BMJ Open

A unified multi-level model approach to assessing patient responsiveness including; return to normal, minimally important differences and minimal clinically important improvement for patient reported outcome measures

Adrian Sayers,[1,2] Vikki Wylde,[1] Erik Lenguerrand,[1] Rachael Gooberman-Hill,[1] Jill Dawson,[3] David Beard,[4] Andrew Price,[4] Ashley W Blom[1]

To cite: Sayers A, Wylde V, Lenguerrand E, *et al*. A unified multi-level model approach to assessing patient responsiveness including; return to normal, minimally important differences and minimal clinically important improvement for patient reported outcome measures. *BMJ Open* 2017;**7**:e014041. doi:10.1136/bmjopen-2016-014041

- ▸ Prepublication history and additional material for this paper are available online. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2016-014041).

Received 26 August 2016
Revised 10 May 2017
Accepted 19 May 2017

For numbered affiliations see end of article.

Correspondence to
Dr. Adrian Sayers; adrian.sayers@bristol.ac.uk

ABSTRACT

Objective This article reviews and compares four commonly used approaches to assess patient responsiveness with a treatment or therapy (return to normal (RTN), minimal important difference (MID), minimal clinically important improvement (MCII), OMERACT-OARSI [Outcome Measures in Rheumatology—Osteoarthris Reseach Society International] (OO)) and demonstrates how each of the methods can be formulated in a multilevel modelling (MLM) framework.

Design Cohort study.

Setting A cohort of patients undergoing total hip and knee replacement were recruited from a single UK National Health Service hospital.

Population 400 patients from the Arthroplasty Pain Experience cohort study undergoing total hip (n=210) and knee (n=190) replacement who completed the Intermittent and Constant Osteoarthritis Pain questionnaire prior to surgery and then at 3, 6 and 12 months after surgery.

Primary outcomes The primary outcome was defined as a response to treatment following total hip or knee replacement. We compared baseline scores, change scores and proportion of individuals defined as 'responders' using traditional and MLM approaches with patient responsiveness.

Results Using existing approaches, baseline and change scores are underestimated, and the variance of baseline and change scores overestimated in comparison with MLM approaches. MLM increases the proportion of individuals defined as responding in RTN, MID and OO criteria compared with existing approaches. Using MLM with the MCII criteria reduces the number of individuals identified as responders.

Conclusion MLM improves the estimation of the SD of baseline and change scores by explicitly incorporating measurement error into the model and avoiding regression to the mean when making individual predictions. Using refined definitions of responsiveness may lead to a reduction in misclassification when attempting to predict who does and does not respond to an intervention and clarifies the similarities between existing methods.

Strengths and limitations of this study

- ▸ Four different approaches to patient responsiveness can be unified into a multilevel model.
- ▸ A multilevel model framework of patient responsiveness highlights the similarities and differences between existing methods.
- ▸ Multilevel models provide a simple framework which incorporates measurement error and non-linear change in trajectories of patient recovery.
- ▸ Multilevel models are technically more demanding than existing formulations of patient responsiveness, and convergence is not guaranteed.
- ▸ Multilevel models does not improve the arbitrary placement of the thresholds that define responsiveness in comparison with existing methods.

INTRODUCTION

Joint replacement is an increasingly common elective procedure worldwide[1–3] and improving patient-reported outcomes after joint replacement is a key research priority due to the high prevalence of poor outcomes after joint arthroplasty.[4] Poor outcomes include continuing pain, functional limitations[5] and increased healthcare utilisation.[6] However, there is some debate on how the efficacy of interventions can be judged due to the variety of different outcomes used in orthopaedic research.[7–18] Traditionally, objective primary outcomes such as prosthetic survivorship and mortality rates were used.[19] However, more recently there has been a shift in focus which ensures that patients' perspective is central to the assessment of intervention success.[20] Many studies now use patient-reported outcome measures (PROMs) as endpoints, and these tools can assess a variety of health outcomes, including pain,[7 21] physical functioning,[7] mental well-being[22] and health-related quality of life.[23]

Although PROMs are widely used,[4] there is still debate in how the results should be interpreted and how to define a clinically meaningful change.[24–35] From a measurement perspective, the ability to estimate if a change has occurred depends on the application of an appropriate statistical model. From a clinical perspective, some authors suggest that the average statistical change is insufficient to 'tell you anything about an individual's chances of improving'.[36] Therefore, the utility of simple statistical analyses are limited when attempting to help patients weigh up the risks and benefits of undergoing surgery.

To supplement simple statistical analysis, many researchers attempt to dichotomise the population into those who have or have not responded to an intervention, creating a two-stage process of defining an outcome. There are a number of different methods (definitions) that can be used to dichotomise the population, and these secondary analyses are collectively referred to as responsiveness analyses.[36] Four substantively different methods of estimating the proportion of individuals who respond to an intervention have been previously identified in orthopaedic research[36]: (1) return to normal (RTN), (2) distribution-based minimally important difference (MID), (3) anchor-based minimal clinically important difference (MCII) and (4) the OMERACT-OARSI (OO) responder criteria. The first three approaches are generic and used in many fields of health research, whereas the fourth approach is specific to orthopaedic research, but in principle could be used in many fields of health research.

Each of these approaches is often thought to be methodologically distinct. However, all of the methods can be shown to be special cases of a multilevel model (MLM). MLM have been used in a wide variety of contexts ranging from growth modelling to modelling educational data. One of the principal reasons to use MLM is to take advantage of the direct estimation of different variance components[37] and provide efficient and unbiased estimates of fixed and random effects.[38]

Despite a number of extensive reviews of patient responsiveness,[31 33 39 40] we will describe these four approaches to calculating responsiveness and highlight the substantively different decisions each method makes. We will then describe how each approach can be translated into a MLM framework, emphasising the benefits of the translation and contrast the approaches using an example from the APEX (Arthroplasty Pain Experience)cohort study.[41]

METHODS

We outline the four existing approaches to patient responsiveness previously used in orthopaedic research[36] and describe their potential limitations and how they can be formulated in an MLM framework.

Review of existing approaches to responsiveness

Return to normal (RTN)[26] suggests that an individual has returned to 'normal' if their score on a postintervention outcome is greater than 2 SDs from the mean baseline response.

The use of 2 SD appears to be justified on theoretical grounds; however, it is quite arbitrary. Assuming scores are normally distributed and measured without error, 2 SDs corresponds to a 95.5% prediction interval for the mean, which is similar to the equally arbitrary and much-criticised significance threshold p=0.05 (type I error=0.05) criterion used throughout medical research.[42 43] However, there is no reason why a 1.6 or a 2.6 SD cut-offs should not be used in preference, which corresponds to 90% and 99% prediction intervals.

The method also assumes the observed change is unlikely to be due to chance alone and does not account for any uncertainty. To alleviate this problem the use of the Relative Change Index (RCI) was proposed to be used in conjunction with the RTN classification.[24 27] The RCI constructs a test of the individual's score at follow-up compared with their baseline, where the SE of the difference is estimated indirectly using the SD of the baseline score and an assumed reliability coefficient from empirical research or a range of reliability values in the spirit of a sensitivity analysis.

A commonly described distribution-based minimally important difference (MID) method classifies individuals as responders if their observed change is greater than a fixed proportion of the SD of the presurgery score.[30] There has been much debate about the exact size, or proportion, of the SD change score to use; however, 0.5 SDs have been reported widely and suggested to be a difference that is minimally perceptible to patients.[30] Any individual with a change score greater than 0.5 SD of the baseline score is defined as responding to the treatment. Similar to the RTN criteria, the decision to use 0.5 is arbitrary and there is no reason why more or less stringent criteria of 0.25, 1 or 2 SDs could not be used. Additionally, there is no reason why a test such as the RCI should not be conducted to check that change is beyond the bounds of measurement error.

Anchor-based minimal clinically important improvement (MCII) is similar to the MID approach, in that it defines an individual as a responder based on their individual change score. However, the cut point is determined in individuals who report themselves as having an outcome which is either good/satisfactory or perceived as improved from baseline using an external anchoring question. The authors proposed using a cut point at the 75th centile of the change score in those who are satisfied.[34] Therefore any individuals, whether they are satisfied or not, who has a change score greater than the 75th centile are defined as responders. A closely related anchor-based metric is the patient acceptable symptom state (PASS),[35] the construction is similar to that of the MCII with the exception that it is based on the final score of patients opposed to change. Conceptually, the PASS is more closely related to the RTN definition of responsiveness, and much of the criticism levied against MCII and RTN can therefore be applied to the PASS.

The OMERACT-OARSI (OO) criteria[32] recognises that a response to an intervention may occur in one or more different measured outcomes, that is, a multivariate response mechanism. In keeping with much of the orthopaedic literature, they assume the proposed score has been rescaled between 0 and 100,[32] and that a responder is defined as any individual with (1) a ≥50% relative change or a ≥20-point absolute change on one or more responses scales or (2) a ≥20% relative change or a ≥10-point absolute change in two or more response scales. Relative change is defined as the ratio of the change to the individual baseline score multiplied by 100. Unlike the RTN, MID or MCII, it is very clear that the thresholds for relative and absolute changes are based on a panel of expert opinions and are fixed.

Despite the variety of existing approaches used to identifying responders, there are a number of problems common to all methods. Common assumptions include: (1) each observed outcome is measured without error and reflects the true underlying patient's response, test–retest reliability studies indicate that this is not a realistic assumption[44]; (2) regression to the mean does not occur and therefore the variance of the change score will not be overestimated; (3) floor and ceiling effects do not bias estimates of the variance of the change score.[45]

Furthermore, in RTN, specific combinations of means and variances may result in a threshold beyond the range of the measurement tool, therefore no individuals would be defined as responding to a therapy. The MCII approach assumes the additional anchoring variable is measured without error and the response trajectory is distinct from those who are unsatisfied.[46] The method also assumes a two-parameter logistic function is an appropriate model for the cumulative proportional rank of patients and change in outcome, and that there is no uncertainty in the calculation of the threshold.[47] Finally, the OO approach considers a response in two or more outcomes. However, it does not explicitly describe how the correlation between the two outcomes is accounted for and fails to recognise that if not modelled appropriately may introduce bias.[48–50]

The four methods identified have a number of other limitations,[25] but they are difficult to compare methods when presented as distinct approaches.

Embedding them in a unified statistical framework makes their underlying assumptions explicit, while highlighting their similarities and differences. In addition, it provides a framework to incorporate non-linear change, measurement error and variability in the timing of measurement occasions, all of which are to be expected in real word data collections and are critical when attempting to asses a patients change at a specified point in time.

MLM approach to responsiveness

We now present a general MLM for patient responsiveness and show how the four approaches described above can be specified as special cases.

Under the assumption of linear change, the measured response (y) at the *i*th occasion for the *j*th individual is modelled as a linear function of time.

$$
\begin{gathered}
y_{ij} = \beta_0 + u_{0j} + \left(\beta_1 + u_{1j}\right) t_{ij} + \varepsilon_{ij} \\
\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u), \qquad \Omega_u = \begin{bmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix} \\
\left[\varepsilon_{ij}\right] \sim N\left(0, \sigma^2_{\varepsilon}\right)
\end{gathered} \tag{1}
$$

where t_{ij} is the time at which measurement was taken on individual j, coded as zero at baseline. β_0 is the baseline population average response and u_{0j} represents the *j*th individual difference from the baseline response. The sum of $\beta_0 + u_{0j}$ is the estimated individual baseline response. β_1 represents the population average change per unit increase in time and u_{1j} represents the *j*th individual difference from the population average change per unit increase in time. The sum of $\beta_1 + u_{1j}$ is the estimated individual average change per unit increase in time. Measurement error in the linear trajectory is represented by ε_{ij}.

The variance in individual deviations from the population average response at baseline and average rate of change are σ^2_{u0} and σ^2_{u1}, respectively. Furthermore, the correlation between baseline measurements and rate of change can be assumed to be independent or correlated by constraining σ_{u01} to be zero or allowing it to be freely estimated. The variances of the shrunken residuals $\hat{u}_{0j}$ and $\hat{u}_{1j}$, also known as empirical Bayes estimates, are typically less than the estimated population variances $\hat{\sigma}^2_{u0}$ and $\hat{\sigma}^2_{u1}$ as they shrink towards the population averages of β_0 and β_1. The extent of the shrinkage depends on the number of measurement occasions and the within-individual variability, with greater shrinkage as the number of measurement occasions decrease and as the within-individual variance increases. A more detailed discussion of MLM can be found in most advanced statistics textbooks.[48 51 52]

We now describe how the four traditional approaches to measuring patient responsiveness can be unified into a MLM framework. General benefits of the MLM over existing approaches include: (1) with more than three measurement occasions, an MLM directly allows for measurement error, ε_{ij}; (2) the use of shrunken residuals $\hat{u_0}_j$ and $\hat{u_1}_j$ allows for regression to the mean when predicting an individual's score[53]; (3) MLM can be extended to include multivariate response models which appropriately model the correlation between two or more outcomes and (4) MLM allows for variability in the timing of measurement occasions. Fundamentally, the MLM approach recognises that observed patient responses are subject to error, and therefore the true patient's response following an intervention must be estimated.

MLM: return to normal

To apply the RTN criteria using an MLM approach, we first estimate the baseline population SD in individuals considered to be abnormal using the model described in equation 1. Assuming y_{ij} is normally distributed at baseline with a population mean β_0 and variance $\sigma^2_{u0}\mathrm{a}100 \cdot \left(1 - \frac{\alpha}{2}\right)$, prediction interval for the baseline measurement can be constructed, that is, $\left[\beta_0 - \sigma_{u0}z_{\left(1-\frac{\alpha}{2}\right)}, \beta_0 + \sigma_{u0}z_{\left(1-\frac{\alpha}{2}\right)}\right]$ where α is the type I error rate and z is the critical value from a standard normal distribution. Importantly, y_{ij} is not assumed to be measured without error, and therefore estimates of σ^2_{u0} are less likely to be biased than using simple methods. However, it is important to note that the choice of α is entirely that of the researcher, and while α=0.05 (leading to $z = 1.96 \approx 2$) is common, more or less stringent criteria could be applied.

The second step is to estimate the score of the individual at time j following surgery and determine if it is within the baseline prediction interval. This prediction is simply calculated by substituting estimates of β_0, β_1 , u_{0j} and u_{1j} into equation 1, to give the empirical best linear unbiased prediction or the jth individual at the ith occasion.[54]

Finally, to determine whether or not the response of the individual following surgery is greater than one would attribute to chance alone, that is, the null hypothesis that the jth individual's slope is not equal to zero, a test statistic similar to RCI should be conducted,

$$\left(\hat{\beta}_1 + \hat{u}_{1j}\right) / SE\left(\hat{\beta}_1 + \hat{u}_{1j}\right), \text{where } SE\left(\hat{\beta}_1 + \hat{u}_{1j}\right)$$
$$= \sqrt{VAR\left(\hat{\beta}_1\right) + VAR\left(\hat{u}_{1j}\right)}$$

MLM: minimally important difference

The threshold of minimally important difference can also be estimated using an MLM. Similar to RTN, a linear model of change is applied, as in equation 1. Then the population SD of the baseline response is estimated by σ_{u0}. By comparing the estimated change for the jth individual $\left(\hat{\beta}_1 + \hat{u}_{1j}\right) t$ with the baseline SD, that is, $\sigma_{u0}/2$, the individual can be classed as a responder or not. The MID approach does not specifically state whether a test of whether an individual's change scores is less than the MID threshold should be conducted, but a test statistic is simply constructed as $\left(\left(\hat{\beta}_1 + \hat{u}_{1j}\right) t - \left(\frac{\hat{\sigma}_{u0}}{2}\right)\right) / \left(SE\left(\hat{\beta}_1 + \hat{u}_{1j}\right) t\right)$.

MLM minimal clinically important improvement

The MLM MCII requires a simple extension of the univariate model presented previously (equation 1). The outcome of interest is stratified using an external criterion. The stratification is achieved by creating dummy variables for those who are unsatisfied/satisfied with some aspect of their treatment, for example, x_{1i} takes the values 0 and 1 representing unsatisfied and satisfied individuals, respectively, and $x_{2i} = 1 - x_{1i}$. These dummy variables are then included as additional explanatory variables, with no overall model intercept, and interacted with t.

$$\begin{gathered}
y_{ij} = \left(\beta_0 + u_{0j}\right) x_{1i} + \left(\beta_1 + u_{1j}\right) t_{ij}x_{1i} + \varepsilon_{1ij}x_{1i} \\
+ \left(\beta_2 + u_{2j}\right) x_{2i} + \left(\beta_3 + u_{3j}\right) t_{ij}x_{2i} + \varepsilon_{2ij}x_{2i} \\
\begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \end{bmatrix} \sim N(0, \Omega_u) : \qquad \Omega_u = \begin{bmatrix} \sigma^2_{u0} & & & \\ \sigma_{u01} & \sigma^2_{u1} & & \\ 0 & 0 & \sigma^2_{u2} & \\ 0 & 0 & \sigma_{u23} & \sigma^2_{u3} \end{bmatrix} \\
\begin{bmatrix} \varepsilon_{1ij} \\ \varepsilon_{2ij} \end{bmatrix} \sim N(0, \Omega_\varepsilon) : \qquad \Omega_\varepsilon = \begin{bmatrix} \sigma_{\varepsilon 1} & \\ 0 & \sigma^2_{\varepsilon 2} \end{bmatrix}
\end{gathered} \tag{2}$$

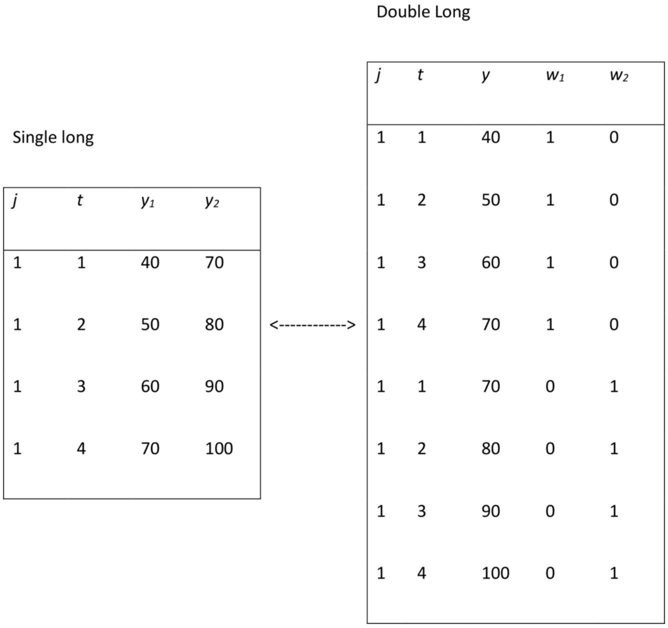
Therefore, β_0 and β_2 are the mean population outcome score at baseline for those who are satisfied and unsatisfied, respectively, and β_1 and β_3 are the corresponding mean population changes per unit of time. Variances and covariances are similarly interpreted for those who are satisfied and unsatisfied, respectively. However, that satisfaction on the external anchoring question is assumed to be known without error, and individual effects and errors for x_{1i} are uncorrelated with those for x_{2i} because the satisfied and unsatisfied categories are mutually exclusive. Whether or not it is desirable to fit a model to both satisfied and unsatisfied individuals simultaneously is debatable, as only those who are satisfied contribute to the definition of MCII. However, we present a simultaneous modelling approach to satisfied and unsatisfied individuals as it make the underlying modelling assumptions explicit. Furthermore, if the stratification on satisfaction status leads to small samples, alternative estimators and degree of freedom can be used in an MLM framework to account for this, that is, restricted maximum likelihood, restricted generalised least squares or adjustments to the denominator df.[55]

Following the prediction of each individual's trajectory, including those unsatisfied with treatment, the second stage in the MCII method requires a threshold for determining responsiveness. Using a similar suggestion to Tubach *et al*,[35] the 75th centile of those who are satisfied could be used to classify all individuals as responding or not. Similar to the MID, there is no suggestion of whether a test against the null value of the 75th centile should be constructed, but this is easily done within the MLM framework.

MLM: OO criteria

The OO criteria can be similarly extended into a multivariate MLM framework by the inclusion of dummy variables and reshaping into a 'double' long format with both responses stored in a single vector. Figure 1 illustrates the data structure for a bivariate model.

Dummy variables, also known as response indicators, are used to denote the response options: w_{1i} is coded 1 for the first measurement outcome (pain) and 0 for the second outcome (function), and $w_{2i} = 1 - w_{1i}$. The response indicators and their interactions with are included as explanatory variables to obtain the following bivariate response model.



Double Long

j	t	y	w₁	w₂
1	1	40	1	0
1	2	50	1	0
1	3	60	1	0
1	4	70	1	0
1	1	70	0	1
1	2	80	0	1
1	3	90	0	1
1	4	100	0	1

Single long

j	t	y₁	y₂
1	1	40	70
1	2	50	80
1	3	60	90
1	4	70	100

<------------>

Figure 1 Illustration of a 'double' long data set-up for creating a bivariate multilevel modelling.

$$
\begin{aligned}
y_{ij} = &\left(\beta_0 + u_{0j}\right) w_{1i} + \left(\beta_1 + u_{1j}\right) t_{ij} w_{1i} + \varepsilon_{1ij} w_{1i} \\
&+ \left(\beta_2 + u_{2j}\right) w_{2i} + \left(\beta_3 + u_{3j}\right) t_{ij} w_{2i} + \varepsilon_{2ij} w_{2i}
\end{aligned}
$$

$$
\begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \end{bmatrix} \sim N(0, \Omega_u) : \qquad \Omega_u = \begin{bmatrix} \sigma^2_{u0} & & & \\ \sigma_{u01} & \sigma^2_{u1} & & \\ \sigma_{u02} & \sigma_{u12} & \sigma^2_{u2} & \\ \sigma_{u03} & \sigma_{u13} & \sigma_{u23} & \sigma^2_{u3} \end{bmatrix}
$$

$$
\begin{bmatrix} \varepsilon_{1ij} \\ \varepsilon_{2ij} \end{bmatrix} \sim N(0, \Omega_\varepsilon) : \qquad \Omega_\varepsilon = \begin{bmatrix} \sigma^2_{\varepsilon 1} & \\ \sigma_{\varepsilon 12} & \sigma_{\varepsilon 2} \end{bmatrix} \tag{3}
$$

With a similar functional form to the univariate MLM, there are separate population and individual intercepts for the first and second outcome (β_0, β_2 and u_{0j}, u_{2j}, respectively), and separate population and individual slopes are estimated for the second outcome (β_1, β_3 and u_{1j}, u_{3j}). Using an MLM approach, the outcomes are modelled jointly, which allows for non-zero covariances between the intercepts and slopes of the two responses (σ_{u02}, σ_{u12}, σ_{u03}, σ_{u13}). The measurement errors for the two responses are not assumed to be independent, with their covariance directly estimated ($\sigma_{\varepsilon 12}$).

Finally, the threshold of response must be decided and individual trajectories estimated and classified. Similar to the other methods, it is relatively simple to construct a test statistic for testing whether individual slopes are significantly different from the chosen threshold.

Limitations of the MLM approach

The MLM approach described by equation 1, equations 2 and 3 assumes that change in the outcome is linearly associated with time. The linearity assumption is imposed for simplicity. Non-linear changes are easily incorporated by including higher order polynomials or using linear or non-linear splines.[56]

The standard MLM approach also fails to directly address the issue of floor and ceiling effects. Mixed-response multilevel Tobit models allow for such effects and provide some adjustment.[45 57] Furthermore, while the MLM described in equation 2 allows for heterogeneity in known groups, they fail to allow for heterogeneity in trajectories when the groups are unknown. The use of group-based trajectory models or growth mixture models in these circumstances may reveal latent (unobserved) classes of individuals with distinct patterns of recovery.[58]

Example: the APEX cohort study

Using a mixed cohort of patients undergoing total hip replacement (THR) and total knee replacement (TKR),[41] we investigated the performance of the existing and MLM approaches using four definitions of responsiveness. A simulated data set and code to fit each of these models are included in the online supplementary material.

Patients in the APEX cohort completed the Intermittent and Constant Osteoarthritis Pain (ICOAP) questionnaire before and after surgery at approximately 0, 3, 6 and 12 months. The date at which the postsurgical questionnaire was completed is recorded in days postsurgery. As the name suggests, the ICOAP questionnaire attempts to measure intermittent and constant pain.[21] The developers of the tool suggest three ways of summarising the scale to generate an intermittent, constant and total pain scores (the sum of the intermittent and constant pain subscales). The tool is scored between 0 and 100 and a full description of the ICOAP scale is provided in the original validation paper.[21] Satisfaction of pain relief following surgery was recorded by asking patients to 'Rate the relief of pain provided by (hip/knee) replacement' using a single-item 5-point scale (none, poor, fair, good, excellent). We categorised good and excellent as a satisfactory outcome following surgery.

Using the three methods of aggregation, we present estimates of pain at baseline and for change at approximately 3 months postsurgery using existing methods (summary statistics) and MLM estimates.

To facilitate comparisons between existing and MLM approaches, we assume that all individuals are measured at exactly 0, 3, 6 and 12 months. While the existing approaches only uses the 0 and 3month measurements, the MLM approach uses a random intercept and random slopes across four measurements occasions, using two linear splines with a knot point at 3 months to estimate the response at 3 months. The inclusion of the second spline and the additional two measurement occasions allows adjustment for measurement error in the MLM approach. Tables 1 and 2 presents results for patients undergoing THR and TKR, respectively. The placement of the knot at 3 months was determined by visually inspecting the data, similar to the methods by Lenguerrand *et al.*[59] With more complex patterns of response an iterative model fitting approach is likely to be required to determine the optimal knot placement. Modelling assumptions were checked using ladder plots and normal plots of residuals.

To describe how the responsiveness classification in patients changed at 3 months, we used an Exact McNemar

Table 1 Mean and SD of baseline and change scores estimated using current and multilevel model approaches to responsiveness in a patient undergoing total hip replacement in the APEX cohort study

			Current approaches to responsiveness				MLM approaches to responsiveness			
			Baseline	Change	Absolute threshold		Baseline	Change	Absolute threshold	
		n	$\beta_0\ \sigma_{u0}$	$\beta_1\ \sigma_{u1}$		P (resp.)	$\beta_0\ \sigma_{u0}$	$\beta_1\ \sigma_{u1}$		P (resp)
Return to normal	Total pain	210	43.71 (22.1)	45.76 (24.0)	87.9	70.5 (63.8, 76.6)	43.71 (20.1)	46.14 (19.7)	83.8	78.1 (71.9, 83.5)
MID		210			11.0	91.9 (87.4, 95.2)			10.0	97.6 (94.5, 99.2)
MCID (satisfied)		185	44.37 (22.0)	48.43 (22.6)	32.6	71.9 (65.3, 77.9)	44.37 (20.3)	48.54 (19.2)	35.8	67.1 (74.5, 85.6)
MCID (unsatisfied)		25	38.77 (22.4)	26.05 (25.4)			38.77 (17.0)	28.43 (16.3)		
Return to normal	Chronic pain	210	49.19 (27.2)	44.23 (27.3)	103.5	0 (0, 1.7)	49.19 (25.6)	44.35 (24.0)	100.3	0 (0, 1.7)
MID		210			13.6	84.3 (78.6, 88.9)			12.8	88.6 (83.5, 92.5)
MCID (satisfied)		185	50.08 (27.4)	46.37 (26.7)	30.0	72.4 (65.8, 78.3)	50.08 (26.3)	46.21 (24.5)	31.0	73.3 (44.2, 58.9)
MCID (unsatisfied)		25	42.60 (24.8)	28.40 (26.9)			42.60 (18.3)	30.60 (12.6)		
OO		210	49.19 (27.2)	44.23 (27.3)	20(10)	92.4 (87.9, 95.6)	49.19 (25.3)	44.35 (23.4)	20(10)	99.5 (54.8, 69)
Return to normal	Intermittent pain	210	39.13 (21.7)	47.06 (26.5)	82.5	70 (63.3, 76.1)	39.13 (18.7)	47.66 (20.5)	76.5	80.5 (90.5, 97.4)
MID		210			10.8	90 (85.1, 93.7)			9.3	97.1 (30, 44.1)
MCID (satisfied)		185	39.60 (21.7)	50.17 (24.9)	37.5	71.4 (64.8, 77.4)	39.60 (19.2)	50.50 (19.1)	40.5	67.1 (84.8, 93.9)
MCID (unsatisfied)		25	35.58 (21.4)	24.08 (26.6)			35.58 (13.9)	26.69 (17.1)		
OO		210	39.13 (21.7)	47.06 (26.5)	20(10)	92.4 (87.9, 95.6)	39.13 (18.5)	47.66 (19.1)	20(10)	99.5 (60.3, 73.5)

Betas represent the population average characteristic and sigma the estimated SD. Baseline is assumed to be the day of surgery, and change is from 0 to 3 months.
MCID, minimal clinically Important difference; MID , minimally important difference; MLM, multilevel model; OO, OMERACT OARSI responder criteria; P (resp.), proportion of responders.

Table 2 Mean and SD of baseline and change scores estimated using current and MLM approaches to responsiveness in patient undergoing total knee replacement in the APEX cohort study

			Current approaches to responsiveness				MLM approaches to responsiveness			
			Baseline	Change	Absolute threshold		Baseline	Change	Absolute threshold	
		n	β_0 σ_{u0}	β_1 σ_{u1}		P (resp.)	β_0 σ_{u0}	β_1 σ_{u1}		P (resp)
Return to normal	Total pain	190	42.86 (19.7)	31.27 (23.2)	82.3	43.2 (36, 50.5)	42.89 (16.7)	32.09 (17.7)	76.3	51.6 (60.3, 73.5)
MID		190			9.9	79.5 (73, 85)			8.3	93.2 (60.3, 73.5)
MCID (satisfied)		138	44.09 (19.7)	38.51 (20.6)	22.7	62.6 (55.3, 69.5)	44.13 (16.7)	38.76 (14.7)	29.9	55.3 (66.8, 79.2)
MCID (unsatisfied)		52	39.62 (19.7)	12.04 (18.0)			39.62 (16.3)	14.28 (11.5)		
Return to normal	Chronic pain	190	47.76 (23.6)	31.61 (25.5)	94.9	44.7 (37.5, 52.1)	47.79 (20.5)	32.46 (19.5)	88.7	36.8 (47.9, 62.5)
MID		190			11.8	74.7 (67.9, 80.7)			10.2	90 (47.9, 62.5)
MCID (satisfied)		138	48.80 (23.4)	38.59 (23.3)	23.7	64.2 (57, 71)	48.88 (20.5)	38.88 (17.7)	30.3	55.3 (47.4, 62)
MCID (unsatisfied)		52	45.00 (24.1)	13.08 (21.9)			45.00 (20.1)	15.26 (13.3)		
OO		190	47.76 (23.6)	31.61 (25.5)	20(10)	81.0 (74.7, 86.3)	47.78 (20.2)	32.50 (18.9)	20(10)	98.4 (47.9, 62.5)
Return to normal	Intermittent pain	190	38.78 (18.2)	30.97 (23.9)	75.3	40.5 (33.5, 47.9)	38.80 (13.8)	31.77 (16.7)	66.4	62.1 (47.9, 62.5)
MID		190			9.1	78.9 (72.5, 84.5)			6.9	94.7 (97.4, 100)
MCID (satisfied)		138	40.15 (18.3)	38.45 (21.2)	24.8	61.6 (54.3, 68.5)	40.20 (14.1)	38.63 (12.8)	31.2	54.7 (97.4, 100)
MCID (unsatisfied)		52	35.14 (17.8)	11.12 (19.0)			35.14 (12.8)	13.40 (10.8)		
OO		190	38.78 (18.2)	30.97 (23.9)	20(10)	81.0 (74.7, 86.3)	38.81 (13.6)	31.74 (15.7)	20(10)	98.4 (95.5, 99.7)

Betas represent the population average characteristic and sigma the estimated SD. Baseline is assumed to be the day of surgery, and change is from 0 to 3 months.
MCID, minimal clinically important difference; MID , minimally important difference; MLM, multilevel model; OO, OMERACT OARSI responder criteria; P (resp.), proportion of responders.

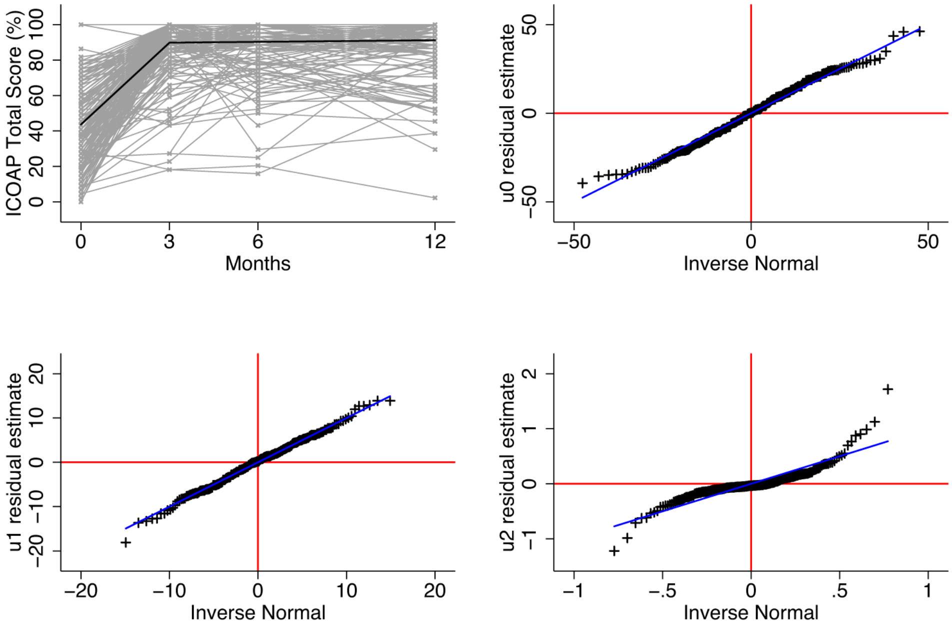


Figure 2 Modelling diagnostic plots. Upper left, ladder plot of observed ICOAP total scores at 0, 3, 6 and 12 months following THR and population average trajectory estimated from a MLM, used in RTN and MID analysis, with two linear splines with a knot at 3 months. Upper right, lower left and right plots are quantile–quantile plots of the residual distribution of random effects estimated from an MLM with two linear splines with a knot at 3 months. ICOAP, intermittent and constant osteoarthritis pain; MID, minimally important difference; MLM, multilevel model; RTN, return to normal; THR, total hip replacement.

test to compare the number of discordant classifications generated by existing and MLM approaches.

The APEX study was approved by Southampton and South West Hampshire Research Ethics Committee (09/H0504/94).

RESULTS

In all subdivisions of the ICOAP questionnaire, for THR/TKR patients, the estimates of the baseline mean and change scores are approximately equal to those from the MLM approaches. In addition, estimates of the SD of baseline and change score are overestimated using existing approaches in THR/TKR patients. The SD of baseline measurements of pain were approximately 3.3 and 3.75 points greater in existing methods compared with MLM methods in THR/TKR patients, respectively, while the corresponding SD of change scores are approximately 6.3 and 7 points greater in existing methods (see tables 1 and 2, respectively). An example of model diagnostics is included in figure 2, which presents the observed ICOAP total scores at 0, 3, 6 and 12 months and the population average response in ICOAP across time. In addition, baseline, change residuals are also presented using quantile–quantile plots.

Return to normal

Using similar baseline score estimates to the conventional RTN approach and different SDs results in a reduction in the threshold of response by approximately five points in THR/TKR patients. The change in threshold is due to smaller estimates of baseline and change SDs. When considering the total ICOAP score, the MLM approach classifies approximately 10% more individuals as responders than existing approaches. It is also interesting to note that the threshold of response using the existing approach when considering total ICOAP score in THR patients is beyond the range of the score.

Minimally important difference

Using similar change score estimates and different SDs results in an approximately 2-point reduction in the MID threshold in THR/TKR patients. The reduced threshold results in more individuals being classified as responders using the MLM approach.

Minimally clinically important difference

Using the MLM approach in satisfied and unsatisfied individuals results in a small increase in the threshold of response in comparison with existing approaches. The increase in threshold is due to shrunken residuals and therefore reduced the variability of predicted change scores. The increase in threshold results in a reduced number of individuals (3% of THR patients and 6% of TKR patients) being identified as responders.

OMERACT-OARSI

The OO approach uses fixed definitions of responsiveness. Individual estimates of change from the bivariate MLM for constant and intermittent pain are very similar to those from the univariate MLM. However, the SD of the change score is reduced by approximately 0.5 and 1 points in constant and intermittent pain comparing the univariate and bivariate MLM, respectively, whereas the

Table 3 Cross-classification of responsiveness status in THR patients using existing and MLM model approaches to responsiveness: RTN, MID, MCII and OO criteria

Total hip replacement ICOAP			Multilevel model							
			RTN		MID		MCII		OO	
			N. resp	Resp	N. resp	Resp	N. resp	Resp	N. resp	Resp
Existing	Total	N. resp	**36**	**26**	**5**	**12**	**52**	**7**	–	–
		Resp	**10**	**138**	**0**	**193**	**17**	**134**	–	–
	Chronic	N. resp	210	0	**24**	**9**	52	6	–	–
		Resp	0	0	**0**	**177**	4	148	–	–
	Intermittent	N. resp	**33**	**30**	**6**	**15**	50	10	–	–
		Resp	**8**	**139**	**0**	**189**	19	131	–	–
	Chronic and intermittent	N. resp	–	–	–	–	–	–	**1**	**15**
		Resp	–	–	–	–	–	–	**0**	**194**

Bold cells indicate significance (p≤0.05) of discordant pairs using Exact McNemar test.
ICOAP, Intermittent and Constant Osteoarthritis Pain; MCII, minimally clinical important improvement; MID, minimally important difference; MLM, multilevel model; N. resp, non-responders; OO, OMERACT OARSI; Resp, responders; RTN, return to normal.

SD of baseline score approximately the same. Despite the larger absolute threshold of 20 and 10 points for changes in one or two items, respectively, that is, larger than MID, there is an increase in the proportion of individuals identified as responding. The increase is partly due to the use of the relative change threshold and the reduced variability in change in comparison with the univariate MLM using MID definition of responsiveness.

Responsiveness classification

The effect of using a MLM approach to defining patient responsiveness compared with existing approaches is presented in tables 3 and 4 for THR and TKR patients,respectively. While the use of MLM provides refined thresholds of responsiveness, it fundamentally changes the way individuals are classified due to adjustment for measurement error, regression to the mean and ability to conduct refined tests. Patients previously defined as non-responding using existing methods are now responders (positive change) in MLM approaches, and similarly, patients defined as responders using existing methods are classified as non-responders (negative change) in MLM (see figure 3 for graphical illustration). MLM MID and OO methods appear to be most consistent in the reclassification of patients increasing the number of patients defined as non-responders using existing methods as responders in MLM approaches, whereas MLM RTN and MCII provide a more fundamental change the classifications of patient responsiveness.

Table 4 Cross-classification of responsiveness status in TKR patients using existing and MLM model approaches to responsiveness: RTN, MID, MCII and OO criteria

TKR ICOAP			Multilevel model							
			RTN		MID		MCII		OO	
			N. resp	Resp	N. resp	Resp	N. resp	Resp	N. resp	Resp
Existing	Total	N. resp	**81**	**27**	**13**	**26**	**64**	**7**	–	–
		Resp	**11**	**71**	**0**	**151**	**21**	**98**	–	–
	Chronic	N. resp	92	13	**19**	**29**	61	7	–	–
		Resp	28	57	**0**	**142**	24	98	–	–
	Intermittent	N. resp	**69**	**44**	**9**	**31**	63	10	–	–
		Resp	**3**	**74**	**1**	**149**	23	94	–	–
	Chronic and intermittent	N. resp	–	–	–	–	–	–	**3**	**33**
		Resp	–	–	–	–	–	–	**0**	**154**

Bold cells indicate significance (p≤0.05) of discordant pairs using Exact McNemar test.
ICOAP, Intermittent and Constant Osteoarthritis Pain; MCII, minimally clinical important improvement; MID, minimally important difference; MLM, multilevel model; N. resp, non-responders; OO, OMERACT OARSI; Resp, responders; RTN, return to normal; TKR, total knee replacement.

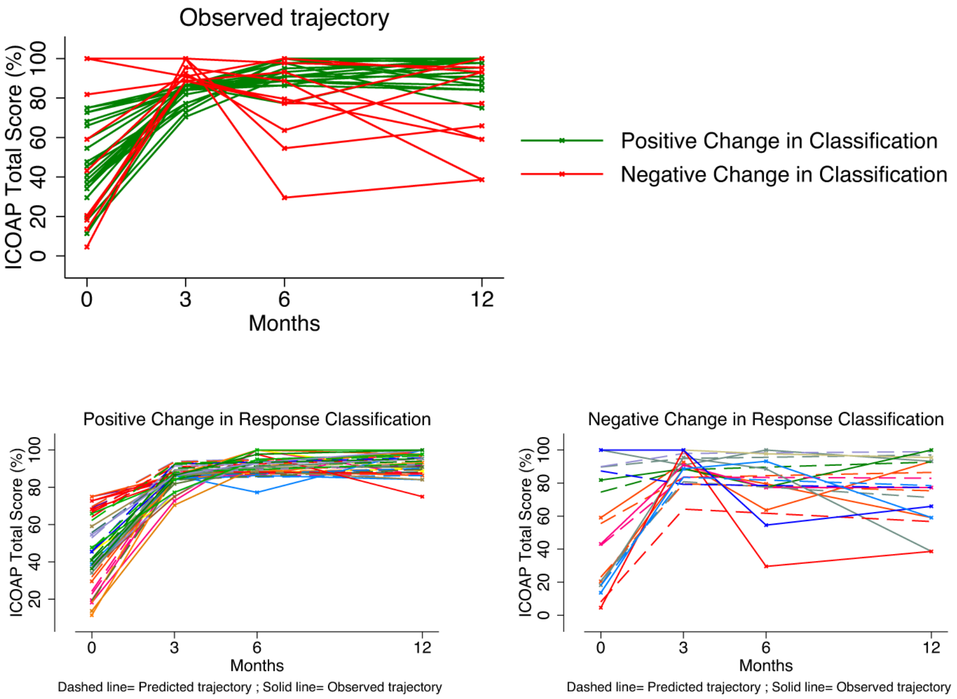


Figure 3 Change in responder classification using an RTN definition comparing existing approaches to MLM approach using the ICOAP total score in patients following THR. Upper left panel illustrates observed trajectories for patients whose responsiveness classification changes using an MLM approach to estimating responsiveness. Lower left panel illustrates the observed and predicted trajectories of ICOAP total score in patients positively reclassified as responders compared with existing approaches. Lower right panel illustrates the observed and predicted trajectories of ICOAP total score in patients negatively reclassified as non-responders compared with existing approaches. ICOAP, Intermittent and Constant Osteoarthritis Pain; MLM, multilevel model; RTN, return to work; THR, total hip replacement.

DISCUSSION

The primary purpose of a responsiveness analysis is to convey the variability of an individual's chances of perceiving an improvement following a treatment. Existing approaches appear to be distinct from one another, and the precise relationship between existing methods was unclear.

We have clearly shown how four commonly used approaches to estimating patient responsiveness can be incorporated into the unified statistical framework of MLM. Their translation into unified framework makes many of the assumption (linearity of response, heterogeneity in the timing of measures, multiple measurements) underpinning existing approaches clear. The application of patient responsiveness models in a cohort of orthopaedic patients illustrates how SDs of baseline and change scores in existing approaches are overestimated in comparison with the MLM approach. Thresholds for defining responders from MLM are lower when based on SD, and therefore existing approaches to RTN and MID may appear to provide a worse case scenario with regards the efficacy of a treatment or therapy. Similarly, responsiveness approaches based on the distribution of predicted change scores (MCII) are higher in MLM, and therefore existing thresholds could be described as a best-case scenario in comparison with existing approaches. However, the reclassification of patients using the MLM is more fundamental than increasing or reducing the threshold to determine responsiveness, the implicit adjustments for measurement error and regression to the mean change which patients are defined as responding or not.

MLM are not the panacea of patient responsiveness methods; however, they do highlight implicit assumptions in existing approaches and provide sensible adjustments for measurement error, regression to the mean and heterogeneity in the timing of measurements in clinical studies.

From a clinical perspective, it is very clear there are differences in the outcomes at 3 months following THR and TKR, while patient's baseline level of pain is similar between THR and TKR, the response to surgery is less and consistently less (lower variability) for all pain domains. Similarly, we have previously observed different patterns of pain, in relation to pain at rest and pain on movement,[60] yet the mechanisms underpinning theses effects are unclear and require more research, but this emphasises the necessity to treat hip and knee osteoarthritis as separate disease states.

Strengths and limitations

One of the key benefits of adopting a MLM approach when defining clinically meaningful change is the improved estimation of individual change by the greater flexibility in the MLM framework. Specifically, MLM do not assume the response is measured without error, they adjust for regression to the mean, the trajectory of recovery is not constrained to be linear and data from multiple measurements and variability in the timing of those measurement occasions can also be incorporated into the model. Furthermore, assuming the underlying MLM adequately represents the true causal mechanism, parameter estimates, SDs and SEs will be unbiased in comparison with existing approaches.

Furthermore, the unification of existing approaches into a MLM framework clearly shows the relationship between the four different approaches. For example, RTN and MID share the same underlying model. MCII is also the same at RTN/MID if you assume the baseline and change scores are the same across strata of unsatisfied/satisfied patients. Similarly, the model underlying OO approach is the same as the RTN/MID approach if you assume independence in the measured outcomes of the two trajectories and the error term.

Despite the numerous benefits of adopting an MLM approach, it is not to say it is without some limitations. MLMs are technically more demanding than existing formulations of patient responsiveness, and while there are no theoretical limits on how large or small samples have to be, model convergence is not guaranteed. The need to use appropriate estimation methods[38] or denominator degrees of freedom[55] when calculating standard errors also requires consideration. Furthermore, it is important to perform model diagnostic to check the data fit with the model. MLM does not improve the arbitrary placement of the thresholds that define responsiveness in comparison with existing methods, and despite the improved trajectory modelling, it is currently unclear if the refined definitions correlate more strongly with patient expectations, functional data, long-term self-reported outcomes or hard endpoints such as mortality and revision. Further research externally validating the classification using patient groups, expert opinion[61] or functional data may demonstrate improved classification of those responding to treatment in comparison with existing methods. In addition, the use of multiple measurements in MLM primarily restricts the method to a research setting.

It is clear the MLMs provide considerable advantages over existing approaches to identifying patients who respond to a treatment. Consequently, the proportion of individuals thought not to be responding to treatment may be smaller than previously thought. Using the redefined definition may reduce the number of individuals misclassified as non-responders and improve the prediction of those individuals who are likely to respond to treatment.

Author affiliations

[1]Musculoskeletal Research Unit, School of Clinical Sciences, University of Bristol, Southmead Hospital, Bristol, UK

[2]School of Social and Community Medicine, University of Bristol, Bristol, UK

[3]Nuffield Department of Population Health, University of Oxford, Oxford, UK

[4]Biomedical Research Unit, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Science, Nuffield Orthopaedic Centre, Oxford, UK

Acknowledgements We thank Professor Fiona Steele for her extensive comments and help preparing this manuscript. The research team acknowledges the support of the National Institute for Health Research (NIHR) through the Comprehensive Clinical Research Network.

Contributors AS: study conception, wrote first draft and revisions and final approval of manuscript; VW: APEX study design acquisition of data, drafting and review of manuscript and final approval of manuscript; EL: APEX study acquisition of data, drafting and review of manuscript and final approval of manuscript; RG-H: APEX study design acquisition of data, drafting and review of manuscript and final approval of manuscript; JD: ACHE study design, drafting and review of manuscript and final approval of manuscript; DB: ACHE study design, drafting and review of manuscript and final approval of manuscript; AP: ACHE study design, drafting and review of manuscript and final approval of manuscript and AWB: APEX/ACHE study design acquisition of data, drafting and review of manuscript and final approval of manuscript.

Funding This work was supported by AS and is funded by an MRC Fellowship MR/L01226X/1 and HTA Project: 11/63/01—'ACHE'. This article presents independent research funded by the NIHR under its Programme Grants for Applied Research programme (RP-PG-0407-10070).

Disclaimer The views expressed in this article are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Data are unavailable to share. CORRECT

REFERENCES

1. Felson DT, Naimark A, Anderson J, *et al*. The prevalence of knee osteoarthritis in the elderly. the Framingham Osteoarthritis Study. *Arthritis Rheum* 1987;30:914–8.
2. Lawrence RC, Felson DT, Helmick CG, *et al*. Estimates of the prevalence of arthritis and other rheumatic conditions in the United States. Part II. *Arthritis Rheum* 2008;58:26–35.
3. National Joint Registry 10th Annual Report 2013.Hemel Hempstead 2013.
4. Beswick AD, Wylde V, Gooberman-Hill R, *et al*. What proportion of patients report long-term pain after total hip or knee replacement for osteoarthritis? A systematic review of prospective studies in unselected patients. *BMJ Open* 2012;2:e000435.
5. Jeffery AE, Wylde V, Blom AW, *et al*. "It's there and I'm stuck with it": patients' experiences of chronic pain following total knee replacement surgery. *Arthritis Care Res* 2011;63:286–92.
6. Kassam A, Dieppe P, Toms AD. An analysis of time and money spent on investigating painful total knee replacements. *British Journal of Medical Practitioners* 2012;5:a526.
7. Bellamy N, Buchanan WW, Goldsmith CH, *et al*. 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;15:1833–40.
8. Klässbo M, Larsson E, Mannevik E. Hip disability and osteoarthritis outcome score. an extension of the Western Ontario and McMaster universities Osteoarthritis index. *Scand J Rheumatol* 2003;32:46–51.
9. Roos EM, Roos HP, Lohmander LS, *et al*. Knee Injury and Osteoarthritis outcome score (KOOS)--development of a self-administered outcome measure. *J Orthop Sports Phys Ther* 1998;28:88–96.
10. Dawson J, Fitzpatrick R, Carr A, *et al*. Questionnaire on the perceptions of patients about total hip replacement. *J Bone Joint Surg Br* 1996;78:185–90.
11. Dawson J, Fitzpatrick R, Murray D, *et al*. Questionnaire on the perceptions of patients about total knee replacement. *J Bone Joint Surg Br* 1998;80:63–9.

12. Focht BC, Rejeski WJ, Ambrosius WT, *et al*. Exercise, self-efficacy, and mobility performance in overweight and obese older adults with knee osteoarthritis. *Arthritis Rheum* 2005;53:659–65.
13. Smith AJ, Dieppe P, Howard PW, *et al*. Failure rates of metal-on-metal hip resurfacings: analysis of data from the National Joint Registry for England and Wales. *Lancet* 2012;380:1759–66.
14. Smith AJ, Dieppe P, Porter M, *et al*. National Joint Registry of England and Wales. Risk of cancer in first seven years after metal-on-metal hip replacement compared with other bearings and general population: linkage study between the National Joint Registry of England and Wales and hospital episode statistics. *BMJ* 2012;344:e2383.
15. Hunt LP, Ben-Shlomo Y, Clark EM, *et al*. 45-day mortality after 467,779 knee replacements for osteoarthritis from the National Joint Registry for England and Wales: an observational study. *Lancet* 2014;384:1429–36.
16. Hunt LP, Ben-Shlomo Y, Clark EM, *et al*. 90-day mortality after 409,096 total hip replacements for osteoarthritis, from the National Joint Registry for England and Wales: a retrospective analysis. *Lancet* 2013;382:1097–104.
17. Riddle DL, Stratford PW, Bowman DH. Findings of extensive variation in the types of outcome measures used in hip and knee replacement clinical trials: a systematic review. *Arthritis Rheum* 2008;59:876–83.
18. Wylde V, Bruce J, Beswick A, *et al*. Assessment of chronic postsurgical pain after knee replacement: a systematic review. *Arthritis Care Res* 2013;65:1795–803.
19. Wylde V, Blom AW. The failure of survivorship. *J Bone Joint Surg Br* 2011;93:569–70.
20. Department of Health. High quality care for all: NHS Next Stage Review final report. 2008.
21. Hawker GA, Davis AM, French MR, *et al*. Development and preliminary psychometric testing of a new OA pain measure--an OARSI/OMERACT initiative. *Osteoarthritis Cartilage* 2008;16:409–14.
22. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand* 1983;67:361–70.
23. Williams A, Kind P. The present state of play about QALYs. Hopkins A, ed. *Measure of the quality of life: the uses to which they may be put*. Chicago, IL: RCP publications, 1992.
24. Christensen L, Mendoza JL. A method of assessing change in a single subject: an alteration of the RC index. *Behav Ther* 1986;17:305–8.
25. Guyatt GH, Osoba D, Wu AW, *et al*. Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002;77:371–83.
26. Jacobson NS, Roberts LJ, Berns SB, *et al*. Methods for defining and determining the clinical significance of treatment effects: description, application, and alternatives. *J Consult Clin Psychol* 1999;67:300–7.
27. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* 1991;59:12–19.
28. Kvien TK, Heiberg T, Hagen KB. Minimal clinically important improvement/difference (MCII/MCID) and patient acceptable symptom state (PASS): what do these concepts mean? *Ann Rheum Dis* 2007;66 (Suppl 3) :iii40–1.
29. Maksymowych WP, Richardson R, Mallon C, *et al*. Evaluation and validation of the patient acceptable symptom state (PASS) in patients with ankylosing spondylitis. *Arthritis Rheum* 2007;57:133–9.
30. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care* 2003;41:582–92.
31. Norman GR, Sridhar FG, Guyatt GH, *et al*. Relation of distribution- and anchor-based approaches in interpretation of changes in health-related quality of life. *Med Care* 2001;39:1039–47.
32. Pham T, van der Heijde D, Altman RD, *et al*. OMERACT-OARSI initiative: osteoarthritis Research Society International set of responder criteria for osteoarthritis clinical trials revisited. *Osteoarthritis Cartilage* 2004;12:389–99.
33. Revicki D, Hays RD, Cella D, *et al*. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008;61:102–9.
34. Tubach F, Ravaud P, Baron G, *et al*. Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. *Ann Rheum Dis* 2005;64:29–33.
35. Tubach F, Ravaud P, Baron G, *et al*. Evaluation of clinically relevant states in patient reported outcomes in knee and hip osteoarthritis: the patient acceptable symptom state. *Ann Rheum Dis* 2005;64:34–7.
36. Judge A, Cooper C, Williams S, *et al*. Patient-reported outcomes one year after primary hip replacement in a European Collaborative Cohort. *Arthritis Care Res* 2010;62:480–8.
37. Goldstein H. *Multilevel statistical models*. London, UK: E. Arnold, 2002.
38. Browne WJ, Draper D. Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Comput Stat* 2000;15:391–420.
39. King MT. A point of minimal important difference (MID): a critique of terminology and methods. *Expert Rev Pharmacoecon Outcomes Res* 2011;11:171–84.
40. Schuck P, Zwingmann C. The 'smallest real difference' as a measure of sensitivity to change: a critical analysis. *Int J Rehabil Res* 2003;26:85–91.
41. Wylde V, Gooberman-Hill R, Horwood J, *et al*. The effect of local anaesthetic wound infiltration on chronic pain after lower limb joint replacement: a protocol for a double-blind randomised controlled trial. *BMC Musculoskelet Disord* 2011;12:53.
42. Altman DG, Gore SM, Gardner MJ, *et al*. Statistical guidelines for contributors to medical journals. *Br Med J* 1983;286:1489–93.
43. Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? *BMJ* 2001;322:226–31.
44. McConnell S, Kolopack P, Davis AM. The Western Ontario and McMaster universities Osteoarthritis Index (WOMAC): a review of its utility and measurement properties. *Arthritis Rheum* 2001;45:453–61.
45. Twisk J, Rijmen F. Longitudinal tobit regression: a new approach to analyze outcome variables with floor or ceiling effects. *J Clin Epidemiol* 2009;62:953–8.
46. Ram N, Grimm KJ. Growth mixture modeling: a method for identifying differences in longitudinal change among unobserved groups. *Int J Behav Dev* 2009;33:565–76.
47. Jones G, Lyons P. Approximate graphical methods for inverse regression. *J Data Sci*;2009:61–72.
48. Snijders TAB, Bosker RJ. *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. 2nd edn. London: Sage Publishers, 2012.
49. Fieuws S, Verbeke G. Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics* 2006;62:424–31.
50. Fieuws S, Verbeke G. Joint modelling of multivariate longitudinal profiles: pitfalls of the random-effects approach. *Stat Med* 2004;23:3093–104.
51. Verbeke G, Molenberghs G. *Linear mixed models for longitudinal data*. USA: Springer, 2000.
52. Rasbash J, Steele F, Browne WJ, *et al*. *A user's guide to MLwIN*. Bristol, UK: Bristol University, 2009.
53. Regression CJB. Prediction and shrinkage. *J R Stat Soc B* 1983;45:311–54.
54. Fitzmaurice GM, Laird NM, Ware JH. *Applied longitudinal analysis*. Hoboken, NJ: Wiley, 2004.
55. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997;53:983–97.
56. Pan H, Goldstein H. Multi-level repeated measures growth modelling using extended spline functions. *Stat Med* 1998;17:2755–70.
57. Rabe-Hesketh S, Skrondal A. Multilevel and latent variable modeling with composite links and exploded likelihoods. *Psychometrika* 2007;72:123–40.
58. Nagin DS, Odgers CL. Group-based trajectory modeling in clinical research. *Annu Rev Clin Psychol* 2010;6:109–38.
59. Lenguerrand E, Wylde V, Gooberman-Hill R, *et al*. Trajectories of pain and function after primary hip and knee arthroplasty: the ADAPT Cohort Study. *PLoS One* 2016;11:e0149306.
60. Sayers A, Wylde V, Lenguerrand E, *et al*. Rest pain and movement-evoked pain as unique constructs in hip and knee replacements. *Arthritis Care Res* 2016;68:237–45.
61. Bellamy N, Carette S, Ford PM, *et al*. Osteoarthritis antirheumatic drug trials. III. Setting the delta for clinical trials--results of a consensus development (Delphi) exercise. *J Rheumatol* 1992;19:451–7.