HEALTH ECONOMICS
Health Econ. **26**(Suppl. 1): 46–69 (2017)
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/hec.3474

A BIAS-ADJUSTED EVIDENCE SYNTHESIS OF RCT AND OBSERVATIONAL DATA: THE CASE OF TOTAL HIP REPLACEMENT

PETRA SCHNELL-INDERST[a]*, CYNTHIA P. IGLESIAS[b,c,d,e], MARJAN ARVANDI[a], ORIANA CIANI[f,g], RAFFAELLA MATTEUCCI GOTHE[a], JAIME PETERS[f], ASHLEY W. BLOM[h], ROD S. TAYLOR[f] and UWE SIEBERT[a,i,j]

[a] *Institute of Public Health, Medical Decision Making and Health Technology Assessment, Department of Public Health, Health Services Research and Health Technology Assessment, UMIT—University for Health Sciences, Medical Informatics and Technology, Eduard Wallnoefer Center I, Hall i.T., Austria*
[b] *Department of Health Sciences, University of York, Heslington, UK*
[c] *Centre for Health Economics, University of York, UK*
[d] *Hull and York Medical School, University of York, UK*
[e] *Luxemboug Institute of Health, Luxembourg*
[f] *Institute of Health Services Research, University of Exeter Medical School, Exeter, UK*
[g] *Centre for Research on Health and Social Care Management, Bocconi University, Milan, Italy*
[h] *Musculoskeletal Research Unit, University of Bristol, Bristol, UK*
[i] *Center for Health Decision Science, Department of Health Policy and Management, Harvard T.H. Chan School of Public Health, Boston, MA, USA*
[j] *Institute for Technology Assessment and Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA*

ABSTRACT

Evaluation of clinical effectiveness of medical devices differs in some aspects from the evaluation of pharmaceuticals. One of the main challenges identified is lack of robust evidence and a will to make use of experimental and observational studies (OSs) in quantitative evidence synthesis accounting for internal and external biases. Using a case study of total hip replacement to compare the risk of revision of cemented and uncemented implant fixation modalities, we pooled treatment effect estimates from OS and RCTs, and simplified existing methods for bias-adjusted evidence synthesis to enhance practical application.

We performed an elicitation exercise using methodological and clinical experts to determine the strength of beliefs about the magnitude of internal and external bias affecting estimates of treatment effect. We incorporated the bias-adjusted treatment effects into a generalized evidence synthesis, calculating both frequentist and Bayesian statistical models. We estimated relative risks as summary effect estimates with 95% confidence/credibility intervals to capture uncertainty.

When we compared alternative approaches to synthesizing evidence, we found that the pooled effect size strongly depended on the inclusion of observational data as well as on the use bias-adjusted estimates. We demonstrated the feasibility of using observational studies in meta-analyses to complement RCTs and incorporate evidence from a wider spectrum of clinically relevant studies and healthcare settings. To ensure internal validity, OS data require sufficient correction for confounding and selection bias, either through study design and primary analysis, or by applying post-hoc bias adjustments to the results. © 2017 The Authors. Health Economics published by John Wiley & Sons, Ltd.

Received 23 February 2016; Revised 2 August 2016; Accepted 23 November 2016

KEY WORDS: medical devices; health technology assessment; generalized evidence synthesis; bias adjustment; expert elicitation

*Correspondence to: Institute of Public Health, Medical Decision Making and Health Technology Assessment, Department of Public Health, Health Services Research and Health Technology Assessment, UMIT—University for Health Sciences, Medical Informatics and Technology Eduard Wallnoefer Center I, A-6060 Hall i.T., Austria. E-mail: petra.schnell-inderst@umit.at

1. INTRODUCTION

In many countries, comparative effectiveness research (CER) is well established as part of health technology assessment (HTA) of pharmaceutical therapies (Panteli, 2016). (Although, there is no consensus on how to optimally implement CER for medical devices (MDs), developing and promoting the use of methodological guidance for the evaluation of MDs within an HTA framework is a goal for the European network for Health Technology Assessment (EUnetHTA) (www.eunethta.eu) in Joint Action 2 (2012–2015) and 3 (2016–2019) (Schnell-Inderst *et al.* 2015). A primary challenge, identified through conducting HTAs of MDs, is the lack of robust evidence on clinical effectiveness and cost-effectiveness (Iglesias, 2015).

MDs typically show rapid and incremental development with product life cycles shorter than three years (Siebert *et al.*, 2002, Schulenburg *et al.*, 2009), which results in frequent technology updates that often demonstrate only minor modifications and market access of similar competing products. The need for new randomized clinical trials (RCTs) to demonstrate the incremental effectiveness of marginal modifications of MDs may be impracticable, limited by insufficient sample size, limited follow-up time and costs (Konstam *et al.*, 2003). RCT designs that can account for the incremental development process, and the additional challenges of MDs, such as patient and clinician preferences, the lack of double-blinding and technology changes over time have been proposed (Bernard *et al.* 2014, Royal Netherlands Academy of Arts and Sciences 2014). Including empirical data on the clinical effectiveness from observational studies (OS) can complement evidence from RCTs. Device- and disease-specific registries have been established to provide long-term data on the effectiveness and safety of MDs in routine clinical practice. These registry data are being used to help guide clinical practice and medical decision making, and are especially relevant in the case of MDs where effectiveness often relies on user proficiency (i.e. a ‘learning curve’) and contextual factors, including the clinical setting in which the MD is being used.

There are some indications of a will to use evidence from RCT and observational studies complementarily. For example, HTA agencies, such as the National Institute for Health and Care Excellence (NICE) in United Kingdom, often require identification of all of the relevant sources of evidence, and do not restrict evidence synthesis to RCTs (NICE 2013). As OSs are prone to selection bias and confounding, the appropriateness of combining experimental and observational evidence quantitatively is the subject of debate (Verde and Ohmann, 2014). The Cochrane Collaboration recommends considering the two types of evidence separately and not pool the different study designs (Higgins and Green, 2011) into a single summary effect estimate. Statistical methods for generalized evidence synthesis approaches perform bias adjustments of observational and randomized evidence are increasingly being published. A recently published review identified 20 unique statistical approaches (in addition to the traditional fixed- or random-effects meta-analytic methods) to combine randomized and non-randomized studies in clinical research (Verde and Ohmann 2014). In 15 of these approaches, there were alternative bias-adjustment approaches, 12 of which used Bayesian methods. Bias correction methods propose either down-weighting studies with a high risk of bias or modelling study-specific biases based on individual study characteristics. Observed treatment effects are typically adjusted at the individual study level prior to synthesizing the evidence (Welton and Ades, 2009, Welton *et al.* 2012). One particular approach, described by Turner *et al.* (2009), allows, at least theoretically, for a complete bias correction by adjusting the observed treatment effects for internal and external biases at the individual study level using expert elicitation, followed by synthesizing across multiple studies. This approach follows standard HTA methods where risk of bias assessments is performed at the individual study level. Despite the advantages of this approach, to our knowledge, it has been rarely implemented in practice (Verde and Ohmann 2014).

In this study, we aim to: (i) illustrate the use of current statistical methods to combine treatment effect estimates from observational studies and RCTs using an illustrative application of total hip replacement (THR) prostheses as a case-study for assessing the clinical effectiveness of MDs, and (ii) simplify existing methods for bias-adjusted evidence synthesis to enhance practical application by HTA practitioners.

As the main objective of this analysis is to illustrate the application of statistical methods, readers are cautioned that our findings should not be considered as definitive evidence to support any claim on the clinical effectiveness associated with THR prostheses which we have used as our case study.

DOI: 10.1002/hec

2. METHODS

2.1. Rationale for the choice of the case example

THR illustrates the life cycle of medical device technology and clinical evidence production/development well. A total hip construct consists of a femoral component that articulates with an acetabular component (see Appendix Table A1 for a classification of prostheses). THR is an MD with a relatively well-supported clinical effectiveness evidence base that includes RCTs as well as data from numerous large national registries, and is well-suited to use as an illustrative application of generalized evidence synthesis approaches that combine RCT and OSs in a meta-analysis (Clarke *et al.* 2013).

2.2. Methodological framework to synthesize RCT and observational evidence

We followed the methodology proposed by Turner *et al.* (2009), to conduct a generalized evidence synthesis to combine observational and RCT evidence on the clinical effectiveness of different fixation methods of THR. The approach used by Turner et al. was considered by members of the MedtecHTA consortium as a methodological framework that: (i) provided a comprehensive approach for bias adjustment in the context of evidence synthesis; (ii) while resource intensive outlined a rationale that was fairly intuitive and easier to follow than that of other statistical approaches for bias adjustment; (iii) explicitly acknowledged the role of expert judgement for bias elicitation in medical device evaluation; and (iv) promoted the use of standard (i.e. simpler) methods for evidence synthesis (i.e. meta-analysis) than other statistical approaches for bias-adjusted evidence synthesis.

Similar to Turner and colleagues, we performed our analysis using the following five steps (for details see original publication (Turner *et al.*, 2009)): (i) framed the clinical target question; (ii) identified the relevant evidence base; (iii) extracted data, assessed bias and transformed reported treatment effect estimates; (iv) elicited expert opinion to determine bias-adjusted treatment effects; and (v) performed quantitative synthesis of RCT and observational data, meta-regression and sensitivity analyses.

2.3. Target question

Our case example target question asked, ‘Which fixation method—cemented or uncemented—is more effective in terms of revision rate for adult patients with end stage hip arthritis undergoing THR?’ We specified this question according to the following PICOS framework:

- Population: Adult population (>18 years) with end stage hip arthritis for whom non-surgical management has failed;
- Intervention: Cemented THR with a polyethylene-metal articulation;
- Comparator: Uncemented THR with a polyethylene-metal articulation;
- Outcomes: Revision risk at ≥5 years follow-up.
- Target setting: THR procedure applicable in a United Kingdom (UK) district general hospital.

2.4. Evidence base

We identified the evidence base for this study using four previously published systematic reviews of THRs undertaken by the HTA Programme in UK (Clarke *et al.*, 2013, Faulkner *et al.*, 1998, Fitzpatrick *et al.*,1998, Vale *et al.*, 2002) and four additional systematic reviews cited in reports (Tsertsvadze *et al.*, 2014, Clement *et al.*,2012, Pakvis *et al.*, 2011, Voigt and Mosier, 2012).

We focused on a subset of THR evidence, that is, all RCTs and OSs (i.e. cohort studies, case-control studies or registries) that directly compared the cemented and uncemented fixation methods for hip implants. Where multiple publications from the same population/study were identified, we selected the most recent publication. We excluded reports where core data for the analyses were not available or where no revisions occurred during follow-up. Duplicate publications and national registry reports from outside the European Union (EU) were excluded because they may not be applicable within the UK setting; however, we included cohort studies conducted outside the EU, if their setting was deemed potentially applicable to the UK setting.

DOI: 10.1002/hec

2.5. Outcome measure

Because our initial review identified few studies that reported hip implant revisions in terms of time-to-event, we used the more commonly reported metric of the proportion of patients who received a revision during follow-up. For this outcome, we calculated (crude) relative risks (RR) and 95% confidence intervals (95%CI) for each study comparing the cemented versus uncemented fixation approach. We selected RR rather than odds ratios (OR) as our preferred metric as they are easier to understand by clinical experts (Froud *et al.* 2012). We also extracted RRs adjusted for baseline covariates (i.e. confounder-adjusted RR) from a subset of studies, if reported.

2.6. Data extraction, bias assessment and transformation of reported treatment effect estimates

For each included study, we extracted study design (e.g. RCT, cohort and registry), duration of follow-up (in years), population characteristics (e.g. mean age, proportion of women) and the proportion of revisions in each treatment arm and/or the confounder-adjusted RR and 95%CI. For each study, internal biases were assessed using the Cochrane risk of bias tool (Sterne *et al.* 2014). External biases were assessed using the framework proposed by Dekkers *et al.* (2010) and Rothwell (2010). Two authors (OC, MA) extracted data, while bias assessment was initially conducted by a single author (OC), and confirmed by a second author (RST).

2.7. Expert elicitation to determine bias-adjusted treatment effects

To improve the feasibility and practical implementation of eliciting bias-adjustment weights suggested by Turner *et al.*, we made four main adaptions that included: (i) eliciting bias-adjusted treatment effects directly rather than for biases per se; (ii) eliciting overall bias at the level of each study rather than for each individual bias separately; (iii) using a modified elicitation tool; and (iv) incorporating a qualitative tool to aid experts when undertaking their bias assessment. The bias elicitation process is summarized in Figure 1. A trained facilitator (OC/RST) introduced the elicitation tasks and invited experts to complete the questions individually. Experts were provided with a qualitative tool to assist them completing the elicitation task. During two separate and consecutive meetings, internal biases (Appendix Table A2) were elicited from methodological experts and external biases (Appendix Table A3) from clinicians. The Turner *et al.* bias-adjustment method recognizes that the impact of biases on treatment effects may differ across different types of biases and suggests individually eliciting bias adjusting weights for each bias type and study.

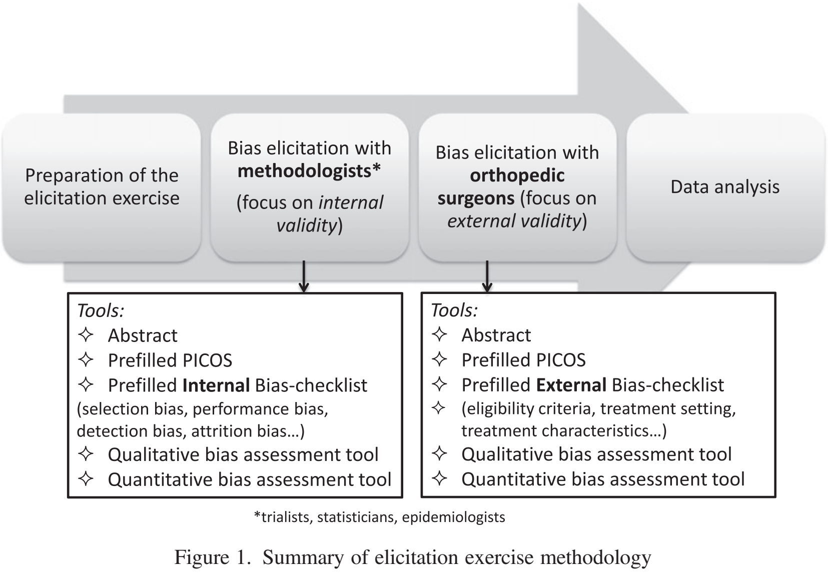

Preparation of the elicitation exercise
Bias elicitation with methodologists* (focus on internal validity)
Bias elicitation with orthopedic surgeons (focus on external validity)
Data analysis
Tools:
✧ Abstract
✧ Prefilled PICOS
✧ Prefilled Internal Bias-checklist (selection bias, performance bias, detection bias, attrition bias…)
✧ Qualitative bias assessment tool
✧ Quantitative bias assessment tool
Tools:
✧ Abstract
✧ Prefilled PICOS
✧ Prefilled External Bias-checklist
✧ (eligibility criteria, treatment setting, treatment characteristics…)
✧ Qualitative bias assessment tool
✧ Quantitative bias assessment tool
*trialists, statisticians, epidemiologists

Figure 1. Summary of elicitation exercise methodology

Preliminary consultation with experts during a pilot session indicated that they may struggle to elicit these unobservable quantities; therefore, we elected to elicit internal and external bias weights aggregately, for each study.

Detailed information on the elicitation exercises are provided in our full report for MedTecHTA project WP3 to the EU, and are available upon request. Briefly, in a first meeting, the impact of internal biases on the THR revision RRs was elicited by methodologists, while a follow-up meeting with orthopaedic surgeons elicited the impact of external biases. For each study, participants were first asked to complete a qualitative tool that asked them to qualitatively assess the likely importance or direction of each bias attribute (e.g. selection bias, performance bias) (Appendix Figure A1). At the first meeting, methodologists were presented with the calculated crude THR RRs and 95% CIs. Methodologists were asked to provide what they considered would be an unbiased treatment effect by indicating the RRs and 95%CI estimates on a grid provided in the elicitation tool for each study after considering the specified internal biases identified for each study (see Appendix Figure A2). At the second meeting for external bias elicitation, the study-specific RR and 95% CIs presented to each surgeon were the 'typical' value estimated from the first methodologists' meeting. We calculated the 'typical' value by taking the median of mean effect, and the upper and lower CIs of these estimates, for the bias-adjusted RR (95% CI) across all assessors of each study (Appendix Figure A3). This was done so that the starting point for assessment of external bias by surgeons was already conditioned by the internal bias adjustment performed by the methodologists, subsequently allowing an *ex post* calculation of the magnitude of the internal and external biases. Surgeons were asked to provide their estimate of the unbiased treatment effect assuming all external biases had been removed.

The two elicitation meetings were held between May and June 2015, each lasting 2 to 3 h. Nine methodologists and eleven orthopedic surgeons attended each workshop. Each expert received five to seven studies to assess, distributed randomly. All nine methodologists were drawn from the Institute of Health Research, University of Exeter Medical School in England and included statisticians, epidemiologists, trialists and health service researchers. The elicitation of external biases was undertaken by eleven orthopedic surgeons, practicing in Bristol in the South West of England.

To accurately quantify a participant's adjusted RR and 95% CI, the quantitative bias-assessment graph for all studies and all participants was digitized using Plot Digitizer 2.6.6© Dice 2014 software. These estimates for the fully bias-adjusted RR (95% CI) were then pooled across participants by taking the median of mean effect, and upper and lower CIs, to obtain a bias-adjusted treatment-effect for a 'typical' assessor. Finally, the log(RR) and standard errors were derived for the observed, methodologist 'typical' assessor and surgeon 'typical' assessor treatment effect for each study, to correct for potentially asymmetrical CIs provided by experts.

2.8. Synthesis of randomized and non-randomized data

2.8.1. Main steps of synthesis. We used different stepwise synthesis models (see below) to pool the reported and bias-adjusted RR of the revision rate that included: (i) RCTs only; (ii) RCTs and registries; and (iii) from all types of study designs (i.e. RCTs, registries and cohorts). In a hierarchical model distinguishing already between study types, all studies were included in a single step. We pre-specified subgroup analyses for variables from the literature that are recognized as influencing treatment effect (i.e. patient age, gender and mobility) and additional variables of interest (i.e. duration of follow-up and study type). Univariate and multivariate meta-regression analyses were conducted for these variables to investigate effect modification.

2.8.2. Analyses and statistical models of meta-analyses. We pooled studies using standard frequentist fixed effect model (FEM) and random effects model (REM) meta-analysis in STATA 14.1 using commands 'metan' and 'admetan' (Stata Corp LP, College Station, TX). Additionally, we undertook two different Bayesian approaches to meta-analysis: (i) Bayesian random effects meta-analysis (BREM) using conjugate, non-informative priors (normal and uniform (0;10) distributions for mean and variance, respectively) and posterior RRs; their standard deviations (SD) were empirically estimated and approximated using Monte Carlo simulation with 100 000 repetitions; (ii). Bayesian hierarchical meta-analysis (BHM) that accounts for the variability between different study types in addition to variability within and between studies (Welton *et al.*, 2012). All Bayesian analyses were performed in WinBUGS software version 1.4.3, 2007 (Lunn *et al.*, 2000) adapting the codes by Welton *et al.* (2012).

For subgroup analyses, missing values were imputed using the mean value from all studies in each arm. Continuous variables were divided into two subgroups with the median as cut-point. Frequentist meta-regression was performed with the command 'metareg' in STATA 14.1 and Bayesian meta-regression with WinBUGS. Because of the limited number of studies and the potential of overfitting, we used forward variable selection from the pre-specified variables. The degree of statistical heterogeneity was assessed by calculating I-square (I^2) and tau-square (τ^2). We calculated the confidence or credibility interval limits ratio (CILR) between the upper and the lower bound as a measure of the relative width of the confidence or credibility intervals (CrIs), respectively.

We performed sensitivity analyses by varying the assumptions for priors in BREM and BHM. We used a t-distribution for the mean and three different weakly informative priors for the variance (i.e. Gamma, Inverse-Gamma, and Half-Cauchy distributions), as suggested by Gelman (2006).

3. RESULTS

3.1. Study selection

We included a total of 15 studies into our evidence synthesis: seven RCTs (Angadi *et al.*, 2012; Bjorgul *et al.*, 2010; Corten *et al.*, 2011; Kim *et al.*, 2011; McCombe and Williams; 2004, Reigstad *et al.*, 1993; Wykman *et al.* 1991), five cohort studies (Clohisy and Harris, 2001; Kim *et al.*, 2003; Kruckhans and Dustmann, 2004; Pospula *et al.*, 2008; Hartofilakidis *et al.*, 2009) and three reports from national arthroplasty registries (Hailer *et al.*, 2010; Makela *et al.*, 2011; Pennington *et al.*, 2013). The study selection process is illustrated in Figure 2, and a full citation listing of the included studies is provided in the Supplementary Web Appendix.

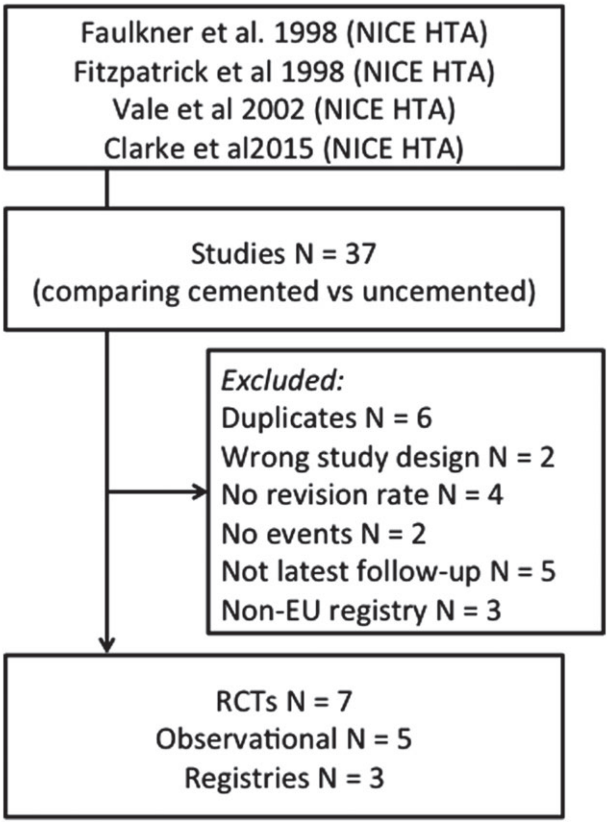

Faulkner et al. 1998 (NICE HTA)
Fitzpatrick et al 1998 (NICE HTA)
Vale et al 2002 (NICE HTA)
Clarke et al2015 (NICE HTA)
Studies N = 37
(comparing cemented vs uncemented)
Excluded:
Duplicates N = 6
Wrong study design N = 2
No revision rate N = 4
No events N = 2
Not latest follow-up N = 5
Non-EU registry N = 3
RCTs N = 7
Observational N = 5
Registries N = 3

Figure 2. Study selection process

3.2. Study characteristics

Studies were published between 1993 and 2013 with a follow-up between 5 and 20 years (Table I). The mean age of study participants varied between 42.5 and 70.2 years, and the proportion of females included in each study ranged from 20 to 80%. Comparable data on the degree of mobility were not available for both study arms. The registry-based studies reported RRs adjusted for different baseline variables, while RCTs and cohort studies did not control for confounding factors.

Table I. Study and patient characteristics for the 15 included studies

Study	Study type	*N* of revisions/ total cemented	*N* of revisions/ total uncemented	Relative risks computed/reported (95% CI)	Overall mean age (years)	Overall proportion of women	Follow-up (years)
Wykman (1991)	RCT	8/90	14/94	0.60 (0.26–1.35)	66.1	0.5	5
Reigstad (1993)	RCT	0/60	2/60	0.20 (0.01–4.08)	64.5	0.7	5
McCombe (2004)	RCT	0/84	4/78	0.10 (0.01–1.89)	67.4	n.r.	8
Bjorgul (2010)	RCT	9/120	4/120	2.25 (0.71–7.11)	65.5	0.7	10
Corten (2011)	RCT	48/124	31/126	1.57 (1.08–2.29)	64.0	0.5	7
Kim (2011)	RCT	17/114	22/114	0.77 (0.43–1.38)	45.1	0.2	20
Angadi (2012)	RCT	17/183	11/104	0.88 (0.43–1.80)	70.2	0.6	10
Hailer (2010)	Registry	9689/161460[a]	1343/8953[a]	0.67 (0.62–0.71)[b]	n.r.	0.6	10
Makela (2011)	Registry	197/1535[a]	60/579[a]	0.85 (0.62–1.16)[b]	65.0	0.8	15
Pennington (2013)	Registry	216/16882[a]	420/18845[a]	0.60 (0.55–0.66)[b]	70.2	0.4	5
Clohisy (2001)	Cohort	1/45	2/45	0.50 (0.05–5.32)	61.5	0.5	10
Kim (2003)	Cohort	7/42	8/45	0.94 (0.37–2.36)	46.8	0.4	10
Kruckhans (2004)	Cohort	10/200	26/200	0.38 (0.19–0.78)	63.7	0.5	5
Pospula (2008)	Cohort	2/87	1/96	2.21 (0.20–23.9)	50.2	0.6	5
Hartofilakidis (2009)	Cohort	14/50	18/51	0.79 (0.40–1.42)	42.5	0.6	10

n.r., not reported; CI, confidence interval; RCT, randomized controlled trial; RR, relative risk.

[a] Numbers of revisions were taken from reported numbers in the study publications (10-year-survival, total number of revisions and 5-year revision rates in Hailer, Makela and Pennington, respectively).

[b] Reported RR (95% CI) adjusted for patient level covariates (Hailer: adjusted for sex, age and primary diagnosis; Makela: adjusted for age and sex; Pennigton: adjusted for age, sex, BMI, ASA grade, surgeon grade, hospital type, Charlson score and date of surgery).

3.3. Bias elicitation exercise

In general, the first meeting with methodologists revealed that for each study they were either uncertain about the impact of internal bias on the revision rates or felt the internal bias had no impact on the outcome (Table II and Appendix Table A4). Their judgment did not appear to be influenced by whether the study was an RCT or an OS. The methodologists appeared to generally agree on the effect of the biases; however, there were exceptions (e.g. qualitative assessment of the impact of selection bias in Kruckhans *et al.* (2004)).

Table II. Qualitative bias-assessment—proportion of methodologists reporting likely impact of internal biases for each study

Study	Study type	Not causing any bias/not likely to favor either implant	Likely to favor cemented implant	Likely to favor uncemented implant	Do not know
Wykman (1991)	RCT	0.13	0.08	0.00	0.79
Reigstad (1993)	RCT	0.04	0.21	0.04	0.71
McCombe (2004)	RCT	0.17	0.17	0.04	0.63
Bjorgul (2010)	RCT	0.21	0.08	0.21	0.50
Corten (2011)	RCT	0.58	0.04	0.04	0.33
Kim (2011)	RCT	0.25	0.13	0.00	0.63
Angadi (2012)	RCT	0.25	0.04	0.21	0.50
Hailer (2010)	Registry	0.25	0.00	0.04	0.71
Makela (2011)	Registry	0.42	0.08	0.04	0.46
Pennington (2013)	Registry	0.17	0.04	0.17	0.63
Clohisy (2001)	Cohort	0.52	0.06	0.04	0.39
Kim (2003)	Cohort	0.04	0.13	0.13	0.71
Kruckhans (2004)	Cohort	0.17	0.08	0.08	0.67
Pospula (2008)	Cohort	0.17	0.08	0.25	0.50
Hartofilakidis (2009)	Cohort	0.29	0.00	0.29	0.42

RCT, randomized controlled trial.

In contrast, the orthopedic surgeons provided stronger opinions about the effects external biases may have on the results of the studies (Table III and Appendix Table A5). According to the surgeons, RCTs were less likely to result in external biases; although, one of the dimensions of external validity (i.e. How was the outcome measured? Who measured the outcome? Adequacy of the length of follow-up?) had the highest degree of variation in judgment across participants' replies.

In general, the bias-adjusted mean treatment effects provided by experts resulted in less effective estimates (Table IV). However, there was a single case where adjusting for bias changed from favoring uncemented implants to favoring cemented implants. The standard error of treatment effect estimate generally increased after bias adjustment indicating increased uncertainty as a result of the bias-adjustment process, although there were some exceptions.

Table III. Qualitative bias-assessment—proportion of orthopaedic surgeons reporting likely impact of external biases for each study

Study	Study type	Not causing any bias/not likely to favor either implant	Likely to favor cemented implant	Likely to favor uncemented implant	Do not know
Wykman (1991)	RCT	0.40	0.25	0.25	0.10
Reigstad (1993)	RCT	0.65	0.20	0.05	0.10
McCombe (2004)	RCT	0.75	0.25	0.00	0.00
Bjorgul (2010)	RCT	0.65	0.10	0.25	0.00
Corten (2011)	RCT	0.65	0.00	0.35	0.00
Kim (2011)	RCT	0.70	0.05	0.20	0.05
Angadi (2012)	RCT	0.65	0.10	0.25	0.00
Hailer (2010)	Registry	0.45	0.35	0.15	0.05
Makela (2011)	Registry	0.55	0.20	0.15	0.10
Pennington (2013)	Registry	0.50	0.25	0.25	0.00
Clohisy (2001)	Cohort	0.40	0.02	0.49	0.09
Kim (2003)	Cohort	0.45	0.10	0.40	0.05
Kruckhans (2004)	Cohort	0.35	0.40	0.20	0.05
Pospula (2008)	Cohort	0.35	0.05	0.40	0.20
Hartofilakidis (2009)	Cohort	0.50	0.15	0.30	0.05

RCT, randomized controlled trial.

Table IV. Elicited bias-adjusted treatment effects by study

Study	Study type	Assessor	Elic RR	Elic LB 95% CI	Elic UB 95%CI
Wykman (1991)	RCT	Typical assessor[a]	0.5	0.1	1.2
		Typical assessor method[b]	0.64	0.32	1.4
		Observed[c]	0.6	0.26	1.35
Reigstad (1993)	RCT	Typical assessor	0.6	0.04	9.69
		Typical assessor method	0.5	0.02	7.89
		Observed	0.2	0.01	4.08
McCombe (2004)	RCT	Typical assessor	0.33	0.1	3.75
		Typical assessor method	0.27	0.02	5.61
		Observed	0.1	0.01	1.89
Bjorgul (2010)	RCT	Typical assessor	1.5	0.7	3.5
		Typical assessor method	1.61	0.52	4.8
		Observed	2.25	0.71	7.11
Corten (2011)	RCT	Typical assessor	1.53	1.08	1.95
		Typical assessor method	1.57	1.08	2.29
		Observed	1.57	1.08	2.29
Kim (2011)	RCT	Typical assessor	0.71	0.42	1.58
		Typical assessor method	0.79	0.44	1.64
		Observed	0.77	0.43	1.38
Angadi (2012)	RCT	Typical assessor	0.8	0.37	1.77
		Typical assessor method	0.83	0.45	1.65
		Observed	0.88	0.43	1.80
Hailer (2010)	Registry	Typical assessor	0.82	0.72	0.96
		Typical assessor method	0.68	0.57	0.76
		Observed	0.67	0.62	0.71
Makela (2011)	Registry	Typical assessor	0.87	0.63	1.27
		Typical assessor method	0.85	0.63	1.12
		Observed	0.85	0.62	1.16
Pennington (2013)	Registry	Typical assessor	0.67	0.55	0.9
		Typical assessor method	0.66	0.48	0.77
		Observed	0.6	0.55	0.66
Clohisy (2001)	Cohort	Typical assessor	0.55	0.05	7.6
		Typical assessor method	0.69	0.06	7.2
		Observed	0.5	0.05	5.32
Kim (2003)	Cohort	Typical assessor	0.63	0.2	3.52
		Typical assessor method	0.86	0.35	2.58
		Observed	0.94	0.37	2.36
Kruckhans (2004)	Cohort	Typical assessor	0.45	0.15	1.45
		Typical assessor method	0.38	0.25	0.6
		Observed	0.38	0.19	0.78
Pospula (2008)	Cohort	Typical assessor	0.78	0.08	8
		Typical assessor method	1	0.15	15.76
		Observed	2.21	0.2	23.91
Hartofilakidis (2009)	Cohort	Typical assessor	0.8	0.43	1.38
		Typical assessor method	0.67	0.35	1.24
		Observed	0.79	0.44	1.42

Elic RR, elicited relative risk; Elic LB 95% CI, elicited lower bound of the 95% confidence interval; Elic UB 95%, elicited upper bound of the 95% confidence interval.

[a] Accounts for internal and external bias elicitation.

[b] Accounts for internal bias elicitation alone.

[c] Empirically observed results from the published study.

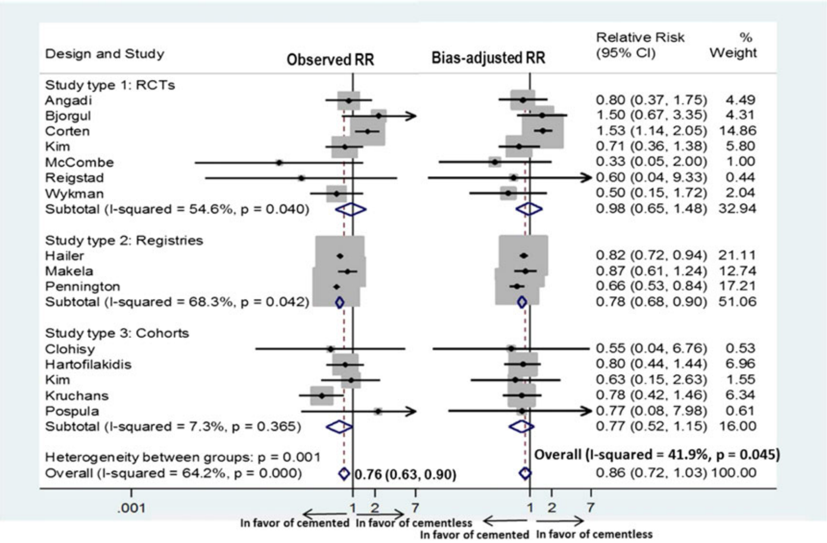
3.4. Evidence synthesis

All bias-unadjusted meta-analytic results from the same study design were included in a stepwise fashion in FEM, REM and BREM starting with RCTs, which were considered to have the highest internal validity, and were repeated for bias-adjusted RRs (Table V). As BHM inherently accounts for the variance of different study designs in the statistical model, a stepwise inclusion is not possible.

3.4.1. Bias-unadjusted effect estimates of stepwise inclusion by study type. When we only considered unadjusted evidence from RCTs in the FEM, REM (Figure 3) and BREM meta-analyses, the pooled effect estimates did not show any favorable statistically significant effect for either treatment modality. After adding the three registry studies with their larger sample sizes, the level of statistical uncertainty decreased and we identified a statistically significant effect for one treatment approach for the FEM and REM analyses, but not in the BREM analysis. After including registry-based studies, the CILR reduced for all three models. Adding the five cohort studies with comparably small sample sizes only slightly changed the effect estimate and CI/CrIs. For BHM, the point estimate was similar to the REM and BREM analyses; however, the CrIs were wider and did not show a differing treatment effect (i.e. still included a RR of 1) (Table V).

3.4.2. Impact of bias adjustment on effect estimates. Using the bias-adjusted effect estimates shifted the pooled effect estimate towards a lower treatment effect, which was no longer statistically significant in any of the models, although the width of the CI/CrI increased slightly, resulting in larger *p*-values (Table V). When all studies were included in the frequentist and Bayesian meta-analyses, adjusting for biases reduced the statistical heterogeneity from $I^2 = 64\%$ to 42%, from $\tau^2 = 0.36$ to 0.28, and from $\tau^2 = 0.80$ to 0.69 for the REM (Figure 4), BREM and BHM models, respectively.

3.4.3. Effect modification and sensitivity analyses. Univariate meta-regression showed no statistically significant effect modification for any variables with either the reported or with the bias-adjusted RRs in REM analysis (Appendix Table A6, Figure 5). REM bivariate meta-regression models identified no statistically significant pair of covariates explaining between-study heterogeneity (Appendix Table A7). This finding was similar for the Bayesian models (Appendix Table A8). Sensitivity analyses substituting non-informative priors for the variance with weakly informative priors in BHM of all studies showed robustness of analyses and resulted in narrower credibility intervals (Appendix Table A9).



RCT: randomized controlled trial, RR: relative risk, 95%CI: 95% confidence interval

Figure 3. Classical random effects meta-analysis of observed and bias-adjusted effect estimates

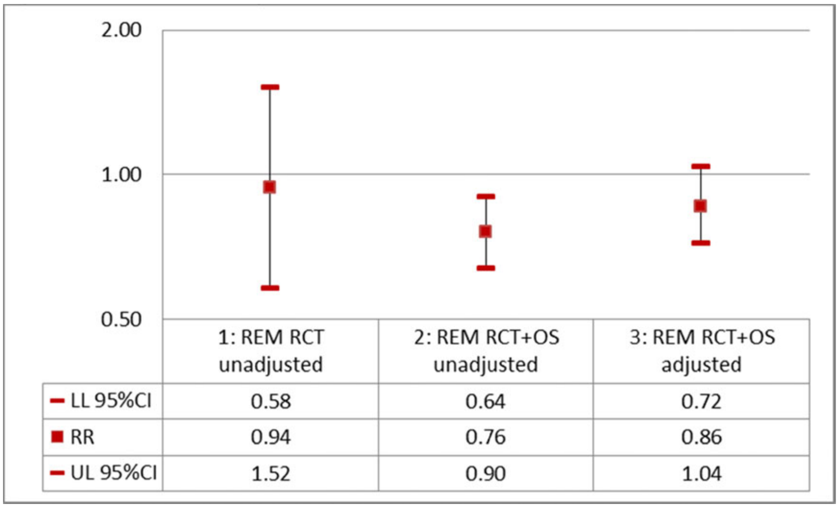
Table V. Results from all meta-analyses for unadjusted and bias-adjusted RRs using frequentist FEM, frequentist REM and Bayesian methods

Frequentist	FEM	REM	FEM	REM
Studies included in meta-analysis	Unadjusted RR (95%CI) [CILR]	Unadjusted RR (95%CI) [CILR]	Bias-adjusted RR (95%CI) [CILR] Inconsistency I^2	Bias-adjusted RR (95%CI) [CILR] Inconsistency I^2
RCTs	1.12 (0.86–1.45) [1.69]	0.94 (0.58–1.52) [2.62] $I^2 = 54.8\%$	1.21 (0.96–1.54) [1.60]	0.98 (0.65–1.48) [2.28] $I^2 = 42.2\%$
RCTs and registries	0.67 (0.64–0.70) [1.09]	0.78 (0.65–0.95) [1.46] $I^2 = 74.1\%$	0.85 (0.77–0.94) [1.22]	0.88 (0.70–1.11) [1.59] $I^2 = 62.0\%$
All 15 studies	0.67 (0.64–0.70) [1.09]	0.76 (0.64–0.90) [1.41] $I^2 = 64.2\%$	0.85 (0.77–0.94) [1.22]	0.86 (0.72–1.04) [1.44] $I^2 = 41.9\%$

Bayesian REM		
Studies included in meta-analysis	Unadjusted posterior RR (95%CrI) Heterogeneity Tau^2	Bias-adjusted posterior RR (95% CrI) Heterogeneity Tau^2
RCTs	0.90 (0.37–1.71) [4.62] $Tau^2 = 0.65$	0.94 (0.46–0.62) [3.52] Tau^2
RCTs and registries	0.80 (0.55–1.17) [2.73] $Tau^2 = 0.43$	0.87 (0.62–1.18) [1.90] $Tau^2 = 0.35$
All 15 studies	0.77 (0.58–1.03) [1.78] $Tau^2 = 0.36$	0.85 (0.66–1.07) [1.62] $Tau^2 = 0.28$
Three-level hierarchical[a]	0.74 (0.16–3.71) [23.19] $Tau^2 = 0.80$	0.82 (0.21–3.31) [15.76] $Tau^2 = 0.69$

FEM, fixed-effects model; REM, random-effects model; RR, relative risk; CILR, confidence/credibility interval limits ratio (i.e. the ratio of upper bound and lower bound of confidence/credibility interval to express the relative confidence interval "width"); RCT, randomized controlled trial.

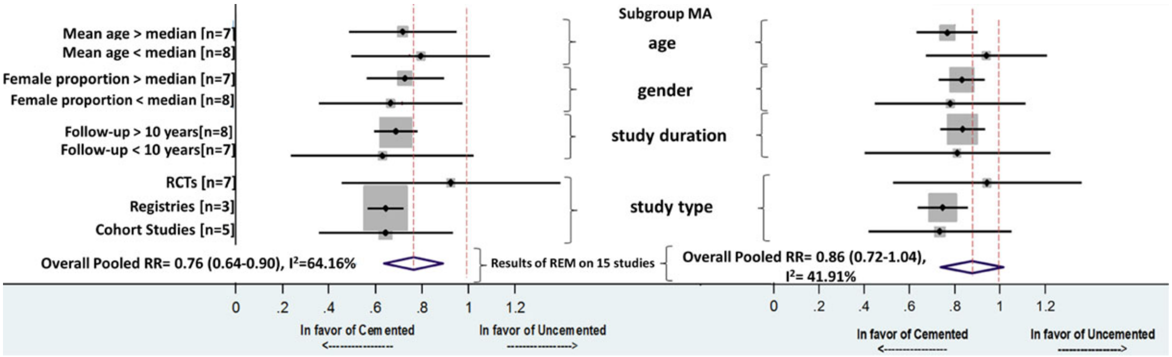
[a] Levels of study type (i.e. RCTs, registries and cohort studies).



	1: REM RCT unadjusted	2: REM RCT+OS unadjusted	3: REM RCT+OS adjusted
▬ LL 95%CI	0.58	0.64	0.72
■ RR	0.94	0.76	0.86
▬ UL 95%CI	1.52	0.90	1.04

REM: random-effects model, RCT: randomized clinical trials, OS: observational studies, RR: relative risk, 95%CI: 95% confidence interval, LL:lower limit, UL: upper limit

Figure 4. Stepwise meta-analysis for the frequentist random effects model



RR: relative risk (with 95% confidence intervals), REM: random effects model, I^2: I-squared (inconsistency)

Figure 5. Subgroups in classical random effects meta-analyses bias-unadjusted and bias-adjusted

4. DISCUSSION

Evaluating the comparative effectiveness of MD technologies often involves evidence from multiple study designs (e.g. RCTs and OSs); raising the question of whether this evidence can be combined into a single meta-analysis (Higgins and Green, 2011). For HTAs, a single combined effect estimate may not only be needed to provide an overall assessment of the relative clinical effectiveness of an MD, but also required to conduct cost-effectiveness analyses. Using THR as an illustrative case, we compared the risk of revision of cemented and uncemented implant fixation modalities by pooling alternative sources of data.

Comparing several frequentist and Bayesian evidence synthesis approaches, we showed that the effect estimates for revision rates were very similar between the frequentist and Bayesian random effects models, with wider 95%CrI than the 95%CI, likely because of the wide-ranging uniform priors. Overall, the pooled results strongly depended on the inclusion of OSs as well as using bias-adjusted estimates. To understand the potential bias introduced by including unadjusted observational data, the point estimates and relative CI or CrI widths between the different analyses should be compared simultaneously. We showed that in the frequentist analyses, statistical significance was achieved only when both shifts in point estimate and a narrowing of the uncertainty bounds occurred simultaneously, which may be a common occurrence for meta-analyses of MD studies. In our study, including bias-adjusted data from OSs may not have influenced a decision based exclusively on statistically significant treatment effects compared to restricting evidence from RCTs only. As bias adjustments using expert elicitation increased the uncertainty because of OS bias, this counterbalanced the gain in statistical power of adding large registry studies. However, if the objective is to identify the best point estimate of treatment effect to inform cost-effectiveness modelling, then a shift in effect size may influence the ultimate results yielded from calculating cost-effectiveness ratios.

Expert bias elicitation is a potentially attractive approach, because bias assessment is a common step in each systematic review of clinical effectiveness. HTA assessors are familiar with concepts of bias assessment such as applied in the Cochrane risk of bias tools (Sterne *et al.* 2014). However, for the purposes of adjusting for biases in statistical modelling analysts need to go beyond qualitative grading (i.e. low or high risk) and quantify the numerical magnitude of the biases. We demonstrated the use of a modified method of bias modelling in evidence synthesis using both Bayesian and standard frequentist meta-analysis. This exercise illustrates how OS data can be included in a meta-analysis, thereby incorporating a larger spectrum of clinically relevant healthcare data.. However, to ensure internal validity, the data from OSs should be sufficiently controlled for confounding or selection bias through the study design and/or primary analysis, or—if this is not possible—bias adjustment post-hoc should be applied to the results to minimize internal biases. As external validity always depends on the context of the decision maker and setting, bias adjustments for external validity should always be considered separately.

As the method of bias-adjustment elicitation inherently relies on judgment, expert opinion may be considered unreliable and overly subjective. To maximize the reliability of the expert opinion, we elicited biases from a group of clinical and methodological experts using a formal approach. Although some judgments on biases differed between the assessors, these differences were generally small, particularly related to the width of the 95% CIs around the RR of revision, which may potentially reflect the different levels of uncertainty among assessors about the effect of the biases.

We modified a method of bias modelling in evidence synthesis developed by Turner and colleagues that allows the meta-analysis of RCT and OS to be adjusted for potential biases formally elicited from experts (Turner *et al.*, 2009). In particular, we implemented some important adaptations to the original elicitation methodology described by Turner *et al.* (2009). We requested each assessor to provide a single bias-adjusted estimate, accounting for the internal (or external) biases of that study. We believe that this adaptation has a number of advantages. First, we were able to reduce the burden on the assessors. For example, providing a bias-adjusted treatment effect estimate for each study for each element of bias, each assessor would have been required to perform approximately 30 elicitations (i.e. 5 studies × 6 dimensions of bias). Second, disentangling the impact of each specific bias on the treatment effect is difficult. We gathered feedback from the methodologists and found that quantifying the impact of each single bias on the treatment effect would have been even more difficult. In the original method (Turner *et al.* 2009), an assumption of independence between each specific bias was made; however, this assumption was not verified and may not be the case in practice. Third, although our approach remains cognitively demanding and challenging for methodological and clinical assessors, we believe that it was more intuitive for them to quantify the importance of bias in terms of the impact of treatment effects, rather than trying to quantify the bias *per se*. Fourth, eliciting an aggregate bias weight per study is in line with the way in which expert judgement is currently used in the context of HTA processes, such as those from the NICE in England and Wales. As part of NICE's deliberative processes, experts are occasionally invited to provide judgement on the direction and potential magnitude of biases associated with published estimates of treatment effects. We acknowledge that this more simplified bias-adjustment approach may be less precise than if all weights were elicited separately. However, we believe a simplifying approach was necessary because of the significant investment required to provide training and to support the learning process of experts to individually elicit different types of bias.

Although our experts agreed that identifying bias weights was conceptually or qualitatively feasible, quantifying the magnitude was quite challenging. Future evidence synthesis studies aiming to bias adjust estimates of treatment effects should: (i) be appropriately resourced to enable individual elicitation of bias adjusting weights for each study and bias type, and (ii) acknowledge the need to fund and develop appropriate training tools to support clinical expert's in the process of drawing forth quantitative expressions (i.e. eliciting judgment) of unobservable quantities.

The key strength of this study was that we sought to apply a previously defined bias elicitation framework to the meta-analysis of randomized experimental and observational evidence. The study also has a number of limitations. First, we purposively sampled our assessors and the sample size was small. We therefore need to be cautious about the generalizability of both bias-adjusted meta-analyses and the experiences of performing this elicitation exercise. Second, while we used an existing framework for study bias, we may not have fully captured all the relevant biases. We observed heterogeneity among studies following bias adjustments that remained approximately 40% in REM meta-analysis. In other words, the variance parameter τ^2, representing unexplained between-study heterogeneity, was non-zero, suggesting that there could be other biases that remain unaddressed. Third, many of the THR studies included in this exercise were relatively old, prior to the publication of CONSORT (Schulz *et al.*, 2010) and STROBE (von Elm *et al.*, 2007) frameworks for the reporting of RCT and OS, respectively. Therefore, in many cases, there was often insufficient detail of study methodology reported to fully assess the nature and level of bias. Furthermore, time-to-event data would have been the preferred outcome measure, but were only available in a limited number of published studies. Integration of individual patient data from registries may have allowed for fitting empirical survival curves but was outside the scope of the current analysis. Importantly, our study is not meant to inform the clinical effectiveness of cemented vs. uncemented fixation modalities; rather, our study was meant to be a methodological exercise.

This bias-adjustment methodology has potentially wide application that includes all situations where there may be caution to pool studies because of high heterogeneity, differences in study design or methodological quality. Another use would be for situations where a single point estimate is needed (e.g. as input in a cost-effectiveness model for an HTA). An alternative and less subjective approach to bias adjustment could involve using published estimates of the magnitude of bias derived from meta-epidemiological research comparing the influence of different design elements within RCTs on intervention effects (Savovic *et al.*, 2012) as well as between study designs (Sacks *et al.*, 1982; Ioannidis *et al.*, 2001; MacLehose *et al.*, 2000; Deeks *et al.*, 2003). However, RCTs currently have only published quantitative estimates of the bias impact for a small number of internal bias attributes e.g. detection bias in terms of outcome blinding, non-random group assignment after imperfect concealment of allocation sequence—and therefore breaking the randomization leading possibly to confounding-by-indication. There is, to our knowledge, no data that quantify the extent and direction of bias in OSs compared to RCTs in the field of MD meta-epidemiological research. In fact, conducting meta-epidemiological research improves the inherent limitations of expert elicitation as the bias is not directly observable. It is also likely that elicited expert opinion from methodologists was influenced by their knowledge of meta-epidemiological research, because the direction of adjustment on average assumed overestimation of treatment effect in most cases. Furthermore, it is preferable to collect information on known and likely confounders within the registry to directly adjust effect estimates for confounding baseline covariates and time-dependent confounding by adequate

DOI: 10.1002/hec

methods (Cox *et al.* 2009; Johnson *et al.*, 2009), so that as less as possible is left to subjective bias adjustment by expert elicitation.

5. CONCLUSION

Combining sources of evidence from RCTs, OSs and large national registries into a single pooled effect estimate may improve the assessment of clinical effectiveness of MDs and the associated decision uncertainty for policy makers. Our case study provides evidence of the feasibility of using data from OSs to complement RCTs and formalizes the use of expert judgement to bias-adjusted outcomes within the context of standard meta-analysis. We recommend considering the use of expert elicited bias-adjustment methods in a sensitivity analysis, when combining evidence from RCT and OSs. Additional MD case studies are needed to demonstrate the acceptability of this approach in the HTA community across other disease areas.

APPENDIX

Table A1. Classification of hip prostheses

Prostheses are usually classified according to:	
1. Fixation	How the prostheses are attached to host bone. Both components can be cemented into place with polymethylmethacrylate (called cemented), both components can be attached to host bone without cement (known as uncemented) or the femoral stem can be cemented and the acetabulum cementless (called hybrid hip replacements).
2. Bearing surface or articulation	This refers to the composition of the femoral head and the inside of the acetabulum, which are the two parts that move against each other. The femoral head can be ceramic or metal and the acetabular bearing can be metal, ceramic or polyethylene.
3. Femoral head size	Commonly between 22.225 mm and 36 mm

Table A2. Listing of domains of internal bias

Internal bias	Description
Selection bias	Refers to when some eligible participants, or the initial follow-up time of some participants are excluded in a way that leads to the association between intervention and outcome differing from the association that would have been observed in complete follow-up of the ideal trial. It relates, for instance, to generation of a randomised sequence or concealment of allocations prior to assignment.
Performance bias	Performance bias is because of knowledge of the allocated interventions by participants and personnel during the study. It relates to blinding of study participants and personnel from knowledge of which intervention a participant received.
Detection bias	Detection bias is because of knowledge of the allocated interventions by outcome assessors.
Attrition bias	Attrition bias is because of amount, nature or handling of incomplete outcome data. It relates to completeness of outcome data for each main outcome, including attrition and exclusions from the analysis.
Confounding	A confounder is an independent risk factor of the outcome of interest that is associated with the exposure/intervention in the study population and is not an intermediate step in the causal pathway between the exposure/intervention and the outcome.
Other bias	This dimension covers any important concerns about bias not addressed in the other domains, for instance the sponsor for the study.

DOI: 10.1002/hec

Table A3. Listing of external biases

External bias	Description
Population-related	Are the eligibility criteria a proper reflection of the study population? Can study results be generalized beyond the eligibility criteria? (Selection of study population, age, comorbidities, exclusion of patients at risk of complications, proportion of patients who declined randomization)
Intervention-related	Are there differences between the study protocol and routine practice? (Study intervention, timing of treatment, prohibition of certain non-trial treatments, therapeutic or diagnostic advances since trial was performed)
Comparator-related	Are there differences between the study protocol and routine practice? (Appropriateness/relevance of control intervention)
Outcomes-related	Who measured the outcome? Adequacy of the length of follow-up?
Setting-related	Do differences in treatment setting translate into possible differences in treatment effects? Do temporal and geographical differences between study population and target populations translate into a limited generalizability? (Health care system, country, participating centres, treatment setting, treating physicians)

Table A4. Qualitative bias-assessment—most likely (i.e. mode) impact of dimensions of internal bias for each study

Study	Study type	Selection	Performance	Detection	Attrition	Confounding	Other
Wykman (1991)	RCT	?	?	?	0/+	?	?
Reigstad(1993)	RCT	?	?	?	+	?	?
McCombe (2004)	RCT	0	?	?	0	?	?
Bjorgul (2010)	RCT	–	?	?/0	0	?	?
Corten (2011)	RCT	0	0	0	?	?	0
Kim (2011)	RCT	0	?	?	0	?	?
Angadi (2012)	RCT	–	?	?	?	?	?/0
Hailer (2010)	Registry	?	?	?	?/0	?/0	?
Makela (2010)	Registry	0	?	?	?/0	?/0	0
Pennington(2013)	Registry	–	?	?	?	0	?
Clohisy (2001)	Cohort	0	?	0	0	?	0
Kim (2003)	Cohort	+/–	?	?	?	?	?
Kruchans (2004)	Cohort	NA	?	?	?	?	?
Pospula (2008)	Cohort	–	?	?	0	?/–	?
Hartofilakidis (2009)	Cohort	–	?	?	0	?/–	0

0, not causing any bias/not likely to favor either implant; +, likely to favor cemented implant; −, likely to favor uncemented implant; ?, Do not know; RCT, randomized clinical trial; NA, not applicable (each assessor gave a different answer).

Table A5. Qualitative bias-assessment—most likely (i.e. mode) impact of dimensions of external bias for each study

Study	Study type	Population	Intervention	Comparator	Outcome	Setting
Wykman (1991)	RCT	0	–	–	0/+	0
Reigstad(1993)	RCT	0	0	0/+	0	0
McCombe (2004)	RCT	0	0	0	0/+	0
Bjorgul (2010)	RCT	–	0	0	0	0
Corten (2011)	RCT	–	0	0	0	0
Kim (2011)	RCT	0	0	0	0	0
Angadi (2012)	RCT	0/–	0	0	1/–	0
Hailer (2010)	Registry	0	0	0	+	+
Makela (2010)	Registry	0	0	+	0	0/–
Pennington (2013)	Registry	0/–	0	+	+	0
Clohisy (2001)	Cohort	0	0/–	–	–	–
Kim (2003)	Cohort	0	–	–	0	0
Kruckhans (2004)	Cohort	+	+	+	NA	0
Pospula (2008)	Cohort	?	0	–	0	0
Hartofilakidis (2009)	Cohort	–	0	0	0	–

0, not causing any bias/not likely to favor either implant; +, likely to favor cemented implant; −, likely to favor uncemented implant; ?, Do not know; RCT: randomized clinical trial; NA, not applicable (each assessor gave a different answer).

Table A6. Results from five univariate meta-regressions using frequentist random effects models and Bayesian random effects models

Frequentist REM

	Unadjusted		Bias-adjusted	
	RR (95%CI)	*p*-value	RR (95%CI)	*p*-value
	Explained Tau^2		Explained Tau^2	
Female proportion	1.36 (0.24–7.81)	0.709	1.60 (0.34–7.51)	0.526
	Tau^2 = 0.11		Tau^2 = 0.06	
Age mean	0.995 (0.97–1.03)	0.729	1.00 (0.98–1.03)	0.826
	Tau^2 = 0.10		Tau^2 = 0.06	
Study type				
RCTs (reference)	1.00		1.00	
Registries	0.67 (0.40–1.13)	0.122	0.70 (0.46–1.06)	0.087
Cohorts	0.65 (0.32–1.30)	0.201	0.69 (0.38–1.25)	0.201
	Tau^2 = 0.05		Tau^2 = 0.02	
Follow-up (per year)	1.02 (0.96–1.08)	0.541	0.999(0.94–1.06)	0.970
	Tau^2 = 0.10		Tau^2 = 0.06	
Unadjusted RR	0.83(0.43 – 1.58)	0.536	0.87 (0.47–1.58)	0.610
	Tau^2 = 0.10		Tau^2 = 0.06	

Bayesian REM

	Unadjusted		Bias-adjusted	
	RR (95%CI)	DIC	RR (95%CI)	DIC
	Explained Tau^2		Explained Tau^2	
Female proportion	1.02 (0.16–7.61)	40.53	1.31 (0.55–4.66)	17.55
	Tau^2 = 0.69		Tau^2 = 0.15	
Age mean	0.98 (0.93–1.03)	40.75	1.00 (0.98–1.02)	19.27
	Tau^2 = 0.66		Tau^2 = 0.16	
Study type				
RCTs (reference)	1.03 (0.53–2.09)	41.05	1.02 (0.71–1.50)	18.80
Registries	1.09 (0.62–2.19)		0.97 (0.68–1.34)	
Cohorts	Tau^2 = 0.70		Tau^2 = 0.16	
Follow-up (per year)	1.03 (0.92–1.15) 0.70	41.44	1.01 (0.96–1.07)	18.94
	Tau^2 = 0.70		Tau^2 = 0.16	

CI, confidence interval; CrI, credibility interval; Tau2, between study variance; DIC, Deviance Information Criterion.

DOI: 10.1002/hec

Table A7. Results from different bivariate meta-regression models of frequentist random effects models

Variable	Unadjusted RR (95%CI)	*p*-value	Bias-adjusted RR (95%CI)	*p*-value
Female proportion	1.51 (0.22–10.34)	0.647	1.59 (0.31–8.29)	0.551
Mean age (per year)	0.99 (0.96–1.03)	0.660	1.00 (0.97–1.03)	0.911
Constants	0.94 (0.13–7.05)	0.949	0.60 (0.09–3.97)	0.569
Explained Tau^2:	0.13		0.07	
Female proportion	2.12 (0.79 – 5.71)	0.122	2.39 (0.89–6.44)	0.07
Study type				
RCTs	1.00 (reference)		1.00 (reference)	
Registries	0.55 (0.37–0.83)	0.008	0.59 (0.43–0.81)	0.003
Cohorts	0.55 (0.28–1.07)	0.074	0.58 (0.34–0.98)	0.044
Constants	0.79 (0.44–1.40)	0.380	0.80 (0.47–1.38)	0.388
Explained Tau^2:	0.004		0.0	
Female proportion	1.44 (0.22–9.26)	0.680	1.62 (0.32–8.24)	0.531
Follow-up (per year)	1.02 (0.95–1.09)	0.538	0.999(0.94–1.06)	0.965
Constants	0.53 (0.15–1.88)	0.298	0.66 (0.23–1.93)	0.420
Explained Tau^2:	0.13		0.07	
Mean age (per year)	0.99 (0.96–1.03)	0.586	1.00 (0.97–1.03)	0.875
Study type				
RCTs	1.00 (reference)	0.234	1.00 (reference)	
Registries	0.71 (0.39–1.29)	0.206	0.72 (0.45–1.15)	0.150
Cohorts	0.61 (0.27–1.37)		0.73 (0.36–1.46)	0.334
Constants	1.74 (0.19–16.05)	0.593	0.94 (0.14–6.31)	0.940
Explained Tau^2:	0.07		0.04	
Mean age (per year)	1.00 (0.96–1.04)	0.963	1.00 (0.97–1.04)	0.811
Follow-up (per year)	1.02 (0.94–1.10)	0.629	1.00 (0.94–1.07)	0.927
Constants	0.69 (0.05–10.06)	0.769	0.66 (0.06–7.69)	0.717
Explained Tau^2:	0.12		0.07	
Study type				
RCTs	1.00 (reference)		1.00 (reference)	
Registries	0.70 (0.39–1.25)	0.202	0.73 (0.45–1.16)	0.162
Cohorts	0.68 (0.321–1.45)	0.289	0.71 (0.37–1.35)	0.262
Follow-up (per year)	1.01 (0.95–1.08)	0.674	0.99 (0.94–1.05)	0.831
Constants	0.88 (0.41–1.91)	0.727	1.12 (0.59–2.12)	0.695
Explained Tau^2:	0.07		0.04	

95%CI, 95% confidence interval; Tau^2, between study variance.

Table A8. Results from different bivariate meta-regression models of Bayesian random effects model

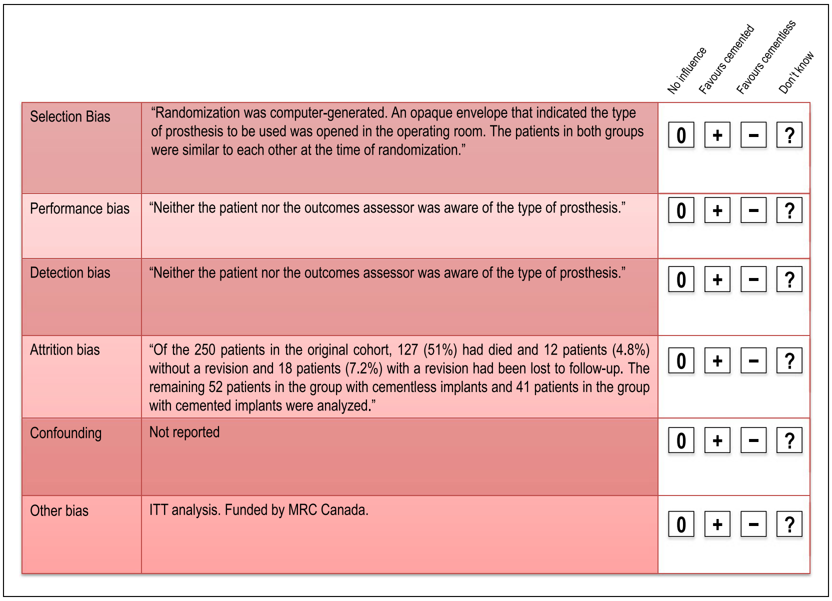
Variable	Unadjusted RR (95%CrI)	DIC	Bias-adjusted RR (95%CrI)	DIC
Female proportion	1.04 (0.47–2.76)	40.81	1.10 (0.65–2.66)	19.28
Mean age (per year)	0.98 (0.93–1.02)		1.00 (0.97–1.02)	
Constants	2.89 (0.15–61.0)		0.75 (0.18–3.42)	
Explained Tau^2:	0.67		0.02	
Female proportion	1.00 (0.40–2.56)	41.15	1.08 (0.66–2.15)	18.87
Study type				
RCTs	1.00 (reference)		1.00 (reference)	
Registries	1.03 (0.55–2.03)		1.02 (0.71–1.47)	
Cohorts	1.08 (0.63–2.12)		0.97 (0.69–1.33)	
Constants	0.66 (0.31–1.31)		0.71 (0.46–1.03)	
Explained Tau^2:	0.70		0.15	
Female proportion	1.02 (0.42–2.65)	41.53	1.12 (0.66–2.80)	18.97
Follow-up (per year)	1.03 (0.93–1.15)		1.01 (0.96–1.07)	
Constants	0.52 (0.15–1.72)		0.62 (0.30–1.17)	
Explained Tau^2:	0.70		0.15	
Mean age (per year)	0.98 (0.93–1.03)	41.25	1.00 (0.97–1.02)	20.21
Study type				
RCTs	1.00 (reference)		1.00 (reference)	
Registries	1.03 (0.67–1.71)		1.02 (0.76–1.40)	
Cohorts	0.99 (0.62–1.57)		0.97 (0.69–1.29)	
Constants	2.95 (0.13–70.0)		0.83 (0.16–4.26)	
Explained Tau^2:	0.67		0.16	
Mean age (per year)	0.98 (0.92–1.03)	42.54	1.00 (0.97–1.03)	21.26
Follow-up (per year)	1.01 (0.90–1.13)		1.01 (0.95–1.07)	
Constants	2.59 (0.06–134)		0.65 (0.09–5.03)	
Explained Tau^2:	0.71		0.17	
Study type		41.96		20.16
RCTs	1.00 (reference)		1.00 (reference)	
Registries	1.01 (0.63–1.65)		1.01 (0.75–1.38)	
Cohorts	1.06 (0.70–1.79)		0.98 (0.73–1.29)	
Follow-up (per year)	1.03 (0.93–1.15)		1.01 (0.96–1.07)	
Constants	0.51 (0.17–1.52)		0.67 (0.39–1.17)	
Explained Tau^2:	0.71		0.16	

CrI, credibility interval; Tau^2, between study variance; DIC, Deviance Information Criterion; RCT, randomized controlled trial; RR, relative risk.

Table A9. Sensitivity analyses for prior distributions in the Bayesian random effects model

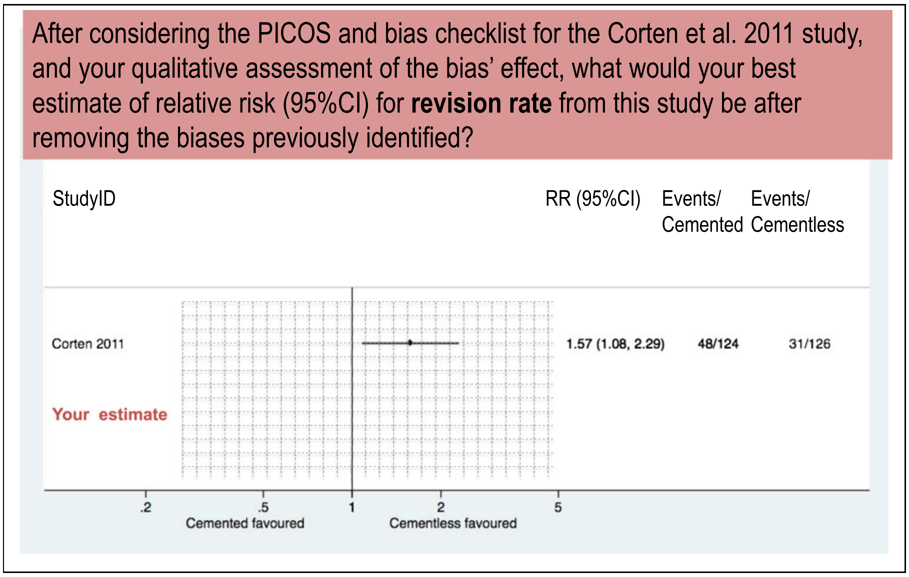
Prior distribution of variance	RR (95%CrI)	RR (95%CrI)	RR (95%CrI)
	RCTs	RCTs and registries	All studies
Uniform (base case)	0.90 (0.37–1.71)	0.80 (0.55–1.17)	0.77 (0.58–1.03)
Gamma	1.11 (0.84–1.40)	0.80 (0.59–1.11)	0.77 (0.60–1.00)
Inverse Gamma	0.99 (0.53–1.51)	0.80 (0.58–1.11)	0.77 (0.60–1.00)
Half-Cauchy	—	—	0.67 (0.28–2.21)

RCT, randomized controlled trial; 95%CrI, 95%credibility interval; Tau^2, between study variance.



		No influence	Favours cemented	Favours cementless	Don't know
Selection Bias	"Randomization was computer-generated. An opaque envelope that indicated the type of prosthesis to be used was opened in the operating room. The patients in both groups were similar to each other at the time of randomization."	0	+	−	?
Performance bias	"Neither the patient nor the outcomes assessor was aware of the type of prosthesis."	0	+	−	?
Detection bias	"Neither the patient nor the outcomes assessor was aware of the type of prosthesis."	0	+	−	?
Attrition bias	"Of the 250 patients in the original cohort, 127 (51%) had died and 12 patients (4.8%) without a revision and 18 patients (7.2%) with a revision had been lost to follow-up. The remaining 52 patients in the group with cementless implants and 41 patients in the group with cemented implants were analyzed."	0	+	−	?
Confounding	Not reported	0	+	−	?
Other bias	ITT analysis. Funded by MRC Canada.	0	+	−	?

Figure A1. An example of a qualitative assessment tool



After considering the PICOS and bias checklist for the Corten et al. 2011 study, and your qualitative assessment of the bias' effect, what would your best estimate of relative risk (95%CI) for **revision rate** from this study be after removing the biases previously identified?

StudyID	RR (95%CI)	Events/ Cemented	Events/ Cementless
Corten 2011	1.57 (1.08, 2.29)	48/124	31/126
Your estimate			

Figure A2. An example of a quantitative assessment too

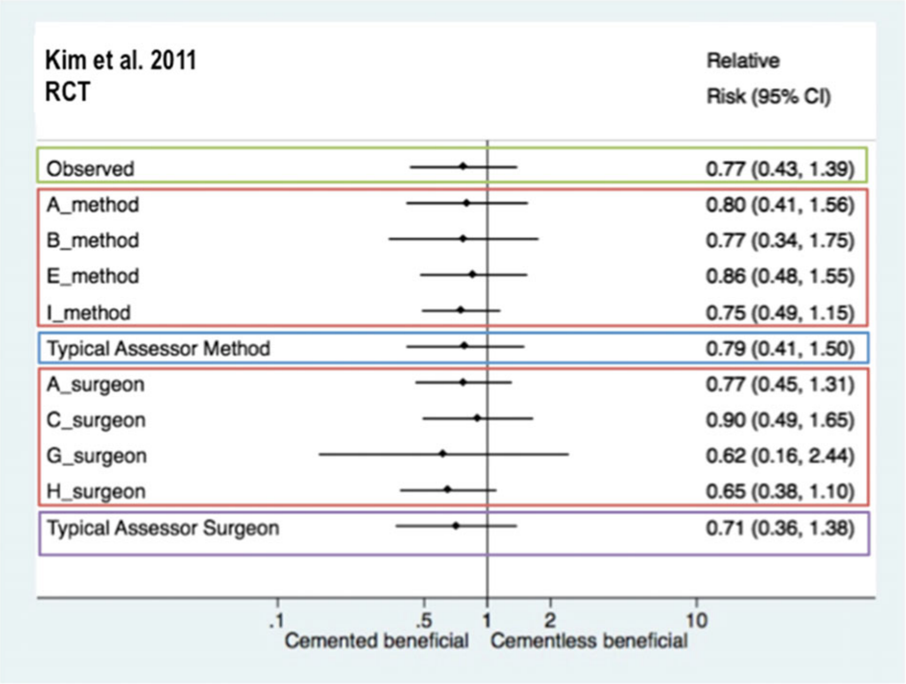


Figure A3. Elicitation of bias-adjusted treatment effects

CONFLICTS OF INTEREST

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement HEALTH-F3-2012-305694 (Project MedtecHTA 'Methods for Health Technology Assessment of Medical Devices: a European Perspective'). All authors declare no conflict of interest.

ETHICAL STATEMENT

This study did not involve patients.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement HEALTH-F3-2012-305694 (Project MedtecHTA 'Methods for Health Technology Assessment of Medical Devices: a European Perspective'). We are extremely grateful to the clinical and methodological experts who took the time from their busy schedules to participate in the study. Without their support, we would have never achieved the results presented in this manuscript. All errors are ours.

REFERENCES

Angadi DS, Brown S, Crawfurd EJ. 2012. Cemented polyethylene and cementless porous-coated acetabular components have similar outcomes at a mean of seven years after total hip replacement: a prospective randomised study. *Journal of Bone and Joint Surgery (British)* **94**: 1604–10.

DOI: 10.1002/hec

Bernard A, Vaneau M, Fournel I, Galmiche H, Nony P, Dubernard JM. 2014. Methodological choices for the clinical development of medical devices. *Medical Devices (Auckl)* **7**: 325–34.

Bjorgul K, Novicoff WM, Andersen ST, Brevig K, Thu F, Wiig M, Ahlund O. 2010. No differences in outcomes between cemented and uncemented acetabular components after 12–14 years: results from a randomized controlled trial comparing Duraloc with Charnley cups. *Journal of Orthopaedics and Traumatology* **11**: 37–45.

Clarke A, Pulikottil-Jacob R, Grove A, Freeman K, Mistry H, Tsertsvadze A, Connock M, Court R, Kandala N-B, Costa M, Suri G, Metcalfe D, Crowther M, Morrow S, Johnson S, Sutcliffe P. 2013. Total hip replacement and surface replacement for the treatment of pain and disability resulting from end stage arthritis of the hip (Review of technology appraisal guidance 2 and 44). *Warwick Evidence* **19**: 1–668.

Clement ND, Biant LC, Breusch SJ. 2012. Total hip arthroplasty: to cement or not to cement the acetabular socket? A critical review of the literature. *Archives of Orthopaedic and Trauma Surgery* **132**: 411–27.

Clohisy JC, Harris WH. 2001. Matched-pair analysis of cemented and cementless acetabular reconstruction in primary total hip arthroplasty. *Journal of Arthroplasty* **16**: 697–705.

Corten K, Bourne RB, Charron KD, Au K, Rorabeck CH. 2011. Comparison of total hip arthroplasty performed with and without cement: a randomized trial. A concise follow-up, at twenty years, of previous reports. *Journal of Bone and Joint Surgery (American)* **93**: 1335–8.

Cox E, Martin B, Van Staa T, Garbe E, Siebert U, Johnson ML. 2009. Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: The International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report-Part II. *Value in Health* **12**: 1053–1061.

Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, Petticrew M, Altman DG, International Stroke Trial Collaborative Group, European Carotid Surgery Trial Collaborative Group. 2003. Evaluating non-randomised intervention studies. *Health Technology Assessment* **7**: 1–173.

Dekkers OM, Von Elm E, Algra A, Romijn JA, Vandenbroucke JP. 2010. How to assess the external validity of therapeutic trials: A conceptual approach. *International Journal of Epidemiology* **39**: 89–94.

Faulkner A, Kennedy LG, Baxter K, Donovan J, Wilkinson M, Bevan G. 1998. Effectiveness of hip prostheses in primary total hip replacement: a critical review of evidence and an economic model. *Health Technology Assessment* **2**: 1–133.

Fitzpatrick R, Shortall E, Sculpher M, Murray D, Morris R, Lodge M, Dawson J, Carr A, Britton A, Briggs A. 1998. Primary total hip replacement surgery: a systematic review of outcomes and modelling of cost-effectiveness associated with different prostheses. *Health Technology Assessment* **2**: 1–64.

Froud R, Underwood M, Carnes D, Eldridge S. 2012. Clinicians' perceptions of reporting methods for back pain trials: A qualitative study. *The British journal of general practice: the journal of the Royal College of General Practitioners* **62**: e151–9.

Gelman A. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**: 515–533.

Hailer NP, Garellick G, Karrholm J. 2010. Uncemented and cemented primary total hip arthroplasty in the Swedish Hip Arthroplasty Register. *Acta Orthopaedica* **81**: 34–41.

Hartofilakidis G, Georgiades G, Babis GC. 2009. A comparison of the outcome of cemented all-polyethylene and cementless metal-backed acetabular sockets in primary total hip arthroplasty. *Journal of Arthroplasty* **24**: 217–25.

Higgins J, Green S. 2011. *Cochrane handbook for systematic reviews of interventions version 5.1.0* [Online]. (Available: www.cochrane-handbook.org, accessed on 21 January 2015).

Iglesias C. 2015. Does assessing the value for money of therapeutic medical devices require a flexible approach? *Expert Review of Pharmacoeconomics & Outcomes Research* **15**: 21–32.

Ioannidis JPA, Haidich A-B, Pappa M, Kokori SI, Tektonidou MG, Contopoulos-Ioannidis DG, Lau J. 2001. Comparison of evidence of treatment effects in randomised and nonrandomised studies. *Journal of the American Medical Association* **286**: 821–830.

Johnson ML, Crown W, Martin BC, Dormuth CR, Siebert U. 2009. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: The ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report—Part III. *Value in Health* **12**: 1062–1073.

Kim YH, Kim JS, Park JW, Joo JH. 2011. Comparison of total hip replacement with and without cement in patients younger than 50 years of age: the results at 18 years. *Journal of Bone and Joint Surgery (British)* **93**: 449–55.

Kim YH, Oh SH, Kim JS, Lee SH. 2003. Total hip arthroplasty for the treatment of osseous ankylosed hips. *Clinical Orthopaedics and Related Research* **414**: 136–48.

Konstam MA, Pina I, Lindenfeld J, Packer M. 2003. A device is not a drug. *Journal of Cardiac Failure* **9**: 155–7.

Kruckhans AR, Dustmann HO. 2004. Indications, methods, and results of cemented, hybrid, and cement-free implantation of THR. *Surgical Technology International* **12**: 253–7.

Lunn DJ, Thomas A, Best N, Spiegelhalter DJ. 2000. WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10**: 325–337.

Maclehose RR, Reeves B, Harvey IM, Sheldon TA, Russell IT, Black AM. 2000. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technology Assessment* **4**: 1–154.

Makela KT, Eskelinen A, Pulkkinen P, Virolainen P, Paavolainen P, Remes V. 2011. Cemented versus cementless total hip replacements in patients fifty-five years of age or older with rheumatoid arthritis. *Journal of Bone and Joint Surgery (American)* **93**: 178–86.

Mccombe P, Williams SA. 2004. A comparison of polyethylene wear rates between cemented and cementless cups. A prospective, randomised trial. *Journal of Bone and Joint Surgery (British)* **86**: 344–9.

National Institute for Health and Care Excellence (Nice) 2013. Guide to the methods of technology appraisal 2013. National Institute for Health and Care Excellence (NICE).

Pakvis D, Van Hellemondt G, De Visser E, Jacobs W, Spruit M. 2011. Is there evidence for a superior method of socket fixation in hip arthroplasty? A systematic review. *International Orthopaedics* **35**: 1109–18.

Panteli D, Nolting A, Eckhardt H, Kulig M, Busse R. 2016. Published and unpublished evidence in coverage decision-making for pharmaceuticals in Europe: existing approaches and way forward. *Health research policy and systems/BioMed Central* **14**: 6.

Pennington M, Grieve R, Black N, Van Der Meulen JH. 2013. Functional outcome, revision rates and mortality after primary total hip replacement—a national comparison of nine prosthesis brands in England. *PLoS One* **8**: e73228.

Pospula W, Abu Noor T, Roshdy T, Al Mukaimi A. 2008. Cemented and cementless total hip replacement. Critical analysis and comparison of clinical and radiological results of 182 cases operated in Al Razi Hospital, Kuwait. *Medical Principles and Practice* **17**: 239–43.

Reigstad A, Rokkum M, Bye K, Brandt M. 1993. Femoral remodeling after arthroplasty of the hip. Prospective randomized 5-year comparison of 120 cemented/uncemented cases of arthrosis. *Acta Orthopaedica Scandinavica* **64**: 411–6.

Rothwell PM. 2010. Commentary: External validity of results of randomized trials: disentangling a complex concept. *International Journal of Epidemiology* **39**: 94–6.

Royal Netherlands Academy of Arts and Sciences (KNAW). 2014. Evaluation of new technology in health care. In *Need of Guidance for Relevant Evidence*, Royal Netherlands Academy of Arts and Sciences: Amsterdam.

Sacks HS, Chalmers TC, Smith H. 1982. Randomized versus historical controls for clinical trials. *American Journal of Medicine* **72**: 233–240.

Savovic J, Jones H, Altman D, Harris R, Juni P, Pildal J, als-nielsen B, Balk E, Gluud C, Gluud L, Ioannidis J, Schulz K, Beynon R, Welton N, Wood L, Moher D, Deeks J, Sterne J. 2012. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. *Health Technology Assessment* **16**: 1–82.

Schnell-Inderst P, Mayer J, Lauterberg J, Hunger T, Arvandi M, Conrads-Frank A, Nachtnebel A, Wild C, Siebert U. 2015. Health technology assessment of medical devices: what is different? An overview of three European projects. *Z. Evid. Fortbild. Qual. Gesundh. wesen* **109**: 309–318.

Schulenburg JM, Mittendorf T, Kulp W, Greiner W. 2009. Health Technology Assessment (HTA) im Bereich der Medizinprodukte—gleiches Spiel mit gleichen Regeln? *Gesundh ökon Qual manag* **14**: 144–155.

Schulz KF, Altman DG, Moher D. 2010. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* **340**: c332.

Siebert M, Clauss LC, Carlisle M, Casteels B, De Jong P, Kreuzer M, Sanghera S, Stokoe G, Trueman P, Lang AW. 2002. Health technology assessment for medical devices in Europe. What must be considered. *International Journal of Technology Assessment in Health Care* **18**: 733–40.

Sterne, J, Higgins, J, Reeves, B. & Acrobat-Nrsi, O. B. O. T.D. G. F. 2014. A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions (ACROBAT-NRSI). *Version 1.0.0* [Online]. [Accessed 16 March 2015].

Tsertsvadze A, Grove A, Freeman K, Court R, Johnson S, Connock M, Clarke A, Sutcliffe P. 2014. Total hip replacement for the treatment of end stage arthritis of the hip: a systematic review and meta-analysis. *PLoS One* **9**: e99804.

Turner RM, Spiegelhalter DJ, Smith GC, Thompson SG. 2009. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society. Series A, Statistics in Society* **172**: 21–47.

Vale L, Wyness L, Mccormack K, Mckenzie L, Brazzelli M, Stearns SC. 2002. A systematic review of the effectiveness and cost-effectiveness of metal-on-metal hip resurfacing arthroplasty for treatment of hip disease. *Health Technology Assessment* **6**: 1–109.

Verde PE, Ohmann C. 2014. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. *Research Synthesis Methods* **6**: 45–62.

Voigt JD, Mosier MC. 2012. Cemented all-polyethylene acetabular implants vs other forms of acetabular fixation: a systematic review and meta-analysis of randomized controlled trials. *Journal of Arthroplasty* **27**: 1544–1553. e10.

Von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. 2007. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Medicine* **4**: e296.

Welton NJ, Ades AE. 2009. Models for potentially biased evidence in meta-analysis using empirically based priors. *Journal of the Royal Statistical Society A* 172.

Welton NJ, Sutton AJ, Cooper NJ, Abrams KR. 2012. *Evidence Synthesis for Decision Making in Health Care*, Wiley: Chichester.

Wykman A, Olsson E, Axdorph G, Goldie I. 1991. Total hip arthroplasty. A comparison between cemented and press-fit noncemented fixation. *Journal of Arthroplasty* **6**: 19–29.