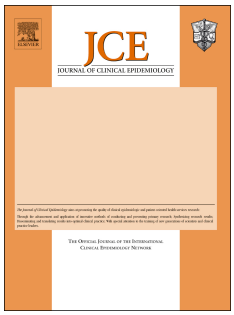
Accepted Manuscript



Retrospectively patient reported pre-event health status showed strong association and agreement with contemporaneous reports

Esther Kwong, Research Fellow in Health Services Research, Nick Black, Professor of Health Services Research

PII: S0895-4356(16)30415-2

DOI: 10.1016/j.jclinepi.2016.09.002

Reference: JCE 9239

To appear in: *Journal of Clinical Epidemiology*

Received Date: 26 October 2015

Revised Date: 30 July 2016

Accepted Date: 5 September 2016

Please cite this article as: Kwong E, Black N, Retrospectively patient reported pre-event health status showed strong association and agreement with contemporaneous reports, *Journal of Clinical Epidemiology* (2016), doi: 10.1016/j.jclinepi.2016.09.002.

Retrospectively patient reported pre-event health status showed strong association and agreement with contemporaneous reports

Esther Kwong

Research Fellow in Health Services Research

Nick Black

Professor of Health Services Research

Department of Health Services Research & Policy

London School of Hygiene & Tropical Medicine

15-17 Tavistock Place

London WS1H 9SH

Keywords: Patient Reported Outcome Measures, Health-related quality of life, Retrospective, Population norms, recall bias, response shift

ABSTRACT

Objective

The unpredictability of the occurrence of illnesses and injuries leading to most emergency admissions to hospital makes it impossible prospectively to collect pre-admission patient reported outcome measures (PROMs). Our aims were to review the evidence for using retrospective PROMs to determine pre-event health status and the validity of using general population norms instead of retrospective PROMs.

Study design and setting

Searches of Medline, PsycINFO, Embase, Global Health, and Health Management information. Six studies met the inclusion criteria for the first aim and 11 studies addressed the second aim. Narrative syntheses were conducted.

Results

Strong associations were found between retrospective and contemporary PROMs in 21 out of 30 comparisons (correlation coefficients over 0.68) and 20 of 24 showed strong agreement for continuous measures (intraclass correlations over 0.75)). Categorical measures revealed only fair to moderate levels of agreement (Kappa 0.3-0.6). Associations were stronger for indices than for individual items and for shorter time intervals. The direction of differences was inconsistent.

Retrospective PROMs reported by elderly patients were similar to the general population but younger adults had been healthier.

Conclusion

Retrospective collection offers a means of assessing PROMs in unexpected emergency admissions. However, further research is needed to establish the best policy for their use.

What is new?

- There is a strong association between PROMs collected retrospectively and contemporaneous collection among patients undergoing elective surgery.
- Agreement is also strong for PROMs that are continuous measures but only fair to moderate for categorical measures.
- Retrospectively collected data suggest that young adult trauma patients are healthier than population norms. The reverse may be true for older patients admitted for medical conditions.
- Retrospective collection offers a means of assessing patient reported outcomes in unexpected emergency admissions. However, further research is needed to establish the best policy for their use.

1. Introduction

The growing acceptance of the importance of patients' views of their outcome when evaluating interventions and assessing the quality of services means that it is necessary to devise ways in which accurate Patient Reported Outcome Measures (PROMs) can be obtained (referred to as PROs in the USA) [1]. PROMs are self-completed questionnaires where patients are asked to report their own state of health (multi-dimensional symptoms, functional status) and health-related quality of life (HRQL) at one point in time. PROMs can be categorised as generic (e.g. EQ-5D, SF36) or disease- specific (Oxford Hip Score or Western Ontario & McMaster Osteoarthritis Index). Generic PROMs capture broad domains on function or HRQL, can be converted into utility scores, and provide the means to compare between conditions and treatments. Disease- specific PROMs have greater sensitivity by incorporating aspects of function and HRQL specific to that condition [1]. By comparing measurements before and after a health care intervention the outcome of care can be determined.

Emergency admissions make up 34% of hospital admissions in England [2]. They can be categorised as either a largely unexpected acute event, such as an acute myocardial

infarction, stroke or injury (about 70% of all emergency admissions) or as an exacerbation of an existing long-term conditions as occur in conditions such as diabetes or chronic obstructive pulmonary disease. While these are not a clear-cut dichotomy, the two categories present different challenges when using PROMs. Unlike for elective admissions when a PROM can be collected before treatment to capture the baseline health at the time (a contemporary PROM), for unexpected emergency admissions this is not possible. (This need not be a problem for emergency admissions due to exacerbations of long-term conditions, such as chronic obstructive pulmonary disease, when PROMs could be collected as part of their routine clinical management i.e. a contemporary PROM). Therefore, for unexpected admissions, other methods must be used to assess patients' pre-admission baseline health status.

There are two possible approaches. First, there is the use of re retrospective PROMs, in which patients are asked to recollect (after their unexpected emergency event, such as an acute myocardial infarction) what their health status and quality of life was like just prior to the emergency event.). This takes the place of contemporaneous collection before the event that can be done when considering planned elective treatments such as hip replacements. Retrospective self-reporting has been extensively used in aetiological case-control studies and in cross-sectional surveys [3] in which respondents are asked to recall characteristics of their health over a specified time frame which may be short (e.g. preceding week) or long (e.g. past year).

Second, and much cheaper than retrospective reporting, is to use age-sex standardised PROMs which have been collected from the general population (or an appropriate comparison group) as part of a cross-sectional survey, as a surrogate measure of a patients' pre-event baseline health [4]. The use of population norms assumes that patients experiencing an emergency admission are typical of the wider population. This assumption could lead to an over or under-estimate of the impact of a health care intervention. If patients are in fact healthier at baseline than the general population, (as might be the case when studying recovery from trauma that occurred while undertaking a dangerous sport such as rock-climbing), using the population norm as a surrogate baseline could lead to an 'overestimate' of the treatment effect. On the other hand if patients were in worse health than their peers beforehand (as might be expected for those suffering a heart attack), an 'underestimation' of the treatment effect will be observed.

Although there has been no review of the strength of association and of agreement between these two approaches in emergency admissions, two systematic reviews have considered other aspects of recall. One considered the length of recall periods for PROMs in clinical

trials and concluded that the optimum depended on two broad categories of factors: characteristics of the phenomenon being recalled (such as how recently it had occurred, its attributes, its complexity) and the context of the recalled phenomenon (such as its salience, the patient's mood) together with the nature of the topic [5]. The second review concluded that recall bias is a concern with PROMs and called for more research to understand and identify situations where the use of recall is acceptable [6].

Our aims were to review systematically the scientific evidence on (i) the extent of association and agreement between PROMs collected retrospectively and contemporaneously to determine pre-event health status and HRQL and (ii) the validity of using general population norms for determining the pre-event health status and HRQL of people with an unexpected emergency admission to hospital.

2. Study design and settings

2.1 Literature search

A search was conducted on studies either (i) comparing retrospective and contemporary PROMs (health status, symptoms, functional status, HRQL) or (ii) comparing retrospective PROMs and population norms. For inclusion, studies had to be: in English; involve self-completed questionnaires; have a recall period of no more than six months. In addition, for comparisons of retrospective and contemporary PROMs, studies had to include a quantitative estimation of the strength of association (Pearson or Spearman rank correlation) or agreement (intra-class correlation coefficient or kappa score). No additional analyses were undertaken to determine missing correlations or levels of agreement.

Our focus was on methods for estimating patients' pre-event health or HRQL that could be used to determine the extent to which treatment restored them to their previous state of health. Many studies ask patients themselves to assess the extent of change in their health (single transitional items) [7][8] but this is a different methodological approach to that of comparing assessments at two points in time and were excluded from this review.

Five databases were searched: Medline, PsycINFO, Embase, Global Health, and Health Management information. A free-text search strategy was employed as subject headings were too broad and non-specific for the research question. The detailed concepts, keywords and search terms are shown in Table 1 and the complete search strategy is shown in Table 2. A forward and backward snowballing strategy was used to complement the free-text search.

Identified articles were exported to a reference manager (Mendeley Desktop version 1.13) and duplicates removed. The title and abstracts were screened by one author (EK) to assess suitability. Studies in children, adolescents, carer proxies and those with cognitive impairments were excluded. The remaining articles were read and forwards and backwards searching of references was conducted (Figure 1).

>>insert Table 1: Literature search: concepts, keywords and search terms<<

>>insert Table 2: Search strategy<<

>> insert Figure 1: Search results<<

2.2 Quality appraisal

For studies comparing retrospective and contemporary PROMs, their methodological quality was appraised by one author (EK) using five relevant items selected from the Quality Appraisal of Diagnostic Reliability (QAREL) Checklist [9]. These items cover the representativeness of participants, time interval between assessments, correct application of assessment, and appropriate statistical analysis. The other items were not applicable in this review: whether participants were blinded to their initial assessment, to other participants' assessments, to any reference standard or to clinical information, or blinded to additional cues that were not part of the test. A simple summation of the five included items was calculated (0 = weak, 5 = strong). Given the heterogeneity of the studies in this review, a narrative synthesis was carried out.

2.3 Definition of strength of association and agreement

Association according to Pearson or Spearman correlation coefficients were classified as: weak (below 0.36), moderate (0.36 to 0.67), strong (0.68 to 0.90) and very strong (above 0.90) [10].

Agreement according to intra-class correlation coefficients were classified as: weak (below 0.36), moderate (0.36 to 0.67), strong (0.68 to 0.90) and very strong (above 0.90). Agreement according to Kappa scores were classified as: slight (<0.20), fair (0.20-0.40), moderate (0.41 to 0.60), substantial (0.61 to 0.80) and almost perfect (0.81 to 1.0) [11].

3. Results

3.1 Search findings

275 articles were identified on Medline, 350 on Embase, 102 on PsycINFO, 18 on Global Health and 2 on Global Management Information (all accessed 22 April 2015). Having removed duplicates, 450 abstracts were reviewed of which four comparing retrospective and contemporary PROMs, and five comparing retrospective PROMs and population norms met the inclusion criteria. The majority of the studies were excluded either because they did not capture a contemporary baseline PROM measurement or there was no statistical assessment of the strength of association or agreement between contemporary and retrospective PROMs. A citation search on PubMed (forward and backward snowballing)

identified two additional studies comparing retrospective and contemporary PROMs and six comparing retrospective PROMs and population norms) (Figure 1). All studies comparing retrospective and contemporary PROMs were methodologically strong according to the QAREL checklist.

3.2 Comparison of retrospective with contemporary PROMs

Of the six studies, one was from the UK [12], one was multinational [13], three were from Canada [14-16], and one from the USA [17] (Table 3). The studies involved 75-177 patients, with one exception with 770 patients [13]. Four involved patients with hip and knee problems and two were based on urological patients. Several reported on the level of agreement between retrospective and contemporary reports for more than one PROM.

Eleven different PROMs were used including the SF-36 or SF-12 (four studies), the Western Ontario and McMaster Osteoarthritis Index (WOMAC) (three studies), the American Urological Association (AUA) Symptom Index (two studies), the Western Ontario Meniscal Evaluation Tool (WOMET), the Knee Injury and Osteoarthritis Outcome Score (KOOS), Oxford Hip Score (OHS), Lower Extremity Functional Scale (LEFS), and the Feeling thermometer. The time period for retrospective reporting was predominantly two weeks to three months though one study reported three days (in addition to longer periods) and one used six months.

All six studies assessed the level of association between retrospective and contemporary PROMs scores using correlation coefficients (four used Pearson and two used Spearman coefficients), all reported on the level of agreement (three used Kappa statistics and three used intra-class coefficients). Most presented analyses of the full index scores though some reported on sub-scales. A total of 30 correlations coefficients of full or sub-scale scores were reported, of which nine were moderate, 18 were strong and three were very strong.

Three studies that each used several PROMs at different time points thus generating 24 comparisons, the level of agreement for continuous data (intra-class correlations) was very strong for eight, strong for 12 and four were moderate. [14,15,16] In contrast, for PROMs that were converted to categorical variables for analysis, Kappa statistics revealed only fair to moderate levels of agreement. [12, 13, 17]

Correlations tended to be stronger, the shorter the time interval; one month or less [14, 15] reported strong or very strong agreement. Intervals of three months or more resulted in only moderate agreement. [12, 13] Another factor associated with the strength of agreement was

the type of patient. The majority of studies that had strong agreement were based on orthopaedic patients suggesting patient characteristics or the type of intervention (e.g. elective surgery rather than medical treatment) may influence the relationship.

There was no consistency in the direction of any difference between retrospective and contemporary accounts. One study found that patients tend to recall better baseline health than what they reported in their contemporary PROMs [17], two studies reported the opposite [13, 15], one found it varied by PROM [16] and two found no difference.[12,14]

The strength of agreement may be limited if the test-retest reliability of the PROM is poor. In Table 4 the reliability estimates for all the measures that were included in studies in Table 3 are presented. Test-retest reliability for all the PROMs used were excellent, and higher than the agreements captured when comparing retrospective to contemporary PROMs. This suggests there are additional reasons that influence recall when retrospective PROMs are used.

>>insert Table 3: Studies comparing retrospective to contemporary PROMs<<

* mean difference or proportions different; p values

>>insert Table 4: Test-retest reliability of PROMs included in literature review<<

3.3 Comparison of retrospective PROMs with population norms

There were 11 studies (Table 5), four from North America [25, 26, 31, 32], four from Australia or New Zealand [29-31, 36] and three from Europe. [27, 33, 34] Eight studies involved fewer than 500 patients (86-472) but three were larger (1500-3000 patients). All the studies involved trauma patients apart from one on patients with acute lung injury [31]. Most studies included adults of all ages. The two exceptions were a study of elderly people who had suffered a fractured neck of femur [27] and a study of young adult trauma victims [28].

All reported on a generic PROM: six used a version of the Short Form (SF-36, SF-12, SF-6); three used the EuroQuol EQ-5D; and two used the Sickness Impact Profile. The time period for retrospective reporting in six studies was less than one week. [26-31, 33] In the other studies it extended from a few weeks to three months.

All but one study used population norms derived from statutory surveys of the general population. The exception used a matched comparison group drawn from the local community [32]. Also, one study of drivers who had suffered trauma in road accidents were compared not only with population norms but also with a sample of uninjured drivers [28].

Of the 10 studies that used general population norms, six found that patients recalled their health as having been better than the general population [28-30, 33-35]. In the four other studies, three found no difference [25-27] and in only one did patients report worse health than the general population [31]. The latter was the only study not focused on trauma patients but on those who had developed acute lung injury who were likely to have been in a poor state of health before being hospitalised. The two studies that compared patients with matched samples rather than the general population reported either no difference [28] or better recalled health [32].

>>insert Table 5: Studies comparing retrospective PROMs with age-sex standardised general population norms<<

[1]also compared with representative sample of drivers; [2] compared with 177 community controls; [3] mean difference; p value

4. Discussion

4.1 Comparison of retrospective and contemporary PROMs

Only six studies have compared retrospective and contemporary PROMs. While the majority of the comparisons (21 of 30) revealed a strong or very strong association (correlation coefficients of over 0.68), the rest were moderate. Levels of agreement for continuous measures were more consistent with 20 out of 24 comparisons being strong or very strong. In contrast, comparisons of categorical measures showed only fair to moderate agreement. Stronger associations were observed for indices (than for individual items), for shorter time periods (one month or less) and for elective surgery patients than for those with medical conditions or treatments. The direction of differences between retrospective and contemporary PROMs also showed no consistent pattern and appeared to be dependent partly on the PROM being used.

Retrospective PROMs may be influenced for three reasons: recall bias; response shift; and lack of validity of the PROM. Recall bias arises because: details may go unnoticed and never be stored; new information may be added to stored memories altering the details; and over time events may be systematically distorted. [6] Recall is influenced by the time interval between the event and the time of its assessment: the longer the interval, the higher the probability of recall bias [37]: 20% of details of an event have been found to be irretrievable after one year and 50% are irretrievable after five years [38].

Response shift refers to the change in perception that can occur when circumstances change [39, 40]. For example, a patient's perception of the severity of a disability or their quality of life may change following treatment. This tends to diminish the assessment of pre-treatment severity and thus underestimate the benefits of the treatment. An example of this is when the term ‘severe’, has a different meaning for the same person in one occasion compared with a previous occasion due to new experiences. This is known as recalibration. Moreover, subjective values may also change over time so that physical, social and psychological aspects of HRQL may be prioritised differently after certain experiences, known as reprioritisation. Patients may also redefine the construct in question and attribute new meanings to it, known as scale reconceptualization [41].

It is possible that the validity of PROMs will be jeopardised when determining retrospective health if the recall interval is lengthy. Most PROMs have been validated for the recall of a person's health over the recent past (between one day to past four weeks). Indeed, many PROMs are based on patients' reports of their health over the preceding few weeks.

However, if patients are required to recall their health for longer periods, the validity of the instrument cannot be assumed.

For comparisons of health care providers or over time, recall bias and response shift will only matter if there is a systematic difference in behaviour between groups of patients being compared (e.g. patients attending different hospitals). There is no evidence that such differences exist within countries though some differences have been demonstrated between countries [42].

4.2 Comparisons of retrospective PROMs and population norms

The studies comparing retrospective PROMs with population norms was inevitably limited to generic instruments because disease-specific PROMs are rarely collected in general population surveys, and hence limits the availability of population data to generic PROMs. The generalizability of the findings is further limited by the focus of all but one study on trauma victims. The finding that most studies observed that trauma patients recalled their pre-injury health as better than average may reflect that patients (mostly car drivers) are fitter and healthier than the general population. [19] While response shift may have contributed, the likelihood that trauma patients were healthier is supported by evidence that rates of sports injuries and gunshot wounds are higher in fitter members of the population. [29-31, 33] This difference is further exaggerated as national population norms are derived from household surveys that include institutionalised individuals. In contrast, the one study of elderly people suffering a stress fracture related to poor bone density found no difference from the general population (age-sex standardised). [27] This is also consistent with the one study in which patients recalled worse health than the general population which focused on acute lung injury [31].

There may be a case for the purposes of estimating pre-event health status that estimates could be adjusted for the presence of long-term conditions to reduce over-estimation. The findings also suggest the potential of underestimating the prior health of patients if population norms are used directly as surrogates in cases where the patient population involved are younger adults. However, this underestimation may be small and may mostly affect studies in this specific cohort of patients. .

4.3 Limitations

There are several limitations to consider. First, only one author (EK) carried out the search, paper selection and quality appraisal. Although uncertainties were discussed and resolved with the other author, the reliability of the review would have been enhanced by double-

reviewing. Second, comparisons of retrospective and contemporary PROMs that have been studied are dominated by orthopaedic surgery (four of six studies) and by studies in North America (four of six). Thus the generalizability of the findings must be treated with caution. Third, many of the studies that investigated retrospective recall were too small to perform subgroup analysis to take into account of clinical characteristics such as severity of illness. Finally, the generalizability of the comparisons of retrospective PROMs and population norms are even more limited with 10 of the 11 studies focused on trauma patients. In addition, only generic PROMs were considered but this is understandable given that population norms are not available for disease-specific PROMs.

4.4 Implications for policy and research

Making judgements as to which of contemporary and retrospective reports is the more valid is unclear. Contemporary reports are usually considered the ‘gold standard’ so if retrospective reports differ, it is the latter that are judged to be 'unreliable'. However, in the context of PROMs, from a patient's point of view the way they recall their previous health may be of greater relevance to them and to assessing the quality of health care than how patients actually assessed it at the time. In this situation, the retrospective report could be viewed as the 'gold standard'. Rather than attach different values to the two types of PROM (in other words, judging whether contemporaneous collection is more or less valid than recalled collection), it is best just to consider the extent to which they differ and the implications both for the use of PROMs in clinical management and in provider comparisons. As long as data are collected in the same way in different providers then comparisons will not be undermined.

Our knowledge of the use of retrospective PROMs in the UK is extremely limited: the relevance of findings in other countries is uncertain given the potential influence of culture and other contextual factors; existing studies of unexpected emergency admissions are limited largely to trauma care; and there have been no published attempts to study both of the issues addressed in this review in a combined study (i.e. retrospective v contemporary v population norms). Until further research has been conducted, the best policy for using PROMs in emergency admissions will remain uncertain.

The key methodological challenges that require further research are: detailed investigation of the relationship between retrospective and contemporary PROMs (inevitably in elective conditions) which should also explore the influence of patient characteristics and of methodological factors on the relationship; determination of the potential use of population norms as a low cost alternative to retrospective PROMs; and testing the feasibility of

retrospective PROMs and population norms in a variety of unexpected emergency hospital admissions.

Acknowledgements

EK is funded by an Economic & Social Research Council doctoral fellowship. Grant Reference: ES/J500021/1

References

1. Black N. Patient reported outcome measures could help transform healthcare. BMJ 2013;346:f167

2. Health & Social Care Information Centre. Hospital Episode Statistics 2013-14. http://www.hscic.gov.uk/hes (accessed 7 October 2015)

3. Hennekens CH, Buring JE. Epidemiology in Medicine. Philadelphia: Lippincott Williams & Wilkins, 1987.

4. McKenzie E. Measuring disability and quality of life postinjury, in Rivara FP, Cummings P, Koepsell PD, Grossman DC, Maier RV (eds). Injury control: a guide to research and program evaluation. Cambridge, UK: Cambridge University Press, 2001.

5. Stull DE, Leidy NK, Parasuraman B et al. Optimal recall periods for patient-reported outcomes: challenges and potential solutions. Curr Med Res Opin 2009;25(4):929–42.

6. Schmier JK, Halpern MT. Patient recall and recall bias of health state and health status. Expert Rev Pharmacoeconomics Outcomes Res 2004;4(2):159–63.

7. Damiano AM, Pastores GM, Ware JE Jr. The health-related quality of life of adults with Gaucher's disease receiving enzyme replacement therapy: results from a retrospective study. Qual Life Res. 1998;7:373–386. 3

8. Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. J Clin Epidemiol. 2002;55(9):900–8.

9. Lucas NP, Macaskill P, Irwig I, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability. Journal of Clinical Epidemiology, 2010; 63:854-861.

10. Taylor R. Interpretation of the correlation coefficient: a basic review. J Diagnostic Medical Sonography 1990;6:35-9

11. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33 (1):159–174

12. Emberton M, Challands A, Styles RA, Wightman JA, Black N. Recollected versus contemporary patient reports of pre-operative symptoms in men undergoing transurethral prostatic resection for benign disease. Journal of Clinical Epidemiology, 1995;48(6):749–56.

13. Lingard EA, Wright EA, Sledge CB. Pitfalls of using patient recall to derive preoperative status in outcome studies of total knee arthroplasty. J Bone Joint Surg Am 2001;83(8):1149–56.

14. Bryant D, Norman G, Stratford P, Marx RG, Walter SD, Guyatt G, et al. Patients undergoing knee surgery provided accurate ratings of preoperative quality of life and function 2 weeks after surgery. J Clin Epidemiol 2006;59(9):984–93.

15. Howell J, Xu M, Duncan CP, Masri BA, Garbuz DS. A comparison between patient recall and concurrent measurement of preoperative quality of life outcome in total hip arthroplasty. J Arthroplasty 2008;23(6):843–9.

16. Marsh J, Bryant D, MacDonald SJ. Older patients can accurately recall their preoperative health status six weeks following total hip arthroplasty [Internet]. Journal of Bone and Joint Surgery - Series A. 2009;91:2827–37

17. Helfand BT, Fought A, Manvar AM et al. Determining the utility of recalled lower urinary tract symptoms. Urology 2010;76(2):442–7

18. Ware JEJ, Snow KK, Kosinski M, Gandek B. SF-36 Health Survey: Manual and interpretation guide. Boston: The Health Institute, New England Medical Centre, 1993

19. Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions ofpatients about total hip replacement. J Bone Joint Surg Br 1996;78(2):185-90

20. Stucki G, Sangha O, Stucki S, Michel BA, Tyndall A, Dick W, et al. Comparison of the WOMAC (Western Ontario and McMaster Universities) osteoarthritis index and a self-report format of the self-administered Lequesne-Algofunctional indexin patients with knee and hip osteoarthritis. Osteoarthritis Cartilage 1998;6(2):79-86

21. Smith S, Cano S, Lamping D, Staniszewska S,Browne J, Lewsey J, van der Meulen J, Cairns J, Black N. Patient-Reported Outcome Measures (PROMs) for routine use in Treatment Centres:recommendations based on a review of the scientific evidence. Report to the Department of Health, 2005.

22. Barry MJ, Fowler FJ Jr, O'Leary MP, et al. The American Urological Association symptom index for benign prostatic hyperplasia. J Urol 1992;148:1549-57

23. Kanakamedala, A. C., Anderson, A. F., & Irrgang, J. J. IKDC Subjective Knee Form and Marx Activity Rating Scale are suitable to evaluate all orthopaedic sports medicine knee conditions: a systematic review. Journal of ISAKOS: Joint Disorders & Orthopaedic Sports Medicine, 2016; 1(2)

24. Mohtadi N. Development and validation of the quality of life out- come measure (questionnaire) for chronic anterior cruciate ligament deficiency. Am J Sports Med 1998;26:350-9.

25. Mock C, MacKenzie E, Jurkovich G, Burgess a, Cushing B, deLateur B, et al. Determinants of disability after lower extremity fracture. J Trauma 2000;49(6):1002–11.

26. Michaels J, Madey SM, Krieg JC, Long WB. Traditional injury scoring underestimates the relative consequences of orthopedic injury. J Trauma 2001;50(3):389–95

27. Tidermark J, Zethraeus N, Svensson O, Törnkvist H, Ponzer S. Femoral neck fractures in the elderly: functional outcome and quality of life according to EuroQol. Qual Life Res 2002;11(5):473–81.

28. Ameratunga SN, Norton RN, Connor JL, Robinson E, Civil I, Coverdale J, et al. A population-based cohort study of longer-term changes in health of car drivers involved in serious crashes. Ann Emerg Med 2006;48(6):729–36.

29. Gabbe BJ, Cameron PA, Graves SE, Williamson OD, Edwards ER. Preinjury status: are orthopaedic trauma patients different than the general population? J Orthop Trauma 2007;21(4):223–8.

30. Watson WL, Ozanne-Smith J, Richardson J. Retrospective baseline measurement of self-reported health status and health-related quality of life versus population norms in the evaluation of post-injury losses. Injury Prevention 2007;13:45–50.

31. Gifford JM, Husain N, Dinglas VD, Colantuoni E, Needham DM. Baseline quality of life before intensive care: A comparison of patient versus proxy responses. Crit Care Med 2010;38(3):855–60

32. Lange RT, Iverson GL, Rose A. Post-concussion symptom reporting and the "good-old-days" bias following mild traumatic brain injury. Arch Clin Neuropsychol 2010;25(5):442–50.

33. Lyons R, Kendrick D, Towner EM, Christie N, Macey S, Coupland C et al. Measuring the population burden of injuries--implications for global and national estimates: a multi-centre prospective UK longitudinal study. PLoS Med 2011;8(12):e1001140

34. Tøien K, Bredal IS, Skogstad L, Myhren H, Ekeberg O. Health related quality of life in trauma patients. Data from a one-year follow up study compared with the general population. Scand J Trauma Resusc Emerg Med 2011;19(1):22.

35. Wilson R, Derrett S, Hansen P, Langley J. Retrospective evaluation versus population norms for the measurement of baseline health status. Health Qual Life Outcomes 2012;10(1):68.

36. Basso O, Olsen J, Bisanti L KW. The performance of several indicators in detecting recall bias. Epidemiology 1997;8(3):269-274

37. Skowronski JJ, Betz AL, Thompson CP, et al. Social memory in every day life: recall of self-events and other-events. J Pers Soc Psychol 1991;64:831-43.

38 Bradburn, N, Rips L, Shevell, S. Answering autobiographical questions: The impact of memory and inference on surveys. Sci New Ser 1987;236:157–61.

39 Sprangers M. Response-shift bias: a challenge to the assessment of patients' quality of life in cancer clinical trials. Cancer Treat Rev 1996;22 Suppl A:55–62.

40. Visser M, Oort F, Sprangers M. Methods to detect response shift in quality of life data: a convergent validity study. Qual Life Res. 2005;14(3):629–39.

41. Howard JS, Mattacola CG, Howell DM, Lattermann C. Response shift theory: An application for health-related quality of life in rehabilitation research and practice. Journal of Allied Health. 2011;40:31–38.

42 Black NA, Glickman ME, Ding J, Flood AB. International variation in intervention rates: what are the implications for patient selection? Int J Tech Ass in Health Care 1995;11:719-735.

Table 1: Literature search: concepts, keywords and search terms

	Search terms			
Concepts	retrospective	population norms	patient reported	outcomes
Keywords	retrospective recall historical bias recollected	population norm$	self-report$ patient report$ patient recall$ self-recall$	outcome$ quality * life H?Q?L EQ-5D function$ SF-36 health status symptom$

Table 2: Search strategy

1. retrospective or recall or historical or recollected
2. bias
3. population norm$
4. self-report$ or patient report$ or patient recall$ or self-recall$
5. outcome$ or quality * life or H?Q?L or EQ-5D or function$ or SF-36 or health status or symptom$
6. 1 OR 2 OR 3
7. 6 ADJ5 4
8. 7 ADJ10 5
9. limit 8 to (humans)
Combined search string:
((retrospective or recall or historical or bias or population norms or recollected) adj5 ((self-report$ or patient report$ or patient recall$ or self-recall$) adj10 (outcome$ or quality * life or H?Q?L or EQ-5D or function$ or SF-36 or health status or symptom$))).mp.

Table 3: Studies comparing retrospective to contemporary PROMs

Author **Country/Year**	**Condition/procedure** **Recall period** **Sample size**	**PROM/s**	**Level of association** (correlation coefficient)	**Level of agreement**	**Retrospective health compared to contemporary report***
Emberton (12) UK 1995	Benign prostatic hyperplasia 3 months n=75	AUA Symptom Index AUA Symptom Impact Index	*Pearson* Symptom Index: 0.6 Symptom Impact Index: 0.6	*Weighted Kappa* Symptom Index: 0.3 Symptom Impact Index: 0.3	No difference
Lingard (13) USA, UK and Australia 2001	Total knee arthroplasty 3 months n= 770	Western Ontario & McMaster Osteoarthritis Index (WOMAC) pain scale SF-36 function scale	*Spearman* WOMAC (pain scale): 0.53 SF-36 (function scale): 0.48	*Weighted Kappa* Individual items: 0.20 - 0.41	Worse for WOMAC pain scale (51.9% no difference, 31.3% recalled more pain, 16.8% recalled less pain) ($p < 0.001$) No consistent difference for SF-36 function scale (75% no difference, 11.8% recalled less limitation, 3.5% recalled more limitation) Patients recalled significantly less limitation for walking >1 mile ($p < 0.001$) but significantly more limitation for walking 100 yards ($p = 0.009$).
Bryant (14) Canada 2006	Knee surgery 2 weeks n=177	SF-36 International Knee Documentation Committee (IKDC) Subjective Form Anterior Cruciate Ligament QoL (ACL-QOL) Western Ontario Meniscal Evaluation Tool (WOMET)	*Pearson* SF-36(PCS): 0.81 SF-36 (MCS): 0.68 IKDC: 0.92 ACL-QOL: 0.86 WOMET: 0.88 KOOS: 0.93	*Intra-class coefficient* SF-36 (PCS): 0.81 SF-36 (MCS): 0.67 IKDC: 0.92 ACL-QOL: 0.86 WOMET: 0.88 KOOS: 0.93	No difference

		Knee Injury and Osteoarthritis Outcome Score (KOOS)			
Howell (15) Canada 2008	Total hip arthroplasty 3 days; 6 and 12 weeks n=104	WOMAC OHS SF-12 (PCS) SF-12 (MCS)	*Spearman* 3 days; 6 weeks; 12 weeks WOMAC: 0.80, 0.78, 0.86 OHS: 0.82, 0.80, 0.92 SF-12 (PCS): 0.66, 0.54, 0.76 SF-12 (MCS): 0.77, 0.71, 0.76	*Intra-class correlation* 3 days; 6 weeks; 12 weeks WOMAC: 0.86, 0.88, 0.93 OHS: 0.91, 0.88, 0.96 SF-12 (PCS): 0.83, 0.77, 0.90 SF-12 (MCS): 0.86, 0.84, 0.93	3 days: Worse (OHSΔ=1.58 p=0.01, WOMAC Δ=-2.21 p=0.029, SF-12 MCS Δ= -4.82 p<0.001) 6 weeks: Worse (SF-12 MCSΔ=-2.79 p=0.01) 12 weeks: No difference
Marsh (16) Canada 2009	Total hip arthroplasty 6 weeks n=174	WOMAC OHS SF-12 (PCS) SF-12 (MCS) Lower Extremity Functional Scale (LEFS) Feeling thermometer)	*Pearson* WOMAC: 0.89 OHS: 0.87 SF-12 (PCS): 0.62 SF-12 (MCS): 0.48 LEFS: 0.86 Feeling thermometer: 0.63	*Intra-class correlation* WOMAC: 0.88 OHS: 0.87 SF-12 (PCS): 0.58 SF-12 (MCS): 0.48 LEFS: 0.86 Feeling thermometer: 0.60	Better (SF-12 PCS Δ= 2.83, p<0.01) No difference (OHS Δ=-0.04, p=0.96; SF-12 MCS Δ=2.04, p=0.10) Worse (WOMAC Δ=2.74, p=0.01 Feeling thermometer Δ= -5.06, p<0.01)
Helfand (17) USA 2010	Urological conditions 6 months n=98	AUA Symptom Index (SI) Quality of life (QoL) scores	*Pearson* AUA SI: 0.73 QoL: 0.73	*Kappa* AUA SI: 0.56 QoL: 0.56	Better: AUA SI (recalled mean score 12.2, contemporary 13.1) No difference: QoL (recalled mean score 2.6, contemporary 2.6)

* mean difference or proportions different; p values

Table 4: Test-retest reliability of PROMs included in literature review

PROM	Test-retest reliability
SF-12	Physical component: ICC 0.83 [16] Mental component 0.91 [16]
SF-36	ICC=0.43-0.90 [18]
Oxford Hip Score	Bland Altman coefficient 7.27 [19]
WOMAC	ICC >0.7 [20]
Lower Extremity Functional Scale	ICC = 0.93 [21]
Feeling thermometer	ICC 0.94 [16]
AUA Symptom Index	r = 0.92 [22]
IKDC subjective Form	ICC=0.85 to 0.99[23]
ACL-QOL	Standard error of measurement (SEM.) is 6% [24]
WOMET	ICC=0.79 [14]
KOOS	ICC =0.75-0.93 [14]

Table 5: Studies comparing retrospective PROMs with age-sex standardised general population norms

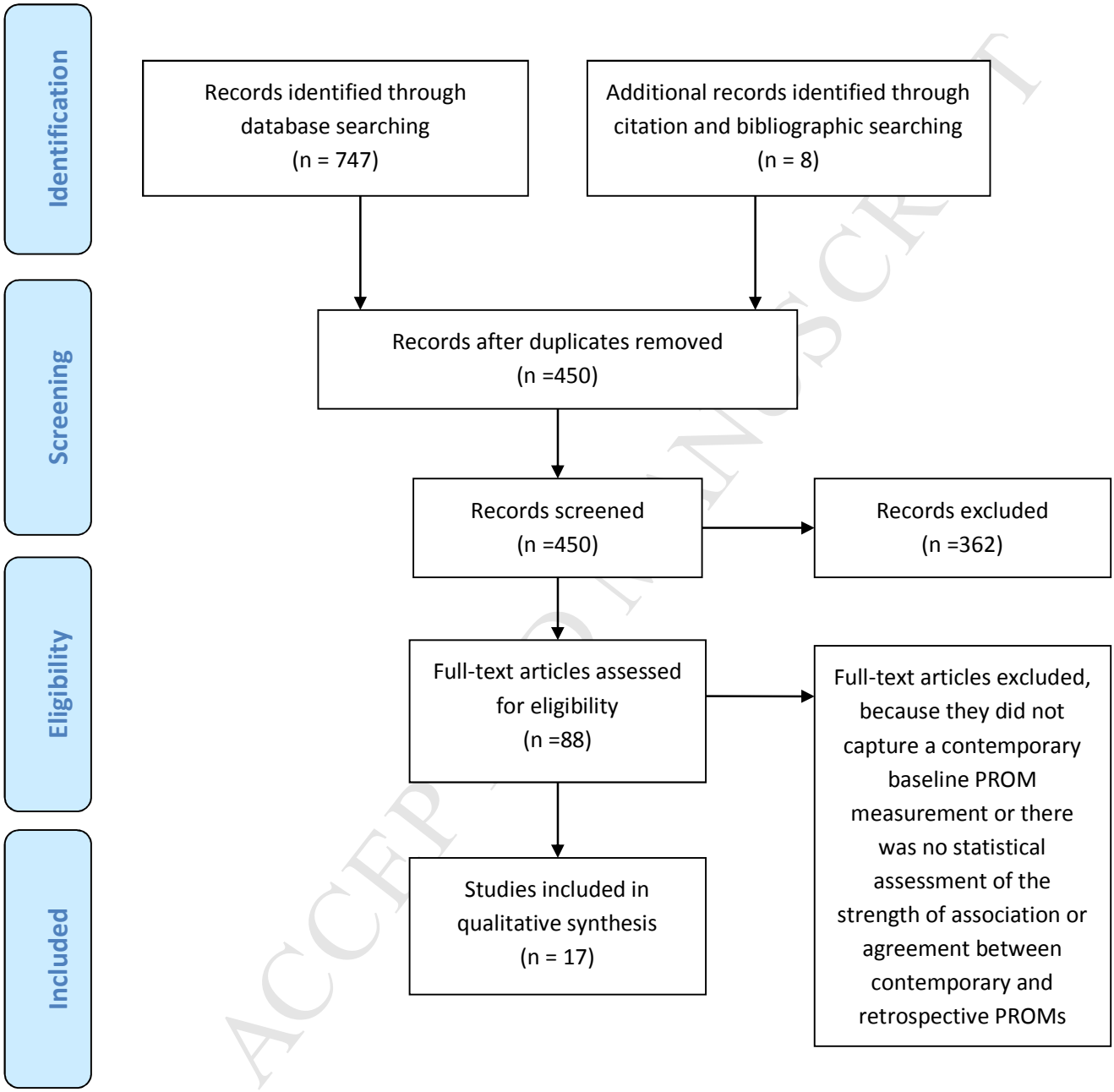
Author **Country/Year**	**Condition/procedure** **Recall period** **Number of patients Patient age and sex**	**PROM/s**	**Retrospective health compared to general population**[3]
Mock (25) USA 2000	Leg injury Weeks (hospital discharge) n=302. Adults (18-64 years)	Sickness Impact Profile (SIP)	No difference
Michaels (26) USA 2001	Trauma (blunt force) Days (early in hospital stay) n=165 Adults (mean age 37 years); 67% male	SF-36 SIP	No difference
Tidermark (27) Sweden 2002	Fractured neck of femur 12-48 hours after admission n=90 Elderly (mean age 80 years)	EQ-5D	No difference
Ameratunga (28) New Zealand 2006	Trauma from motor vehicle accident [1] One day n=472 Young adults (70% 15-44 years); 63% male	SF-36	Better than general population No difference from representative sample of drivers
Gabbe (29) Australia 2007	Trauma (mixed) Median 6 days (IQR 3-12 days) n=2388. Adults	SF-12	Better: SF-12 (physical) mean 50.9 vs. 48.9 ($p < 0.001$) SF-12 (mental) mean 54.5 vs. 52.4 ($p<0.001$), Differences confined to men and under 55 years.
Watson (30) Australia 2007	Trauma (mixed) 4 days (median) n=186. Adults (18-74 years)	SF-6D SF-36 AQoL	Better: AQoL population norm mean utility 0.83, recalled 0.95 SF-6D population norm mean utility 0.78, recalled 0.92 Better for all age groups ($p<0.05$).

Gifford (31) USA 2010	Acute lung injury Days-weeks (as soon as patient regained capacity) n=136. Adults (median age 49 years; IQR 40-60)	SF-36	Worse: mean paired difference for all SF-36 domains (mean paired differences ranged from 2.6-17.9) Mean paired difference were significantly better in population norm for all SF-36 domains ($p<0.01$) except for Vitality ($p=0.12$) Mean retrospective domain scores ranged 56.4-75.6, mean population norm domains scores ranged 58.9-87.6
Lange (32) Canada 2010	Mild traumatic brain injury[2] Median 1.8 months (0.2-8.0) n=86 Adults (mean age 37 years; SD 13.7)	British Columbia Post-Concussion Symptom Inventory	Better: overall score ($p < 0.01$) and in 6 of the 13 individual items ($p < 0.05$)
Lyons (33) UK 2011	Trauma (mixed) Within 7 days n=1517 Adults (median age 37 years; IQR 21-61)	EQ-5D	Better: mean score 3.3% (95% CI 1.9%–4.7%) higher
Toien (34) Norway 2011	Trauma (mixed) 17 days (non ICU) and 44 days (ICU patients) n= 242 Adults (mean age 42 years)	SF-36	Better: mean score higher ($p < 0.001$).
Wilson (35) New Zealand 2012	Trauma (mixed) 3 months n=2856. Adults (18-64 years)	EQ-5D	Better: Both the recovered and non-recovered groups had significantly better recalled than the population norm Recovered at 5-months: retrospective mean (SD) 0.98 (0.97-0.99) v norm 0.85 (0.84-0.86) Not Recovered at 5-months: retrospective mean (SD) 0.93 (0.92-0.94) v norms 0.85 (0.84-0.87) Recovered at 12-months: retrospective mean (SD) 0.96 (0.96-0.97)v norms 0.86 (0.85-0.87) Not Recovered at 12-months: retrospective mean (SD) 0.93(0.93-0.94) v norms 0.85 (0.83-0.86)

[1]also compared with representative sample of drivers; [2] compared with 177 community controls; [3] mean difference; p value

Figure 1: Search results

PRISMA 2009 Flow Diagram



From Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097