MAPPING UTILITY SCORES: HOW DOES IT COMPARE TO ACTUAL TRIAL RESULTS?

Garry Barton,[1] Tracey Sach,[2,3] Claire Jenkinson,[3] Anthony Avery,[3] Michael Doherty,[4] Kenneth Muir.[3]

[1] Health Economics Group, School of Medicine, Health Policy and Practice, University of East Anglia, Norwich, UK. e-mail: g.barton@uea.ac.uk

[2] School of Chemical Sciences and Pharmacy, University of East Anglia, Norwich, UK

[3] School of Community Health Sciences, University of Nottingham, Nottingham, UK

[4] Academic Rheumatology, University of Nottingham, Nottingham UK.

Aims: First, to develop (mapping) models that can be used to predict EQ-5D scores from scores on the Western Ontartio and McMaster Universities Osteoarthritis Index (WOMAC). Second, to compare actual quality adjusted life year (QALY) gains, and cost per QALY estimates, to those predicted using mapping models.

Methods: Within a study which compared 4 different interventions 389 individuals were asked to complete the EQ-5D and WOMAC at baseline, 6, 12, and 24 months post-intervention. Using baseline data various linear regression mapping models were developed. These mapping models were then used to predict the EQ-5D scores, at all time points, for individuals who had complete study data for the WOMAC and EQ-5D. The mean absolute error (MAE), between the predicted and actual EQ-5D score, was calculated for all EQ-5D post-intervention scores, and used to determine the preferred model. The area under the curve method was used to calculate the mean QALY gain associated with each intervention using both actual and predicted EQ-5D scores. Along with previously estimated costs the QALY gains were used to estimate the cost per QALY gain associated with each of the four interventions.

Results: At baseline 348 individuals completed both the EQ-5D and the WOMAC, and 259 had complete study data. The MAE in the preferred model was 0.129. Based on the predicted scores of this model the mean QALY gains for the four interventions were estimated to be 0.006, 0.058, 0.058, and 0.136 respectively, compared to the actual QALY gains of 0.089, 0.081, 0.120, and 0.149. The most effective intervention was also estimated to be associated

with a cost per QALY of £5,626, according to the same mapping model, but was estimated to have a cost per QALY of £11,170 when actual data was used.

Conclusion: Actual QALY gains differed from QALY gains predicted on the basis of mapping. This was also true for the cost per QALY estimates, suggesting that mapping should not be considered as a substitute for actual utility measurement.

INTRODUCTION

Given that health care resources are scarce there is a need to evaluate the cost-effectiveness of different health care interventions. Within such studies economists generally seek to measure the benefits in terms of utility, a scale where 0 represents death and 1 is equivalent to full health, in order for the benefits of many interventions to be compared on a common scale [1-3]. Collecting information on these measures of utility does however place a burden on both the patient and researchers. In an attempt to reduce this burden, an increasing amount of research has now been conducted on mapping, where results from condition-specific (non preference-based) measures are 'converted' into utility (preference-based) measures using a pre-defined formulae [4]. Mapping thereby presents the possibility of estimating the cost-utility (i.e. the cost per quality adjusted life year (QALY) [1]) of interventions that have previously only been evaluated using condition-specific measures. However, in spite of the increasing number of mapping models that have been developed [5-19], and the use of mapping in cost per QALY calculations [20, 21], we are not aware of any studies that have sought to compare either i) the actual QALY gain associated with different interventions to those predicted on the basis of mapping models, or ii) the cost per QALY estimated on the basis of actual results to the cost per QALY derived from mapping models, as such this constitutes the main focus of this paper.

METHODS

Participants

All participants were taking part in the Lifestyle Interventions for Knee Pain (LIKP) study, which was designed to compare the effectiveness and cost-effectiveness of four different interventions. The four interventions were receipt of a leaflet, advice on knee strengthening exercises, dietary advice, and both dietary and exercise advice (hereafter these interventions are referred to as 1, 2, 3 and 4 as the main focus of this paper is methodological). Ethical approval for this study was granted by the UK Nottingham Research Ethics Committee. In order to recruit people into the LIKP study all registered patients in five Nottingham general practices who were aged ≥45 years, and deemed (by their general practitioner) to be well enough to complete a questionnaire, were sent an ascertainment questionnaire, and a local media campaign was also conducted. Responding individuals were recruited into the LIKP study if they reported that they had had knee pain on most days of the

last month, were aged ≥45 years, had a body mass index (BMI) >28.0 kg/m^2, and gave consent to be randomised to one of the four interventions.

Outcome Measures

At both pre-intervention (baseline) and post-intervention (at 6, 12 and 24 months) participants in the LIKP study were asked to complete both the WOMAC (Western Ontario and McMaster Universities Osteoarthritis Index) and EQ-5D.

The WOMAC contains 24 questions and measures the amount of pain (5 questions), stiffness (2 questions), and difficulty in physical functioning (17 questions), where the response options are none (0), mild (1), moderate (2), severe (3) or extreme (4) [22]. Scores can thereby range between 0 and 20 on the pain sub-scale (pain), 0 and 8 on the stiffness sub-scale (stiffness), 0 and 68 on the functioning sub-scale (functioning), and sum to between 0 to 96 (total), where higher scores denote a worse response [23, 24]. Previous evidence of the adequate performance of the WOMAC has been shown for construct validity [25] and responsiveness [26, 27].

The EQ-5D has five questions, where the respondent is asked to report the level of problems they have (no problems, some/moderate problems, and severe/extreme problems) with regard to mobility, self-care, usual activities, pain, and anxiety/depression [28]. Responses to these five dimensions are converted into one of 243 different EQ-5D health state descriptions, which range between no problems on all five dimensions (11111) and severe/extreme problems on all five dimensions (33333). A utility score was assigned to each of these 243 health states using the York A1 tariff [29], which was based on the preferences elicited from a survey of 3395 UK residents – EQ-5D scores range between -0.594 and 1 (full health).

Costs

As described elsewhere [30], levels of resource use, for each individual, were combined with unit cost data to estimate the cost over the two year study period for all individuals in the LIKP study. Bootstrapping was subsequently used to estimate the mean cost associated with each of the four interventions.

Statistical Analyses

Model Specification

Linear regression analysis was used to develop mapping models which could predict EQ-5D scores. Five models were developed, starting with the most parsimonious. In each of the models different baseline WOMAC scores took the form of independent variables and the baseline EQ-5D score acted as the dependent variable. The predictor variables in each of the five models were as follows.

Model A: total;

Model B: pain, stiffness, functioning;

Model C: total, total^2

Model D: pain, stiffness, functioning, pain*stiffness, pain*functioning, stiffness*functioning, pain^2, stiffness^2, functioning^2;

Model E: best of above models plus patient characteristics of age and sex.

Model performance

We sought to identify the 'best' of the five aforementioned models by comparing actual EQ-5D scores to EQ-5D scores predicted on the basis of the five mapping models. This comparison was performed at 6, 12, and 24 months post-intervention for individuals who had complete study data (i.e. completed both the EQ-5D and each of the WOMAC sub-scales at all of the four time points within our study). Baseline data was not used within these comparisons as, in line with previous studies [6, 8], we sought to assess the performance of the model on a different sample of data to that used within the model. We inferred the 'best' model to be the one with the lowest Mean Absolute Error (MAE), where the MAE was calculated by taking the average value of each absolute prediction error (the prediction error equals the difference between actual EQ-5D score and the EQ-5D score predicted on the basis of the mapping model). For each model we also report the adjusted r^2 and the root mean square error (RMSE) (the RMSE is the positive square root of the average squared prediction error). Finally, in order to assess how well the model performs across the range of EQ-5D scores, we also plot the actual EQ-5D scores against the prediction errors.

In line with the mapping models which are developed here, a previous study has also attempted to predict utility scores using scores on the WOMAC [6]. The study differed from ours in that it measured utility using the Health Utilities Index [31], rather than the EQ-5D, and whilst acknowledging that there is an argument that utilities derived from different

instruments should not be compared [32], we also sought to compare the utility scores predicted by the mapping models of Grootendorst et al. [6] to our actual EQ-5D scores. Grootendorst et al. [6] developed four models, but we were only able to predict utility scores using two of their models (here referred to as G1 and G2) as the other two models used independent variables (e.g. duration of osteoarthritis) which were not available for the subjects within our study. Models G1 and G2 had the same independent variables as our Model D, and G2 also included the variables of age and sex. The performance of these two models was again assessed by calculating the MAE and RMSE, and by comparing the prediction errors to the actual EQ-5D scores.

Validation: Comparing actual trial results to those predicted using mapping models

QALY gain

The LIKP study sought to estimate the effectiveness of four different interventions. We thereby used the following methods to compare the QALY gain of each of these four interventions, as estimated by actual data, to that predicted on the basis of the mapping models. First, actual data, for each individual who had complete study data, was used to calculate the mean QALY gain associated with each of the four interventions (over the 24 month trial period) using the area under the curve (AUC) method, with adjustment for baseline scores [33]. No discounting of future benefits was undertaken within this calculation. Second, each of the five mapping models we had previously estimated, and models G1 and G2, were used to predict EQ-5D scores at baseline, 6, 12, and 24 months post-intervention for individuals who had complete study data (EQ-5D scores were predicted at each point of follow-up, including baseline, as this is the method that would most likely be used to predict utility scores in another study which had not measured outcomes in terms of utility). Third, these predicted EQ-5D scores were used to estimate the mean QALY gain associated with each of the four interventions, again using the same AUC method. Finally, we compared the mean QALY gain for each of the four interventions according to actual EQ-5D data to that predicted on the basis of our five mapping models, and models G1 and G2. The paired t-test was also used to assess whether the actual mean QALY gains differed significantly ($p<0.05$) from that predicted on the basis of each of the five models, and models G1 and G2.

Cost per QALY gain

For those individuals who had complete study data it was possible to use their cost data to estimate the cost per QALY gain, for each of the four interventions, based on both i)

actual EQ-5D data, and ii) EQ-5D scores estimated using the various mapping models. Cost per QALY calculations were made after excluding those interventions which either dominated (had a higher cost and lower effect) or subject to extended dominance (combinations of other interventions could provide a higher benefit at lower/equivalent cost).

RESULTS

Participants

Across the five general practices 12,500 individuals were sent an ascertainment questionnaire, and 8,044 (64.4%) were returned. Subsequently, 318 individuals met the entry criteria for the LIKP study, and gave consent to be randomised to one of the four interventions. An additional 71 participants were recruited via the media campaign. The mean age of these 389 participants was 62.0 years, 66.0% were female, and 23.4% were classified as overweight (BMI 25 to <30 kg/m^2), 50.4% as class I obese (30 to <35 kg/m^2), 16.9% as class II obese (35 to <40 kg/m^2), and 9.9% as class III obese (≥40 kg/m^2). At baseline 348 participants fully completed both the EQ-5D and the WOMAC, and data for these individuals were used to develop the five mapping models. The mean score (95% confidence interval) for these 348 individuals was 0.557 (0.528 to 0.587) on the EQ-5D, 7.76 (7.39 to 8.13) on the pain sub-scale, 3.91 (3.74 to 4.07) on the stiffness scale, 27.89 (26.54 to 29.23) on the physical functioning scale, and 39.55 (37.77 to 41.34) on the total WOMAC scale.

Statistical Analyses

Model Specification and performance

The parameter estimates for each of the five models that we developed to predict the baseline EQ-5D scores are summarised in Table 1, where it should be remembered that a higher WOMAC score denotes a worse response. When these models were used to predict the EQ-5D scores for the 259 individuals who had complete study data, it can be seen that Model C had the lowest MAE (0.140) out of the first four models when the actual scores at 6, 12 and 24 months were compared to those predicted on the basis of these models. As such, Model E used the same independent variables as model C, with the additional variables of age and sex. Model E had an MAE of 0.129, and was thus deemed to be our preferred model. Figure 1 shows how the prediction errors vary according to the actual EQ-5D scores. By way of an example of how these models are used to estimate EQ-5D scores, our preferred model would predict that a male with the aforementioned mean baseline characteristics (age=62 years; total=39.55) would have an EQ-5D score of 0.577 (-0.3474012785279 + (-

0.00059777089424435 2*39.55) + (-0.00010815602476990 3*39.55^2) + (0.0326027536450507*62) + (-0.0002352456360253 87*62^2) + (0.0475889686593746*0)), the actual mean baseline EQ-5D score was 0.566 (95% confidence interval 0.532 to 0.600). As for models G1 and G2, though the MAE of these two models was higher than that predicted by Model E, they did have a lower MAE than Model D which used the same WOMAC predictor variables. Additionally, for both models G1 and G2, the relationship between the prediction error and the actual EQ-5D score was near identical to that depicted in Figure 1 (analysis not shown, but available from author).

Validation: Comparing actual trial results to those predicted using mapping models

QALY gain

The WOMAC and EQ-5D were fully completed at baseline, 6, 12 and 24 months post-intervention by 259 individuals (66.6% of trial participants). Based on the responses for these individuals, over the trial period, the mean QALY gain, for each of the four interventions, was estimated to be 0.089, 0.081, 0.120 and 0.149, respectively, according to the AUC method. In Table 2 these values are compared to the QALY gains predicted using each of the five mapping models that we developed (A-E), and the QALY gains predicted on the basis of models G1 and G2. It can be seen that the QALY gains derived from the preferred model (Model E: 0.006, 0.058, 0.058 and 0.136) were consistently lower than the actual estimated QALY gains, and that this was also generally the case for the four other models we developed, and models G1 and G2. That said, the estimated mean QALY gain for each of the four interventions, based on actual results, did not differ significantly from the mean levels that were predicted on the basis of our preferred model, and significant differences were only found between the actual results for intervention 4 and those predicted by Models C, G1 and G2 (see Table 2).

Cost per QALY gain

Based on the responses for the 259 individuals who had complete study data the mean costs for each of the four interventions (1-4) were estimated to be –£18.84, £269.31, £775.18 and £702.27, respectively. The mean cost reduction can be explained by the fact that, on average, for those individuals receiving intervention 1, the reduced analgesic costs (over the two year study period) outweighed the cost of the intervention [30]. These four mean costs were subsequently combined with the estimated QALY gains, based on both actual data and the mapping models (as shown in Table 2), to give the cost per QALY estimates which are

reported in Table 3. Based on both actual and predicted EQ-5D scores intervention 4 had both a higher mean effect, and lower mean cost, than intervention 3 – intervention 4 thereby dominated intervention 3. When actual data was used to estimate the cost-effectiveness of intervention 2, this intervention was estimated to be subject to extended dominance as it had a higher successive incremental cost-effectiveness ratio (ICER) than intervention 4 (see [34] for further information on how to determine when an intervention is subject to extended dominance). Conversely, when each of the mapping models which were developed in this paper (A-E) were used to estimate the cost-effectiveness of intervention 2, this intervention was not estimated to be subject to extended dominance – the estimated ICER ranged between £4,118 and £5,903 per QALY (see Table 3). That said, when the two mapping models of Grootendorst et al. [6] were used to estimate the QALY gains, and, in turn, cost per QALY of intervention 2, in line with the estimates based on actual data, this intervention was estimated to be subject to extended dominance.

The above meant that the intervention to which intervention 4 was compared, in order to calculate the cost per QALY, varied between the calculations based on actual data and mapping models A-E. The cost per QALY associated with intervention 4 was estimated to be £11,170 when actual data were used (when compared to intervention 1), whereas the ICER was estimated to range between £5,863 and £7,953 (when compared to intervention 2) when the mapping models (A-E) were used. Finally, when the two models of Grootendorst et al. [6] were used (G1 and G2), intervention 4 was estimated to have an ICER of £5,915 and £5,914, respectively, (when compared to intervention 1).

DISCUSSION

Within this paper we have shown how mapping models can be used to predict the QALY gain associated with different interventions, and in turn calculate the ICER associated with different interventions. When these predicted results are compared to actual results we found that our preferred model consistently underestimated the mean QALY gain associated with the four compared interventions. The ICER of each of the four interventions, based on actual data, also differed from that based on the mapping models (see Table 3) – intervention 4 was estimated to be more cost-effective according to the mapping models (ICER= £5,863 to £7,953), compared to when actual data was used (ICER=£11,170). This result can be explained by the fact that the QALY gain of the most effective intervention (intervention 4) was the most closely estimated by the mapping models. Conversely, the QALY gains

associated with interventions 1 and 2 were particularly underestimated by the mapping models.

One possible explanation for the above QALY differences is as follows. Figure 1 shows that that the prediction errors of our preferred model tend to be increasingly positive for lower EQ-5D scores and increasingly negative for higher EQ-5D scores. This suggests that the regression would tend to over predict the EQ-5D score for those at low levels of utility, and under predict the EQ-5D score for those at high levels of utility. As the EQ-5D scores tend to increase post-intervention (baseline mean EQ-5D score = 0.566 and 24 month mean = 0.639, for N=259), the consequence of this is that the final EQ-5D scores tend to be underestimated and thus the QALY gain associated with each of the four interventions also tends to be underestimated.

One possible implication of the results presented here is that utility estimates that are based on mapping models should not be seen as a substitute for actual utility measurement. Moreover, prospective clinical trials should seek to measure outcomes with a utility measure, rather than using a condition specific measure and a mapping model to estimate the utility gain associated with an intervention. This also concurs with others who have pointed out that mapping models can only encompass the gains that are detected by the condition-specific measure [14].

Strengths and weaknesses

We consider the main potential limitation of our study to be that the results may not be generalizable. This arises because we only used the WOMAC to predict the EQ-5D score, and other studies which use a different condition-specific measure, or utility measure, may find the actual study results are more similar to those predicted on the basis of the mapping models.

The main strength of our paper is that we believe it is the first paper to compare the actual QALY gain associated with particular interventions to the QALY gain that would have been predicted, for the same interventions, on the basis of mapping. This comparison has been undertaken using both mapping models developed from baseline data, and the mapping models previously developed by Grootendorst et al. [6]. The former approach of using both a condition-specific measure and a utility measure at baseline, in order to develop a mapping model, and then only using a condition-specific measure thereafter, was adopted as we considered that future trials might use this methodology. Here, we have shown that both this

approach, and the mapping of utility scores at all time points, can underestimate the 2 year QALY gain associated with an intervention by as much as 0.09 to 0.11 QALYs.

Comparisons with other studies

We are aware of three papers which have compared actual utility scores to those predicted by a mapping model [8, 13, 16]. Bansback et al. [13] estimated the mean EQ-5D utility gain for those who were deemed to have responded (according to the Health Assessment Questionnaire) to be 0.10, compared to a prediction of 0.13 according to the mapping model. This mean difference of 0.03 (N=76), and other mean differences, were considered to be small [13]. Similarly, Buxton et al. [8] found that the mean difference between the observed and predicted SF-6D scores was 0.0011 (N=802), and Dobrez et al. [16] found the mean difference between actual and predicted time trade off (TTO) scores to be 0.027 (N=717). We found that the mean utility gain on the EQ-5D (across all four interventions at 24 months post-intervention) was 0.068, compared to a predicted mean gain of 0.038 (mean difference = 0.03, N=259), though this higher mean difference did not seem to arise because of a higher prediction error – the MAE (0.140) and RMSE (0.185) of our preferred model was comparable to the MAE of Dobrez et al. [16] (0.19) and the RMSE in the preferred EQ-5D model developed by Bansback et al. [13] (0.183).

CONCLUSIONS

We have shown how mapping can be used to estimate both the QALY gain, and cost per QALY, associated with different interventions, and compared these predictions to actual results. In our study the mapping models developed from the WOMAC tended to underestimate the QALY gain associated with each of four interventions, compared to that which was derived from actual EQ-5D scores. Similarly, the cost per QALY estimates based on the mapping models also differed from those based on actual data. This suggests that mapping should not be considered as a direct substitute for actual utility measurement.

REFERENCES

1. Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. Methods for the Economic Evaluation of Health Care Programmes (3rd Edition). New York: Oxford University Press; 2005.
2. Barton GR, Bankart J, Davis AC. A comparison of the quality of life of hearing-impaired people as estimated by three different utility measures. Int J Audiol. 2005;44:157-163.
3. Sach TH, Barton GR, Doherty M, Muir K, Jenkinson C, Avery AJ. The relationship between BMI and health related quality of life: comparing the EQ-5D, EuroQol VAS, and SF-6D. Int J Obes Relat Metab Disord. 2007;31:189-196.
4. Brazier JE, Ratcliffe J, Salomon JA, Tsuchiya A. Measuring and Valuing Health Benefits for Economic Evaluation. New York: Oxford University Press Inc.; 2007.
5. Sengupta N, Nichol MB, Wu J, Globe D. Mapping the SF-12 to the HUI3 and VAS in a managed care population. Med Care. 2004;42:927-937.
6. Grootendorst P, Marshall D, D. P, Bellamy N, Feeny D, Torrance GW. A model to estimate health utilities index mark 3 utility scores from WOMAC index scores in patients with osteoarthritis of the knee. J Rheumatol. 2007;34:534-542.
7. Brennan DS, Spencer AJ. Mapping oral health related quality of life to generic health state values. BMC Health Serv Res. 2006;6:96.
8. Buxton MJ, Lacey LA, Feagan BG, Niecko T, Miller DW, Townsend RJ. Mapping from disease-specific measures to utility: an analysis of the relationships between the Inflammatory Bowel Disease Questionnaire and Crohn's Disease Activity Index in Crohn's disease and measures of utility. Value Health. 2007;10:214-220.
9. Sullivan PW, Ghushchyan V. Mapping the EQ-5D index from the SF-12: US general population preferences in a nationally representative sample. Med Decis Making. 2006;26:401-409.
10. Franks P, Lubetkin EI, Gold MR, Tancredi DJ. Mapping the SF-12 to preference-based instruments: convergent validity in a low-income, minority population. Med Care. 2003;41:1277-1283.
11. Franks P, Lubetkin EI, Gold MR, Tancredi DJ, Jia H. Mapping the SF-12 to the EuroQol EQ-5D Index in a national US sample. Med Decis Making 2004;24:247-254.

12. Gray AM, Rivero-Arias O, Clarke PM. Estimating the association between SF-12 responses and EQ-5D utility values by response mapping. Med Decis Making. 2006;26:18-29.
13. Bansback N, Marra C, Tsuchiya A, et al. Using the health assessment questionnaire to estimate preference-based single indices in patients with rheumatoid arthritis. Arthritis Rheum. 2007;15:963-971.
14. Brazier JE, Kolotkin RL, Crosby RD, Williams GR. Estimating a Preference-Based Single Index for the Impact of Weight on Quality of Life-Lite (IWQOL-Lite) Instrument from the SF-6D. Value Health 2004;7:490-498.
15. Yang M, Dubois D, Kosinski M, Sun X, Gajria K. Mapping MOS Sleep Scale scores to SF6D utility index. Curr Med Res Opin. . in press;
16. Dobrez D, Cella D, Pickard AS, Lai JS, Nickolov A. Estimation of patient preference-based utility weights from the functional assessment of cancer therapy - general. Value Health. 2007;10:266-272.
17. Lawrence WF, Fleishman JA. Predicting EuroQoL EQ-5D preference scores from the SF-12 Health Survey in a nationally representative sample. Med Decis Making. 2004;24:160-169.
18. Longworth L, Buxton MJ, Sculpher M, Smith DH. Estimating utility data from clinical indicators for patients with stable angina. Eur J Health Econ. 2005;6:347-353.
19. Nichol MB, Sengupta N, Globe DR. Evaluating quality-adjusted life years: estimation of the health utility index (HUI2) from the SF-36. Med Decis Making. 2001;21:105-112.
20. Brennan A, Bansback N, Reynolds A, Conway P. Modelling the cost-effectiveness of etanercept in adults with rheumatoid arthritis in the UK. Rheumatology 2004;43:62–72.
21. Barton P, Jobanputra P, Wilson J, Bryan S, Burls A. The use of modelling to evaluate new drugs for patients with a chronic condition: the case of antibodies against tumour necrosis factor in rheumatoid arthritis. Health Tech Assess. 2004;8 (11):1-104.
22. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol. 1988;15:1833-1840.
23. Bellamy N. WOMAC: a 20-year experiential review of a patient-centered self-reported health status questionnaire. J Rheumatol. 2002;29:2473-2476.

24. McCarthy CJ, Mills PM, Pullen R, et al. Supplementation of a home-based exercise programme with a class-based programme for people with osteoarthritis of the knees: a randomised controlled trial and health economic analysis. Health Technol Assess. 2004;8 (46):1-76.
25. Miller GD, Rejeski WJ, Williamson JD, et al. The Arthritis, Diet and Activity Promotion Trial (ADAPT): design, rationale, and baseline results. Control Clin Trials. 2003;24:462-480.
26. Davies GM, Watson DJ, Bellamy N. Comparison of the responsiveness and relative effect size of the western Ontario and McMaster Universities Osteoarthritis Index and the short-form Medical Outcomes Study Survey in a randomized, clinical trial of osteoarthritis patients. Arthritis Rheum. 1999;12:172-179.
27. Theiler R, Bischoff-Ferrari HA, Good M, Bellamy N. Responsiveness of the electronic touch screen WOMAC 3.1 OA Index in a short term clinical trial with rofecoxib. Osteoarthritis Cartilage. 2004;12:912-916.
28. Brooks R. EuroQol: the current state of play. Health Policy. 1996;37:53-72.
29. Dolan P, Gudex C, Kind P, Williams A. A social tariff for the EuroQol: results from a UK general population survey (Discussion Paper 138). York, UK: Centre for Health Economics, University of York; 1995.
30. Barton GR, Sach TH, Avery AJ, Doherty M, Jenkinson C, Muir KR. Lifestyle Interventions for Knee Pain for persons aged ≥45 and overweight or obese: Cost-effectiveness analysis. Paper submitted for publication.
31. Feeny D, Furlong W, Torrance GW, et al. Multi-attribute and single attribute utility functions for the Health Utilities Index Mark 3 system. Med Care 2002;40:113-128.
32. Lenert L, Kaplan RM. Validity and interpretation of preference-based measures of health-related quality of life. Med Care. 2000;38 (Supplement 2):138-150.
33. Manca A, Hawkins N, Sculpher MJ. Estimating mean QALYs in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. Health Econ. 2005;14:487-496.
34. UK BEAM Trial Team. United Kingdom back pain exercise and manipulation (UK BEAM) randomised trial: cost effectiveness of physical treatments for back pain in primary care. BMJ. 2004;329:1381.

Table 1. Parameter estimates for the five models (A-E) which were used to predict the baseline EQ-5D scores (based on data for 348 individuals), and the two previous models previously developed by Grootendorst et al. (2007). Levels of significance are reported for Models A-E (* p<0.05, † p<0.01, and ‡ p<0.001).

	Model						
	A	B	C	D	E	G1	G2
Intercept	0.900‡	0.886‡	0.747‡	0.691‡	-0.347	0.823	0.590
total	-0.009‡		0.001		-0.001		
pain		-0.012		-0.009		0.010	0.010
stiffness		0.006		0.092†		0.010	0.007
functioning		-0.009‡		-0.005		-0.007	-0.007
pain* stiffness				-0.005		0.003	0.003
pain*functioning				-0.001		0.001	0.001
stiffness*functioning				0.001		0.000	0.000
$pain^2$				0.002		-0.003	-0.003
$stiffness^2$				-0.011		-0.003	-0.003
$functioning^2$				0.000		0.000	0.000
$total^2$			-0.000‡		-0.000†		
age					0.033*		0.009
age^2					0.000		-0.000
sex (if Female)					0.048		-0.025
MAE	0.148	0.147	0.140	0.146	0.129	0.142	0.144
Adjusted r^2	0.275	0.274	0.296	0.299	0.313		
RMSE	0.189	0.187	0.185	0.190	0.180	0.201	0.203

MAE = Mean Absolute Error; RMSE = Root Mean Squared Error

Table 2. The mean estimated QALY gain associated with each of the four interventions based on both actual data and mapping models. The estimated mean difference and results of the paired t-test are also reported in brackets (* $p<0.05$, † $p<0.01$, and ‡ $p<0.001$).

	Intervention			
	1 (N=58)	2 (N=47)	3 (N=82)	4 (N=72)
Actual results	0.089	0.081	0.120	0.149
Model A	0.023 (-0.066)	0.085 (0.004)	0.073 (-0.047)	0.148 (0.000)
Model B	0.017 (-0.072)	0.084 (0.003)	0.072 (-0.049)	0.144 (-0.004)
Model C	0.001 (-0.088)*	0.048 (-0.033)	0.050 (-0.070)	0.123 (-0.025)
Model D	-0.002 (-0.091)	0.051 (-0.030)	0.045 (-0.075)	0.120 (-0.028)
Model E	0.006 (-0.083)	0.058 (-0.023)	0.058 (-0.062)	0.136 (-0.013)
Model G1	-0.002 (-0.098)*	0.027 (-0.054)	0.062 (-0.059)	0.114 (-0.034)
Model G2	-0.011 (-0.100)*	0.026 (-0.055)	0.059 (-0.061)	0.113 (-0.036)

Table 3. The estimated cost per QALY gain associated with each of the four interventions based on both actual data and mapping models.

	Intervention			
	1 (N=58)	2 (N=47)	3 (N=82)	4 (N=72)
Actual results	N/A	Subject to ED	D by 4	£11,170
Model A	N/A	£4,463	D by 4	£7,541
Model B	N/A	£4,118	D by 4	£7,953
Model C	N/A	£5,903	D by 4	£6,017
Model D	N/A	£5,360	D by 4	£6,604
Model E	N/A	£5,333	D by 4	£5,863
Model G1	N/A	Subject to ED	D by 4	£5,915
Model G2	N/A	Subject to ED	D by 4	£5,914

N/A = Not applicable (this intervention is the least costly and least effective)

ED= Extended dominance

D=Dominated

Figure 1. Comparison of the actual EQ-5D scores and the prediction errors of Model E. Scores were available for 259 individuals at 6, 12 and 24 months post-intervention (N=777 data points).