# Doctor, why does my hand hurt? The nature, course and treatment of pain in hand osteoarthritis

Meulen, C. van der

# CHAPTER 7

## IN CLINICAL HAND OSTEOARTHRITIS RESEARCH, SELF-REPORTED PAIN QUESTIONNAIRES DO NOT REFLECT THE PATIENT EXPERIENCE

Coen van der Meulen, Lotte A. van de Stadt, Féline P.B. Kroon, Marieke Niesters, Frits R. Rosendaal, Margreet Kloppenburg

Rheumatology (Oxford University Press). 2025 June;64(6): 3492–3499

*doi: 10.1093/rheumatology/keaf029*

# ABSTRACT

## Objective

Pain in hand osteoarthritis (OA) is evaluated with repeated pain questionnaires. It is unclear whether these questionnaires adequately capture changes in pain recalled by patients. This study investigated whether changes on pain questionnaires (real-time evaluation) correspond to recalled pain.

## Methods

Data from hand OA patients from the HOSTAS cohort (four one-yearly intervals) and HOPE trial (one six-week interval) were used. Pain was measured with the Australian/Canadian hand Osteoarthritis Index (AUSCAN, range 0-20) and a recall question (how is the pain compared to your last visit). Changes in AUSCAN pain were categorized into improved (≤-1), stable or worsened pain (≥1) and compared with the recall question using Cohen's kappa and percentage agreement. We determined concordance between measurement methods, and investigated associations of mental wellbeing and illness perceptions with concordance using generalized estimating equations (GEE).

## Results

Of 708 intervals from HOSTAS (307 patients, 82% women, mean age 61.0 years, mean AUSCAN 9.1), AUSCAN changes and recall were concordant in 42% (Cohen's kappa 0.13). There was concordance in 47% of 86 intervals (Cohen's kappa 0.14) from the HOPE trial (86 patients, 80% women, mean age 63.5, mean AUSCAN 10.7). The most frequent recall answer was worsened pain in the HOSTAS (60%), improved pain in the HOPE trial (76%). In both studies, AUSCAN pain most frequently improved. Depression and anxiety showed no association with concordance.

## Conclusion

Changes in repeatedly measured AUSCAN pain often differ from the recalled course of pain over the same period. This has profound implications for evaluating patient reported pain in clinical trials.

## Key messages

- Changes in patient-reported pain scores on questionnaire and patient's recalled pain are often discordant
- This discordance is seen in both cohort and trial settings
- No associations between this discordance and mental wellbeing were seen

# INTRODUCTION

Hand osteoarthritis (OA) is a common subtype of OA with pain as its primary symptom. As no curative treatment currently exists, guidelines advocate symptomatic treatment and patient education. (1) Pain is recognized as one of the core domains for hand OA studies by the Outcome Measures in Rheumatology (OMERACT) hand OA working group. (2) Furthermore, the OMERACT pain working group has described the need for standardized measures to assess pain in trials across rheumatic musculoskeletal diseases, to correctly assess the multidimensionality of pain. (3)

Pain in hand OA is commonly investigated with validated questionnaires such as the Australian/Canadian hand Osteoarthritis Index (AUSCAN), (4) and using pain scales such as the visual analogue scale (VAS) or numeric rating scale (NRS). Changes in pain over time are often evaluated by the differences between measurements at different time points. These changes can be presented with absolute values (e.g. the difference between two measurements) or a cut-off (e.g. minimal clinical important improvement (MCII)). (5, 6) However, various factors may influence the differences over time captured in this way, not just changes in pain. Methodological issues such as response shift, including changes in the way patients answer a questionnaire following changes in their internal standard of pain, their perception of what constitutes their pain or what consequences their pain has, as well as the context in which they evaluate their pain, including intercurrent positive or negative life events, can also play a role. (7) Similarly, patient expectations and perceptions have been described to affect patients experience from medical treatment, and thus how they answer pain questionnaires. (8)



In clinical practice, changes in pain are usually assessed by asking the patient how they recall the change in their pain compared with the last visit. Treatment decisions are based on the changes in pain and current pain reported by the patient. This clinical approach can be approximated in studies with a recall question asking the patient whether their pain has worsened, improved, or remained stable. Again, however, factors other than changes in pain may influence the reported changes, for example bias due to recall. (9)

It is unclear how changes on validated questionnaires correspond to changes recalled by patients. If these do not correspond, the outcomes of trials in hand OA may not adequately reflect the course of hand pain we wish to investigate. Thus, in this study we aimed to compare changes on a validated pain questionnaire with a recall question regarding the pain course, and whether influencing factors could be identified.

# METHODS

## Study design

Data from two previously published studies were used. The first data set involves data from the HOSTAS (Hand OSTeoArthritis in Secondary care) cohort study. HOSTAS is an observational cohort consisting of 538 consecutively included patients with primary hand OA diagnosed by a rheumatologist, collected from the Leiden University Medical Center (LUMC) rheumatology outpatient clinic between June 2009 and October 2015. Data from baseline to year four were used. Additionally, data from the Hand Osteoarthritis Prednisolone Efficacy (HOPE) trial, a randomized clinical trial (RCT) in which patients were treated for 6 weeks with prednisolone or placebo, were analysed. The HOPE trial included 92 patients with primary hand OA fulfilling the American College of Rheumatology criteria, (10) with at least 30 mm finger pain on a 100 mm VAS with a flare of 20 mm on NSAID washout, and signs of inflammation on ultrasound. The trial was conducted between December 2015 and October 2018. Full details about in- and exclusion criteria for both studies have been published previously. (11, 12)

Both studies were approved by the medical ethics committee at the LUMC (P09.004 and P15.096) and conducted in accordance with Good Clinical Practice guidelines, and with the Helsinki Declaration of 1975, as revised in 2000. All patients provided written informed consent.

## Outcomes

Pain was measured with the AUSCAN pain subscale (range 0-20) at baseline and annually thereafter in the HOSTAS study, and at baseline and after six weeks in the HOPE trial. (4) The AUSCAN pain subscale consists of five questions (how much pain did you have in the hands in rest, when gripping items, when lifting items, when turning items over and when squeezing items during the previous 48 hours), each scored 0-4, summed to reach a total score with a range of 0-20, where higher scores represent more pain. (4) Patients were unaware of previously submitted scores on the questionnaire at each study visit. AUSCAN pain scores were regarded as missing if two or more questions were missing.

From 2014 onwards, an annual recall question was added to the HOSTAS study: "Think only of the pain you experienced in your hands during the past 48 hours due to your hand osteoarthritis. How was the pain during the last 48 hours, compared with the last study visit?", with the answer options "Worsened – more pain", "No change", "Improved - less pain" and "I have never had this symptom". The HOPE trial collected the recall question at week 6. Patients were included in the current analysis when both change in AUSCAN pain (calculated from AUSCAN at the beginning and end of the interval) and

the recall question (collected at the end of the interval) were available for at least one time interval.

In both studies, age and sex were collected at baseline. Further baseline measurements included patient recalled symptom duration, hand examination including tender joint count, and hand radiographs. Radiographs were scored by trained readers blinded for clinical data using the Kellgren-Lawrence system and the Verbruggen-Veys method to assess erosive disease. (13, 14) Reliability of scoring was good in both studies. (11, 12)

In HOSTAS, the hospital anxiety and depression scale (HADS), consisting of separate 0-21 scales for anxiety and depression was collected annually. (15) Higher scores indicate more symptoms of anxiety or depression. No missing values were allowed for the calculation of the HADS scores. The illness perception questionnaire (IPQ) was collected at baseline, year 2 and year 4. (16) For all domains measured, higher scores indicate a stronger belief in the investigated construct. For details on the scales and calculation thereof, see appendix 2. For the IPQ scales, 1 or 2 missing values were accepted and treated with mean imputation, based on the scale, as per the IPQ instruction.

## Statistical analysis

Change in pain measured with AUSCAN between visits was categorized according to two methods. First, any change in pain was categorized as worsening, improvement or stable (i.e, change scores ≤-1 were categorized as improvement, scores of 0 as stable, and scores ≥1 as worsening). Secondly, the minimal clinical important improvement (MCII) of 1.6 (according to work by Bellamy et. al.) was used as a cutoff. (5) As no minimal clinical important deterioration or similar value is available, this cutoff was used in both directions. As the AUSCAN only allows for changes in discrete numbers, a cut-off of 1.6 functions as a cut-off for ≥2. (i.e, change scores ≤-2 were categorized as improvement, scores between -2 and 2 as stable, and scores ≥2 as worsening).

Unweighted Cohen's kappa indicated the overall agreement between the recall question and AUSCAN pain change. As in the HOSTAS data from multiple years were available, these were first calculated separately for each year, and afterwards with data from all years pooled. The analysis was repeated after splitting the cohort into low pain at start of interval and high pain at start of interval, based on median pain at start of the interval, being 9 for the HOSTAS and 11 for the HOPE. This yielded the best-balanced groups in terms of size. Kappa values <0 were regarded as poor, 0 – 0.20 as slight, 0.21 – 0.40 as fair, 0.41 – 0.60 as moderate, 0.61 – 0.80 as substantial and 0.81 – 1.00 as almost perfect. (17) To address potential effects of the unbalanced groups, we also calculated the Prevalence Adjusted Bias Adjusted Kappa (PABAK). (18) Percentage agreement between the

AUSCAN change and the recall question were calculated overall and stratified by the AUSCAN change categories (stable, improvement and deterioration).

Variables hypothesized to influence the concordance between AUSCAN changes and the recall question were assessed in the HOSTAS population. To this end, the change in AUSCAN pain was used as the index parameter compared with which patients overestimated or underestimated the recall question. Patients were categorized as being concordant, overestimating (recall question answered as improved pain when the change in AUSCAN reflects equal or worsened pain, or recall question answered as stable pain when the change in AUSCAN reflects worsened pain) or underestimating (recall question answered as worsened pain when the change in AUSCAN reflects equal or improved pain, or recall question answered as stable pain when the change in AUSCAN reflects improved pain).

Variables investigated consisted of age, sex, body mass index (BMI) measured at baseline, the HADS scales and the IPQ scales. The factors were chosen based on previous literature describing effects of illness and treatment perceptions and mental wellbeing on answers provided to pain questionnaires. (19-21) AUSCAN scores from the beginning of the investigated interval were used. HADS domain scores were dichotomized into presence of depression or anxiety using a cutoff of 8 or higher. (22) As no cutoffs were available for the IPQ, the scales were divided into tertiles to investigate trends between patients scoring low, middle or high on the scales, as the continuous scores violated the assumption of a linear association with the log odds of being concordant in the two questionnaires. The HADS and IPQ scores collected at the end of the interval investigated were used, for example if the recall question at the second visit was used, the HADS of the second visit was used.

Associations between being concordant versus overestimating or underestimating and variables of interest were assessed using logistic generalized estimating equations (GEE) to adjust for clustering within a patient, with robust standard errors and the correlation matrix specified as exchangeable. All analyses were performed with R version 4.1.3 and STATA version 16 (for the GEE analyses).

## RESULTS

Of the patients in the HOSTAS study, 307 had both an AUSCAN change score and recall question for at least one year. The mean age amongst these patients was 61.0 years, with

82% women. The mean BMI was 27.3 kg/m$^2$, and the mean AUSCAN pain at baseline 9.1 (standard deviation (SD) 4.2)) (table 1).

**Table 1. Patient characteristics**

| | HOSTAS N=307 | HOPE N=86 |
|---|---|---|
| **Patient characteristics** | | |
| Female sex; N (%) | 252 (82) | 69 (80) |
| Age, years | 61.0 (8.2) | 63.5 (8.7) |
| BMI, kg/m2 | 27.3 (4.8) | 27.3 (4.7) |
| **Disease characteristics** | | |
| ACR criteria fulfilled; N (%) | 276 (90) | 86 (100) |
| KL sum score (range 0-120); median (IQR) | 16 (8-28.5) | 37.5 (27.3-45) |
| Symptom duration, years; median (IQR) | 5.9 (2.3-13.5) | 10.1 (4.1-16.3) |
| Tender joint count (range 0-30); median (IQR) | 3 (1- 6) | 4 (2-7) |
| **Patient reported outcome measures** | | |
| AUSCAN | | |
| Pain (range 0-20) | 9.1 (4.2) | 10.7 (3.2) |
| Function (range 0-36) | 15.0 (8.3) | 18.7 (7.1) |

Data are mean (SD) unless indicated otherwise. BMI = Body mass index. HADS = Hospital Anxiety and Depression scale. ACR = American college of Rheumatology criteria for hand OA. KL = Kellgren-Lawrence. AUSCAN = Australian/Canadian osteoarthritis hand index. VAS = Visual analog scale. Percentage of missing data was lower than 5%, unless indicated otherwise. Currently working n = 326.



In the data from the HOSTAS study, 708 annual intervals with both change in AUSCAN and recall questions were available. Of the 307 patients, 95 provided one interval, 74 provided two intervals, 87 provided three and 51 provided four intervals. Results are described in table 2. The most frequent answer on the recall question was a worsening of pain (422/708 intervals, 60%), whereas AUSCAN pain worsened in only 279 (39%) of intervals. The mean (SD, 95% confidence interval, range) change in AUSCAN pain score in patients stating their pain had worsened was +0.43 (3.04, 0.14 to 0.72, -12 to 8). For no change and less pain the AUSCAN change scores were -0.68 (2.93, -1.07 to -0.28, -12 to 5) and -2.36 (3.71, -3.23 to -1.49, -8 to 10), respectively (table 3, supplementary figure A1). Yearly kappa's were all low (year 1: 0.12, year 2: 0.06, year 3: 0.18, year 4: 0.12). When pooling the years, the recall question was in accordance with the AUSCAN pain in 295 out of 704 (42%) of the intervals, with a Cohen's kappa of 0.13 and a PABAK of -0.16 (table 2). Of the discordant intervals, patients answered the recall more negatively than the change in AUSCAN reflected in 322 intervals (underestimate group) and more positively in 86 intervals (overestimate group). Stratifying for high or low baseline pain did not yield different results (Kappa 0.19 for low baseline pain, 0.08 for high baseline pain) (data not shown). Expressed in percentage agreement, improvement was recalled in 16% of

intervals with improvement scored on the AUSCAN, a stable level of pain was recalled in 35% of intervals with a stable AUSCAN score and a worsening of pain was recalled in 72% of intervals with a worsening indicated by AUSCAN change score.

**Table 2. Change in AUSCAN pain over one year compared with recalled pain over the last year in the HOSTAS**

| | | Change in AUSCAN pain (any) | | | |
|---|---|---|---|---|---|
| | | **Improved (≤-1)** | **Stable (=0)** | **Worsened (≥1)** | **Total** |
| **Recall question** | Worsened – more pain | 143 (20) | 78 (11) | **201 (28)** | *422 (60)* |
| | No change | 101 (14) | **47 (7)** | 63 (9) | *211 (30)* |
| | Better – less pain | **47 (7)** | 8 (1) | 15 (2) | *70 (10)* |
| | Never had this symptom | 3 (0) | 2 (0) | 0 (0) | *5 (1)* |
| | Total | *294 (41)* | *135 (19)* | *279 (39)* | *708 (100)* |

The number and % of concordant answers in bold. In the HOSTAS study, the answer "I have never had this symptom" was given in 5 intervals.

**Table 3. AUSCAN baseline, follow-up and change scores per recall answer in the HOSTAS data**

| | | **N** | **Baseline AUSCAN pain** | **Follow-up AUSCAN pain** | **Change in AUSCAN pain** |
|---|---|---|---|---|---|
| **Recall question** | Worsened – more pain | 422 | 9.7 (3.9) | 10.1 (3.9) | 0.4 (3.0) |
| | No change | 211 | 7.7 (4.1) | 7.1 (3.7) | -0.7 (2.9) |
| | Better – less pain | 70 | 8.1 (4.5) | 5.7 (4.3) | -2.4 (3.7) |
| | Never had this symptom | 5 | 1.8 (1.8) | 1.0 (1.0) | -0.8 (0.8) |

Data are mean (SD)

As some patients provided multiple intervals, stability of concordance over time within patients was assessed visually. Nearly all patients varied between overestimating on the recall question, underestimating on the recall question and answering the recall question concordantly during four years of follow-up in the HOSTAS (figure 1).

Comparing age, sex and BMI between the three groups with the concordant group as index, no differences were found (table 4). The subdomains anxiety and depression of the HADS and the IPQ domains similarly showed no associations with concordance (table 4 and supplementary table A1).

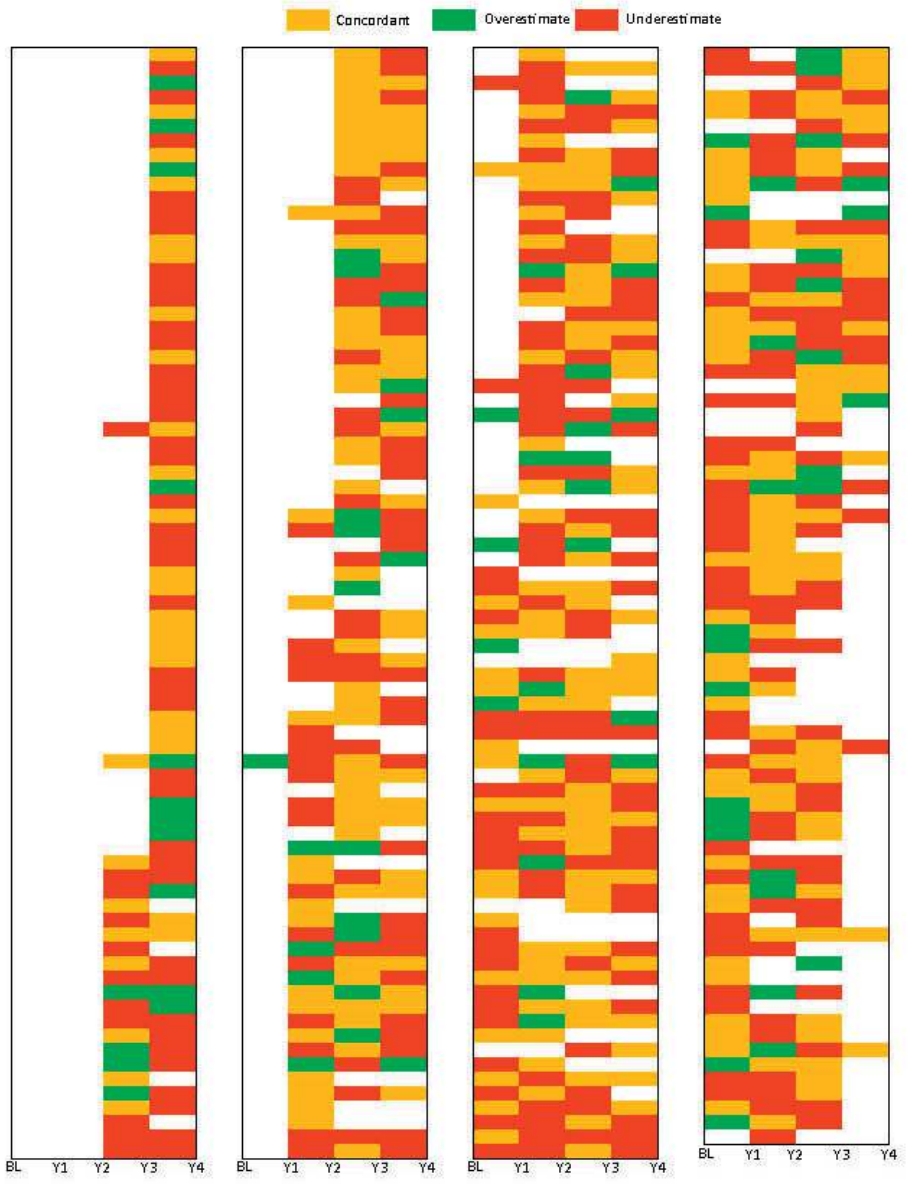


**Figure 1.**
Overview of concordance between recall questions and AUSCAN changes. Every row of four blocks represents a single patient. White blocks represent years of which no information is available. Orange blocks represent years in which measurements were concordant (n=295), red blocks represent years in which the measurement was an underestimation (recall question more negative than the AUSCAN change, n=322) and green blocks represent overestimated years (recall question more positive than the AUSCAN change, n=86).

**Table 4. Association between concordance and age, sex, BMI and HADS scores**

| | Concordant n=295 | Underestimate n=322 | | Overestimate n=86 | |
|---|---|---|---|---|---|
| | | | Odds ratio (95% CI) | | Odds ratio (95% CI) |
| Age at baseline visit; mean (SD) | 60.8 (7.9) | 61.2 (8.0) | 1.00 (0.99-1.02) | 60.8 (7.4) | 1.00 (0.97-1.03) |
| BMI at baseline visit; mean (SD) | 27.6 (4.9) | 27.2 (4.8) | 0.99 (0.96-1.02) | 27.8 (5.2) | 1.01 (0.96-1.06) |
| Sex; N (%) | | | | | |
| Female | 241 (81.7%) | 262 (81.4%) | 1.05 (0.74-1.49) | 67 (77.9%) | 0.76 (0.40-1.44) |
| Male | 54 (18.3%) | 60 (18.6%) | | 19 (22.1%) | |
| Depression; N (%) | | | | | |
| Yes | 34 (11.8%) | 29 (9.2%) | 0.74 (0.48-1.16) | 5 (5.8%) | 0.48 (0.19-1.21) |
| No | 255 (88.2%) | 288 (90.9%) | | 81 (94.2%) | |
| Anxiety; N (%) | | | | | |
| Yes | 47 (16.4%) | 45 (14.2%) | 0.83 (0.55-1.23) | 7 (8.2%) | 0.47 (0.21-1.02) |
| No | 240 (83.6%) | 271 (85.8%) | | 78 (91.8%) | |

Associations with overestimation or underestimation of change in pain. Overestimate = recall question answered more positively than the change in AUSCAN. Underestimate = recall question answered more negatively than the change in AUSCAN. CI = Confidence Interval. HADS = Hospital Anxiety and Depression Scale, scale 0-21, using cutoffs at 8 or higher.

The concordance results from the observational HOSTAS cohort were compared with data from the HOPE trial. Of patients in the HOPE trial, 86 had AUSCAN change scores and recall questions available (mean age 63.5 years, 80% female, mean BMI 27.3 kg/$m^2$, mean AUSCAN pain at baseline 10.7 (SD 3.2)). For details see table 1. In the HOPE trial, improvement of pain was the most frequent answer on the recall question after six weeks (43%). The mean (SD, 95% confidence interval) change in AUSCAN pain score in the group stating their pain had worsened was -1.38 (3.73, -3.41 to +0.64). For no change and improvement of pain the AUSCAN change scores were -1.53 (2.30, -2.28 to -0.78) and -4.89 (4.20, -6.24 to -3.54), respectively (supplementary figure A2). Of 86 intervals, 40 (47%) were in accordance, with a Cohen's kappa of 0.14 and a PABAK of -0.06 (table 5). Splitting the trial into the intervention and placebo arms yielded values of 0.23 and -0.06 for Cohen's kappa, respectively. Stratifying for high or low baseline pain did not yield different results (Kappa 0.10 for low baseline pain, 0.17 for high baseline pain) (data not shown). Expressed in percentage agreement, improvement was recalled in 51% of intervals with improvement scored on the AUSCAN, a stable level of pain was recalled in 33% of intervals with a stable AUSCAN score and a worsening of pain was recalled in 33% of intervals with a worsening indicated by AUSCAN change score.

**Table 5. Change in AUSCAN pain over six weeks compared with change in pain on recall question over the last six weeks in the HOPE study**

| | | Change in AUSCAN pain (any) | | | |
|---|---|---|---|---|---|
| | | **Improved (≤-1)** | **Stable (=0)** | **Worsened (≥1)** | **Total** |
| **Recall question** | Worsened – more pain | 5 (6) | 3 (3) | **5 (6)** | *13 (15)* |
| | No change | 27 (31) | **2 (2)** | 7 (8) | *36 (42)* |
| | Better – less pain | **33 (38)** | 1 (1) | 3 (3) | *37 (43)* |
| | Never had this symptom | 0 | 0 | 0 | *0* |
| | Total | *65 (76)* | *6 (7)* | *15 (17)* | *86 (100)* |

The number and % of concordant answers in bold.

For both the cohort and the trial data, comparing the recall questions with categories of AUSCAN responses classified based on the MCII yielded even lower concordance than comparing the recall questions answers to any change in AUSCAN pain, and lower Cohen's kappa's (supplementary tables A2 and A3)

# CONCLUSION



In this study, annual changes on the AUSCAN pain scale and a recall question asking patients how they recalled the course of their pain between two study visits were compared. Patients in the HOSTAS cohort most frequently regarded the course of their pain as worsening, which was not reflected by changes in AUSCAN pain scores. On average the AUSCAN pain score improved more in the group of patients indicating pain had improved than in the group of patients indicating pain had worsened. There was little concordance between changes in AUSCAN pain scores and the recall question when comparing to absolute change in AUSCAN pain or changes in AUSCAN pain categorized based on the MCII. Similarly, little concordance was found in the HOPE trial data, although in this trial patients most frequently regarded the course of their pain as improving. These data are of great importance for evaluating patient reported pain in trials, which is nearly always the primary outcome measure in hand OA research.

The AUSCAN was previously described to be responsive to change, (23) and the recall question used in this study closely mirrored the anchor question used by the investigator that developed the AUSCAN and the associated MCII. (5) When looking at average change scores per group of the recall question answers, the improve group showed an average decrease, and the deterioration group showed an average increase in pain scores. The largest difference was seen in the group that indicated their pain had im-

proved, with an average decrease of 2.4 points on the AUSCAN. However, categorical changes on the AUSCAN were frequently incongruent with the recall question, both for any change or the MCII as a cutoff. HOSTAS patients most frequently answered that their pain had worsened over the past year, in 422 out of 708 (60%) of the intervals whereas the changes between yearly AUSCAN pain questionnaires indicated worsened pain in 39% of intervals. The other answer categories, stable and decreasing pain, were similarly incongruent. Patients almost always differed in being concordant, overestimating or underestimating the answer to the recall questions compared with the change in AUSCAN pain between measured intervals. Although data from the HOPE trial showed improved pain more often on both the recall question and the change in AUSCAN scores, concordance was still low. In line with the overall concordance data, the percentage agreement between changes in the AUSCAN with the recall question were low and had a wide range.

This discordance could be caused by the different structures of the questions used, with the AUSCAN questions centered around specific ways of using the hands, whereas the recall question only concerns pain, leaving interpretation to the patient. The question can be interpreted as maximal pain level, average pain level, or even as the frequency with which patients are forced to alter behavior due to pain. However, since both the HOSTAS and HOPE studies focus on hand pain, this will be the most important determinant of the recalled pain. Another important difference between the questionnaires is that the AUSCAN has a maximal score, whereas the recall question does not. Reaching this ceiling, the worst possible score, precludes increasing pain scores on the AUSCAN, but patients can still experience increasing pain and state so on the recall question. Imprecise recall may also play a role, as recalling the level of pain experienced a year ago is very difficult. As the recall question is explicitly based on a recalled experience and the AUSCAN is not, this could further drive the discordance. The comparison between the recall question and the categorized MCII change in AUSCAN could also be distorted by the fact that the cutoff was determined solely for improvement. Patients may regard deterioration differently, necessitating a different cutoff for deterioration compared to improvement. Similar differences have been found for fatigue. (24) To properly study pain development, either increases or decreases, it may thus be valuable to establish a minimal clinical difference for deterioration as well. Finally, pain in hand OA may fluctuate on a narrow scale within a patient. This may potentially further drive the discordance, as the AUSCAN is more likely to respond to small fluctuations than the recall question, due to the difference in time scales. This difference in responsiveness to these fluctuations could hypothetically drive the discordance further. However, repeating the analysis with the MCII as a cutoff rather than any change yielded similar results. This makes it unlikely that the small fluctuations had an effect in our analysis.

There also are several known methodological drawbacks of questionnaires that may have caused discordance. It has been described for global perceived effect scales that patients base their answers on a current state when asked how a symptom has developed over a period of time, rather than on the change. (25) Should a patient experience the condition as equally negative at two consecutive visits, this could lead to stable AUSCAN scores but worsened pain on the recall question. Both questionnaires may also be affected by response shift and changes in behavior, which affects the pain answer that patients provide. (7)

Previous work has described that treatment perceptions can also be influenced to change outcomes in pain relief, emphasizing the clinical importance of these perceptions. (19) Prior to the analyses, we thus hypothesized that more negative illness perceptions, such as attributing more consequences to the hand OA, would be associated with more negative recall compared to the measured change in pain. However, we did not find any associations between the various IPQ domains and the concordance between the recall question and change in AUSCAN pain, when measuring the IPQ at moment the recall question was collected. Mental wellbeing, measured with the HADS subscales anxiety and depression, was similarly expected to be associated with more negative recall. This association was not found either. It should be noted that the low number of patients per stratum and thus the low power of this analysis necessitates confirmation in future studies. Stratification on high or low baseline pain did not affect the found agreement between the recall question and change in AUSCAN.



The setting in which a study is conducted may also influence the concordance between pain measurement by subsequent questionnaires or a recall question. As such, we compared trial data with cohort data in this study. Most patients in the HOPE stated improved pain on both questions, causing most concordant patients to also be in this group, as opposed to the HOSTAS data. However, there was still low overall concordance in the clinical trial. This might be explained in various ways. Firstly, previous literature described the effects of patient expectations on treatment outcomes and the placebo effect, supporting this. (8, 26, 27) If we take the recall question to be true, this would mean that the decreases in pain seen in trials are potentially overstated (as seen with the placebo effect, and accounted for by using control groups) and that in cohorts it may be the increases in pain that are overstated. Secondly, there was a large difference in the interval time between the studies (1 year vs 6 weeks). This could have affected pain recall. More importantly, one might expect the concordance between pain questionnaires and recall to be better over a shorter period of time. However, the Cohen's kappa values obtained were similar. It may thus be that the discrepancy between pain recall and repeated questionnaires is already present after a shorter time period than

6 weeks. Based on this, future studies should investigate the optimal amount of time to study pain recall, since many questionnaires include an inherent recall component (e.g. the AUSCAN measures pain over the past 48 hours). This knowledge could also aid in establishing cutoffs for questionnaires using anchoring methods, which is often done, and recommended in rheumatology by the Outcome Measures in Rheumatology (OMERACT). (5, 6, 28-30) Thirdly, the AUSCAN and recall question may capture different elements of the pain experience, as described earlier in the discussion. Differences between the AUSCAN and the recall can become more visible depending on the setting, in case these elements differ between the study settings.

Our study has various strengths. The large cohort allowed for the investigations of numerous intervals, illustrating the performance of both a pain questionnaire and a recall question in a natural setting over multiple years. The trial on the other hand allowed us to study the influence the placebo effect has on these pain measurements and their concordance. Using both allowed us to compare between the different settings, contrasting long follow-up with short follow-up and a cohort setting with a trial setting. There are also some weaknesses. As stated, the study was not powered to investigate the associations between mental wellbeing and concordance of pain measurements. We also did not have the data to further investigate the optimal recall interval for pain, which would be of great additional value.

To conclude, these findings indicate that research on pain development may not reliably estimate changes in pain recalled by the patient, and that recalled changes in pain and changes in pain need not be the same. This discordance has important consequences for the chance of finding successful interventions to alleviate patient reported symptoms, as it highlights the difficulty in accurately measuring pain. These mechanisms may contribute to the large amount of negative trial results in the OA field. A new pain assessment tool, specifically aimed at assessing long term pain changes, may be required. Alternatively, a recall question and a change score may both be needed to more properly assess changes in pain.

***Author contributions***

CvdM, LAvdS and MK designed the study. FPBK and MK collected the data. CvdM, LAvdS, and MK analysed the data. CvdM, LAvdS, FPBK, MN, FRR and MK interpreted the data and wrote and reviewed the report. All authors approved the final version of the manuscript.

***Role of the funding source***

For the current study, MK reports funding from the SKMS, paid to the institution.

***Competing interest statement***

MK reports the following, all outside the current study: Grants from IMI-APPROACH and the Dutch Arthritis Society, paid to the institution. Royalties or licences from Wolters Kluwer and Springer Verlag, paid to the institution. Fees for consulting/advisory boards by Abbvie, Kiniksa, Galapagos, CHDR, Novartis, UCB, GSK, all paid to the institution. Payment or honoraria for lectures or presentations from Galapagos, Novartis and Jansen, paid to the institution. Roles on the OARSI board (member), EULAR council (member advocacy committee EULAR) and presidency of the Dutch Society for Rheumatology. For the current study, MK reports funding from SKMS, paid to the institution. MN reports honoraria for lectures by Amgen and Grunenthal. The other authors report no competing interests.

***Data availability statement***

No data are available.

# REFERENCES


1. Kloppenburg M, Kroon FP, Blanco FJ, Doherty M, Dziedzic KS, Greibrokk E, et al. 2018 update of the EULAR recommendations for the management of hand osteoarthritis. Ann Rheum Dis. 2019;78(1):16-24.
2. Kloppenburg M, Bøyesen P, Visser AW, Haugen IK, Boers M, Boonen A, et al. Report from the OMERACT Hand Osteoarthritis Working Group: Set of Core Domains and Preliminary Set of Instruments for Use in Clinical Trials and Observational Studies. J Rheumatol. 2015;42(11):2190-7.
3. Chiarotto A, Kaiser U, Choy E, Christensen R, Conaghan PG, Cowern M, et al. Pain Measurement in Rheumatic and Musculoskeletal Diseases: Where To Go from Here? Report from a Special Interest Group at OMERACT 2018. J Rheumatol. 2019;46(10):1355-9.
4. Bellamy N, Campbell J, Haraoui B, Buchbinder R, Hobby K, Roth JH, et al. Dimensionality and clinical importance of pain and disability in hand osteoarthritis: Development of the Australian/Canadian (AUSCAN) Osteoarthritis Hand Index. Osteoarthritis Cartilage. 2002;10(11):855-62.
5. Bellamy N, Hochberg M, Tubach F, Martin-Mola E, Awada H, Bombardier C, et al. Development of multinational definitions of minimal clinically important improvement and patient acceptable symptomatic state in osteoarthritis. Arthritis Care Res (Hoboken). 2015;67(7):972-80.
6. Kvien TK, Heiberg T, Hagen KB. Minimal clinically important improvement/difference (MCII/MCID) and patient acceptable symptom state (PASS): what do these concepts mean? Ann Rheum Dis. 2007;66 Suppl 3(Suppl 3):iii40-1.
7. Vanier A, Oort FJ, McClimans L, Ow N, Gulek BG, Böhnke JR, et al. Response shift in patient-reported outcomes: definition, theory, and a revised model. Qual Life Res. 2021;30(12):3309-22.
8. Laferton JA, Kube T, Salzmann S, Auer CJ, Shedden-Mora MC. Patients' Expectations Regarding Medical Treatment: A Critical Review of Concepts and Their Assessment. Front Psychol. 2017;8:233.
9. Coughlin SS. Recall bias in epidemiologic studies. J Clin Epidemiol. 1990;43(1):87-91.
10. Altman R, Alarcón G, Appelrouth D, Bloch D, Borenstein D, Brandt K, et al. The American College of Rheumatology criteria for the classification and reporting of osteoarthritis of the hand. Arthritis Rheum. 1990;33(11):1601-10.
11. Damman W, Liu R, Kroon FPB, Reijnierse M, Huizinga TWJ, Rosendaal FR, et al. Do Comorbidities Play a Role in Hand Osteoarthritis Disease Burden? Data from the Hand Osteoarthritis in Secondary Care Cohort. J Rheumatol. 2017;44(11):1659-66.
12. Kroon FPB, Kortekaas MC, Boonen A, Böhringer S, Reijnierse M, Rosendaal FR, et al. Results of a 6-week treatment with 10 mg prednisolone in patients with hand osteoarthritis (HOPE): a double-blind, randomised, placebo-controlled trial. The Lancet. 2019;394(10213):1993-2001.
13. Verbruggen G, Veys EM. Numerical scoring systems for the anatomic evolution of osteoarthritis of the finger joints. Arthritis Rheum. 1996;39(2):308-20.
14. Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis. Ann Rheum Dis. 1957;16(4):494-502.
15. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand. 1983;67(6):361-70.
16. Moss-Morris R, Weinman J, Petrie K, Horne R, Cameron L, Buick D. The Revised Illness Perception Questionnaire (IPQ-R). Psychology & Health. 2002;17(1):1-16.
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.
18. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. 1993;46(5):423-9.

19. Bingel U, Wanigasekera V, Wiech K, Ni Mhuircheartaigh R, Lee MC, Ploner M, et al. The effect of treatment expectation on drug efficacy: imaging the analgesic benefit of the opioid remifentanil. Sci Transl Med. 2011;3(70):70ra14.
20. Neogi T. The epidemiology and impact of pain in osteoarthritis. Osteoarthritis and cartilage. 2013;21(9):1145-53.
21. Mulrooney E, Neogi T, Dagfinrud H, Hammer HB, Pettersen PS, Gaarden TL, et al. The associations of psychological symptoms and cognitive patterns with pain and pain sensitization in people with hand osteoarthritis. Osteoarthr Cartil Open. 2022;4(2):100267.
22. Bjelland I, Dahl AA, Haug TT, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale. An updated literature review. J Psychosom Res. 2002;52(2):69-77.
23. Visser AW, Bøyesen P, Haugen IK, Schoones JW, van der Heijde DM, Rosendaal FR, et al. Instruments Measuring Pain, Physical Function, or Patient's Global Assessment in Hand Osteoarthritis: A Systematic Literature Search. J Rheumatol. 2015;42(11):2118-34.
24. Schwartz AL, Meek PM, Nail LM, Fargo J, Lundquist M, Donofrio M, et al. Measurement of fatigue. determining minimally important clinical differences. J Clin Epidemiol. 2002;55(3):239-44.
25. 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. J Clin Epidemiol. 2010;63(7):760-6.e1.
26. Ongaro G, Kaptchuk TJ. Symptom perception, placebo effects, and the Bayesian brain. Pain. 2019;160(1):1-4.
27. Price DD, Finniss DG, Benedetti F. A comprehensive review of the placebo effect: recent advances and current thought. Annu Rev Psychol. 2008;59:565-90.
28. Tubach F, Wells GA, Ravaud P, Dougados M. Minimal clinically important difference, low disease activity state, and patient acceptable symptom state: methodological issues. The Journal of Rheumatology. 2005;32(10):2025-9.
29. Wells G, Anderson J, Beaton D, Bellamy N, Boers M, Bombardier C, et al. Minimal clinically important difference module: summary, recommendations, and research agenda. J Rheumatol. 2001;28(2):452-4.
30. Wells G, Beaton D, Shea B, Boers M, Simon L, Strand V, et al. Minimal clinically important differences: review of methods. J Rheumatol. 2001;28(2):406-12.

# APPENDIX 1. SUPPLEMENTARY METHODS

Illness perceptions and attributions were investigated using the Illness Perception Questionnaire (IPQ). The IPQ consists of 9 domains: Identity (how many other symptoms are present and if the patient associates these with their hand OA, scored 0-14), timeline acute/chronic (whether the disease is regarded as chronic, 6-30), timeline cyclical (whether the disease is experienced as fluctuating, 4-20), consequences (perceived severity of consequences of the disease, 6-30), personal control (perceived personal control over the disease, 6-30), treatment control (perceived control the treatment has on the disease, 5-25), emotional representations (amount and severity of negative emotions experienced due to the disease, 6-30), illness coherence (how well the patient understands the disease, 5-25) and attributions (which factors patients think caused their disease, further divided into psychological, risk factors, immunity and chance domains). For all domains, higher scores indicate a stronger belief in the investigated construct. For illness coherence, a higher score indicates better understanding. (1)

# REFERENCES

1. Moss-Morris R, Weinman J, Petrie K, Horne R, Cameron L, Buick D. The Revised Illness Perception Questionnaire (IPQ-R). Psychology & Health. 2002;17(1):1-16.

# APPENDIX 2. SUPPLEMENTARY RESULTS

**Table A1. Association between illness perceptions and concordance in recall and changes in AUSCAN**

| | Concordant n=148 | Underestimate n=190 | | Overestimate n=42 | |
|---|---|---|---|---|---|
| **Identity** | **N (%)** | **N (%)** | **OR (95% CI)** | **N (%)** | **OR (95% CI)** |
| 0-3 | 44 (29.9) | 60 (31.9) | - | 15 (36.6) | - |
| 4-5 | 61 (41.5) | 77 (41.0) | 0.92 (0.54-1.58) | 15 (36.6) | 0.50 (0.20-1.27) |
| 6-14 | 42 (28.6) | 51 (27.1) | 0.89 (0.51-1.55) | 11 (26.8) | 0.77 (0.32-1.82) |
| **Timeline Chronic** | | | | | |
| 6-25 | 41 (28.1) | 60 (31.6) | - | 16 (39.0) | - |
| 26-29 | 45 (30.8) | 64 (33.7) | 0.97 (0.56-1.68) | 14 (34.2) | 0.73 (0.29-1.86) |
| 30 | 60 (41.1) | 66 (34.7) | 0.75 (0.44-1.28) | 11 (26.8) | 0.43 (0.18-1.04) |
| **Consequences** | | | | | |
| 6-13 | 42 (28.8) | 68 (35.8) | - | 17 (41.5) | - |
| 14-17 | 45 (30.8) | 58 (30.5) | 0.79 (0.46-1.38) | 12 (29.3) | 0.47 (0.18-1.24) |
| 18-30 | 59 (40.4) | 64 (33.7) | 0.66 (0.40-1.10) | 12 (29.3) | 0.51 (0.22-1.20) |
| **Personal control** | | | | | |
| 6-16 | 49 (33.6) | 61 (32.1) | - | 8 (19.5) | - |
| 17-19 | 54 (37.0) | 78 (41.1) | 1.16 (0.70-1.93) | 16 (39.0) | 1.64 (0.69-3.92) |
| 19.2-30 | 43 (29.5) | 51 (26.8) | 0.95 (0.55-1.64) | 17 (41.5) | 2.28 (0.92-5.65) |
| **Treatment control** | | | | | |
| 5-11 | 40 (27.6) | 54 (28.6) | - | 14 (34.2) | - |
| 12-14 | 62 (42.8) | 73 (38.6) | 0.87 (0.51-1.51) | 12 (29.3) | 0.62 (0.29-1.35) |
| 15-20 | 43 (29.7) | 62 (32.8) | 1.07 (0.61-1.86) | 15 (36.6) | 0.84 (0.36-2.00) |
| **Illness coherence** | | | | | |
| 5-18 | 50 (34.0) | 68 (35.8) | - | 13 (31.7) | - |
| 19-20 | 48 (32.7) | 51 (26.8) | 0.78 (0.46-1.34) | 13 (31.7) | 1.08 (0.43-2.73) |
| 21-25 | 49 (33.3) | 71 (37.4) | 1.07 (0.64-1.79) | 15 (36.6) | 1.14 (0.43-3.06) |
| **Timeline cyclical** | | | | | |
| 4-12 | 55 (37.4) | 69 (36.3) | - | 15 (36.6) | - |
| 13-15 | 49 (33.3) | 61 (32.1) | 0.99 (0.57-1.71) | 16 (39.0) | 1.17 (0.46-2.99) |
| 16-20 | 43 (29.3) | 60 (31.6) | 1.11 (0.67-1.84) | 10 (24.4) | 1.00 (0.41-2.41) |
| **Emotional representations** | | | | | |
| 6-11 | 42 (28.6) | 64 (33.7) | - | 15 (36.6) | - |
| 12-14 | 55 (37.4) | 70 (36.8) | 0.83 (0.50-1.40) | 14 (34.2) | 0.69 (0.32-1.49) |
| 15-30 | 50 (34.0) | 56 (29.5) | 0.73 (0.43-1.26) | 12 (29.3) | 0.71 (0.30-1.64) |

Association with overestimation or underestimation of change in pain, reported on the recall questions, with change in AUSCAN as comparison. Overestimate = recall question answered more positively than the change in AUSCAN. Underestimate = recall question answered more negatively than the change in AUSCAN. CI = Confidence Interval. IPQ = Illness Perception Questionnaire.

**Table A2. Recall questions versus AUSCAN pain changes categorized according to MCII (1.6) in the HOSTAS cohort**

| | | Change in AUSCAN pain | | | |
|---|---|---|---|---|---|
| | | **Improved (≤-2)** | **Stable (>-2 & <2)** | **Worsened (≥2)** | **Total** |
| **Recall question** | Worse – more pain | 98 (14) | 173 (24) | **151 (21)** | 422 (60) |
| | No change | 74 (10) | **96 (14)** | 41 (6) | 211 (30) |
| | Better – less pain | **39 (6)** | 20 (3) | 11 (2) | 70 (10) |
| | Never had this symptom | 1 (0) | 4 (1) | 0 (0) | 5 (1) |
| | Total | 212 (30) | 293 (41) | 203 (29) | 708 (100) |

The number and % of concordant answers in bold.

Yearly kappa's were all very low (year 1: 0.06, year 2: 0.07, year 3: 0.23, year 4: 0.09). When pooling the years, the recall question was in accordance with the AUSCAN pain in 286 out of 704 (40%) of the intervals, with a Cohen's kappa of 0.12.

**Table A3. Recall questions versus AUSCAN pain changes categorized according to MCII (1.6) in the HOPE trial**

| | | Change in AUSCAN pain | | | |
|---|---|---|---|---|---|
| | | **Improved (≤-2)** | **Stable (>-2 & <2)** | **Worsened (≥2)** | **Total** |
| **Recall question** | **Worse – more pain** | 5 (6) | 7 (8) | **1 (1)** | 13 (15) |
| | **No change** | 20 (23) | **11 (13)** | 5 (6) | 36 (42) |
| | **Better – less pain** | **30 (35)** | 5 (6) | 2 (2) | 37 (43) |
| | **Never had this symptom** | 0 | 0 | 0 | 0 |
| | **Total** | 55 (64) | 23 (27) | 8 (9) | 86 (100) |

The number and % of concordant answers in bold.

Cohen's kappa of 0.15. Concordance in 42/86 intervals (46%). When split by treatment arm, the prednisolone group yielded a Cohen's kappa of 0.19, the placebo group of -0.15.

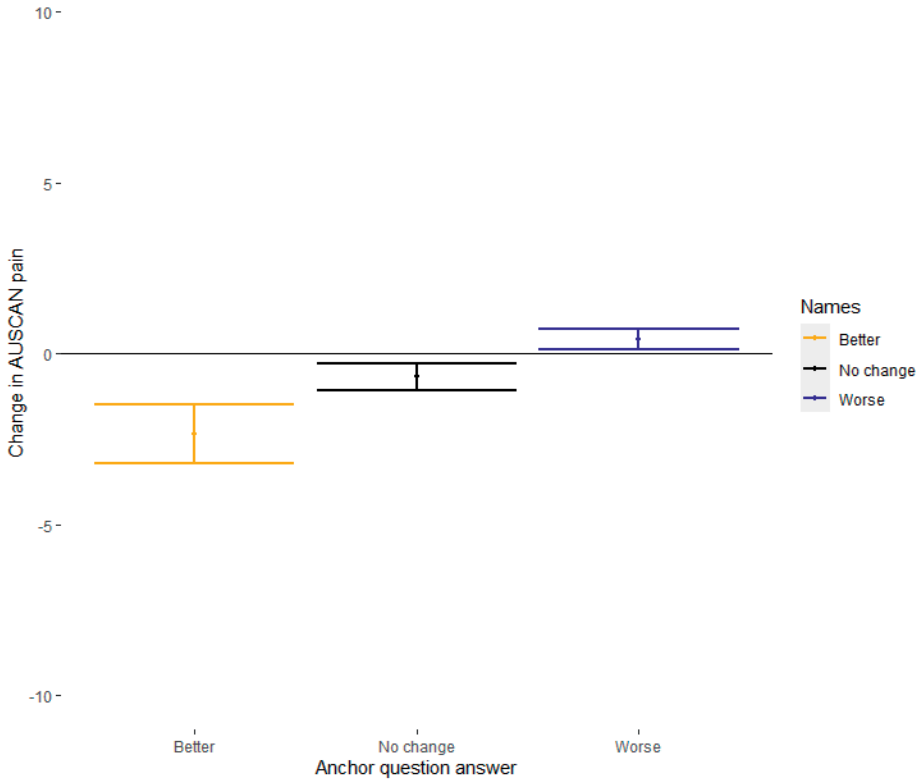


**Figure A1. Change scores of AUSCAN pain per recall question answer category, from the HOSTAS study**

Mean change in AUSCAN pain (dots) with 95% confidence interval (plungers). Data pooled from all years, with data from patients stating pain had improved on the recall question in yellow, patients stating pain remained stable in black and patients stating pain had worsened in blue.

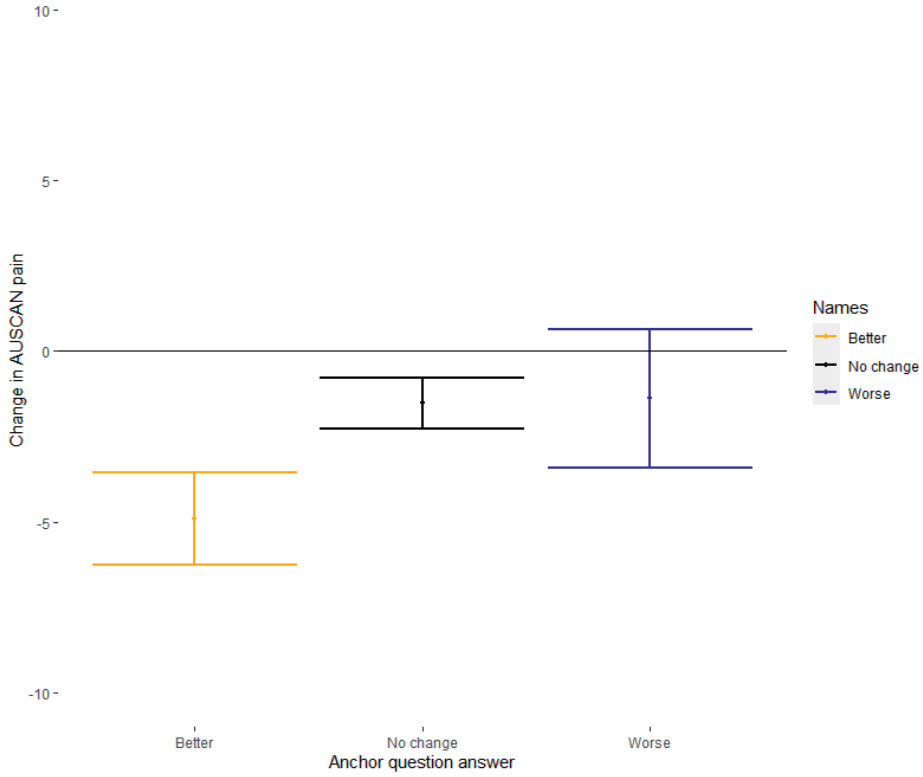


**Figure A2. Change scores of AUSCAN pain per recall question answer category, from the HOPE trial**
Mean change in AUSCAN pain (dots) with 95% confidence interval (plungers). Data from patients stating pain had improved on the recall question in yellow, patients stating pain remained stable in black and patients stating pain had worsened in blue.