**Low**	**Low**	**Low**	**Low**
Neck Pain	**Low**	**Low**	**Low**	**Low**
Neck Disability	**Mod**	**Mod**	**Low**	**Mod**
ROC Curve				
Global Change in Health	-	-	-	-
Global Change in Neck Symptoms	+	-	-	+
Neck Pain	-	-	-	-
Neck Disability	+	+	-	+

Criteria: ICC: "+" ≥0.70, "-" <0.70;
Limits of agreement: "green" = uniform variability, "red" = non-uniform variability;
Construct validity: "+" = Monotonically decreasing, "-" = non-monotonically decreasing;
Correlation "Very Low" = very low (0.0 to 0.29), "Low" = low (0.30 to 0.49), "Mod" = moderate (0.50 to 0.69);
ROC Curve: "+" ≥0.70, "-" <0.70

Figure 1. Overview of Methods

Quality Adjusted Life Year (QALY) Measures

Primary Aim: QALY measures valued by the U.S. population

SF-6D; EQ-5D; EQ VAS

Secondary Aim: QALY measures derived from the SF-6D

Discrete choice experiment; Standard gamble; Standard gamble (Bayesian model); Ordinal ranking

Psychometric Properties

Reliability & Agreement methods

Intraclass Correlation Coefficient (ICC)

Limits of Agreement

Smallest Detectable Change

Construct Validity

Known Group Validity

Responsiveness

Correlation

Area under the ROC curve

External Criteria for Construct Validity & Responsiveness Analyses

Global Change in Health

Global Change in Neck Symptoms

Neck Pain

Neck Disability

Figure 2. Examples of QALY profiles

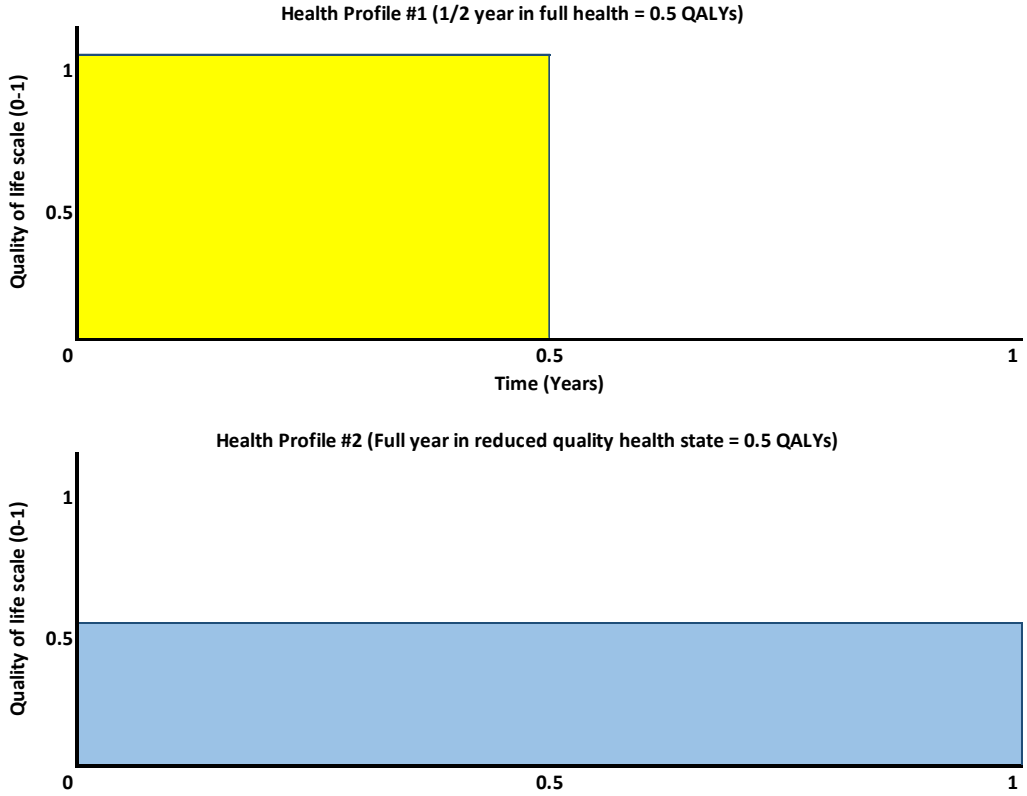


Figure 3. Baseline distributions of US QALY measures

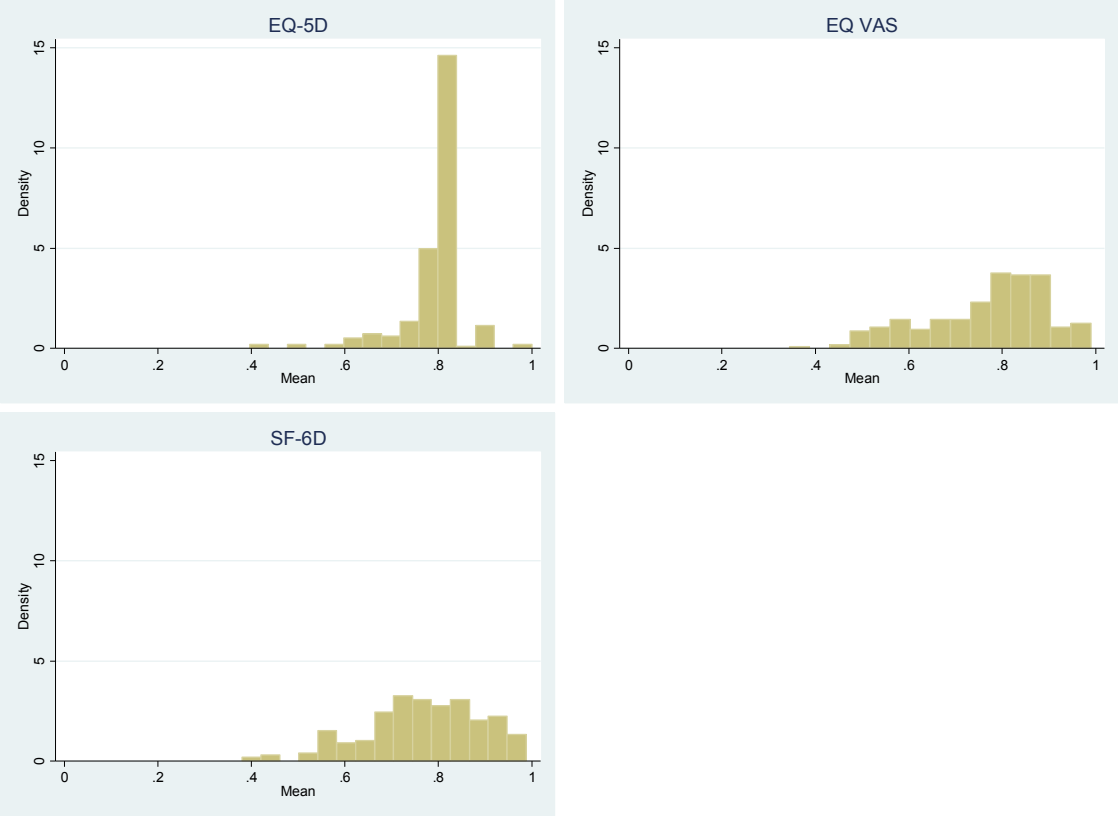
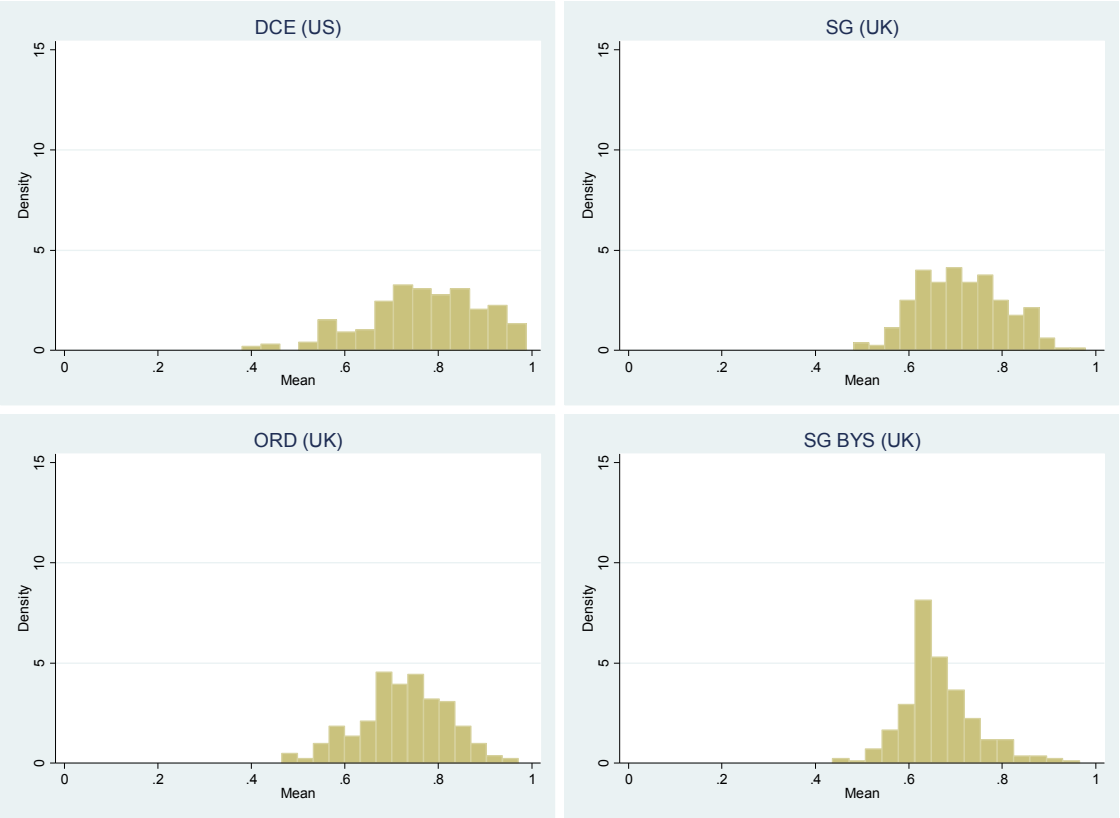


Figure 4. Baseline distributions of SF-6D based QALY measures



DCE = discrete choice experiment; SG = standard gamble; BYS = Bayesian model; ORD = ordinal ranking

Figure 5. Bland-Altman Limits of Agreement Plots – U.S. measures

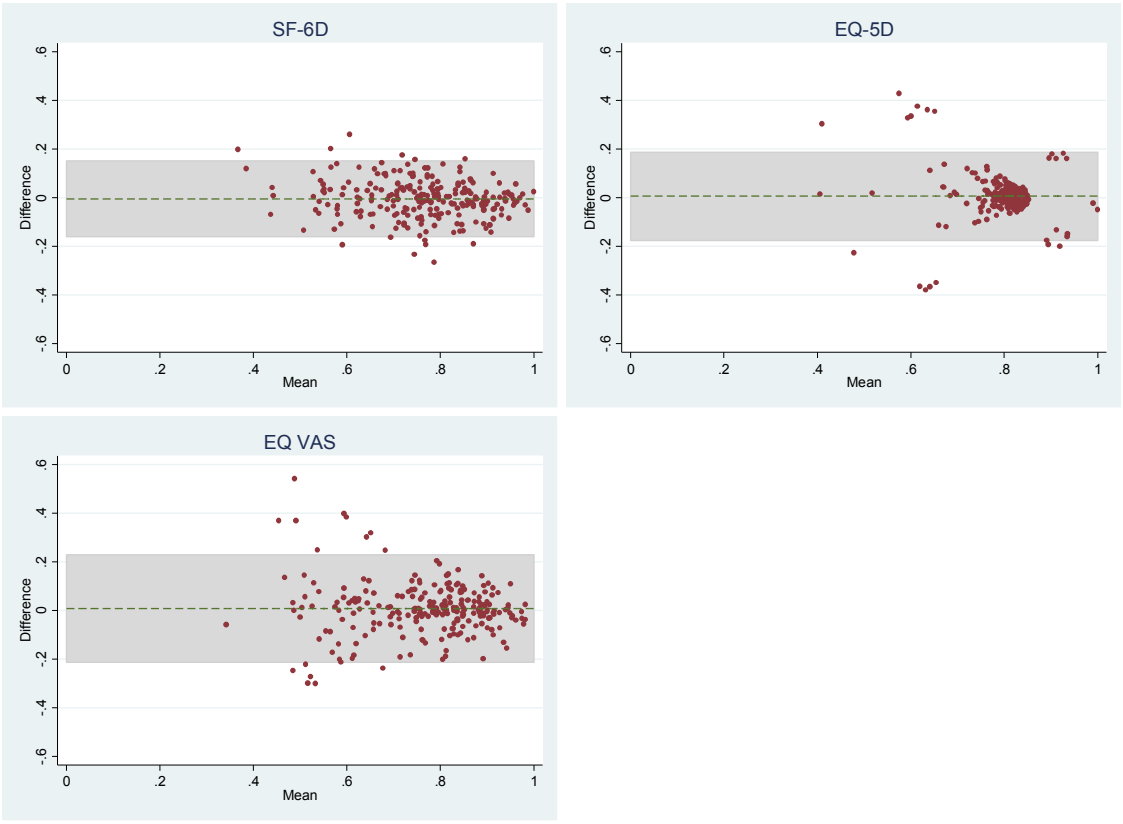
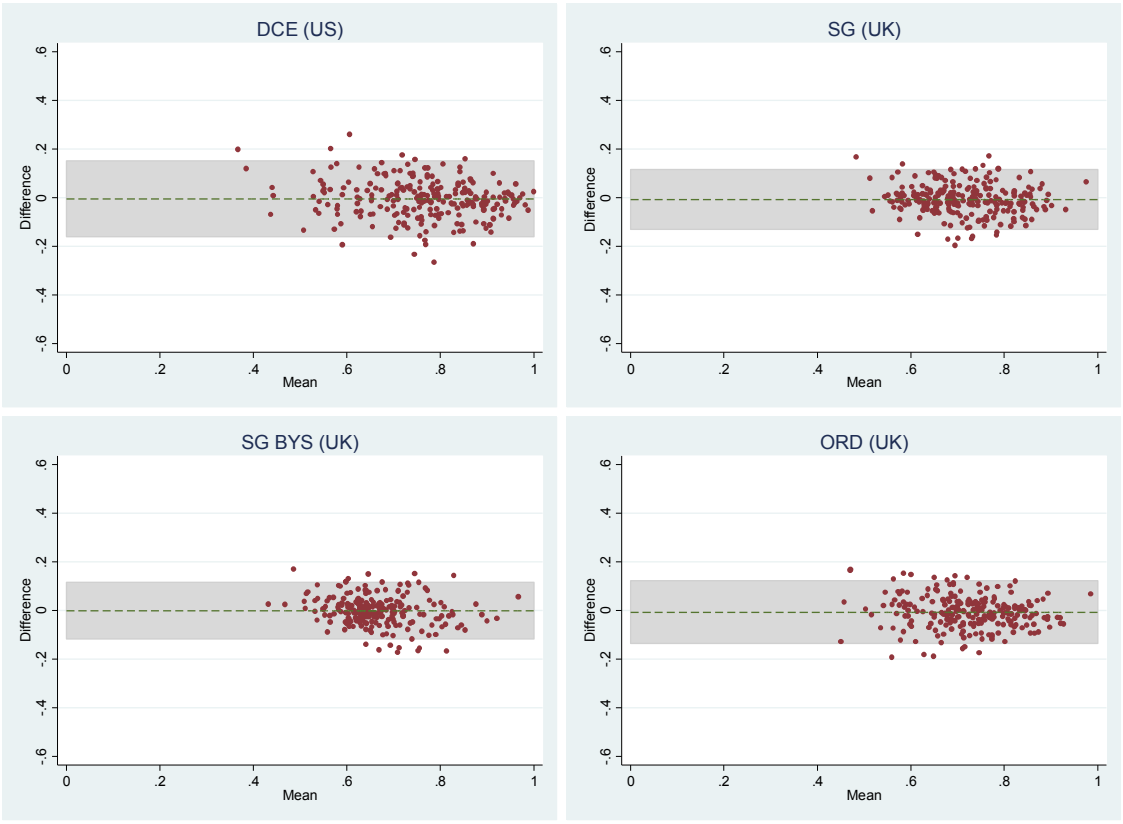
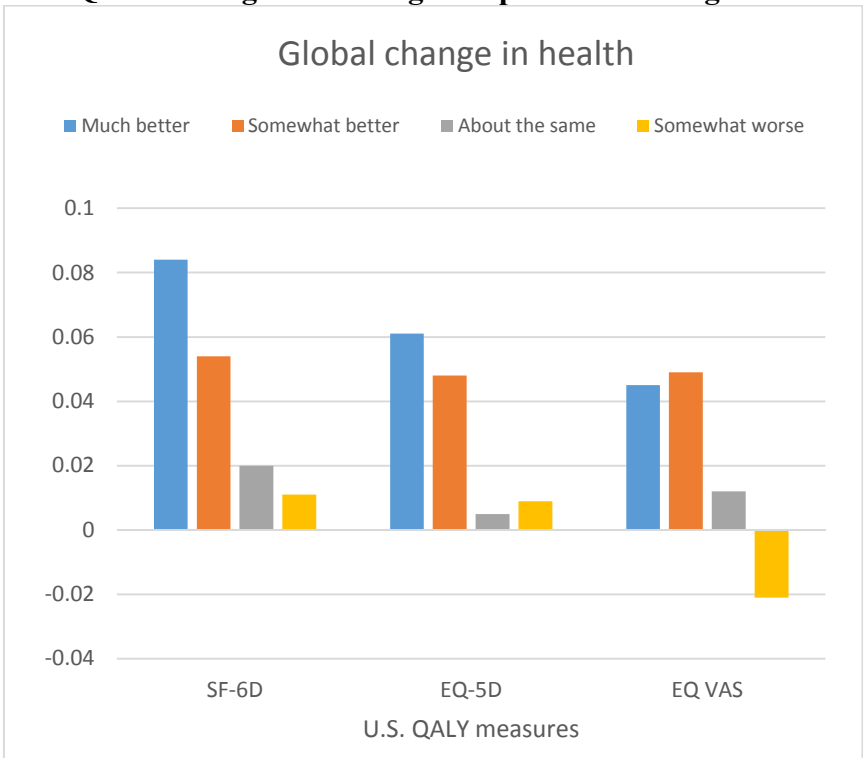


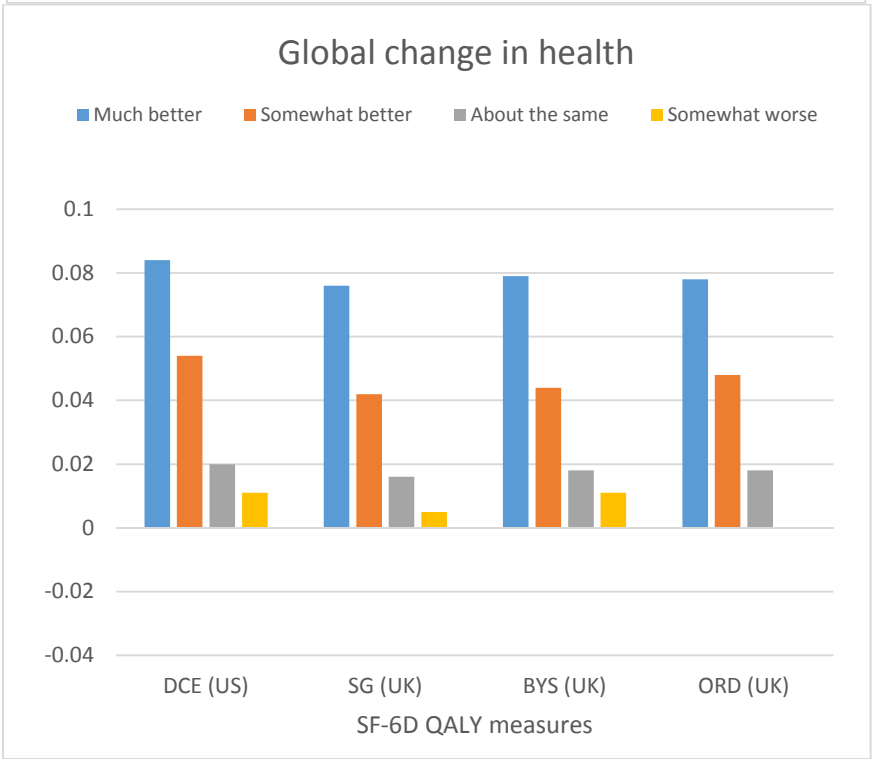
Figure 6. Bland-Altman Limits of Agreement Plots – SF-6D based measures



DCE = discrete choice experiment; SG = standard gamble; BYS = Bayesian model; ORD = ordinal ranking

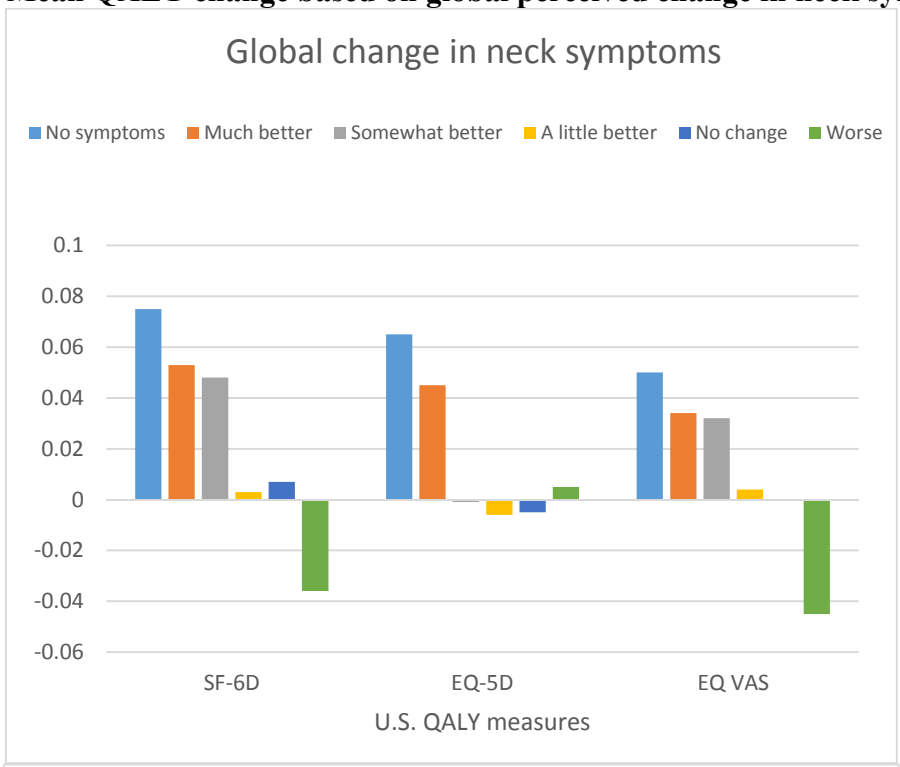
Figure 7. Mean QALY change based on global perceived change in health

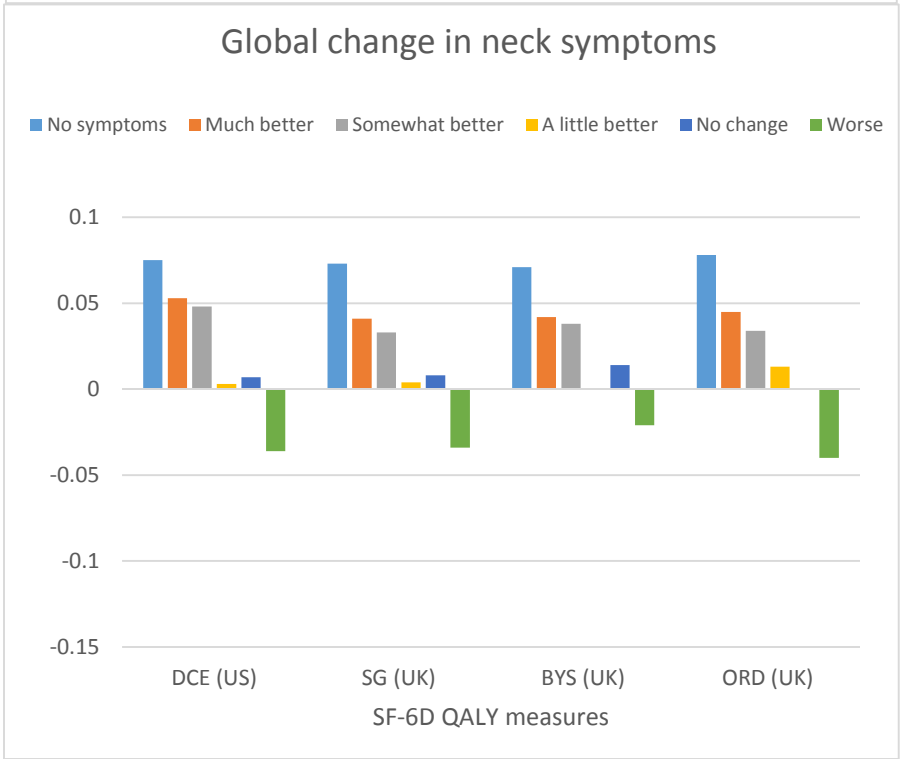




DCE = discrete choice experiment; SG = standard gamble; BYS = Bayesian model; ORD = ordinal ranking

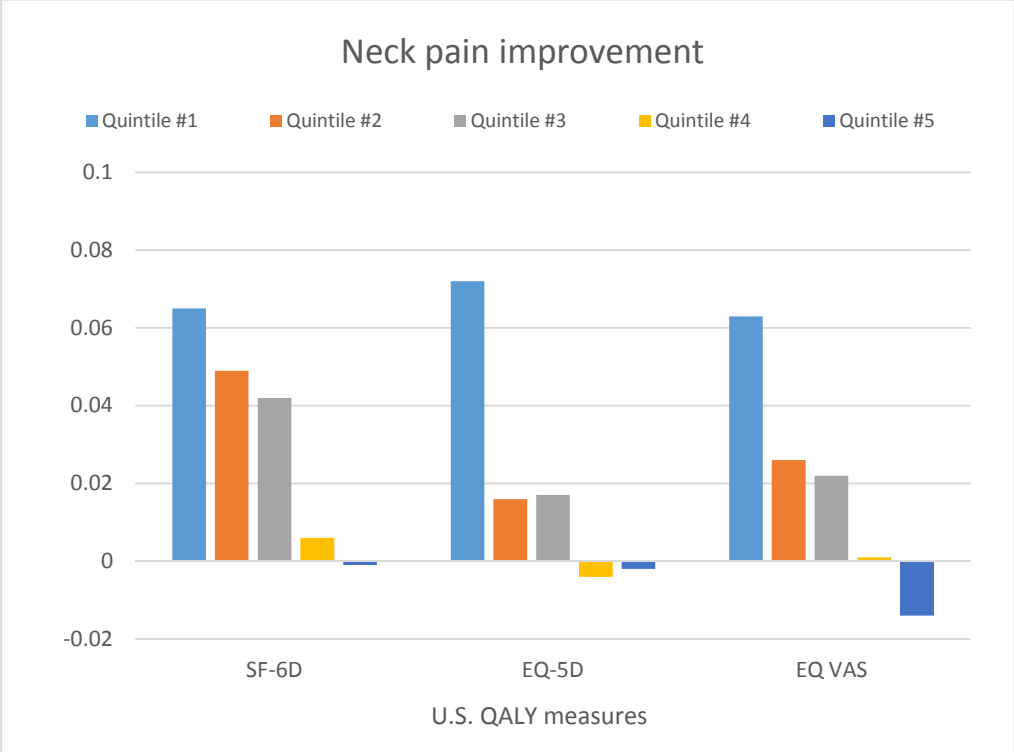
Figure 8. Mean QALY change based on global perceived change in neck symptoms

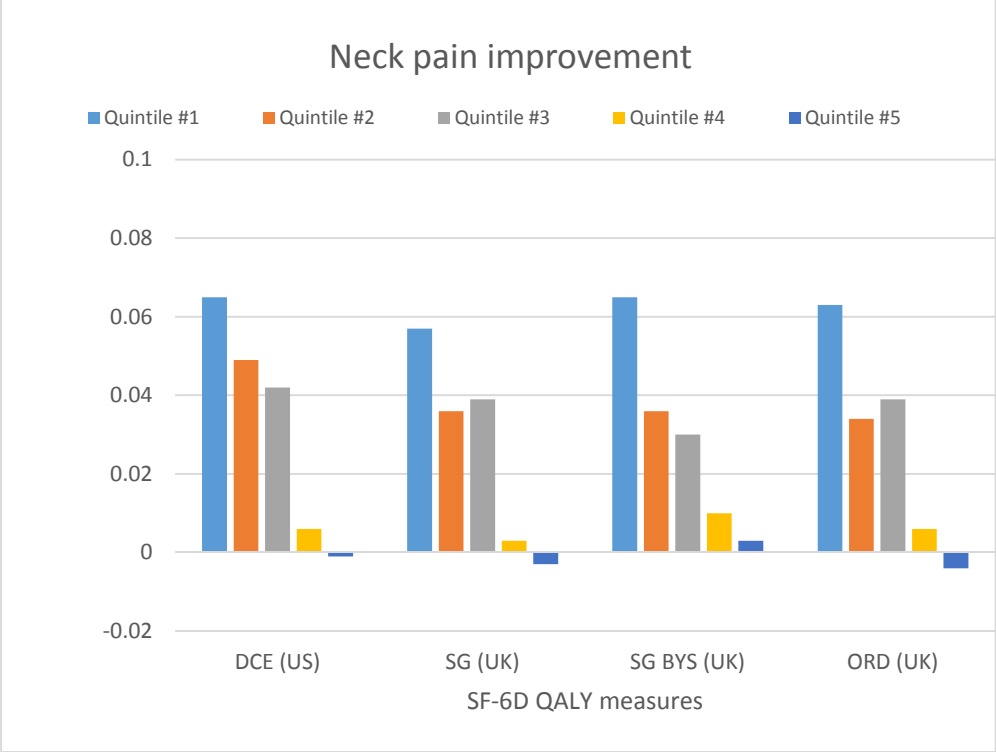




DCE = discrete choice experiment; SG = standard gamble; BYS = Bayesian model; ORD = ordinal ranking

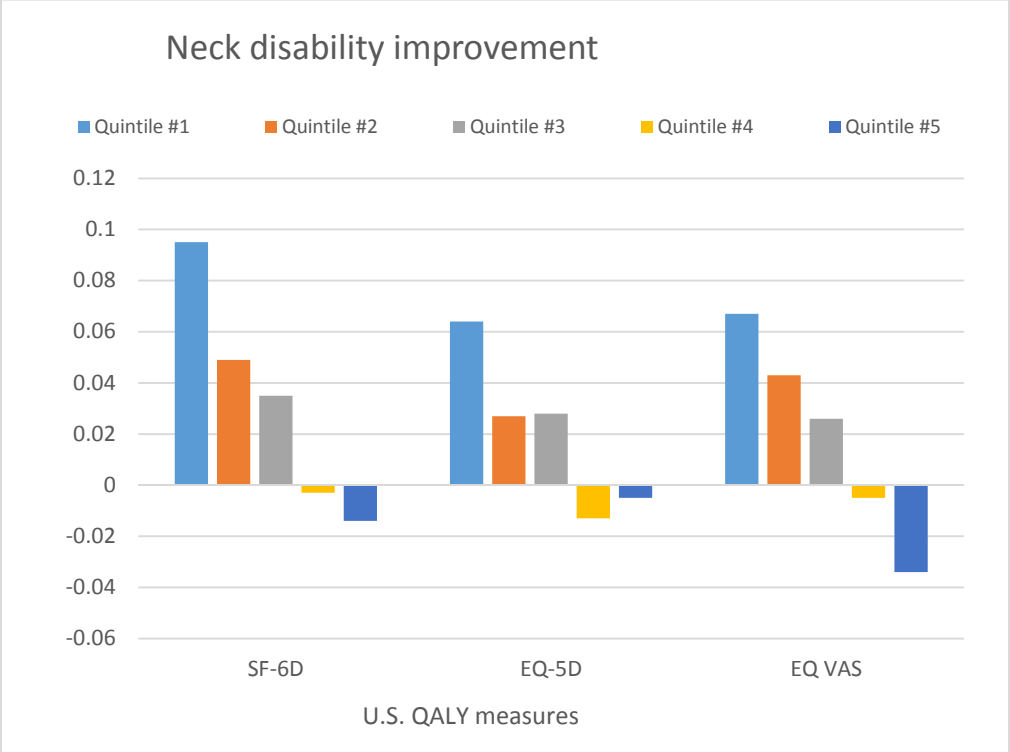
Figure 9. Mean QALY change for quintiles of neck pain improvement

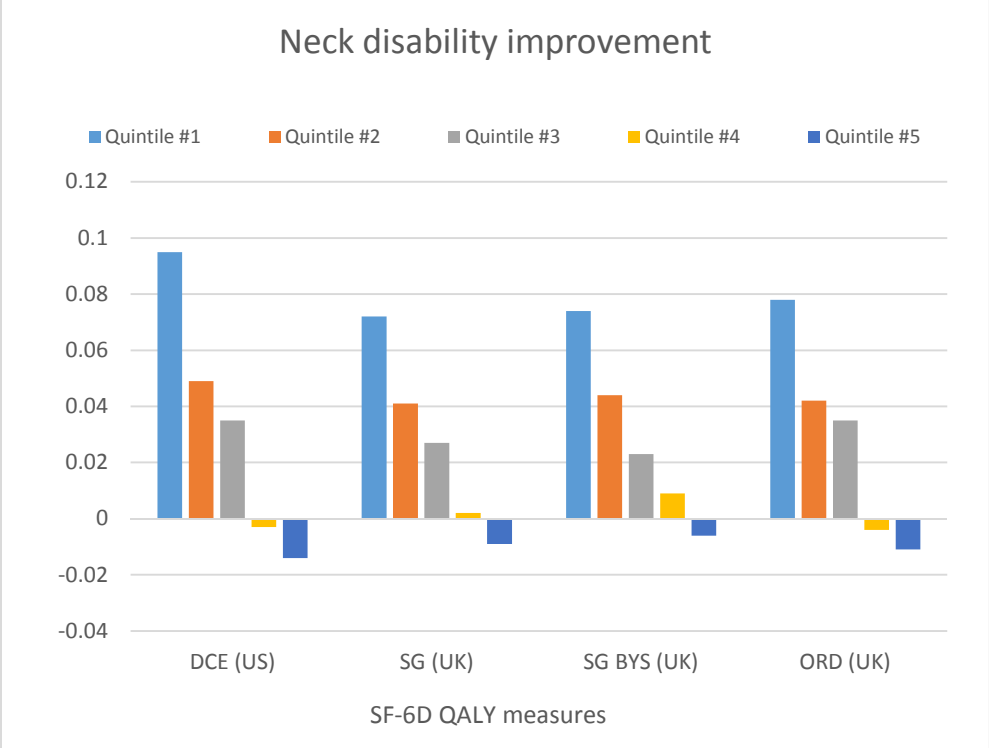




DCE = discrete choice experiment; SG = standard gamble; BYS = Bayesian model; ORD = ordinal ranking

Figure 10. Mean QALY change for quintiles of neck disability improvement





DCE = discrete choice experiment; SG = standard gamble; BYS = Bayesian model; ORD = ordinal ranking

Figure 11. Area under the ROC curve for U.S. QALY measures

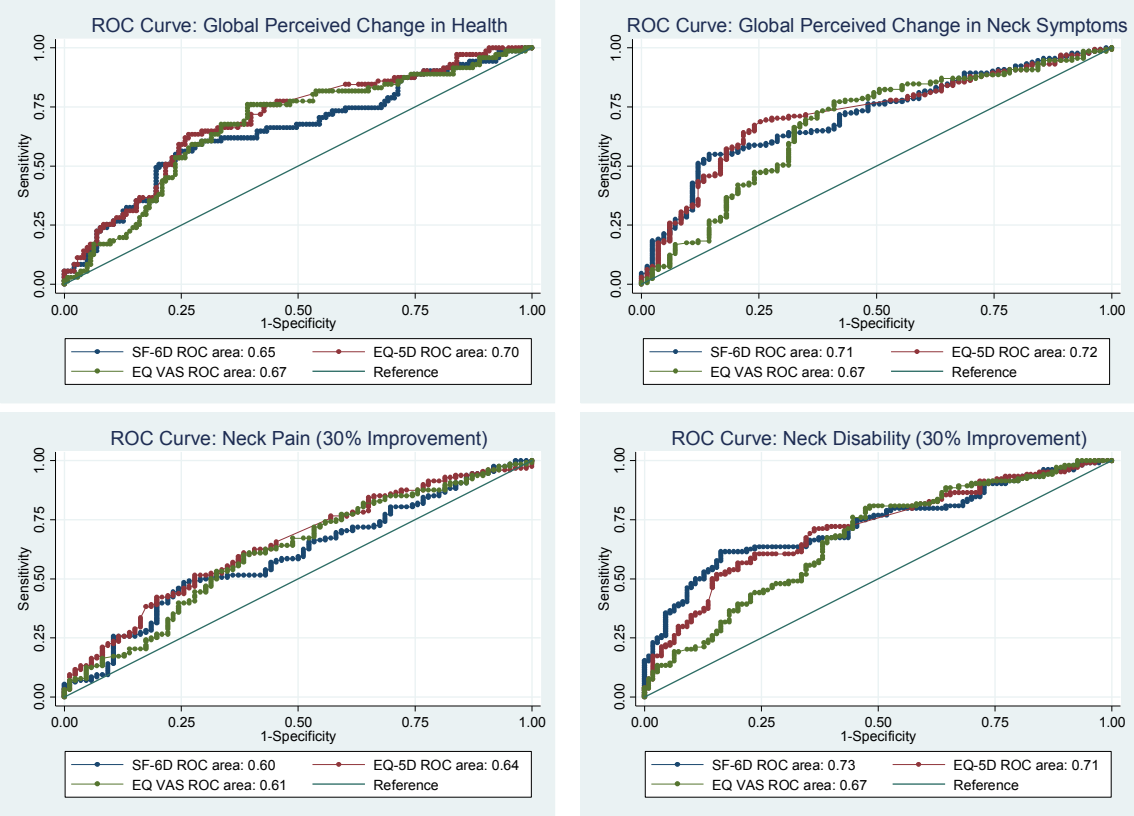
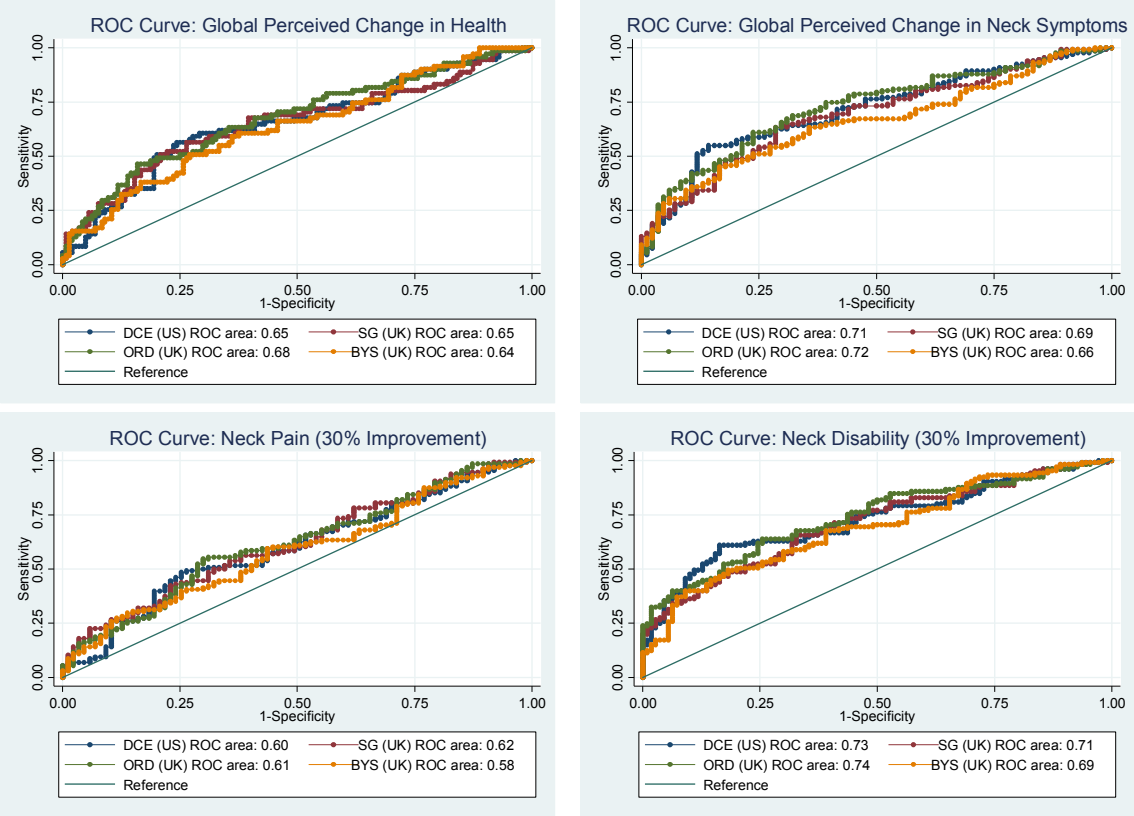


Figure 12. Area under the ROC curve for SF-6D-based QALY measures



DCE = discrete choice experiment; SG = standard gamble; BYS = Bayesian model; ORD = ordinal ranking

References

1. Sculpher M, Drummond M, O'Brien B. Effectiveness, efficiency, and NICE. *BMJ.* 2001;322(7292):943-944.
2. Neumann PJ, Sullivan SD. Economic evaluation in the US: what is the missing link? *PharmacoEconomics.* 2006;24(11):1163-1168.
3. Owens DK, Qaseem A, Chou R, Shekelle P. High-value, cost-conscious health care: concepts for clinicians to evaluate the benefits, harms, and costs of medical interventions. *Ann Intern Med.* 2011;154(3):174-180.
4. Gold MRS, J.E.; Russell, L.B.; Weinstein M.C. *Cost-effectiveness in Health and Medicine.* New York, NY: Oxford University Press; 1996.
5. Excellence NIfHaC. *Guide to the methods of technology appraisal 2013.* April 2013 2013.
6. EuroQol G. EuroQol--a new facility for the measurement of health-related quality of life. *Health Policy.* 1990;16(3):199-208.
7. Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI): concepts, measurement properties and applications. *Health Qual Life Outcomes.* 2003;1:54.
8. Brazier J, Usherwood T, Harper R, Thomas K. Deriving a preference-based single index from the UK SF-36 Health Survey. *Journal of clinical epidemiology.* 1998;51(11):1115-1128.
9. Sintonen H. The 15D instrument of health-related quality of life: properties and applications. *Ann Med.* 2001;33(5):328-336.
10. Hawthorne G, Richardson J, Osborne R. The Assessment of Quality of Life (AQoL) instrument: a psychometric measure of health-related quality of life. *Qual Life Res.* 1999;8(3):209-224.
11. Kaplan RM, Anderson JP. A general health policy model: update and applications. *Health Serv Res.* 1988;23(2):203-235.
12. Gray A, Clarke PM, Wolstenholme JL, Wordsworth S. *Applied Methods of Cost-effectiveness Analysis in Healthcare.* Oxford University Press; 2010.
13. Rasanen P, Roine E, Sintonen H, Semberg-Konttinen V, Ryynanen OP, Roine R. Use of quality-adjusted life years for the estimation of effectiveness of health care: A systematic literature review. *Int J Technol Assess Health Care.* 2006;22(2):235-241.
14. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Econ.* 2004;13(9):873-884.
15. Whitehurst DG, Bryan S. Another study showing that two preference-based measures of health-related quality of life (EQ-5D and SF-6D) are not interchangeable. But why should we expect them to be? *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research.* 2011;14(4):531-538.
16. McDonough CM, Grove MR, Tosteson TD, Lurie JD, Hilibrand AS, Tosteson ANA. Comparison of EQ-5D, HUI, and SF-36-derived societal health state values

among Spine Patient Outcomes Research Trial (SPORT) participants. *Quality of Life Research.* 2005;14(5):1321-1332.
17. Tsuchiya A, Brazier J, Roberts J. Comparison of valuation methods used to generate the EQ-5D and the SF-6D value sets. *J Health Econ.* 2006;25(2):334-346.
18. Doctor JN, Bleichrodt H, Lin HJ. Health utility bias: a systematic review and meta-analytic evaluation. *Medical decision making : an international journal of the Society for Medical Decision Making.* 2010;30(1):58-67.
19. Drummond MFS, M.J.; Torrance, G.W., O'Brien, B.J.; Stoddart GL. *Methods for the Economic Evaluation of Health Care Programmes.* Oxford University Press; 2005.
20. Peeters Y, Stiggelbout AM. Health state valuations of patients and the general public analytically compared: a meta-analytical comparison of patient and population health state utilities. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research.* 2010;13(2):306-309.
21. Martin BI, Deyo RA, Mirza SK, et al. Expenditures and health status among adults with back and neck problems. *JAMA.* 2008;299(6):656-664.
22. Johannes CB, Le TK, Zhou X, Johnston JA, Dworkin RH. The prevalence of chronic pain in United States adults: results of an Internet-based survey. *The journal of pain : official journal of the American Pain Society.* 2010;11(11):1230-1239.
23. Murray CJ, Atkinson C, Bhalla K, et al. The state of US health, 1990-2010: burden of diseases, injuries, and risk factors. *JAMA.* 2013;310(6):591-608.
24. Deyo RA, Mirza SK, Turner JA, Martin BI. Overtreating chronic back pain: time to back off? *Journal of the American Board of Family Medicine : JABFM.* 2009;22(1):62-68.
25. Chapman JR, Norvell DC, Hermsmeyer JT, et al. Evaluating common outcomes for measuring treatment success for chronic low back pain. *Spine (Phila Pa 1976).* 2011;36(21 Suppl):S54-68.
26. Chotai S, Parker SL, Sivaganesan A, Godil SS, McGirt MJ, Devin CJ. Quality of Life and General Health After Elective Surgery for Cervical Spine Pathologies: Determining a Valid and Responsive Metric of Health State Utility. *Neurosurgery.* 2015.
27. Johnsen LG, Hellum C, Nygaard OP, et al. Comparison of the SF6D, the EQ5D, and the oswestry disability index in patients with chronic low back pain and degenerative disc disease. *BMC musculoskeletal disorders.* 2013;14:148.
28. McDonough CM, Tosteson TD, Tosteson AN, Jette AM, Grove MR, Weinstein JN. A longitudinal comparison of 5 preference-weighted health state classification systems in persons with intervertebral disk herniation. *Medical decision making : an international journal of the Society for Medical Decision Making.* 2011;31(2):270-280.
29. Sogaard R, Christensen FB, Videbaek TS, Bunger C, Christiansen T. Interchangeability of the EQ-5D and the SF-6D in long-lasting low back pain.

Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. 2009;12(4):606-612.
30. Soer R, Reneman MF, Speijer BL, Coppes MH, Vroomen PC. Clinimetric properties of the EuroQol-5D in patients with chronic low back pain. *The spine journal : official journal of the North American Spine Society.* 2012;12(11):1035-1039.
31. Solberg TK, Olsen JA, Ingebrigtsen T, Hofoss D, Nygaard OP. Health-related quality of life assessment by the EuroQol-5D can provide cost-utility data in the field of low-back surgery. *Eur Spine J.* 2005;14(10):1000-1007.
32. Suarez-Almazor ME, Kendall C, Johnson JA, Skeith K, Vincent D. Use of health status measures in patients with low back pain in clinical settings. Comparison of specific, generic and preference-based instruments. *Rheumatology (Oxford).* 2000;39(7):783-790.
33. Obradovic M, Lal A, Liedgens H. Validity and responsiveness of EuroQol-5 dimension (EQ-5D) versus Short Form-6 dimension (SF-6D) questionnaire in chronic pain. *Health Qual Life Outcomes.* 2013;11:110.
34. Barton GR, Sach TH, Avery AJ, et al. A comparison of the performance of the EQ-5D and SF-6D for individuals aged >or= 45 years. *Health Econ.* 2008;17(7):815-832.
35. McDonough CM. *Valuing Health for Economic Evaluation in Spine Disorders*, Dartmouth College; 2007.
36. Maiers M, Bronfort G, Evans R, et al. Spinal manipulative therapy and exercise for seniors with chronic neck pain. *The spine journal : official journal of the North American Spine Society.* 2014;14(9):1879-1889.
37. Maiers MJ, Hartvigsen J, Schulz C, Schulz K, Evans RL, Bronfort G. Chiropractic and exercise for seniors with low back pain or neck pain: the design of two randomized clinical trials. *BMC musculoskeletal disorders.* 2007;8:94.
38. Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care.* 1992;30(6):473-483.
39. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *J Health Econ.* 2002;21(2):271-292.
40. Craig BM, Pickard AS, Stolk E, Brazier JE. US valuation of the SF-6D. *Medical decision making : an international journal of the Society for Medical Decision Making.* 2013;33(6):793-803.
41. Brooks R. EuroQol: the current state of play. *Health Policy.* 1996;37(1):53-72.
42. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. *Medical care.* 2005;43(3):203-220.
43. Kharroubi SA, Brazier JE, Roberts J, O'Hagan A. Modelling SF-6D health state preference data using a nonparametric Bayesian method. *J Health Econ.* 2007;26(3):597-612.
44. McCabe C, Brazier J, Gilks P, et al. Using rank data to estimate health state utility models. *J Health Econ.* 2006;25(3):418-431.

45. Glick HA, Doshi JA, Sonnad SS, Polsky D. *Economic Evaluation in Clinical Trials.* Oxford University Press; 2014.
46. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of clinical epidemiology.* 2007;60(1):34-42.
47. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *Journal of clinical epidemiology.* 2006;59(10):1033-1039.
48. Ramasundarahettige CF, Donner A, Zou GY. Confidence interval construction for a difference between two dependent intraclass correlation coefficients. *Stat Med.* 2009;28(7):1041-1053.
49. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;1(8476):307-310.
50. Streiner DL, Norman GR, Cairney J. *Health Measurement Scales: A Practical Guide to Their Development and Use.* Oxford University Press; 2014.
51. Dworkin RH, Turk DC, Farrar JT, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. *Pain.* 2005;113(1-2):9-19.
52. Wood BM, Nicholas MK, Blyth F, Asghari A, Gibson S. Assessing Pain in Older People With Persistent Pain: The NRS Is Valid But Only Provides Part of the Picture. *The Journal of Pain.* 2010;11(12):1259-1266.
53. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. *J Manipulative Physiol Ther.* 1991;14(7):409-415.
54. McCarthy MJH, Grevitt MP, Silcocks P, Hobbs G. The reliability of the Vernon and Mior neck disability index, and its validity compared with the short form-36 health survey questionnaire. *Eur Spine J.* 2007;16(12):2111-2117.
55. MacDermid JC, Walton DM, Avery S, et al. Measurement properties of the neck disability index: a systematic review. *J Orthop Sports Phys Ther.* 2009;39(5):400-417.
56. Hinkle DE, Wiersma W, Jurs SG. *Applied Statistics for the Behavioral Sciences.* Houghton Mifflin; 2003.
57. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC medical research methodology.* 2010;10:22.
58. Steiger JH. Tests for comparing elements of a correlation matrix. *Psychological Bulletin.* 1980;87:245-251.
59. Hoerger M. ZH: An updated version of Steiger's Z and web-based calculator for testing the statistical significance of the difference between dependent correlations. 2013. Accessed 11/30/2015.
60. Dworkin RH, Turk DC, Wyrwich KW, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *The journal of pain : official journal of the American Pain Society.* 2008;9(2):105-121.
61. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988;44(3):837-845.

62. Kamper SJ, Ostelo RW, Knol DL, Maher CG, de Vet HC, Hancock MJ. Global Perceived Effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. *Journal of clinical epidemiology.* 2010;63(7):760-766.e761.
63. Lin CW, Haas M, Maher CG, Machado LA, van Tulder MW. Cost-effectiveness of guideline-endorsed treatments for low back pain: a systematic review. *European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society.* 2011;20(7):1024-1038.
64. Driessen MT, Lin CW, van Tulder MW. Cost-effectiveness of conservative treatments for neck pain: a systematic review on economic evaluations. *Eur Spine J.* 2012;21(8):1441-1450.
65. Team UBT. United Kingdom back pain exercise and manipulation (UK BEAM) randomised trial: cost effectiveness of physical treatments for back pain in primary care. *BMJ (Clinical research ed.).* 2004;329(7479):1381.
66. Rivero-Arias O, Gray A, Frost H, Lamb SE, Stewart-Brown S. Cost-utility analysis of physiotherapy treatment compared with physiotherapy advice in low back pain. *Spine.* 2006;31(12):1381-1387.
67. Whitehurst DG, Lewis M, Yao GL, et al. A brief pain management program compared with physical therapy for low back pain: results from an economic analysis alongside a randomized clinical trial. *Arthritis and rheumatism.* 2007;57(3):466-473.
68. Williams NH, Edwards RT, Linck P, et al. Cost-utility analysis of osteopathy in primary care: results from a pragmatic randomized controlled trial. *Family practice.* 2004;21(6):643-650.
69. Critchley DJ, Ratcliffe J, Noonan S, Jones RH, Hurley MV. Effectiveness and cost-effectiveness of three types of physiotherapy used to reduce chronic low back pain disability: a pragmatic randomized trial with economic evaluation. *Spine.* 2007;32(14):1474-1481.
70. Bosmans JE, Pool JJ, de Vet HC, van Tulder MW, Ostelo RW. Is behavioral graded activity cost-effective in comparison with manual therapy for patients with subacute neck pain? An economic evaluation alongside a randomized clinical trial. *Spine.* 2011;36(18):E1179-1186.
71. Witt CM, Jena S, Selim D, et al. Pragmatic randomized trial evaluating the clinical and economic effectiveness of acupuncture for chronic low back pain. *American journal of epidemiology.* 2006;164(5):487-496.
72. Hollinghurst S, Sharp D, Ballard K, et al. Randomised controlled trial of Alexander technique lessons, exercise, and massage (ATEAM) for chronic and recurrent back pain: economic evaluation. *Bmj.* 2008;337:a2656.
73. Tosteson AN, Lurie JD, Tosteson TD, et al. Surgical treatment of spinal stenosis with and without degenerative spondylolisthesis: cost-effectiveness after 2 years. *Ann Intern Med.* 2008;149(12):845-853.
74. Tosteson AN, Skinner JS, Tosteson TD, et al. The cost effectiveness of surgical versus nonoperative treatment for lumbar disc herniation over two years: evidence

from the Spine Patient Outcomes Research Trial (SPORT). *Spine (Phila Pa 1976).* 2008;33(19):2108-2115.
75. Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. *Journal of clinical epidemiology.* 2002;55(9):900-908.
76. Hellum C, Johnsen LG, Storheim K, et al. Surgery with disc prosthesis versus rehabilitation in patients with low back pain and degenerative disc: two year follow-up of randomised study. *BMJ.* 2011;342:d2786.
77. Nahin RL. Estimates of Pain Prevalence and Severity in Adults: United States, 2012. *The journal of pain : official journal of the American Pain Society.* 2015;16(8):769-780.
78. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. *Ann Med.* 2001;33(5):337-343.