Author/s:
Schilling, C;Petrie, D;Dowsey, MM;Choong, PF;Clarke, P

Title:
The Impact of Regression to the Mean on Economic Evaluation in Quasi-Experimental Pre–Post Studies: The Example of Total Knee Replacement Using Data from the Osteoarthritis Initiative

Date:
2017-12-01

Citation:
Schilling, C., Petrie, D., Dowsey, M. M., Choong, P. F. & Clarke, P. (2017). The Impact of Regression to the Mean on Economic Evaluation in Quasi-Experimental Pre–Post Studies: The Example of Total Knee Replacement Using Data from the Osteoarthritis Initiative. Health Economics United Kingdom, 26 (12), pp.e35-e51. https://doi.org/10.1002/hec.3475.

Persistent Link:
https://hdl.handle.net/11343/292366

Title: The impact of regression to the mean on economic evaluation in quasi-experimental pre-post studies: the example of total knee replacement using data from the Osteoarthritis Initiative

Running head: Regression to the mean in economic evaluation

Authors: C Schilling[a,b,d], D Petrie[a], MM Dowsey[b,c], PF Choong[b,c], P Clarke[a].

a) Centre for Health Policy, School of Population and Global Health, The University of Melbourne, Victoria 3051

b) The University of Melbourne Department of Surgery, St Vincent's Hospital, Victoria 3065

c) Department of Orthopaedics, St Vincent's Hospital Melbourne, Victoria 3065

d) Corresponding author: C Schilling, 207 Bouverie St, Carlton, Victoria, Australia 3051. Chris.schilling@unimelb.edu.au; +61 3 9035 3965.

Key words: regression to the mean, health-related quality of life, quasi-experimental design, economic evaluation, total knee replacement

Word count: 4,456

Funding: This research was supported by the University of Melbourne, the Australian Research Council's Discovery Early Career Awards funding scheme (Project DE150100309), and the National Health and Medical Research Council (APP1093229). The views expressed herein are those of the authors and are not necessarily those of the Australian or National Health and Medical Research Councils.

The OsteoArthritis Initiative (OAI) is a public-private partnership comprised of five contracts (N01-AR-2-2258; N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human

Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health. This manuscript was prepared using an OAI public use data set and does not necessarily reflect the opinions or views of the OAI investigators, the NIH, or the private funding partners.

Conflict of interests: Each author certifies that he or she has no commercial associations that might pose a conflict of interest in connection with the submitted article.

Title: The impact of regression to the mean on economic evaluation in quasi-experimental pre-post studies: the example of total knee replacement using data from the Osteoarthritis Initiative

Running head: Regression to the mean in economic evaluation

ABSTRACT

Introduction: Many treatments are evaluated using quasi-experimental pre-post studies susceptible to regression to the mean (RTM). Ignoring RTM could bias the economic evaluation. We investigated this issue using the contemporary example of total knee replacement (TKR), a common treatment for end-stage osteoarthritis of the knee.

Methods: Data (n= 4,796) were obtained from the Osteoarthritis Initiative database, a longitudinal observational study of osteoarthritis. TKR patients (n=184) were matched to non-TKR patients, using propensity score matching on the predicted hazard of TKR, and exact matching on osteoarthritis severity and health-related quality of life (HrQoL). The economic evaluation using the matched control group was compared to the standard method of using the pre-surgery score as the control.

Results: Matched controls were identified for 56% of the primary TKRs. The matched control HrQoL trajectory showed evidence of RTM accounting for a third of the estimated QALY gains from surgery using the pre-surgery HrQoL as the control. Incorporating RTM into the economic evaluation significantly reduced the estimated cost-effectiveness of TKR

and increased the uncertainty. A generalized ICER bias correction factor was derived to account for RTM in cost-effectiveness analysis.

Conclusion: RTM should be considered in economic evaluations based on quasi-experimental pre-post studies.

1. INTRODUCTION

Identifying patients for treatment on the basis of a relatively poor health score is common practice in healthcare (Linden, 2013). However health symptoms can fluctuate over time and one-off measurements can be prone to error. On average, patients identified by poor scores improve upon re-measurement, even in the absence of treatment (Linden, 2013, Barnett et al., 2005). This concept is known as regression to the mean (RTM), first documented in 1886 by Francis Galton who noticed that children from taller than average parents were typically shorter than their parents, while children from shorter than average parents were typically taller (Galton, 1886). Milton Friedman has lamented the economics profession for ignoring biases driven by RTM (Friedman, 1992).

Pre-post quasi-experimental studies are particularly susceptible to RTM due to the difficulties in establishing appropriate controls (Gardner and Heady, 1973, Davis, 1976). Where randomized control trials (RCTs) explicitly measure control groups and therefore account for RTM in the calculation of the treatment effect (placebo responses include RTM (Morton and Torgerson, 2003)), pre-post studies typically use the pre-treatment score as the control (Harris et al., 2006). Economic evaluations based on these studies are therefore susceptible to bias from RTM, yet there has been little empirical research into the implications of RTM on cost-

effectiveness analysis. To date, studies have explained RTM and provided approaches to identify it (Barnett et al., 2005), quantified RTM in simulated datasets (Linden, 2013), and briefly discussed its impact on decision-making in health (Morton and Torgerson, 2003), but none have estimated the bias in incremental cost-effectiveness ratios caused by RTM in pre-post studies.

In this paper we investigate this issue using the contemporary example of total knee replacement (TKR), a common treatment for end-stage osteoarthritis (OA) performed over 1.5 million times each year across OECD countries[1] (OECD Indicators, 2015). Economic evaluations of TKR are typically based on pre-post study designs (see for example Waimann et al. (2014), Jenkins et al. (2013), Schilling et al. (2016), or Dakin et al. (2012)). These studies have shown that TKR improves health-related quality of life (HrQoL) relative to the preoperative score, yet surprisingly little is known about the trajectory of patients should they not have surgery. While OA is typically thought of as a degenerative disease, recent longitudinal studies have revealed, on average, stable or improved physical function, and the presence of RTM (Collins et al., 2014, Zou et al., 2016, Øiestad et al., 2015).

We used a longitudinal observational dataset of almost 5,000 OA patients, some of whom received TKR, to estimate a 'without surgery' control group for TKR using an exact and propensity score matching analysis. We traced the HrQoL trajectory of the matched control group to empirically estimate the prevalence of RTM in this cohort, and assessed the implications of RTM on the cost-effectiveness of TKR. We then calculated a general RTM

[1] Author calculations from OECD knee replacement surgery rates per 100,000 inhabitants from the OECD's Health at a Glance 2015 [4] and OECD population rates.

bias correction factor for economic evaluations based on pre-post studies because the ramifications of our findings are applicable to many common treatments that are evaluated using study designs susceptible to RTM. To aid the reader, Section 2 provides a background into the current methods of economic evaluations of TKR, how RTM can be estimated, and the derivation of the RTM bias correction factor, before proceeding as normal with *Methods*, *Results* and *Discussion* sections.

2. BACKGROUND

2.1 Current methods of economic evaluation of TKR

Economic evaluation of TKR is challenging. Ethical and practical difficulties with sham surgeries have limited RCTs (Macklin, 1999, Miller, 2003). Joint replacement registries have been used successfully to identify the techniques and prostheses that deliver better outcomes (Graves and Wells, 2006), but typically do not capture data on patients who *do not* have surgery. Researchers have analysed waiting times to investigate how outcomes vary as a result of the timing of surgery, however such research has not provided a comprehensive picture of the HrQoL trajectory in the absence of surgery (see for example Ostendorf et al. (2004), or Nikolova et al. (2015)). Economic evaluations of TKR are therefore typically one-group pretest-posttest designs (Harris et al., 2006), that use the patient's preoperative HrQoL utility score as the control (see for example recent studies by Jenkins et al. (2013), Dakin et al. (2012), Schilling et al. (2016) or Losina et al. (2009)). The Quality Adjusted Life Year (QALY) gain from TKR is then calculated as the area between the actual HrQoL utility curve after TKR and this hypothetical constant 'without surgery' utility. Such an assumption for the control could significantly bias economic evaluations. While the HrQoL trajectory of OA

patients is still under considerable debate (Collins et al., 2014), significant deterioration in pain and functional outcomes has been clearly observed in end-stage OA patients in the period immediately preceding the surgery (Collins et al., 2014, Riddle et al., 2013). Assuming a constant 'without surgery' control at preoperative levels ignores the prior trend of deterioration, possibly underestimating the 'area between the curves' that represent the QALY gain from TKR (Dakin et al., 2012). Alternatively, the presence of RTM could mean that patients identified for treatment by their relatively poor health state could experience improved outcomes upon future re-measurement even without treatment (Linden, 2013). This would result in the current practice overestimating the true QALY gains from TKR.

2.2 Regression to the mean

RTM is sometimes considered a statistical concept (Barnett et al., 2005), because it highlights the implications of random fluctuations in patient outcomes, but there are 'real' reasons why those fluctuations occur, such as the episodic nature of a disease, patient adaptation, variability with self-reported measurements, or simple randomness. These fluctuations should not be attributed to the treatment itself. The impact of RTM can be estimated by having or deriving an explicit control group for the treated and tracking their improvement over time, or where no potential control group is available, by using a formula derived in the literature based on assumptions about the data generating process and the cut-off used to decide on treatment. From Barnett et al. (2005), the formula for calculating expected RTM in simple normally distributed repeated measured data is given by:

$$RTM = \frac{\sigma_w^2}{\sqrt{\sigma_w^2+\sigma_b^2}}\, C(z) \qquad (1)$$

$$= \sigma_t(1-\rho)C(z), \qquad -1 \leq \rho \leq 1$$

Where the total variance of the outcome in period t is given by $\sigma_t^2 = \sigma_w^2 + \sigma_b^2$, the within-subject variance is given by $\sigma_w^2 = (1-\rho)\sigma_t^2$, the between subject variance is given by $\sigma_b^2 = \rho\sigma_t^2$, ρ is the correlation between outcomes across time (a measure of the persistence of outcomes over time at the individual level), and $C(z) = \emptyset(z)/\aleph(z)$ where $z = (\mu - c)/\sigma_t$ if the selection for treatment occurs when the individual's score is less than the cut-off *c*. $\emptyset(z)$ and $\aleph(z)$ are the standard normal probability density and cumulative distribution functions respectively. This formula shows that higher levels of RTM is expected when the cut-off is more extreme (those within the selected treatment group are more likely to have been at an extreme outcome prior to treatment) and when there is little persistence in individual outcomes over time (individuals currently at extreme outcomes are expected to quickly return to the mean).

While the above formula has been shown to accurately estimate RTM in a simulated example for a simple data generating process, the results can be sensitive to the assumptions made (Linden, 2013). In practice the decision to treat is often based on more than one variable, may be assessed at multiple points in time and may also be based on recent changes in outcomes rather than the current outcome level. In addition, as in the case of TKR, there may be an expectation of continued decline in health in the absence of treatment which may be heterogeneous across individuals. In these instances, using a comparable control group can

provide a more compelling view of the HrQoL trajectory in the absence of treatment and the presence of any RTM. The ideal case for a comparable control group is where treatment is randomized such that the control group's outcomes are unrelated to treatment assignment. In non-randomized studies such as quasi-experiments, comparable control groups can be constructed by propensity score or exact matching, and the post-treatment trajectory of the control group compared to the trajectory of the treated.

2.3 Estimating the impact of regression to the mean on cost-effectiveness

While it is known that economic evaluations based on pre-post quasi-experimental studies are susceptible to bias from RTM, there has been little research into the implications of RTM on cost-effectiveness analysis. To help fill this gap, we generalized the impact of RTM on cost-effectiveness by developing a RTM bias correction factor based on the level of RTM. In pre-post quasi-experimental studies such as those used in the economic evaluations of surgery, the observed QALY gain from treatment $QALY_o$ is typically estimated as function of the observed change in utility scores ΔU_o (the postoperative utility score minus the preoperative utility score), and the expected duration of the gain.

$$QALY_o = \Delta U_o \times duration \quad (2)$$

The presence of RTM has the effect of reducing the utility gain from treatment by some factor α (the ratio of RTM/observed treatment effect) such that the gain in QALYs after adjusting for RTM now becomes:

$$QALY_r = (1 - \alpha) \times QALY_o \quad (3)$$

Assuming that RTM does not affect costs (a patient is not usually classified for treatment by their poor cost outcomes), the impact of RTM on the observed incremental cost-effectiveness ratio (ICER) can then be calculated as a function of α, the ratio of RTM/observed treatment effect:

$$ICER_r = \frac{1}{1-\alpha} ICER_o \quad (4)$$

where $ICER_o$ is the observed ICER, $ICER_r$ is the RTM-adjusted ICER, and $\frac{1}{1-\alpha}$ is the RTM bias correction factor (see full derivation in the Appendix).

3. METHODS

3.1 Data source

Data were obtained from the Osteoarthritis Initiative (OAI) database, which is publicly accessible at http://www.oai.ucsf.edu/. The OAI is a multi-center, longitudinal, prospective observational study of knee OA that collects a range of socio-economic, healthcare access, self-reported and clinical data on 4,796 patients with OA aged 45-79 enrolled between 2004 and 2006. Specific datasets used are baseline, 12-month, 24-month, 36-month, 48-month, 60-month, 72-month, 84-month and 96 month clinical and medical history data (release versions 0.2.2, 1.2.1, 3.2.1, 5.2.1, 6.2.1, 7.2.1, 8.2.1, 9.2.1 and 10.2.1 respectively), enrollee demographic information (release version 22). The OAI has approval from the Committee on Human Research, the University of California, San Francisco (Approval number 10-00532).

3.2 Treatment, outcomes and associated factors

The treatment for this study was primary total knee replacement. The primary outcomes were the HrQoL utility scores, calculated from the Short-Form Health Survey (SF-12) patient-reported outcomes using the standard Brazier algorithm (Brazier and Roberts, 2004). The factors associated with the likelihood of TKR treatment were age, gender, race, annual income, access to health insurance, body mass index (BMI), smoking status, comorbidities as measured by the Charlson Comorbidity Index (CCI) (Hall et al., 2004), OA severity measured radiographically by the Kellgren Lawrence (KL) scale (Kellgren and Lawrence, 1957) and subjectively by the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) patient-reported outcome (Bellamy, 1988), mental component summary (MCS), physical component summary (PCS) and HrQoL scores. BMI was classified as moderately obese (30-34.9 kg/m^2), severely obese (35-39.9kg/m^2) and morbidly obese (> 40kg/m^2), comorbidities as none, 1, and 2 or more, race as white and non-white, smoking status as current smoker versus non-smoker.

3.3 Statistical methods for estimating a control group for TKR to evaluate RTM

Constructing appropriate control groups using propensity scoring is a common method for inferring causal treatment effects in the absence of randomized control studies or valid instruments, but most propensity score analyses are based on cross-sectional data where a single observation is observed for each individual (Lu, 2005). However the OAI is a longitudinal cohort dataset that follows individuals and tracks the progression of their OA over time including TKR as an end-stage treatment for some patients. We therefore followed Li et al (Li et al., 2001) and Lu (Lu, 2005) and calculated a proportional hazard for TKR, based on age, gender, race, radiographic OA at baseline; ongoing BMI, comorbidities,

smoking status, HrQoL, PCS, MCS and WOMAC scores at each follow-up. We then performed matching without replacement using the patient's hazard of treatment at each point in time. Note that we did not consider the cumulative hazard: recent research has found that the anticipated general deterioration in OA patients over long periods of time is not observed in practice (Collins et al., 2014), but that relatively sudden deterioration in pain and function may trigger TKR (Riddle et al., 2013). We therefore included annual change in HrQoL, PCS and WOMAC in the model for predicting TKR.

Matching occurred such that a patient treated at time *t* could be matched against any patient not yet treated at time *t*. The treated could also act as a control if treatment occurred after time *t+2* years so that if selected as a control the non-surgery trajectory could be observed for 2 years. Once treated, a patient was censored and could no longer act as a control (Li et al., 2001). To reduce the possibility that a matched control patient might have commenced some other treatment, we censored controls if they commenced frequent medication or injection treatments during the comparator period. The dataset contains patients at different levels of OA progression, so we performed an exact match on the patient's KL score, to ensure that we were comparing patients with the same level of radiographic OA. As radiographic and symptomatic progression have been shown to be nonparallel (Collins et al., 2014), we also performed an exact match on the patient's HrQoL (rounded to two decimal places), to ensure we were comparing patients with consistent self-reported symptomatology. The propensity score was then used to further match patient characteristics, using a caliper (or a maximum allowable difference) of 0.05, or approximately 0.4 standard deviations. We chose HrQoL for the exact match rather than a WOMAC or PCS score because the focus of this research was

on economic evaluations derived from HrQoL scores, however we included an interaction term between HrQoL and WOMAC to capture any non-linearities between overall quality-of-life and knee-specific OA (for example to account for when HrQoL change may have been due to an unrelated condition).

We evaluated the balance between the treated and control covariates using standardized differences between means (Austin, 2009), with a preferred cut-off of 10% for those covariates shown to be significant predictors of TKR as recommended by Ho et al. (2007). To ensure similar distributions as well as means between the treated and control, variance ratios for the propensity score and continuous individual covariates were also evaluated as per Rubin (2001). To assess the presence of RTM, we then compared the average HrQoL trajectories of those who have a TKR with their control matches who did not have a TKR. In some cases, an appropriate match could not be found for TKR patients with particularly high propensity scores. No control group could be estimated for this group, but to provide some indications of possible trends for the unmatched group, we repeated the analysis for subset of TKR patients that could be matched who had TKR propensity scores in the top 5th percentile (were most likely to receive a TKR and therefore the most similar to the unmatched group). The matching analysis was carried out in Stata 13.1 IC (Stata Corp, College Station, USA) using the psmatch2 version 4.0.11 program (Leuven and Sianesi, 2015).

3.4 Data for estimating a control group for TKR to evaluate RTM

We considered patients that had been observed for at least three years: one year prior to TKR to provide evidence of a consistent deterioration between treated and controls prior to

surgery, and two years post TKR to show differences in the treated and control group after surgery. Baseline, 84- and 96-month data were therefore censored due to the lack of prior and post data respectively. 60-month data were censored due to the lack of HrQoL, PCS or MCS collection during this wave. For other missing data, linear interpolation was used for PCS, MCS, BMI and HrQoL where observations were available before and after the missing observation; last value carried forward was used for missing years for OA severity, comorbidities and smoking status. The data availability of the key variables is shown in the Appendix (Table 5).

As some patients had their last follow-up many months before TKR, we needed to estimate their condition just prior to surgery (Øiestad et al., 2015). We used regression imputation methods to estimate measured deterioration in HrQoL prior to TKR as a function of the time between the last follow-up measurement prior to TKR and patient covariates discussed previously. This regression was used to predict HrQoL at the time of TKR for recipients whose last follow-up measurement was more than three months prior to surgery. Postoperatively, the gains from TKR typically stabilize after six months (Naylor et al., 2009). A measurement at a minimum of six months after surgery is therefore deemed an appropriate proxy for the 12-month outcome. For patients whose first follow-up was less than six months after TKR, the measurement of HrQoL was unlikely to reflect outcomes at 12 months. We therefore used their second follow-up post-surgery to represent their 12 month follow-up from TKR. In this way, the 12-month follow-up was measured on average 12 months after surgery, but between 6-18 months depending on the individual patient. The variation in outcomes between 6-18 months post-surgery is minimal relative to the variation between 0

and 12 months post-surgery (Naylor et al., 2009). A more detailed description of this data adjustment procedure is provided in the supplementary analysis.

3.5 Robustness analysis for control group for TKR to evaluate RTM

To evaluate the uncertainty of the matching analysis, we bootstrapped the matching process, creating 1,000 new datasets from the original dataset by sampling with replacement, and repeated the matching and calculation of treatment and control HrQoL trajectories. Confidence intervals were calculated from the percentiles of the bootstrapped results. Additionally, we parameterized the formula from the literature (equation 1) using the OAI data to provide a comparative estimate of RTM.

In supplementary analyses, we also completed a range of further checks to test the robustness of our results. We considered exact matching on WOMAC rather than HrQoL to investigate if matching on a knee-specific rather than generic patient reported outcome would provide a different result. To test the impact of the timing of follow-up measurements relative to the TKR, we repeated the analysis using just those TKRs that occurred within 3 months of a follow-up.

3.6 Implications for cost-effectiveness analysis

To investigate the implications of our findings on the outcomes from TKR, we estimated the QALY gains in the two years post-surgery, first using the standard preoperative utility score 'without surgery' control, and second using the control trajectory estimated from the matching exercise. We then assumed that utility scores measured at 2 years post-surgery

continue to 15 years post-surgery, and that the cost of TKR is a constant $20,000 per surgery, broadly in line with previous economic evaluations (Losina et al., 2009), to estimate the impact of our findings on overall cost-effectiveness. We present these results as indicative cost-acceptability curves for TKR when derived from the assumed control, versus when derived from the matching exercise. Finally, we confirmed the impact of RTM on cost-effectiveness by estimating the RTM bias correction factor from equation (4).

4. RESULTS

4.1 Sample characteristics

After censoring, the number of primary TKRs observed was 263 out of a cohort of 4,768. Of these, 184 (70%) had full data available to complete the matching analysis. Of the missing, 32 had insufficient length of follow-up (less than 3 years), and 47 had missing covariate data. Sample characteristics are shown in Table 1. Of note is that over 96% of the cohort had access to some form of health insurance.

[INSERT TABLE 1]

4.2 Matching

The results from the multivariate proportional hazards model of TKR are shown in Table 2. The most significant independent predictors of TKR were severe OA, both radiographic (KL) and symptomatic (WOMAC), and deterioration in WOMAC. Non-whites, smokers, low income earners and those with significant comorbidity were significantly less likely to have a TKR. Good mental health was positively associated with likelihood of TKR.

[INSERT TABLE 2]

103 (56%) treated patients were able to be matched. The balance of the covariates between the matched treatment and 'without surgery' control groups are shown in Table 3. The standardized differences were small with the majority of covariates falling under the 10% target, and a good balance achieved on the key predictors significantly associated with TKR: OA severity (both KL and WOMAC), MCS, WOMAC deterioration, and race. The matched control group were more comorbid and obese, and had a wider income distribution. The variance ratio between the matched treatment and control groups for the propensity score is 1.00, and the variance ratios for the covariates are close to one indicating both means and variance have been well-balanced across the two groups.

[INSERT TABLE 3]

81 (44%) treated patients were unable to be matched. These patients had a higher treatment propensity score versus treated patients who were matched (mean 0.29±0.10 versus 0.10±0.09) and were significantly different across almost all patient characteristics. In particular, the unmatched had significantly poorer WOMAC, PCS and HrQoL scores and experienced significantly larger deterioration in these scores prior to surgery, relative to matched TKR recipients (Table 4).

[INSERT TABLE 4]

4.3 Matched treated and control HrQoL trajectory

Figure 1 shows the mean of the HrQoL trajectory for the matched treated and control patients and 95% confidence intervals from bootstrapping. Prior to the surgery there was a decline in

HrQoL of almost 0.05 for both treated and control groups. After surgery, the treated group had a statistically significant increase in HrQoL of 0.05 resulting in a gain of 0.07 QALYs over the two years after surgery, relative to the preoperative HrQoL. However the 'without surgery' control group also experienced a mean improvement in HrQoL of 0.02 and a gain of 0.02 QALYs relative to the preoperative HrQoL. The increase in HrQoL for the treated group is no longer statistically significant when compared to this 'without surgery' control group, and the estimated QALY gains from surgery reduce from 0.07 to 0.05 QALYs (33% reduction, 95% confidence intervals -5:78%).

[INSERT FIGURE 1]

4.4 A proxy for the unmatched

An analysis of the matching showed that those with the highest propensity for TKR were least likely to be matched, because these patients invariably received a TKR. Thus, a sub-cohort of the matched group with the top 5th percentile propensity scores was selected to investigate if the results would change based on the propensity for TKR. The 5th percentile sub-cohort of the matched cohort therefore acted as a proxy to highlight the potential impact of RTM in the unmatched group (see Appendix Table 6 for comparison of these two groups).

Figure 2 shows the HrQoL trajectory for the matched treated and control patients for a sub-cohort with the top 5th percentile propensity scores. This sub-cohort had a larger deterioration in HrQoL of 0.10 prior to surgery, and a larger improvement in HrQoL of 0.14 for the treated group after surgery, relative to the full cohort. This resulted in a QALY improvement of 0.13 QALYs relative to the preoperative HrQoL. However the control group for this sub-cohort

also experienced a mean improvement in HrQoL of 0.06 and a gain of 0.06 QALYs relative to the preoperative HrQoL. Using this 'without surgery' counterfactual would reduce the estimated QALY gains from surgery from 0.13 to 0.07 QALYs (47% reduction, 95% CI -26:126%), and the HrQoL gains from surgery are no longer statistically significant.

[INSERT FIGURE 2]

4.5 Impact on cost-effectiveness of TKR

Equation (4) showed how RTM impacts on the ICER for economic evaluations that use a pre-post quasi-experimental study design. The bias correction factor is increasing in α (Figure 3), which suggests that cost-effectiveness will be increasingly affected as RTM increases: RTM equal to 20% of the observed treatment effect ($\alpha = 0.20$) causes the ICER to increase by 25% (ICER bias correction factor = 1.25); RTM equal to 50% of the observed treatment effect ($\alpha = 0.50$) causes the ICER to increase by 100% (ICER bias correction factor = 2).

[INSERT FIGURE 3]

Using the matched controls, we estimated RTM equal to 39% of the observed treatment effect ($\alpha = 0.39$), resulting in an ICER bias correction factor of 1.64. Figure 4 shows indicative cost-acceptability curves for the matched TKR recipients when using the standard assumption of a constant preoperative HrQoL in the absence of surgery (blue), and when using the matched control counterfactual from this analysis (red). Incorporating RTM in the matched control group significantly reduced the likelihood of cost-effectiveness and increased the uncertainty around the cost-effectiveness of TKR as illustrated by the flatter

acceptability curve. At a willingness to pay threshold of $50,000/QALY, this suggests that TKR is likely to be cost-effective for 50% of the matched treatment group rather than almost 100% suggested under the standard control assumptions.

[INSERT FIGURE 4]

4.6 Robustness

The 95% confidence intervals for the control group HrQoL trajectory shown in Figure 1, derived from the bootstrapping exercise, suggests the findings of RTM are robust to sampling variation of the potential matches. The supplementary analyses contain the results of further checks to test the robustness of our results. We found no significant differences from using WOMAC instead of HrQoL for the exact matching, or from analyzing only the sub-cohort of patients with ideal follow-up relative to TKR.

Using equation (1) and data from the OAI dataset (HrQoL population mean = 0.76; high risk threshold = 0.70 and between year correlation = 0.75), we estimated RTM for this OA population cohort of 0.03, indicating that the mean of those with HrQoL $\leq$0.70 can be expected to improve by 0.03 simply due to RTM. This is similar in magnitude to that estimated from the matched control group and shown in Figure 1.

5. DISCUSSION

5.1 Estimating regression to the mean in total knee replacement

Patients who receive TKR typically experience a significant drop in HrQoL in the year preceding the TKR, which might have suggested an ongoing decrease in the absence of TKR, but we found little evidence to support this. We identified matches to act as a control group

for 56% of the primary TKRs in the OAI dataset. The average trajectory of this control group showed evidence of natural remission in HrQoL without the surgery intervention. This remission accounted for a third of the QALY improvements from surgery in the matched treatment group, and significantly reduced the cost-effectiveness of TKR, relative to the traditional assumption that in the absence of surgery, patients' HrQoL would have remained at preoperative levels.

We discounted the possibility that remission in the control group was associated with the commencement of some other treatment as we censored controls if they commenced frequent medication or injections. Further investigating the wider OA literature, we found support for the theory that our findings were due to RTM. Researchers have showed that at the average level knee pain changes little over time, and 'is characterized by persistent rather than inexorably worsening symptoms' (Collins et al., 2014). At the individual level, pain and functional impairment is intermittent and variable (Gooberman-Hill et al., 2007; Zhang et al., 2011; Øiestad et al., 2015); patients actively adapt their behavior to mitigate the impact of OA on the lives (Gooberman-Hill et al., 2007; Wright et al., 2008); there is evidence of non-parallel progression between radiographic and symptomatic OA (Collins et al., 2014); and placebo effects, which include RTM, are common in OA-related RCTs: a recent meta-analysis of such RCTs found a placebo effect size of 75% of the total effect size (Zou et al., 2016). Combined, this suggested that a poor WOMAC or HrQoL score in one year might not be indicative of a persistent poor score in the next, and therefore that RTM was a plausible explanation of the improvement observed in our matched control group. This was supported

empirically by our findings of RTM when the parameters from the OAI dataset were employed in equation (1).

5.2 Generalizability of findings

The TKR patients for which we were able to find appropriate non-TKR matches had relatively high preoperative HrQoL, and reported only modest gains from the surgery. By contrast, the unmatched TKR group had significantly poorer preoperative HrQoL, and subsequently reported larger gains from surgery. It is the unmatched group that appears more representative of TKR patients across the United States (see Appendix Table 6), which posed some questions about the generalizability of our findings. However in the analysis we further explored the unmatched cohort by examining the results for a sub-cohort of the matched cohort who were most similar to the unmatched. This sub-cohort had comparable pre and post-surgery HrQoL to the unmatched group, and also showed evidence of RTM (Figure 2), suggesting that our findings are generalizable to the wider United States.

In Australia and the United Kingdom, patients who receive TKR typically have a lower preoperative HrQoL and in turn achieve a higher gain in HrQoL from the surgery (see Appendix Table 6), and the rates of TKR across the population are much lower in these countries. It is therefore difficult to categorically conclude that RTM would be present in these settings, however we note that 1) the sub-cohort of the matched with the lowest preoperative HrQoL and the highest gains from surgery showed evidence of RTM (Figure 2); and 2) that in the formula-based method for assessing RTM, the magnitude of RTM increases as the cut-off threshold for treatment becomes more extreme, as described in Section 2.

5.3 Implications for economic evaluations of total knee replacement

These results have implications for economic evaluations of TKR and related issues. In the absence of randomized controlled trials, many cost-effectiveness analyses use observational data, and adopt methods such as regression or matching to help minimize bias (Kreif et al., 2013). Economic evaluations in hip and knee surgery tend to ignore such methods, perhaps because of a lack of appropriate data; such evaluations tend to be based on surgery registries capturing only treated patients, rather than observational studies such as the OAI that track the OA disease over time. Our findings of RTM in the OA cohort suggest that cost-effectiveness analysis of TKR may be biased if an adequate control group is not established. This is also true for model-based economic evaluations that assume no expected improvement without treatment. A recent systematic review of joint replacement in the US highlighted 12 studies that had completed cost-utility analysis around TKR or TKR-related topics (Nwachukwu et al., 2015). All used some form of modelling (predominantly Markov or decision-tree), but none afforded non-operative patients any opportunity to improve. Similarly, Mather et al. (2014) found TKR without delay was a cost-effective treatment when compared to waiting for TKR, but assumed that waiting was at a much lower utility (0.60 for end-stage OA, versus 0.90 after successful primary TKR) and did not allow for RTM while on the waiting list. Our results suggest that, at least for some of the TKR cohort, waiting might be accompanied by an improvement in HrQoL. The systematic review also evaluated the quality of each paper using the Quality of Health Economic Studies instrument and found on average the papers were of high quality with a mean score of 89.9 out of 100 (Nwachukwu et al., 2015), but the evaluations of quality appear to ignore the potential impacts of RTM.

44% of TKR recipients were unable to be matched. These patients were significantly different in the key predictive attributes with poorer WOMAC, PCS, MCS and HrQoL scores and larger deterioration prior to surgery. The inability to find matches for this group suggests that patients who had such a progression invariably had surgery, and meant we were unable to estimate an appropriate control trajectory. We analysed the sub-cohort of the control group most similar to the unmatched group: those with a propensity score for TKR in the top 5th percentile. We found that the trajectory of this sub-cohort of the matched control group also experienced RTM, consistent with the theory that RTM increases as the cut-off threshold for classification becomes more extreme. However, perhaps reassuringly, this highly likely to be treated sub-group experienced larger QALY gains from surgery relative to the full cohort, even after accounting for RTM. This suggests that in populations like Australia and the United Kingdom, where preoperative utility scores are relatively low and gains from surgery relatively large, TKR is still likely to remain cost-effective, on average, even in the presence of RTM.

5.4 Wider implications for quasi-experimental studies

Our findings are generalizable to other quasi-experimental studies that use a pre-post study design. Such studies are likely to be susceptible to RTM just as we have shown with TKR. There are a range of mainstream treatments that fall under this category, including cardiovascular treatments such as statins (Ble et al., 2016), infection control (Stone et al., 2007, Eliopoulos et al., 2005) and public health initiatives (Cummins et al., 2005). Economic evaluation of these treatments should consider RTM. Some steps noted in the literature are first to acknowledge RTM and where possible adjust for it (Morton and Torgerson, 2003).

Second, increasing the number of preoperative measurements will add certainty around selection into treatment (Stone et al., 2007, Davis, 1976). Finally, we have provided a simple ICER bias correction factor based on the level of RTM.

5.5 Limitations

Several limitations warrant mention. First, the use of propensity score and exact matching is unlikely to be as effective as RCTs at eliminating bias. There is the possibility that unknown or unobservable factors are important in selection for surgery, that there are some underlying differences between the treated and control group, and/or that control group patients were treated for a non-related condition during the follow-up. It would be useful to replicate this study in conditions where both RCT and observational data are available, to provide a gold standard for adjusting observational data for the impact of RTM.

Nonetheless the matching covered key socio-demographic, healthcare access and disease-specific factors that have been identified in the literature associated with TKR (Apold et al., 2014), and considered important in tools being developed to aid decision-making (for example, the nomogram tool developed in Dowsey et al., 2016). For the known, observable covariates, balancing was imperfect and there was a trade-off between the degree of covariate balance and the proportion of the treated that could be adequately matched. Our preferred matching delivered balance on key predictive covariates, but resulted in 44% of treated being unmatched. We found evidence of RTM in the sub-cohort of the matched most similar to the unmatched group, but the exact size of the RTM in the unmatched cohort could not be determined in the absence of a control group.

Second, there was large uncertainty in the expected trajectory of our control group. The OAI dataset is the largest long-term sample of OA progression available with almost 5,000 patients followed over 9 years and counting[2], but a larger sample would have allowed us to provide a more certain view of the average trajectory of patients who do not have surgery. Nonetheless, current economic evaluations typically assume with perfect certainty that the control group HrQoL remains at the preoperative utility score (for example Jenkins et al. (2013), Schilling et al. (2016) and Dakin et al. (2012)). Our work provided some evidence to help inform more realistic control assumptions, even though further data and analysis would help to improve the certainty around the results.

A third limitation is the timing of follow-up relative to the TKR. We made adjustments to the data to account for this based on the literature and observed trends in this dataset. The supplementary analysis of the sub-cohort of patients who had TKR within three months of their last follow-up indicated that these adjustments had not biased the results but ideally, follow-up would have aligned closely with the TKR event for all patients.

5.6 Conclusions

Despite the broad awareness of RTM, and the known susceptibility of common quasi-experimental study designs to RTM, many economic evaluations ignore it. This can lead to a bias in economic evaluation such that the cost-effectiveness of treatment will be overstated. We found evidence of RTM in OA that accounted for a third of the previously predicted QALY gain from surgery and significantly reduced the cost-effectiveness of TKR. This

[2] The Swedish 'Better management of patients with OsteoArthritis' study may surpass it in time.

common example highlights the ramifications of ignoring RTM in cost-effectiveness analyses. Quasi-experimental pre-post studies are likely to be susceptible to RTM and should adjust for it in economic evaluations.

6. REFERENCES

APOLD, H., MEYER, H. E., NORDSLETTEN, L., FURNES, O., BASTE, V. & FLUGSRUD, G. B. 2014. Risk factors for knee replacement due to primary osteoarthritis, a population based, prospective cohort study of 315,495 individuals. BMC musculoskeletal disorders, 15, 1.

AUSTIN, P. C. 2009. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity escore matched sa in Medicine, 28, 3083-3107.

BARNETT, A. G., VAN DER POLS, J. C. & DOBSON, A. J. 2005. Regression to the mean: what it is and how to deal with it. International journal of epidemiology, 34, 215-220.

BELLAMY, N. 1988. Validation study of WOMAC: a health status instrument for measuring clinically-important patient-relevant outcomes following total hip or knee arthroplasty in osteoarthritis. J Orthop Rheumatol, 1, 95-108.

BLE, A., HUGHES, P. M., DELGADO, J., MASOLI, J. A., BOWMAN, K., ZIRK-SADOWSKI, J., MUJICA MOTA, R., HENLEY, W. E. & MELZER, D. 2016. Safety and Effectiveness of Statins for Prevention of Recurrent Myocardial Infarction in 12 156 Typical Older Patients: A Quasi-Experimental Study. The journals of gerontology. Series A, Biological sciences and medical sciences.

BRAZIER, J. E. & ROBERTS, J. 2004. The estimation of a preference-based measure of health from the SF-12. Medical Care, 42, 851-859.

COLLINS, J. E., KATZ, J. N., DERVAN, E. E. & LOSINA, E. 2014. Trajectories and Risk Profiles of Pain in Persons with Radiographic, Symptomatic Knee Osteoarthritis: Data

from the Osteoarthritis Initiative. Osteoarthritis and cartilage / OARS, Osteoarthritis Research Society, 22, 622-630.

CUMMINS, S., PETTICREW, M., HIGGINS, C., FINDLAY, A. & SPARKS, L. 2005. Large scale food retailing as an intervention for diet and health: quasi-experimental evaluation of a natural experiment. Journal of Epidemiology and Community Health, 59, 1035-1040.

DAKIN, H., GRAY, A., FITZPATRICK, R., MACLENNAN, G., MURRAY, D. & GROUP, K. T. 2012. Rationing of total knee replacement: a cost-effectiveness analysis on a large trial data set. BMJ Open, 2, e000332.

DAVIS, C. 1976. The effect of regression to the mean in epidemiologic and clinical studies. American Journal of Epidemiology, 104, 493-498.

DOWSEY, M.M., SPELMAN, T., CHOONG, P.F. 2016. Development of a Prognostic Nomogram for Predicting the Probability of Nonresponse to Total Knee Arthroplasty 1 Year After Surgery. The Journal of arthroplasty.

ELIOPOULOS, G. M., HARRIS, A. D., LAUTENBACH, E. & PERENCEVICH, E. 2005. A systematic review of quasi-experimental study designs in the fields of infection control and antibiotic resistance. Clinical Infectious Diseases, 41, 77-82.

FRIEDMAN, M. 1992. Do old fallacies ever die? Journal of Economic Literature, 30(4) (1992), 2129-2132.

GALTON, F. 1886. Regression towards mediocrity in hereditary stature. The Journal of the Anthropological Institute of Great Britain and Ireland, 15, 246-263.

GARDNER, M. & HEADY, J. 1973. Some effects of within-person variability in epidemiological studies. Journal of Chronic Diseases, 26, 781-795.

GRAVES, S. & WELLS, V. 2006. A review of joint replacement surgery and its outcomes: appropriateness of prostheses. Australian Centre for Health Research.

HALL, W. H., RAMACHANDRAN, R., NARAYAN, S., JANI, A. B. & VIJAYAKUMAR, S. 2004. An electronic application for rapidly calculating Charlson comorbidity score. BMC cancer, 4, 94.

HARRIS, A. D., MCGREGOR, J. C., PERENCEVICH, E. N., FURUNO, J. P., ZHU, J., PETERSON, D. E. & FINKELSTEIN, J. 2006. The use and interpretation of quasi-experimental studies in medical informatics. Journal of the American Medical Informatics Association, 13, 16-23.

HO, D. E., IMAI, K., KING, G. & STUART, E. A. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political analysis, 15, 199-236.

JENKINS, P. J., CLEMENT, N. D., HAMILTON, D. F., GASTON, P., PATTON, J. T. & HOWIE, C. R. 2013. Predicting the cost-effectiveness of total hip and knee replacement: A health economic analysis. Bone & Joint Journal, 95-B, 115-121.

KELLGREN, J. & LAWRENCE, J. 1957. Radiological assessment of osteo-arthrosis. Annals of the Rheumatic Diseases, 16, 494.

KREIF, N., GRIEVE, R. & SADIQUE, M. Z. 2013. Statistical Methods For Cost - Effectiveness Analyses That Use Observational Data: A Critical Appraisal Tool And Review Of Current Practice. Health economics, 22, 486-500.

LEUVEN, E. & SIANESI, B. 2015. PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Statistical Software Components.

LI, Y. P., PROPERT, K. J. & ROSENBAUM, P. R. 2001. Balanced Risk Set Matching. Journal of the American Statistical Association, 96, 870-882.

LINDEN, A. 2013. Assessing regression to the mean effects in health care initiatives. Bmc Medical Research Methodology, 13, 119.

LOSINA, E., WALENSKY, R. P., KESSLER, C. L., EMRANI, P. S., REICHMANN, W. M., WRIGHT, E. A., HOLT, H. L., SOLOMON, D. H., YELIN, E. & PALTIEL, A. D. 2009. Cost-effectiveness of total knee arthroplasty in the United States: patient risk and hospital volume. Archives of internal medicine, 169, 1113.

LU, B. 2005. Propensity Score Matching with Time-Dependent Covariates. Biometrics, 61, 721-728.

MACKLIN, R. 1999. The ethical problems with sham surgery in clinical research. New England Journal of Medicine, 341, 992-996.

MATHER, R. C., HUG, K. T., ORLANDO, L. A., WATTERS, T. S., KOENIG, L., NUNLEY, R. M. & BOLOGNESI, M. P. 2014. Economic evaluation of access to musculoskeletal care: the case of waiting for total knee arthroplasty. BMC musculoskeletal disorders, 15, 1.

MILLER, F. G. 2003. Sham surgery: an ethical analysis. American journal of Bioethics, 3, 41-48.

MORTON, V. & TORGERSON, D. J. 2003. Effect of regression to the mean on decision making in health care. British Medical Journal, 326, 1083.

NATIONAL HEALTH SERVICE. 2016. Patient Reported Outcome Measures (PROMs). https://www.england.nhs.uk/statistics/statistical-work-areas/proms/#. Accessed 7th October 2016.

NAYLOR, J. M., HARMER, A. R., HEARD, R. C. & HARRIS, I. A. 2009. Patterns of recovery following knee and hip replacement in an Australian cohort. Australian Health Review, 33, 124-135.

NWACHUKWU, B. U., BOZIC, K. J., SCHAIRER, W. W., BERNSTEIN, J. L., JEVSEVAR, D. S., MARX, R. G., & PADGETT, D. E. (2015). Current status of cost utility analyses in total joint arthroplasty: a systematic review. Clinical Orthopaedics and Related Research®, 473(5), 1815-1827.

NIKOLOVA, S., HARRISON, M. & SUTTON, M. 2015. The Impact of Waiting Time on Health Gains from Surgery: Evidence from a National Patient -reported Outcome Dataset. Health economics.

OECD INDICATORS 2015. Health at a Glance 2015.

ØIESTAD, B. E., WHITE, D. K., BOOTON, R., NIU, J., ZHANG, Y., TORNER, J., LEWIS, B., NEVITT, M., LAVALLEY, M. & FELSON, D. T. 2015. The longitudinal course of physical function in people with symptomatic knee osteoarthritis: Data from the MOST study and the OAI. Arthritis Care Res (Hoboken).

OSTENDORF, M., BUSKENS, E., VAN STEL, H., SCHRIJVERS, A., MARTING, L., DHERT, W. & VERBOUT, A. 2004. Waiting for total hip arthroplasty: avoidable

loss in quality time and preventable deterioration. The Journal of arthroplasty, 19, 302-309.

RIDDLE, D. L., PERERA, R. A., STRATFORD, P. W., JIRANEK, W. A. & DUMENCI, L. 2013. Progressing Toward, and Recovering From, Knee Replacement Surgery: A Five-Year Cohort Study. Arthritis & Rheumatism, 65, 3304-3313.

RUBIN, D. B. 2001. Using propensity scores to help design observational studies: application to the tobacco litigation. Health Services and Outcomes Research Methodology, 2, 169-188.

SCHILLING, C. G., DOWSEY, M. M., PETRIE, D. J., CLARKE, P. M., & CHOONG, P. F. 2016. Predicting the Long-Term Gains in Health-Related Quality of Life After Total Knee Arthroplasty. The Journal of Arthroplasty.

STONE, S. P., COOPER, B. S., KIBBLER, C. C., COOKSON, B. D., ROBERTS, J. A., MEDLEY, G. F., DUCKWORTH, G., LAI, R., EBRAHIM, S. & BROWN, E. M. 2007. The ORION statement: guidelines for transparent reporting of outbreak reports and intervention studies of nosocomial infection. Journal of Antimicrobial Chemotherapy, 59, 833-840.

WAIMANN, C. A., FERNANDEZ - MAMZA, R. J., CANTOR, S. B., LOPEZ - OLIVO, M. A., ZHANG, H., LANDON, G. C., SIFF, S. J. & SUAREZ - ALMAZOR, M. E. 2014. Cost - Effectiveness of Total Prospective Cohort Study. Arthritis Care Res (Hoboken), 66, 592-599.

WRIGHT, L. J., ZAUTRA, A. J. & GOING, S. 2008. Adaptation to early knee osteoarthritis: the role of risk, resilience, and disease severity on pain and physical functioning. Annals of Behavioral Medicine, 36, 70-80.

ZOU, K., WONG, J., ABDULLAH, N., CHEN, X., SMITH, T., DOHERTY, M. & ZHANG, W. 2016. Examination of overall treatment effect and the proportion attributable to contextual effect in osteoarthritis: meta-analysis of randomised controlled trials. Annals of the Rheumatic Diseases, annrheumdis-2015-208387.

7. TABLES

Table 1 Sample characteristics at baseline

Variable	TKR (n = 184)	Full sample (n = 4,768)
Age	67.2 ± 8.5	61.2 ± 9.2
Female	105 (57%)	2804 (58.5%)
Race: white	161 (87.5%)	3790 (79.1%)
Income bands		
< $10,000	2 (1.1%)	160 (3.6%)
$10,000 to < $25,000	16 (8.7%)	454 (10.2%)
$25,000 to < $50,000	51 (27.7%)	1135 (25.6%)
$50,000 to < $100,000	70 (38%)	1610 (36.3%)
$100,000+	45 (24.5%)	1075 (24.2%)
Access to health insurance	182 (98.9%)	4579 (96.4%)
BMI: kg/m^2	29.8 ± 4.7	28.6 ± 0.048
Moderately obese	62 (33.7%)	1230 (25.7%)
Severely obese	26 (14.1%)	416 (8.7%)
Morbidly obese	3 (1.6%)	84 (1.8%)
Comorbidities		
0	128 (69.6%)	3565 (75.4%)
1	35 (19%)	724 (15.3%)
2+	21 (11.4%)	441 (9.3%)
OA severity: KL scale		
0	3 (1.6%)	1,265 (28.1%)
1	1 (0.5%)	691 (15.3%)
2	17 (9.2%)	1,365 (30.3%)
3	56 (30.4%)	892 (19.8%)
4	107 (58.2%)	293 (6.5%)
Smoking status: smoker	4 (2.2%)	313 (6.6%)

WOMAC	38 ± 12.8	12 ± 0.153
MCS	55.7 ± 8.6	53.6 ± 8.1
PCS	37.3 ± 7.5	48.8 ± 9.2
HrQoL	0.70 ± 0.15	0.80 ± 0.12

Mean ± standard deviation are reported for continuous variables; counts (percentages) for discrete variables.

Table 2: Multivariate proportional hazards model of TKR

Variable	Odds ratio	p-value	Lower 95% CI	Upper 95% CI
Age	1.02	0.052	1.00	1.04
Gender				
Male	ref	ref	ref	ref
Female	1.21	0.230	0.89	1.66
Race				
White	ref	ref	ref	ref
Non-white	0.40	<0.001	0.25	0.66
Income bands				
< $10,000	ref	ref	ref	Ref
$10,000 to < $25,000	0.17	0.008	0.04	0.63
$25,000 to < $50,000	0.59	0.065	0.33	1.03
$50,000 to < $100,000	0.74	0.142	0.50	1.10
$100,000+	1.30	0.168	0.90	1.88
Access to health insurance	1.31	0.747	0.26	6.60
Obesity				
Non-obese	ref	ref	ref	ref
Moderately obese	1.29	0.128	0.93	1.78
Severely obese	1.05	0.858	0.63	1.74
Morbidly obese	1.05	0.921	0.42	2.61
Smoking				
Non-smoker	ref	ref	ref	ref
Smoker	0.36	0.041	0.13	0.96
Comorbidities				
0	ref	ref	ref	ref
1	0.82	0.358	0.55	1.24
2+	0.65	0.080	0.40	1.05

Osteoarthritis				
KL 0	ref	ref	ref	ref
KL 1	3.19	0.101	0.80	12.80
KL 2	5.54	0.005	1.66	18.47
KL 3	17.47	<0.001	5.35	57.07
KL 4	57.07	<0.001	17.36	187.55
WOMAC	1.05	<0.001	1.04	1.06
Quality of life				
MCS	1.06	0.003	1.02	1.10
PCS	1.00	0.929	0.97	1.03
*HRQOL**	0.60	0.030	0.38	0.95
Osteoarthritis change (KL) from previous year				
< 1	ref	ref	ref	ref
1+	1.17	0.502	0.74	1.85
WOMAC	1.02	0.005	1.01	1.03
Quality of life change from previous year				
PCS	0.98	0.071	0.95	1.00
*HRQOL**	0.87	0.359	0.66	1.17
* Standardised to improve interpretability				

Table 3: Balance of covariates between treatment and control after matching

Variable	Matched Treatment	Matched Control	Standardized difference %	Variance ratio^
Age	66.4 ± 8.6	66.8 ± 8.7	-0.046	0.99
Female	58 (56.3%)	64 (62.1%)	-0.118	
Race: white	87 (84.5%)	84 (81.6%)	0.077	
Income bands				
< $10,000	1 (1%)	4 (3.9%)	-0.188	
$10,000 to < $25,000	11 (10.7%)	15 (14.6%)	-0.118	
$25,000 to < $50,000	27 (26.2%)	29 (28.2%)	-0.045	
$50,000 to < $100,000	46 (44.7%)	33 (32%)	0.263	
$100,000+	18 (17.5%)	22 (21.4%)	-0.099	
Access to health insurance	101 (98.1%)	102 (99%)	-0.075	
BMI: kg/m^2	29.7 ± 5.1	30.4 ± 5.3	-0.135	0.98
Moderately obese	30 (29.1%)	31 (30.1%)	-0.022	
Severely obese	16 (15.5%)	11 (10.7%)	0.143	
Morbidly obese	1 (1%)	6 (5.8%)	-0.267	
Smoking status: smoker	4 (3.9%)	2 (1.9%)	0.119	
Comorbidities				
0	68 (66%)	63 (61.2%)	0.100	
1	23 (22.3%)	20 (19.4%)	0.071	
2+	12 (11.7%)	20 (19.4%)	-0.214	
Osteoarthritis				
KL 0	3 (2.9%)	3 (2.9%)	0.000	
KL 1	0 (0%)	0 (0%)	0.000	
KL 2	13 (12.6%)	13 (12.6%)	0.000	
KL 3	30 (29.1%)	30 (29.1%)	0.000	
KL 4	57 (55.3%)	57 (55.3%)	0.000	
WOMAC	34.4 ± 11.5	33.9 ± 20	0.031	0.76

Quality of life				
MCS	55.7 ± 8.1	55.9 ± 10.3	-0.022	0.89
PCS	38.7 ± 6.6	37.4 ± 8.9	0.166	0.86
HrQoL	0.72 ± 0.13	0.72 ± 0.13	0.000	1.00
Osteoarthritis change (KL) from previous year				
< 1	91 (88.4%)	97 (94.2%)	-0.207	
1+	12 (11.7%)	6 (5.8%)	0.210	
WOMAC	7.2 ± 14.2	7.3 ± 13.7	-0.007	1.02
Quality of life change from previous year				
PCS	-3.6 ± 7.0	-4.6 ± 8.2	0.131	0.92
HrQoL	-0.05 ± 0.075	-0.04 ± 0.107	-0.108	0.84

Mean ± standard deviation are reported for continuous variables; counts (percentages) for discrete variables; ^Variance ratios are only available for continuous variables.

Table 4: Comparison of covariates between matched and unmatched treated group

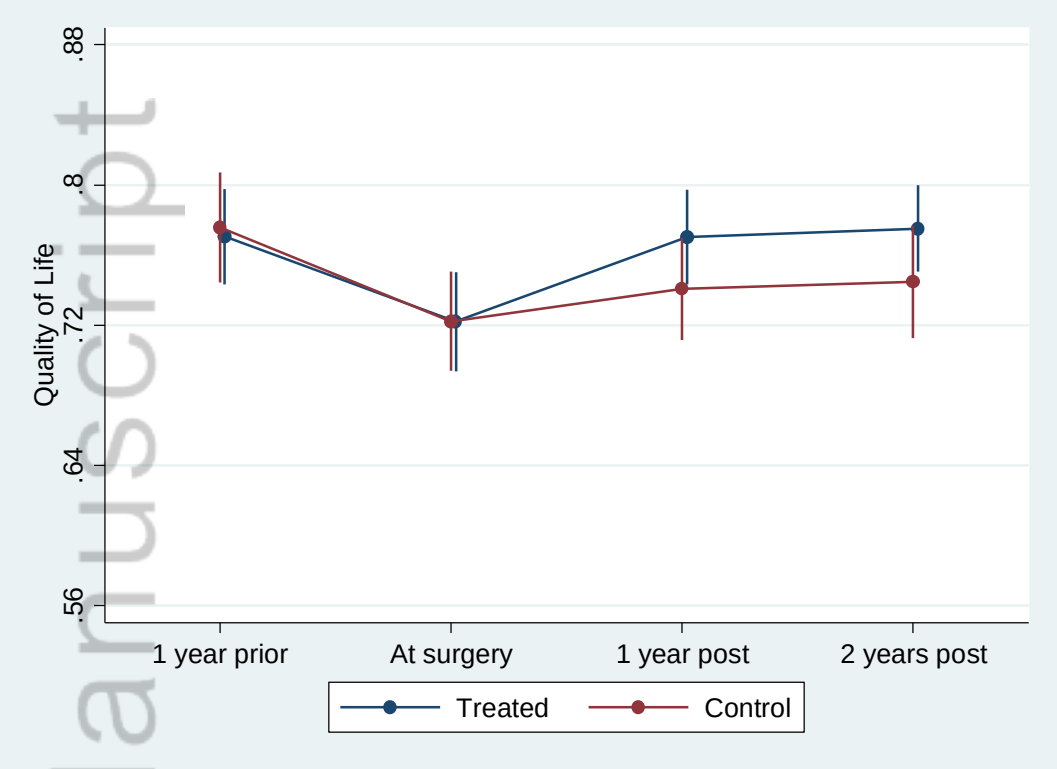
Variable	Matched	Unmatched	Standardized difference %	Variance ratio^
Age	66.4 ± 8.6	68.2 ± 8.3	0.213	0.98
Female	58 (56.3%)	47 (58%)	0.034	
Race: white	87 (84.5%)	74 (91.4%)	0.213	
Income bands				
< $10,000	1 (1%)	1 (1.2%)	0.019	
$10,000 to < $25,000	11 (10.7%)	5 (6.2%)	-0.162	
$25,000 to < $50,000	27 (26.2%)	24 (29.6%)	0.076	
$50,000 to < $100,000	46 (44.7%)	24 (29.6%)	-0.316	
$100,000+	18 (17.5%)	27 (33.3%)	0.369	
Access to health insurance	101 (98.1%)	81 (100%)	0.197	
BMI: kg/m^2	29.7 ± 5.1	29.9 ± 4.2	0.043	0.91
Moderately obese	30 (29.1%)	32 (39.5%)	0.220	
Severely obese	16 (15.5%)	10 (12.4%)	-0.090	
Morbidly obese	1 (1%)	2 (2.5%)	0.115	
Smoking status: smoker	4 (3.9%)	0 (0%)	-0.285	
Comorbidities				
0	68 (66%)	60 (74.1%)	0.178	
1	23 (22.3%)	12 (14.8%)	-0.194	
2+	12 (11.7%)	9 (11.1%)	-0.019	
Osteoarthritis				
KL 0	3 (2.9%)	0 (0%)	-0.244	
KL 1	0 (0%)	1 (1.2%)	0.156	
KL 2	13 (12.6%)	4 (4.9%)	-0.275	
KL 3	30 (29.1%)	26 (32.1%)	0.065	
KL 4	57 (55.3%)	50 (61.7%)	0.130	
WOMAC	34.4 ± 11.5	42.5 ± 13	0.660	1.06

Quality of life				
MCS	55.7 ± 8.1	55.8 ± 9.5	0.011	1.08
PCS	38.7 ± 6.6	35.6 ± 8.3	-0.413	1.12
HrQoL	0.72 ± 0.13	0.68 ± 0.16	-0.274	1.11
Osteoarthritis change (KL) from previous year				
< 1	91 (88.4%)	66 (81.5%)	-0.194	
1+	12 (11.7%)	15 (18.5%)	0.191	
WOMAC	7.2 ± 14.2	15.2 ± 18.2	0.490	1.13
Quality of life change from previous year				
PCS	-3.6 ± 7.0	-6.7 ± 8	-0.412	1.07
HrQoL	-0.05 ± 0.075	-0.08 ± 0.088	-0.367	1.08

Mean ± standard deviation are reported for continuous variables; counts (percentages) for discrete variables; ^Variance ratios are only available for continuous variables.

8. FIGURES

Figure 1: Mean HrQoL for treated (TKR) and control matches



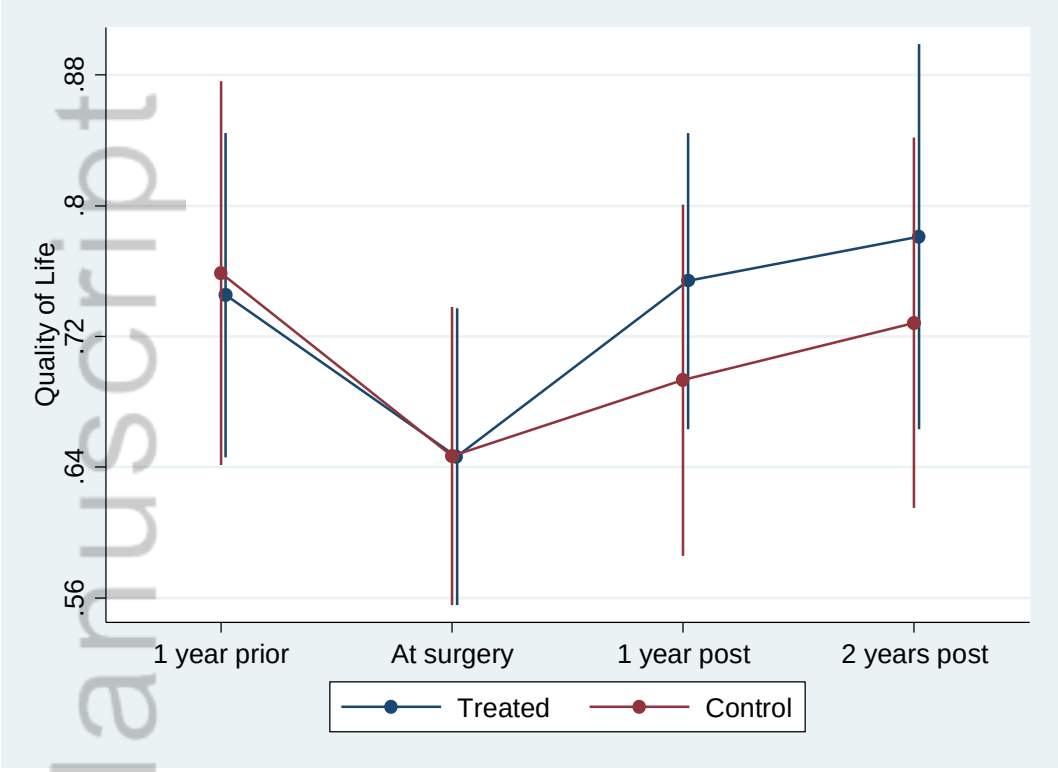
Supplementary table to Figure 1

	1 year prior	At surgery	1 year post	2 years post
Treated utility value (standard error)	0.771 (0.016)	0.722 (0.015)	0.770 (0.014)	0.775 (0.012)
Control utility value (standard error)	0.775 (0.015)	0.722 (0.015)	0.740 (0.015)	0.744 (0.015)
Treated vs pre-test[1] *Utility difference* *t-test p-value*	 0.049 0.001***	 0.000 N/A	 0.048 0.001***	 0.053 <0.001***
Treated vs matched control *Utility difference* *t-test p-value*	 -0.004 0.862	 0.000 >0.999	 0.030 0.147	 0.031 0.110

[1] Comparison with pre-test "at surgery" utility value is current practice in economic evaluation of many surgical interventions. *** statistically significant at the 1 per cent level.

Legend 1: Mean HrQoL trajectory for treated and control matches. When using current methods of economic evaluation, treatment significantly improves HrQoL. However when compared to the estimated control which shows evidence of RTM, treatment does not statistically improve HrQoL.

Figure 2: Mean HrQoL for treated (TKR) and control matches: top 5th percentile TKR propensity score



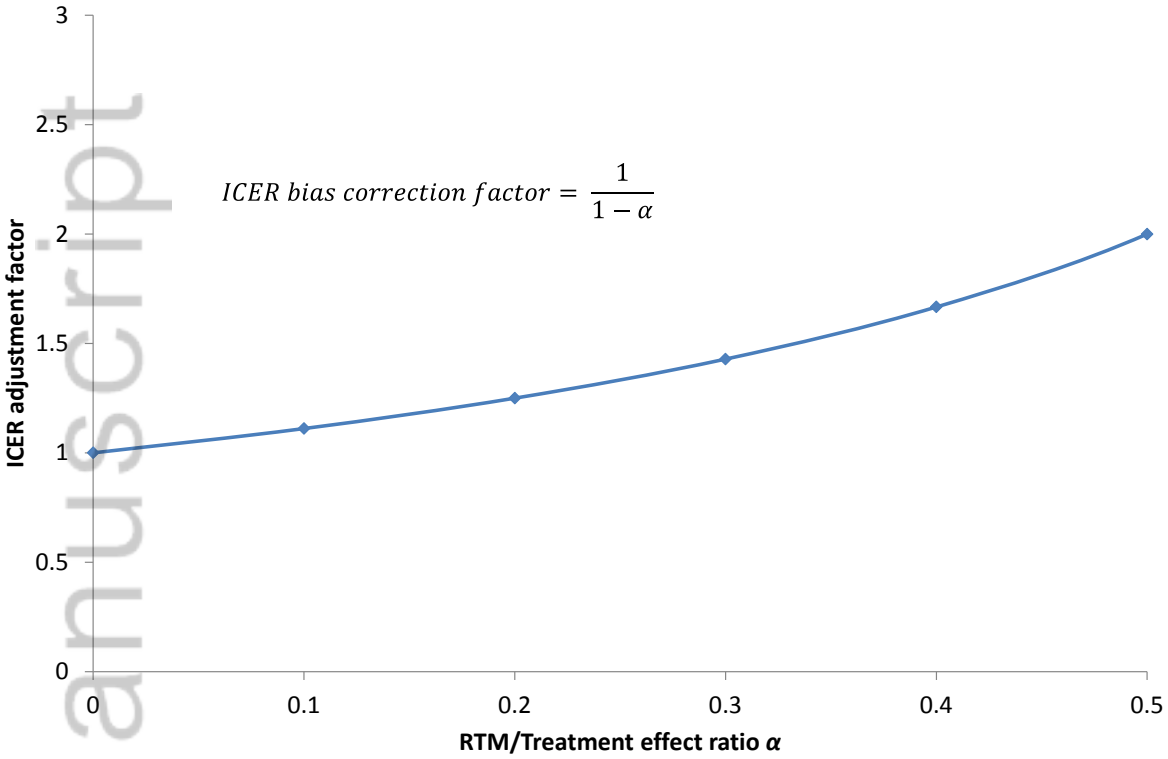
Supplementary table to Figure 2

	1 year prior	At surgery	1 year post	2 years post
Treated utility value	0.745	0.646	0.754	0.781
(standard error)	(0.050)	(0.046)	(0.046)	(0.060)
Control utility value	0.759	0.647	0.693	0.728
(standard error)	(0.060)	(0.046)	(0.055)	(0.057)
Treated vs pre-test[1]				
Utility difference	0.099	0.000	0.108	0.135
t-test p-value	0.052*	N/A	0.021**	0.027**
Treated vs matched control				
Utility difference	-0.013	-0.001	0.061	0.053
t-test p-value	0.865	0.994	0.397	0.528

[1] Comparison with pre-test "at surgery" utility value is current practice in economic evaluation of many surgical interventions. ** statistically significant at the 5 per cent level; * statistically significantly at the 10 per cent level.

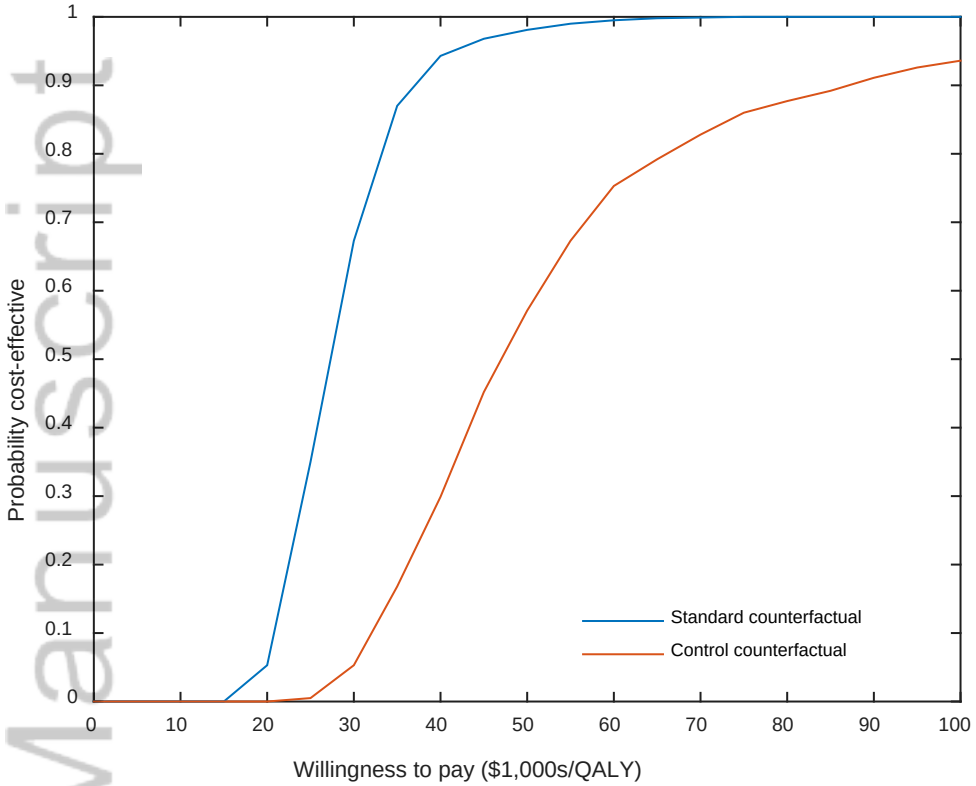
Legend 2: Mean HrQoL trajectory for treated and control matches for those most likely to receive treatment (top 5th percentile TKR propensity score). Relative to the average, this cohort has larger deteriorations in HrQoL prior to treatment, larger gains from treatment, and larger RTM in the absence of treatment.

Figure 3: ICER RTM bias correction factor as a function of RTM/Treatment effect ratio



Legend 3: ICER bias correction factor for RTM. The relationship between RTM and the ICER bias correction factor is increasing. As RTM increases, the bias of the cost-effectiveness of the treatment using pre to post changes grows increasingly worse.

Figure 4: Indicative cost-acceptability curve using matched treatment and control groups



Legend 4: Indicative cost-acceptability curves of using standard economic methods (blue), and after adjusting for RTM using an explicit matched control group (red). After adjusting for RTM, the probability that TKR is cost-effective at a willingness to pay threshold of $50,000/QALY drops from almost 100% to around 50%.