Hernia (2023) 27:41–54
https://doi.org/10.1007/s10029-022-02689-5

ORIGINAL ARTICLE

Recall bias in pain scores evaluating abdominal wall and groin pain surgery

W. A. R. Zwaans[1,2,3] · J. A. de Bruijn[1] · J. P. Dieleman[4] · E. W. Steyerberg[5] · M. R. M. Scheltinga[1,2] · R. M. H. Roumen[1,2]

Received: 18 July 2022 / Accepted: 15 September 2022 / Published online: 18 October 2022

Abstract

Purpose To determine whether levels of pre-operative pain as recalled by a patient in the post-operative phase are possibly overestimated or underestimated compared to prospectively scored pain levels. If so, a subsequent misclassification may induce recall bias that may lead to an erroneous effect outcome.

Methods Data of seven retrospective cohort studies on surgery for chronic abdominal wall and groin pain using three different pain scores were systematically analyzed. First, it was assessed whether retrospectively acquired pre-operative pain levels, as scored by the patient in the post-operative phase, differed from prospectively obtained pre-operative pain scores. Second, it was determined if errors associated with retrospectively obtained pain scores potentially lead to a misclassification of treatment outcome. Third, a meta-analysis established whether recall misclassifications, if present, affected overall study conclusions.

Results Pain data of 313 patients undergoing remedial surgery were evaluated. The overall prevalence of misclassification due to a recall error was 13.7%. Patients not benefitting from surgery ('failures') judged their pre-operative pain level as more severe than it actually was. In contrast, patients who were pain free after remedial surgery ('successes') underestimated pre-operative pain scores. Recall misclassifications were significantly more present in failures than in successful patients (odds ratio 2.4 [95% CI 1.2–4.8]).

Conclusion One in seven patients undergoing remedial groin surgery is misclassified on the basis of retrospectively obtained pre-operative pain scores (success instead of failure, or vice versa). Misclassifications are relatively more present in failures after surgery. Therefore, the effect size of a therapy erroneously depends on its success rate.

Keywords Post-operative pain · Chronic pain · Pain measurement · Bias · Mental recall

Introduction

Bias in epidemiological studies can lead to incorrect or inconsistent conclusions. A widely accepted bias classification recognizes three main types: selection bias, information bias, and confounding bias [1]. Correction during data analysis is only possible for the latter type, and thus, identification and avoidance of selection bias and information bias in the scientific research are crucial.

Recall bias is a subtype of information bias which commonly arises in retrospective studies but may also occur in prospective cohort studies and even randomized controlled trials [2]. This type of bias alludes on the fact that study participants recall information either inaccurately or incompletely. If distributed unevenly across study groups, it can affect the study's internal validity [2–5]. The distribution of recall errors determines the direction of bias [4]. Factors

✉ W. A. R. Zwaans
willemzwaans@gmail.com

1 Department of General Surgery, Máxima Medical Centre, De Run 4600, P.O. Box 7777, 5500 MB Veldhoven, Eindhoven, The Netherlands

2 SolviMáx Center of Excellence for Abdominal Wall and Groin Pain, Eindhoven, The Netherlands

3 NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University Medical Centre+, Maastricht, The Netherlands

4 Máxima Medical Centre Academy, Máxima Medical Centre, Veldhoven, The Netherlands

5 Department of Public Health, Erasmus Medical Center, Rotterdam, The Netherlands

known to affect the reliability of recalled information include the state of memory [6–8], duration of the retention interval [9, 10], patient demographics [11], and the occurrence of events [2, 4].

Nowadays, self-reported pain intensity is used as an outcome measure in research more often than before [11]. Given the importance of self-reported pain scores in clinical research, its reliability is critical. Literature on the influence of recall bias on the assessment of pain scores in clinical studies is limited and contradictory [8, 12–19]. Currently, it is unclear if recall error just results in inaccurate measurements or may also lead to a significantly altered outcome of retrospective studies on pain.

The main objective of the present study was to determine the influence of recall bias in surgical studies having pain intensity as the primary outcome. We compared retrospectively collected pre-operative pain scores with prospectively collected pain scores to estimate the risk of recall error, recall bias, and erroneous conclusions due to recall bias.

Methods

Setting

The analysis was performed at Máxima Medical Centre (MMC), a teaching hospital in Veldhoven/Eindhoven, The Netherlands. In recent years, a sub-department of general surgery (SolviMáx) has specialized on the treatment of patients with chronic abdominal wall pain and groin pain syndromes. The number of evaluated patients has expanded over the years from around 250 in 2012 to more than 1200 in 2021. The present study did not require permission from a medical ethics committee, since it involved evaluation of previously collected anonymous data.

Study design

Study eligibility criteria

We included data from all retrospective MMC cohort studies reporting results of surgical interventions for abdominal wall pain or groin pain up to 2015. A study was considered eligible when pre-operative self-reported pain scores were collected retrospectively using questionnaires or structured interviews (recalled data). These studies used retrospectively obtained pain scores, because prospective pain scores were often missing from the patients' electronic hospital files. The original study databases were used for analysis. Studies that only used prospective data were excluded.

Eligibility criteria for participants

A subset of patients of the included studies was used for the current study. Only patients operated for chronic abdominal wall or groin pain in MMC were eligible. For individual patients, both pre-operative and post-operative prospectively registered self-reported pain scores had to be retrievable from routine electronic patient records. We thereby excluded patients whose prospective pain scores were missing. Pre-operative pain scores had to be collected retrospectively after surgery using questionnaires or structured interviews, as part of the study protocol. Furthermore, the prospectively and retrospectively applied pain scales had to be identical. Assembly of the treatment outcome (success or failure, Table 1) had to be reproducible from the documented post-operative and pre-operative pain scores. If patients participated in multiple studies, only data from the first study were used. Losses to follow-up were excluded.

Data collection process

Retrospectively obtained pre-operative pain scores were extracted from the original study databases and were considered as 'potentially biased scores'. Pre-operative and post-operative prospectively obtained pain scores were extracted from electronic patient records and were considered as 'unbiased scores'. Patient characteristics, type of pain treatment, and effectiveness of pain surgery were collected from the original study databases. For the purpose of the present analysis, the original study databases were combined into a new database.

Table 1 Definitions of severe pain as defined by popular pain scores and its relation with outcome following surgical intervention

Pain score	Successful treatment	Failed treatment
Numerical rating scale (NRS) (0–10)	Post-operative NRS ≥ 50% reduction compared to pre-operative NRS	Results not meeting the criteria for successful treatment
Visual analog scale (VAS) (0–100)	Post-operative VAS reduction of at least 50% compared to pre-operative VAS	Results not meeting the criteria for successful treatment
5-point verbal rating scale (VRS) (1–5)	A minimal 2-point reduction using VRS at the post-operative time point compared to pre-operative VRS	Results not meeting the criteria for successful treatment

Pain scales

Three different pain scales were used in the selected studies. The numerical rating scale (NRS) instructs individuals to score pain on a 0 (no pain) to 10 (unbearable pain) scale. It is a commonly used one-dimensional pain scale that is easy to use for both clinical and research purposes [20]. The visual analog scale (VAS) uses a horizontal line of 100 mm in length, the left end point (0 mm) representing absence of pain, and the right end point (100 mm) indicating unbearable pain [21, 22]. The patient is asked to place a mark on the line that corresponds to the intensity of the experienced pain. The verbal rating scale (VRS) consists of a five-point [1–5] categorical Likert-like scale that uses commonly used words to describe pain (Fig. 1). Patients are asked to choose the words that best describe their pain [23–25].

Outcome definitions

The primary outcome was the magnitude and direction of recall bias. Recall bias was defined as a systematic difference in overall treatment effect (i.e. study conclusion) between analyses using retrospective data and analyses using prospective data. Definitions of success or failure after a surgical intervention are described in Table 1. A recall error was defined as a discrepancy between prospectively and retrospectively obtained, pre-operative pain scores. Recall misclassification was defined as a recall error leading to misclassification of treatment outcome. A negative recall misclassification indicates that treatment success was falsely classified as failure based on retrospective scores, while prospective scores indicated a success. Conversely, a positive recall misclassification indicated that treatment failure was misclassified as treatment success.

Stepwise approach

A stepwise approach was used. First, accuracy of retrospectively obtained self-reported pre-operative pain scores was assessed by comparing these values with prospective *pre*-operative pain scores (recall error) within studies. Second, individual treatment outcomes were dichotomized as success or failure. Treatment outcome was classified as success or failure using the retrospective or prospective *pre*-operative scores for all individual patients, as compared to post-operative pain scores (Table 1).

Misclassification of the treatment outcome due to the use of retrospective pain scores in individual patients was identified. The prevalence of recall misclassification within studies was calculated as the proportion of patients with misclassification of the treatment outcome. This was performed for positive and negative misclassifications, both separate as well as together. In addition, the net direction of the misclassification was presented as the difference in proportion of negative recall misclassification and positive recall misclassification. Third, a meta-analysis was performed to investigate the difference in risk of misclassification between failures and successes. Significant differences point at factors leading to differential misclassification, resulting in an actual recall bias. Finally, differences in recall bias between different pain scales (NRS, VAS, and VRS) were analyzed.

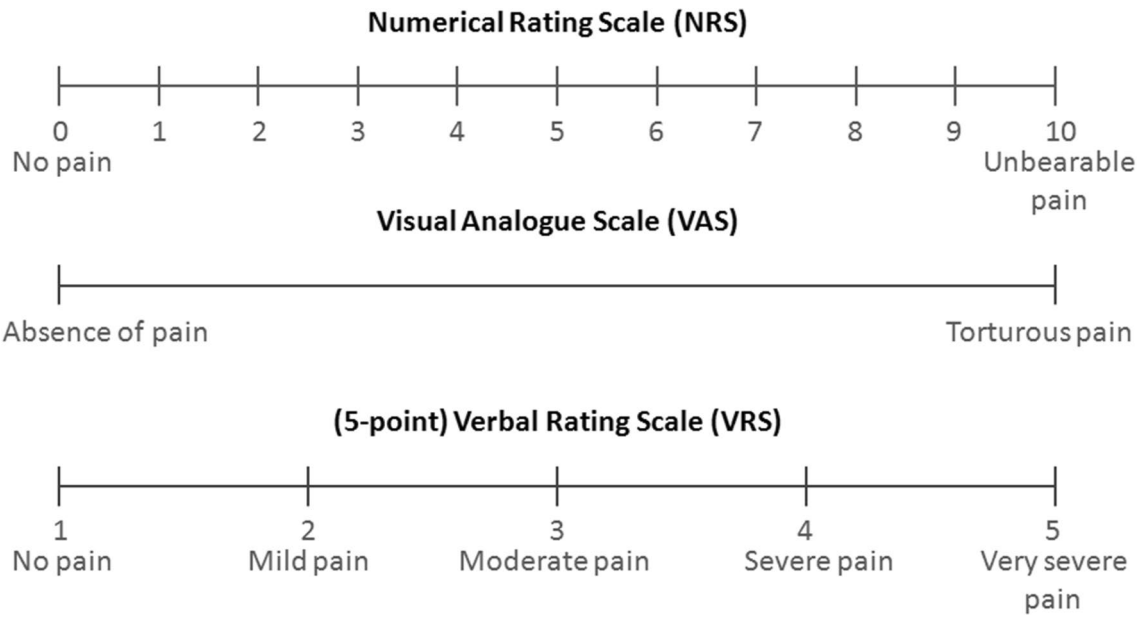

Fig. 1 Various pain scales that were used in the present analysis

Statistical methods

Data were analyzed using IBM SPSS Statistics 22 software (SPSS Inc., Chicago, Illinois, United States). Mean ± SD or median [interquartile range; IQR] of the prospective and retrospective pre-operative pain scores was calculated per study, as appropriate. Recall errors were assessed by comparing means (or medians) of these pain scores within studies. A paired Student's t test (normal distribution) or Wilcoxon signed-rank test (non-normal distributed pain scores) was used to test statistical differences. Statistical significance was accepted at a two-sided p value of ≤ 0.05 and confirmed the presence of recall errors within studies. In addition, the absolute intraclass correlation coefficient (ICC) was calculated per study using the two-way mixed model.

To assess the influence of treatment effect on the risk of recall bias due to retrospective relative to prospective collection of pain scores, odds ratios (ORs) and corresponding 95% confidence intervals (95%CI) were calculated using Review Manager version 5.3 (The Cochrane Collaboration, London, UK). The number of negative and positive recall misclassifications and the total number of included successes and failures were entered, so that an OR > 1 points a higher risk of bias in failures than in successes. Hence, an OR > 1 at a bias toward a more positive treatment result with recalled pain scores than with prospective pain scores. ORs and 95% CIs were depicted in a forest plot. A subgroup analysis was performed per pain scale. The random-effects model was used for pooling the results. Statistical heterogeneity of studies regarding the recall bias was evaluated by Chi-square test and calculation of the inconsistency (I^2).

To explore the presence of selection bias in our analysis, we compared baseline characteristics of included and excluded patients. Bivariate and continuous data were tested using the Chi-square test and independent t test, respectively.

Results

Studies and participants

Seven studies on surgical treatment of patients with chronic abdominal wall or groin pain fulfilled the inclusion criteria [23–29]. Patient recruitment for these studies occurred between December 2015 and August 2000 (Table 2). Four patients who did not undergo an intervention were excluded [26]. Sixty-six overlapping patients were excluded, as well. Prospective, retrospective, or post-operative pain scores were missing in 291 patients, leading to analysis of 313 patients having complete pain data sets were analyzed in the present study (Fig. 2). Median patient follow-up was 21 months [IQR 12–30].

Main results

Is a recall error present in studies?

Recall errors were present in all seven studies, but statistical significant differences between prospective and retrospective pain scores were found in only four studies (Table 3). Agreement between pre-operative prospective and retrospective pain scores within patients as expressed by ICC was fair in three studies, whereas a moderate agreement was found in the four remaining studies.

What is the actual misclassification of recall and is it differential?

The overall prevalence of recall misclassification was 13.7% [95% CI 10.3–18.0, range 3.3–36.4%] (Table 3). Figure 3 represents the percentage of recall misclassifications per study. The net amount of recall misclassification varied considerably (ranging from 1 to 33%) but all directed toward more positive misclassification if using retrospective pain scores.

In general, patients failing treatment tended to recall pre-operative NRS and VRS pain scores as significantly higher than they actually were as indicated by the prospectively obtained pain scores (NRS prospective median 7.5 [IQR 6.9–8.0] vs. recall median 8.0 [IQR 7.4–9.0]; VRS prospective median 4.0 [IQR 3.0–4.0] vs. VRS recall median 4.0 [IQR 4.0–4.5]). In contrast, the reverse was found regarding the VAS pain scale (VAS prospective mean 70.7 ± 13.9 vs. VAS recall mean 61.8 ± 12.9).

On the other hand, patients with a successful treatment outcome recorded lower pre-operative pain scores if assessed retrospectively (VAS prospective mean 71.1 ± 15.4 vs. VAS recall mean 55.8 ± 15.4; VRS prospective median 5.0 [IQR 4.0–5.0] vs. VRS recall median 4.0 [IQR 4.0–5.0]), with exception of the NRS pain scale (NRS prospective median 7.5 [IQR 7.0–8.0] vs. recall median 8.0 [IQR 7.5–9.0]).

What is the actual recall bias in studies?

ORs of recall misclassification in studies are presented in Fig. 4. An OR > 1 indicates more positive recall misclassification (i.e., patients failing treatment if based on the prospective pre-operative pain score but having a successful outcome if based on the retrospective pre-operative pain score). On the contrary, an OR < 1 demonstrates more negative recall misclassification (i.e., patients successfully treated using prospective pre-operative pain score but having an unsuccessful outcome based on retrospective pre-operative pain score). ORs varied considerably among the studies. The overall OR of 2.4 [95% CI 1.1–4.8] was significant,

Table 2 Characteristics of reports ($n = 7$) that were analyzed in the present study

Study	Year	*n* (pt)	Objective	Age (year)	M/F ratio[a]	Duration of pain prior to intervention (mo)	Patients suffering from severe pain (%)[b]	In-/exclusion criteria	Intervention	Success rate (%)	Period of intervention	Study period	Pain scale	Recall method
Loos [23]	2008	22 (22)	To assess long-term pain relief after these treatment modalities in patients suffering from chronic pain because of Pfannenstiel-induced nerve entrapment	43 [22–67]	0/27	24 [4–384]	70	Incl: incisional pain after Pfannenstiel procedure > 3 months, treated for neuralgia of the iliohypogastric and/or ilioinguinal nerve after a Pfannenstiel incision Excl: other known causes of pain suspected, CNS sensitization	Neurectomy of ilioinguinal and/or iliohypogastric nerve	72.7	2000–2007	09/2007–11/2007	VRS	Questionnaire
Loos [24]	2010	56 (54)	To assess the long-term efficacy of surgical neurectomy for chronic, postherniorrhaphy groin neuralgia	50 [18–88]	43/11	30 [3–300]	67.3	Incl: unilateral inguinal hernia repair and > 3 months of pain associated with previous operative procedure Excl: Patients treated for non-neuropathic pain syndromes exclusively, follow-up < 3 months	Inguinal neurectomy	52	2003–2008	10/2008	VRS	Questionnaire
Boelens [26]	2011	139 (139)	To evaluate the efficacy of a diagnostic workup protocol and treatment regimen in patients with suspected ACNES	47 ± 17	32/107	6 [3–29]	80.4	Incl: Abdominal wall pain qualified by a constant site of tenderness that is superficially located with a small (< 2 cm^2) area of maximal tenderness, the most intense pain localized by the tip of 1 finger, and tenderness increased by Carnett's test. Patients qualified for surgery if level of pain after the injection regimen was unacceptable or grossly interfered with daily activities Excl: Diagnosis of ACNES not probable	Injection of 1% lidocaine followed by subsequent therapeutic injections including corticosteroids. Occasionally followed by a neurectomy	71	2003–2008	08/2008–12/2010	VAS and VRS	Questionnaire

Table 2 (continued)

Study	Year	*n* (pt)	Objective	Age (year)	M/F ratio[a]	Duration of pain prior to intervention (mo)	Patients suffering from severe pain (%)[b]	In-/exclusion criteria	Intervention	Success rate (%)	Period of intervention	Study period	Pain scale	Recall method
van Assen [27]	2015	181 (154)	To determine the long-term success rate of surgery in an extensive ACNES population	47 [33–61]	27/127	NA	90.7	Incl: Adult ACNES§ patients (≥ 18 year) having undergone a primary anterior neurectomy Excl: Other interventions than a neurectomy or other diagnosis than ACNES	Primary anterior neurectomy	61	2004–2012	01/2011–01/2013	NRS	Telephone interview
Zwaans [28]	2015	153 (148)	To identify potential patient or surgery related factors predicting the effectiveness of surgery for inguinodynia after Lichtenstein repair	55 ± 13	138/15	32 [0–213]	65.6	Incl: open groin pain surgery via original incision, previously used for Lichtenstein repair, ≥ 3 months inguinodynia after the primary repair, age ≥ 18 year and presence of operative report Excl: Patients who underwent groin pain surgery more than once, pain following other types of hernia surgery, malignancy, cognitive impaired individuals, meralgia paresthetica and signs of groin infection at the time of surgery	Inguinal neurectomy and/or (partial) mesh removal	54.4	2000–2013	01/2014–03/2014	NRS	Telephone interview
Verhagen [25]	2018	101 (101)	To report on long-term pain relief after a surgical neurectomy in a group of patients with chronic pain due to Pfannenstiel-induced nerve entrapment	52 [49–54]	0/101	24 [3–384]	87	Incl: treated for neuralgia of the iliohypogastric and/ or ilioinguinal nerve following a Pfannenstiel incision Excl: other known causes of pain suspected or diagnosed previously	Inguinal neurectomy	65.0	2000–2015	12/2010–12/2015	VRS	Questionnaire

Table 2 (continued)

Study	Year	*n* (pt)	Objective	Age (year)	M/F ratio[a]	Duration of pain prior to intervention (mo)	Patients suffering from severe pain (%)[b]	In-/exclusion criteria	Intervention	Success rate (%)	Period of intervention	Study period	Pain scale	Recall method
Siawash [29]	2022	22 (22)	To investigate the long-term effects of treatment for ACNES on amelioration of pain and quality of life in children between 12 and 18 years of age	15 [11–18]	4/18	14 [1–48]	100	Incl: children from 12 to 18 years diagnosed with ACNES Excl: Any visceral pathology explaining the abdominal discomfort, ACNES§ workup being the reason for referral or insufficiency of necessary data	Injection therapy (lidocaine 1%), occasionally followed by a neurectomy	72.7	2011–2012	06/2013	NRS	Questionnaire

CNS Central Nervous System, *ACNES* Anterior Cutaneous Nerve Entrapment Syndrome, *VRS* 5-point Verbal Rating Scale, *VAS* Visual Analog Scale, *NRS* Numerical Rating Scale, *NA* Not Available

[a]M/F ratio = male/female ratio

[b]Severe pain = NRS ≥ 7.0, VRS ≥ 4, or VAS ≥ 70 mm

indicating predominance of positive misclassification over negative misclassification among studies. Therefore, an overall actual recall bias was present due to differential misclassification in successes and failures.

Heterogeneity as determined by the Chi-square test was absent ($p = 0.78$). The I^2 was 0%, indicating no important inconsistencies between different studies (Fig. 4).

Does recall bias differ per pain score?

The prevalence of recall misclassification differed per type of pain score. Prevalence of recall misclassification in NRS was 6.3% [95% CI 3.5–10.8], VAS 26.0% [95 %CI 17.3–37.2], and VRS 24.5% [95% CI 14.5–38.2]. The OR of the recall misclassification also varied per pain score (Fig. 4) being 2.0 for NRS, [95% CI 0.6–6.7], 3.6 for VAS [95% CI 1.1–11.6], and 1.6 for the VRS pain scale [95% CI 0.4–6.3].

Selection bias

Characteristics of the population with incomplete data sets ($n = 255$) were similar to the population with sufficient data ($n = 313$, Table 4) reducing the likelihood of selection bias in the present study.

Discussion and conclusions

The present study demonstrates that retrospectively collected pain scores of studies on efficacy of surgery for chronic groin pain result in erroneous measurement of pain intensities. It shows that misclassification due to recall errors affect both patients with successful surgery and patients with unsuccessful surgery, with an overall prevalence of 13.7%. Positive recall misclassification is more likely to occur than negative recall misclassification with an overall pooled OR of 2.4 [95% CI 1.2–4.8]. Patients with an unsuccessful outcome recalled their pre-operative pain scores as being higher than they actually were as indicated by pre-operatively obtained pain scores. Conversely, patients with successful surgery demonstrated lower pain intensities when recalled. It may be concluded that using recalled pain scores has a significant impact on the measurement of surgical outcomes of patients suffering from abdominal wall or groin pain, depending upon the success rate. Hence, recall bias does indeed exist in this patient population.

The present study demonstrated significant recall bias if relying on retrospectively acquired pain scores. A schematic version illustrating how the present results should be interpreted in the context of other studies using recalled pain scores is depicted in Fig. 5. In the first example, a hypothetical retrospective study is performed using recalled pre-intervention pain scores. The hypothetical study included

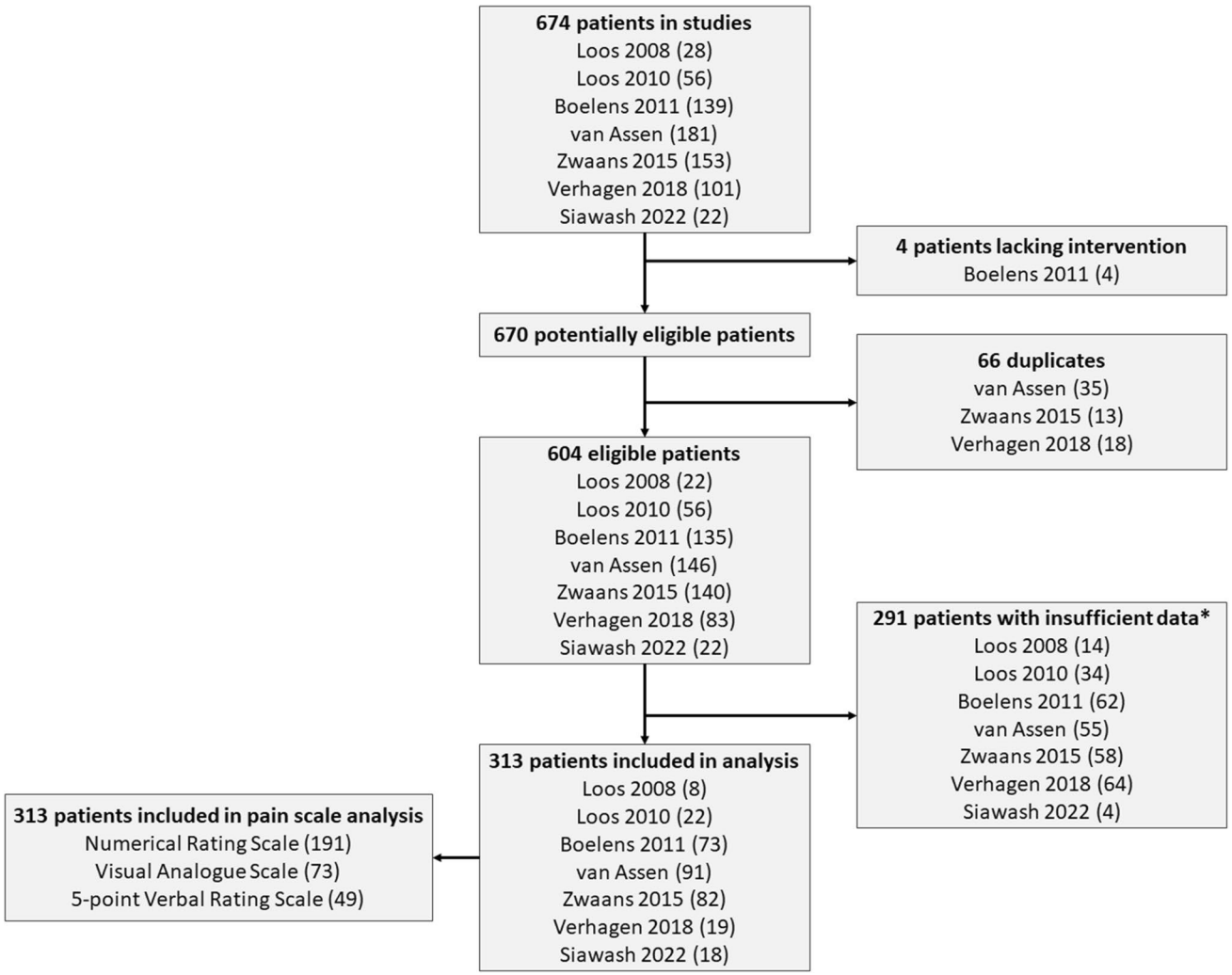

Fig. 2 Flowchart of the selection process. *Data were considered insufficient if the prospectively obtained, retrospectively obtained or post-operative pain score was missing

100 patients and assumes that the intervention is successful in 80 patients (80%). As demonstrated by the present study, recall misclassification affected 13.7% of patients (Table 3). For the purpose of clarity, let us assume that in 15% of the patients, treatment effect is misclassified due to recall. Recall misclassification was about twice more likely in failures than in successfully treated patients (OR 2.35; Fig. 4). As there are fewer failures ($n=20$) than successes ($n=80$), the absolute number of misclassified patients in the success group is higher. As a consequence, 5 of the 15 hypothetical misclassified patients were actually failures (based on prospective pre-intervention pain scores). The other 10 misclassified patients were actually successes. The net number of misclassifications is 5 patients (10 minus 5). Following this line of thought, the percentage of successfully treated patients decreased from 80 to 75% and the number of failures increased by 5%.

A second example is illustrated in Fig. 5. Since the number of successes is now lower, the absolute number of misclassified patients in the successful group is lower. This leads to a 5% success overestimation in this hypothetical example if retrospective pre-intervention data are used. Using a similar calculation based on true data from the present study indicated that the net recall bias is nullified if the success rate is 67% (Fig. 6). As a consequence, the net direction will go toward an underestimation of the treatment effect in highly successful treatment (i.e., > 67%). Conversely, it will go toward an overestimation of the treatment effect in a less successful treatment (< 67%).

Previous studies on total knee arthroplasty [17], total hip replacement [16], or treatment for lower back pain [19] reported significantly higher pain levels if using recalled data. Additional literature on recall bias also confirms our finding that pain is often remembered as more intense by

Table 3 Recall errors and recall misclassification in studies reporting on pain attenuation following (surgical) interventions

Study	n in analysis	Pain scale	Prospective pre-operative pain score	Recalled pre-operative pain score	Recall error (*p* value)*	ICC[a] [95%CI[b]]	Prevalence recall misclassification (*n* (%) [95%CI[b]])
Loos 2008 [23] *n* = 22	8; 6 success, 2 failure	VRS	4.5 [3.0–5.0]	4.0 [3.8–4.0]	*p* = 1.00	0.42 [− 0.46 to 0.85]	2 (25 [6.3–59.9])
Loos 2010 [24] *n* = 56	22; 12 success, 10 failure	VRS	4.0 [4.0–5.0]	4.0 [3.0–4.0]	*p* = 0.32	0.35 [− 0.07 to 0.66]	8 (36.4 [19.6–57.1])
Boelens 2011 [26] *n* = 135	73; 56 success, 17 failure	VAS	70 ± 14	55 ± 16	p < 0.01	0.21 [− 0.03–0.43]	19 (26.0 [17.3–37.2])
vAssen 2015[27] *n* = 146	91; 46 success, 45 failure	NRS	8.0 [7.0–8.0]	8.0 [8.0–9.0]	*p* < 0.01	0.46 [0.24 to 0.63]	3 (3.3 [0.7–9.7])
Zwaans 2015 [28] *n* = 140	82; 39 success, 43 failure	NRS	7.2 ± 1.5	7.7 ± 1.3	*p* < 0.01	0.45 [0.24 to 0.62]	7 (8.5 [3.9–16.9])
Verhagen 2018 [25] *n* = 47	19; 10 success, 9 failure	VRS	4.0 [3.0–4.0]	4.0 [4.0–5.0]	*p* = 0.3	0.42 [− 0.02 to 0.73]	2 (10.5 [1.7–32.6])
Siawash 2022 [29] *n* = 22	18; 12 success, 6 failure	NRS	7.5 [7.5–8.0]	8.5 [8.0–9.0]	*p* < 0.01	0.31 [− 0.11 to 0.68]	2 (11.1 [1.9–34.0])
Total	313; 181 success, 132 failure						43 (13.7 [10.3–18.0])

*Significance of differences in recalled versus prospective pain scores, calculated by paired t test or Wilcoxon signed-rank test as appropriate

VRS Verbal Rating Scale, *VAS* Visual Analog Scale, *NRS* Numerical Rating Scale

[a]Intraclass correlation coefficient

[b]95% CI 95% Confidence Interval

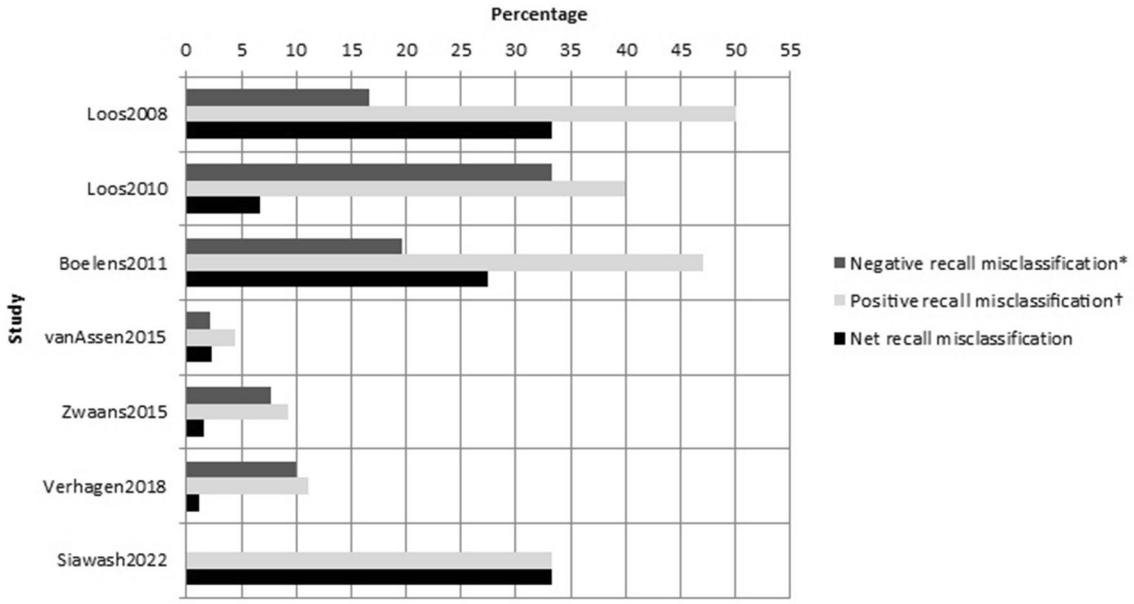

Fig. 3 Number of patients with positive and negative recall misclassifications by study. *Negative values indicate negative recall misclassification (a shift from success to failure by the recall error). †Positive values indicate positive recall misclassification [shift from failure group (determined by prospective pain score) to the successful group (using retrospective pain score)]

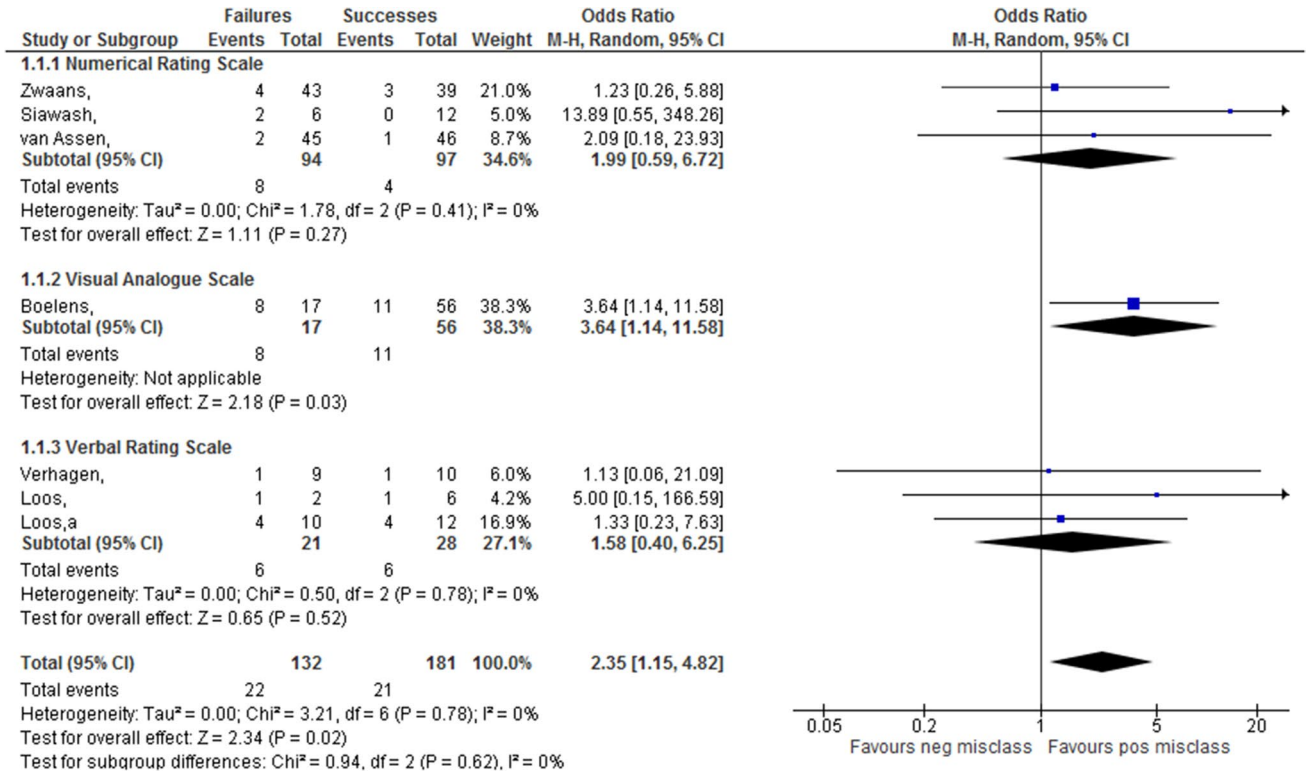

Fig. 4 Forest plot of the pooled odds ratios of the recall bias by pain score. *Neg misclass* negative recall misclassification, indicating a shift from the failure group (determined by prospective pain score) to the successful group (using retrospective pain score); *Pos misclass* positive recall misclassification, indicating a shift from the successful group (based on the prospective pain score) to the failure group (as determined by the retrospective pain scores). Events are the number of misclassified cases if retrospective pre-operative pain scores were used

patients suffering from pain after treatment whereas pain intensity is underestimated after success [12, 30–32]. Others suggested that errors in recalling pain intensity are generally non-differential [8, 10, 13, 15, 33].

Most researchers would argue that current state of mood influences pain recollection [7–9, 11, 12, 32, 34, 35]. A clear example illustrating this theory is the recall of pain intensity in a postnatal stage [36]. Women who just gave birth underrate previously experienced pain during labor due to an overwhelming feeling of happiness caused by carrying their healthy newborn. In other words, patients become accustomed to improvements in their condition, a term that is referred to as 'satisfaction treadmill' [6, 37]. Results of the present study also demonstrate that pain-free patients do (probably unintentionally) underestimate their pre-operative pain, possibly as a result of the positive emotions experienced during recall. A similar theory may, vice versa, hold true for a failure group. Their negative emotions will modulate memory processing and, consequently, recall of pain in the past [35]. It may be concluded that recall pain intensities are likely congruent with pain, emotions, and interference of daily activities of the pain at the time of recall. These phenomena may lead to higher and lower recalled pre-operative pain scores in failure and successes, respectively.

Nowadays, it is recognized that patient outcomes for (chronic pain after) hernia surgery cover more than just

Table 4 Baseline characteristics of excluded (missing data sets) and included patients (complete sets of pain scores). Data are presented as means ± standard deviation or ratios

Baseline Characteristic	Excluded ($n = 255$)	Included ($n = 313$)	p value	Statistical test
Sex ratio (male:female)	108:147	118:195	0.26	Pearson X^2 test
Age (years)	46.8 ± 16.2	47.0 ± 17.7	0.86	Independent t test
% successful treatment outcome[a]	58%	58%	0.95	Pearson X^2 test

[a]Treatment outcome ratio based on retrospectively obtained pain scores, as prospective pain scores were lacking

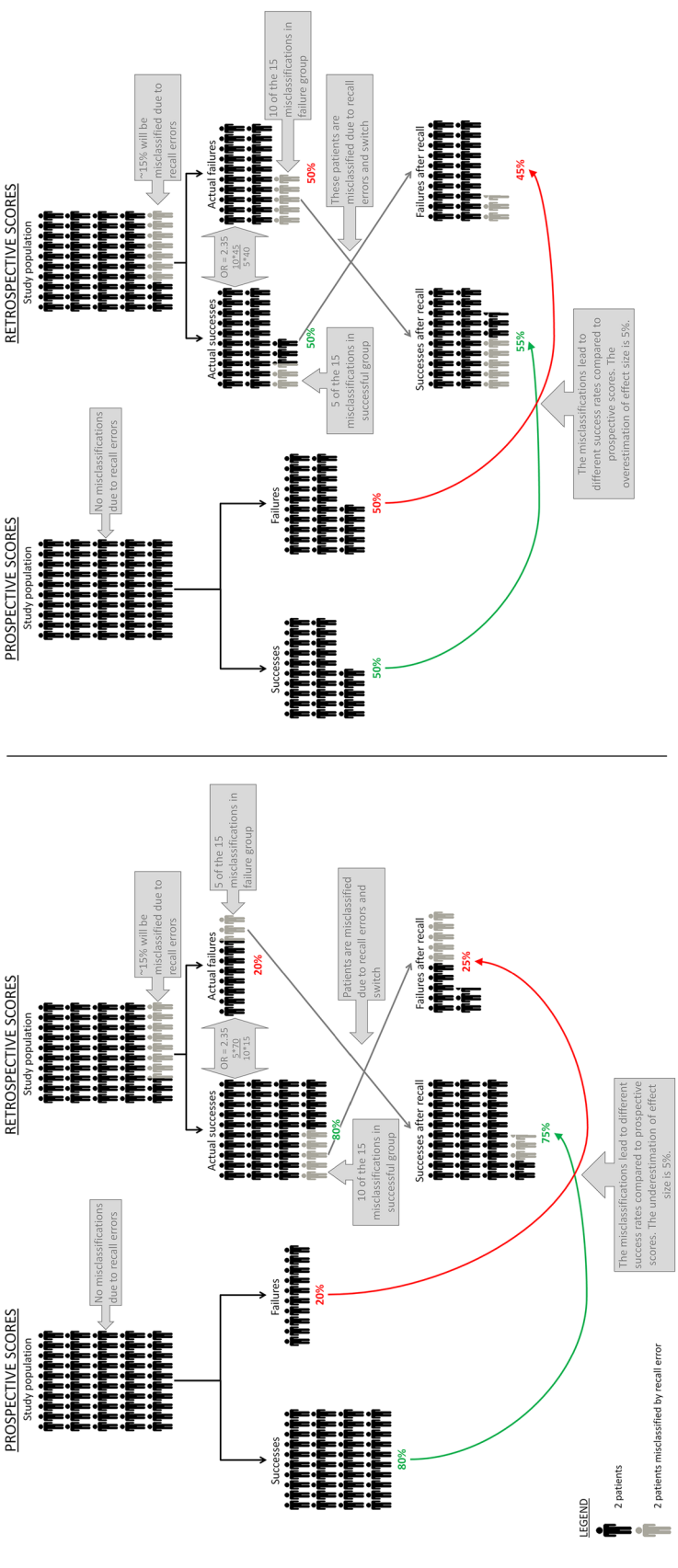

Fig. 5 Recall bias in retrospective studies. Recall errors are more present in the actual failures leading to disproportional numbers of recall misclassifications in failures and successes. Due to the disproportional numbers of misclassifications, recall bias occurs resulting in an overestimation of beneficial effect size

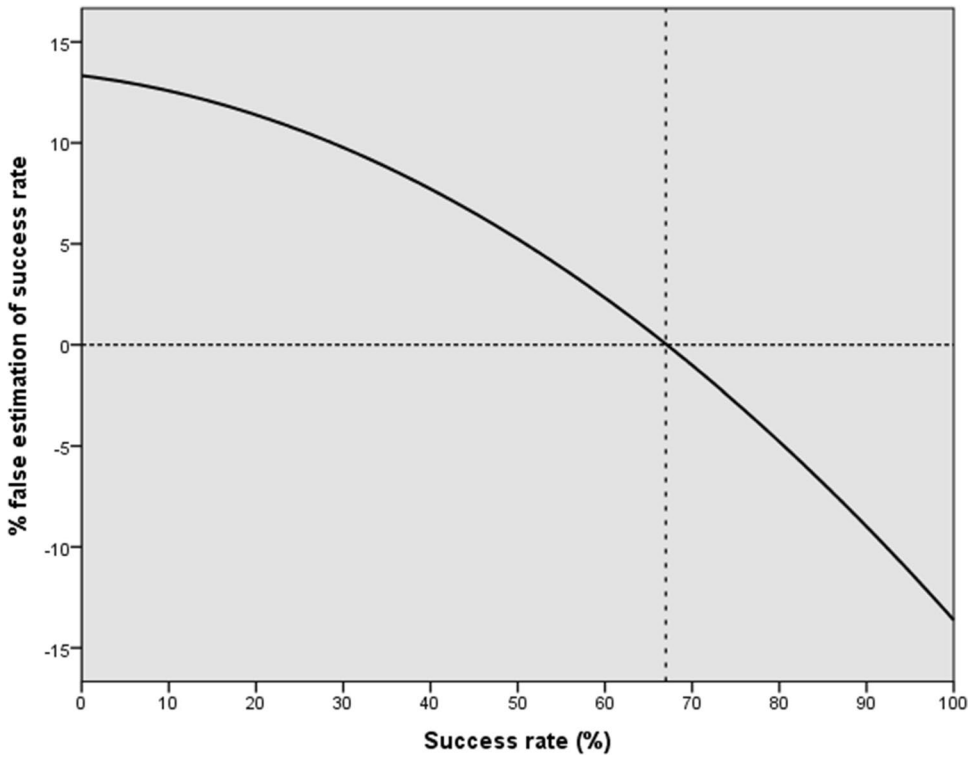

Fig. 6 Percentage of false estimation, based on the retrospective pain scores in relation to the actual success rate, as based on prospective pain scores

unilateral, simplified pain scores. More dedicated and multidimensional instruments to assess outcomes have been developed over the years, including the Carolinas Comfort Scale [38], the short-form Inguinal Pain Questionnaire [39], and the Activity Assessment Scale [40]. Unfortunately, no consensus has been reached on what outcome measurements are preferred to date [41]. Chronic pain is multifactorial, and has been linked to worse mental health and both psychosocial and functional factors are known to have an impact on the pain experienced by patients [42]. The Hospital Anxiety Depression Scale and Pain Catastrophizing Scale may give additional insight in these confounding factors and how psychological factors influence the results of pain scores. It is likely that the phenomenon of recall bias is also present when more comprehensive scales are completed by patients in a retrospective manner. The extent of bias in studies using these scales may be less, compared to the conventional pain scores used in the present paper, as more specific activities and functions are assessed. Although it is hypothesized that recall bias also occurs when using the more extensive outcome measurements, results of the present paper cannot be extrapolated with certainty. Additional research is desired to assess the bias in other outcome measures following hernia surgery.

Potential study limitations

Our study has several potential limitations. A possible limitation is the fact that the analysis of recall error is likely to be underpowered, since some of the seven studies included less than 20 eligible patients. However, the main issue is whether prospectively obtained pain scores differ from retrospectively obtained pain scores. Independently of the statistical analysis used to assess this issue, the prospective and recalled pain scores indicate a significant difference.

Selection bias may have been created as populations with incomplete data were excluded. Since the basic characteristics of included and excluded patients were similar, selection bias is less likely to play a role. Publication bias was avoided, since all eligible MMC studies were included irrespective of publication status.

The present study demonstrates that recall bias varied between pain scales. VAS seemed most susceptible to recall error, but this observation relied on one single study. Therefore, no firm conclusions can be drawn and further research is required.

Conclusions

Surgery outcomes in one in seven patients undergoing remedial surgery are misclassified on the basis of retrospectively obtained pre-operative pain scores (success instead of failure, or vice versa). Misclassification is more likely in unsuccessful surgery than successful surgery. Therefore, the estimated effect size in studies using recalled pre-operative pain scores depends upon the actual success rate. Success rates exceeding 67% are underestimated, whereas effect sizes are overestimated when success rates are below 67%. Detailed pain scales seem to be more susceptible for recall errors, but this issue needs further investigation.

Author contributions WZ: study design, data collection, data analysis, and writing up of the first draft of the paper, JB: writing up of the first draft of the paper, and critical revision of manuscript, JD: study design, data analysis, writing up of the first draft of the paper, and critical revision of manuscript, ES: writing up of the first draft of the paper, and critical revision of manuscript, MS: study design, writing up of the first draft of the paper, and critical revision of manuscript, RR: study design, writing up of the first draft of the paper, and critical revision of manuscript.

Funding None.

Declarations

Conflict of interest All other authors declare that they have no conflict of interest.

Ethical approval The present study did not require permission from a medical ethics committee since it involved evaluation ofpreviously collected anonymous data that was obtained for clinical purposes. All patients consent in writing for the use of datafor research.

Human and animal rights The study was approved by the institutional ethics committee. This article does not contain any studies with human-participants or animals performed by any of the authors.

Informed consent Informed consent was obtained from the patient.

References

1. Miettinen OS. Principles of Epidemiologic Research. Cambridge: School of Public Health, Harvard University, 1975
2. Hassan E (2005) Recall Bias can be a threat to retrospective and prospective research designs. Int J Epidemiol. https://doi.org/10.5580/2732
3. Last JM (2000) A Dictionary of Epidemiology, 4th edn. Oxford University Press, Oxford
4. Lippman A, Mackenzie SG (1985) What is "recall bias" and does it exist? Prog Clin Biol Res 163C:205–209
5. Sedgwick P (2012) What is recall bias? BMJ 344:e3519
6. Gendreau M, Hufford MR, Stone AA (2003) Measuring clinical pain in chronic widespread pain: selected methodological issues. Best Pract Res Clin Rheumatol 17(4):575–592
7. Redelmeier DA, Kahneman D (1996) Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. Pain 66(1):3–8
8. Salovey P, Smith AF, Turk DC, Jobe JB, Willis GB (1993) The accuracy of memory for pain: not so bad most of the time. APS Journal 2(3):184–191
9. Eich E (1993) On the accuracy of memory for pain. APS Journal 2(3):192–194
10. Middel B, Goudriaan H, de Greef M, Stewart R, van Sonderen E, Bouma J et al (2006) Recall bias did not affect perceived magnitude of change in health-related functional status. J Clin Epidemiol 59(5):503–511
11. Williams DA, Park KM, Ambrose KR, Clauw DJ (2007) Assessor status influences pain recall. J Pain 8(4):343–348
12. Eich E, Reeves JL, Jaeger B, Graff-Radford SB (1985) Memory for pain: relation between past and present pain intensity. Pain 23(4):375–380
13. Rofe Y, Algom D (1985) Accuracy of remembering post-delivery pain. Percept Mot Skills 60(1):99–105
14. Raphael K (1987) Recall bias: a proposal for assessment and control. Int J Epidemiol 16(2):167–170
15. Wright J, Morley S (1995) Autobiographical memory and chronic pain. Br J Clin Psychol Br Psycholo Soc 34(Pt 2):255–265
16. Mancuso CA, Charlson ME (1995) Does recollection error threaten the validity of cross-sectional studies of effectiveness? Med Care 33(4 Suppl):AS77-88
17. Lingard EA, Wright EA, Sledge CB (2001) Pitfalls of using patient recall to derive preoperative status in outcome studies of total knee arthroplasty. J Bone Joint Surg Am Vol 83-A(8):1149–1156
18. Stone AA, Broderick JE, Shiffman SS, Schwartz JE (2004) Understanding recall of weekly pain from a momentary assessment perspective: absolute agreement, between- and within-person consistency, and judged change in weekly pain. Pain 107(1–2):61–69
19. Pellise F, Vidal X, Hernandez A, Cedraschi C, Bago J, Villanueva C (2005) Reliability of retrospective clinical data to evaluate the effectiveness of lumbar fusion in chronic low back pain. Spine 30(3):365–368
20. Ho K, Spence J, Murphy MF (1996) Review of pain-measurement tools. Ann Emerg Med 27(4):427–432
21. Loos MJ, Houterman S, Scheltinga MR, Roumen RM (2008) Evaluating postherniorrhaphy groin pain: visual analogue or verbal rating scale? Hernia 12(2):147–151
22. Price DD, McGrath PA, Rafii A, Buckingham B (1983) The validation of visual analog scales as ratio scale measures for chronic and experimental pain. Pain 17(1):45–56
23. Loos MJ, Scheltinga MR, Roumen RM (2008) Surgical management of inguinal neuralgia after a low transverse Pfannenstiel incision. Ann Surg 248(5):880–885
24. Loos MJ, Scheltinga MR, Roumen RM (2010) Tailored neurectomy for treatment of postherniorrhaphy inguinal neuralgia. Surgery 147(2):275–281
25. Verhagen T, Loos MJ, Mulders LG, Scheltinga MR, Roumen RM (2018) A step up therapeutic regimen for chronic post-Pfannenstiel pain syndrome. Eur J Obstet Gynecol Reprod Biol 231:248–254
26. Boelens OB, Scheltinga MR, Houterman S, Roumen RM (2011) Management of anterior cutaneous nerve entrapment syndrome in a cohort of 139 patients. Ann Surg 254(6):1054–1058
27. van Assen T, Boelens OB, van Eerten PV, Perquin C, Scheltinga MR, Roumen RM (2015) Long-term success rates after an anterior neurectomy in patients with an abdominal cutaneous nerve entrapment syndrome. Surgery 157(1):137–143
28. Zwaans WA, Verhagen T, Roumen RM, Scheltinga MR (2015) Factors determining outcome after surgery for chronic groin pain following a Lichtenstein hernia repair. World J Surg 39(11):2652–2662

29. Siawash M, Mol FM, Perquin CW, van Eerten PV, Roumen RM, Scheltinga MR. Management of anterior cutaneous nerve entrapment syndrome (ACNES) in adolescents. Unpublished data. 2022.
30. Choi BC, Noseworthy AL (1992) Classification, direction, and prevention of bias in epidemiologic research. J Occup Med 34(3):265–271
31. Jamison RN, Sbrocco T, Parris WC (1989) The influence of physical and psychosocial factors on accuracy of memory for pain in chronic pain patients. Pain 37(3):289–294
32. Tasmuth T, Estlanderb AM, Kalso E (1996) Effect of present pain and mood on the memory of past post-operative pain in women treated surgically for breast cancer. Pain 68(2–3):343–347
33. Hunter M, Philips C, Rachman S (1979) Memory for pain. Pain 6(1):35–46
34. Bower GH (1981) Mood and memory. Am Psychol 36(2):129–148
35. Gedney JJ, Logan H (2006) Pain related recall predicts future pain report. Pain 121(1–2):69–76
36. Norvell KT, Gaston-Johansson F, Fridh G (1987) Remembrance of labor pain: how valid are retrospective pain measurements? Pain 31(1):77–86
37. Kahneman D (1999) Objective happiness. In: Kahneman D, Deiner E, Schwarz N (eds) Well-being: the foundations of hedonic psychology. Russell Sage Foundation, New York City, pp 3–25
38. Heniford BT, Lincourt AE, Walters AL, Colavita PD, Belyansky I, Kercher KW et al (2018) Carolinas comfort scale as a measure of hernia repair quality of life: a reappraisal utilizing 3788 international patients. Ann Surg 267(1):171–176
39. Olsson A, Sandblom G, Franneby U, Sonden A, Gunnarsson U, Dahlstrand U (2019) The short-form inguinal pain questionnaire (sf-IPQ): An instrument for rating groin pain after inguinal hernia surgery in daily clinical practice. World J Surg 43(3):806–811
40. McCarthy M Jr, Jonasson O, Chang CH, Pickard AS, Giobbie-Hurder A, Gibbs J et al (2005) Assessment of patient functional status after surgery. J Am Coll Surg 201(2):171–178
41. HerniaSurge G (2018) International guidelines for groin hernia management. Hernia 22(1):1–165
42 Miller BT, Scheman J, Petro CC, Beffa LRA, Prabhu AS, Rosen MJ et al (2022) Psychological disorders in patients with chronic post-operative inguinal pain. Hernia. https://doi.org/10.1007/s10029-022-02662-2