INTERNATIONAL JOURNAL OF SURGERY

Evidence-based objective performance criteria for the evaluation of hip and knee replacement devices and technologies

Marc J. Nieuwenhuijse, BEng, MEpi, MD, PhD[a,c,*], Per-Henrik Randsborg, MD, PhD[b], Jensen H. Hyde, MD PhD[d], Wenna Xi, PhD[c], Patricia Franklin, MS, PhD[e], Limin Sun, MS, PhD[f,g], Xinyan Zheng, MS[c], Samprit Banerjee, PhD[c], Jialin Mao, MD, MS[c], Suvekshya Aryal, MPH[c], Priscilla Chan, MD[i], Amanda Chen, MS PhD[c], Alexander Liebeskind, BS[c], Pablo Bonangelino, PhD[f,g], Paul Voorhorst, MS, MBA[h], Laura E. Gressler, MS, PhD[f,g], Vincent Devlin, MD[f,g], Raquel Peat, PhD, MPH[f,g], Danica Marinac-Dabic, MD, PhD[f,g], Elizabeth Paxton, PhD[i], Art Sedrakyan, MD, PhD[c]

Background: Objective performance criteria (OPC) is a novel method to provide minimum performance standards and improve the regulated introduction of original or incremental device innovations in order to prevent patients from being exposed to potentially inferior designs whilst allowing timely access to improvements. We developed 2-year safety and effectiveness OPC for total hip and knee replacement (THR and TKR).

Methods: Analyses of large databases were conducted using various data sources: a systematic literature review; a direct data analysis from The Functional Outcomes Research for Comparative Effectiveness in Total Joint Replacement and Quality Improvement Registry (FORCE-TJR) and the Kaiser Permanente Implant Registry (KPIR); and claims data analyses from longitudinal discharge data in New York and California states. The literature review included U.S. patients ($\geq$ 18 years) who received THR or TKR for primary end-stage osteoarthritis and prospectively collected data on patient-reported outcome measures (PROMs) from at least 100 subjects and/or 2-year implant survival for at least 250 implants. Random effects models were used for meta-analysis.

Results: Data were available from a total of 951 100 patients. After screening of 7979 abstracts, 294 studies underwent full-text review and 31 studies contributed to the evidence synthesis (333 995 implants). Direct data analysis of FORCE-TJR contributed 9223 joint replacement patients to the construction of OPC for effectiveness; KPIR contributed 262 044 patients for the construction of OPC for safety. Claims database analysis contributed 345 838 patients to the construction of safety OPC. OPC for safety were constructed for cumulative incidences of 2-year all-cause and septic revision (THR/TKR 2.0%/1.6% and 0.6%/0.7%), and OPC for effectiveness were constructed based on four disease-specific and three general health-related quality of life PROMs (HOOS/KOOS 87.1/80.6; HSS/KSS function 94.4/90.6; SF-12/SF-36, PCS 46.5/41.9, EQ-5D 0.88/0.84).

Conclusion: This study is the first to construct a 2-year OPC for the safety and effectiveness of THR and TKR based on U.S. real-world data. Based on these OPC, potential benchmarks for (single-arm study) evaluation of new device innovations are suggested for a regulated and safe introduction to the (commercial) market.

Keywords: objective performance criteria, orthopedic surgery, patient-reported outcome measures, real-world evidence, total joint arthroplasty

[a]Department of Orthopedic Surgery, Amphia Hospital, Breda, The Netherlands, [b]Department of Orthopaedic Surgery, Akershus University Hospital, Lørenskog, Norway, [c]Department of Population Health Sciences, Weill Cornell Medical College, New York, New York, [d]Internal Medicine, University of Tennessee, Chattanooga, Tennessee, [e]Department of Medical Social Sciences Northwestern University Feinberg School of Medicine, [f]Orthopedics Outcomes Research, FORCE-TJR, Chicago, Illinois, [g]Center for Devices and Radiological Health (CDRH), FDA, Silver Spring, Maryland, [h]Worldwide Clinical Research, DePuy Synthes Companies, a Johnson & Johnson Company, Fort Wayne, Indiana and [i]Surgical Outcomes and Analysis, Kaiser Permanente, San Diego, California, USA

The lead authors, Marc J. Nieuwenhuijse and Art Sedrakyan, affirm that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

Corresponding author. Address: Department of Population Health Sciences, Weill Cornell Medical College, 570 Lexington Ave, 10th Floor, New York, New York 10022, USA. Tel.: +1 646-962-8001. E-mail address: marcusjnieuwenhuijse@gmail.com (M.J. Nieuwenhuijse).

International Journal of Surgery (2023) 109:1125–1135

Received 21 August 2022; Accepted 28 December 2022

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.journal-surgery.net.

Published online 10 April 2023

http://dx.doi.org/10.1097/JS9.0000000000000169

Introduction

Symptomatic osteoarthritis (OA) affects over 10% of the United States (U.S.) population[1,2] and significantly contributes to health-related disability and expenditures[3]. Currently, over one million total hip and knee replacements (THR and TKR) are performed annually in the U.S. for primary end-stage OA. This number is expected to exponentially increase to ~3 million procedures annually by 2030[4]. These surgical procedures involve the implantation of regulated devices and the performance of the devices is of great interest to the patient, orthopedic, and public health as well as from the commercial and regulatory perspective.

There is a continuous strive to improve implant performance. Innovation is often incremental, aiming to improve patient outcomes by continuously modifying the design of existing devices but sometimes by introducing new implant concepts. It is essential to balance the desire to introduce novel and potentially improved implant designs while continuously ensuring their safety and effectiveness. Even incremental device innovations, like changes in surface finish, bone cement viscosity, or articulation surface (e.g. the recent metal-on-metal articulation), can have catastrophic results (i.e. large-scale early failure) and a robust and reliable system for the regulated introduction of device innovation is required[5-8].

Most innovations of joint replacement devices are introduced by premarket notification through the 510(k) program, which requires sponsors to demonstrate that their medical device is substantially equivalent to another similarly legally marketed device – a so-called predicate device – in terms of intended use, technological characteristics, and performance testing, as needed. Although clinical data are not typically included in 510(k)s, clinical data may be requested by the U.S. Food and Drug Administration (FDA) when nonclinical performance data are not adequate to support a substantial equivalence determination[9]. However, this system is not without its weaknesses[10], and it is crucial to continuously assess device performance throughout the total product life-cycle (TPLC) to avoid the adoption of sub-optimal technologies.

HIGHLIGHTS

- We developed evidence-based guidelines for hip and knee devices and replacement for 2-year evaluations using literature, registries, and administrative databases. The developed estimates include both device revision estimates as well as patient-reported outcome measures (PROMs)
- These evidence-based guidelines are intended to use in clinical trials and real-world evidence-based evaluations by scientists, manufacturers, payors, and regulators.
- As objective performance criteria (OPC) evolves, stakeholders globally can define margins for superiority and noninferiority assessments based on their country's requirements and recommendation.

From a TPLC device evaluation perspective, it is crucial to recognize that theoretical improvements in new implant designs do not necessarily translate to patient benefits[11]. Novel and pragmatic methodologies are needed to evaluate devices using less burdensome but robust approaches. Developing and using objective performance criteria (OPC) is one of the approaches that can fit this purpose[12]. OPC are numerical target values derived from clinical studies or real-world data (RWD) and may be used in single-arm studies to evaluate the safety and effectiveness of joint replacement devices[12]. OPC can be generated using clinical studies, registries, and other data sources. In particular, RWD is an attractive option to generate OPC because subsequent evaluations of devices can also be conducted using these data sources[13].

Given the overall maturity of THR and TKR devices and the continuous (incremental) innovation for their potential improvement, generating OPC for these devices is long overdue. The objective of this study was to develop OPC for the assessment of the 2-year safety and effectiveness of THR and TKR devices using U.S. data sources. The OPC can serve as performance targets and aid the evaluation of new implant designs using single-arm studies within a TPLC framework.

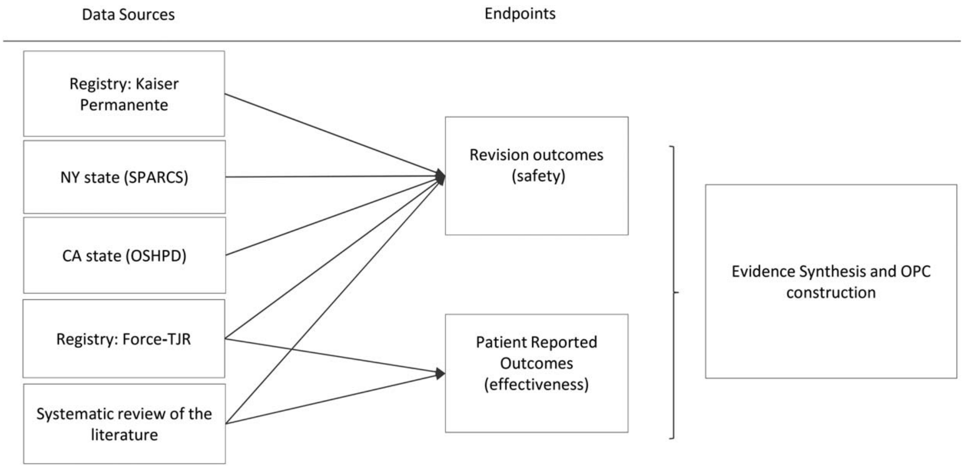
Data Sources
Endpoints
Registry: Kaiser Permanente
NY state (SPARCS)
CA state (OSHPD)
Registry: Force-TJR
Systematic review of the literature
Revision outcomes (safety)
Patient Reported Outcomes (effectiveness)
Evidence Synthesis and OPC construction

Figure 1. Flowchart of data origin. Force-TJR, The Functional Outcomes Research for Comparative Effectiveness in Total Joint Replacement and Quality Improvement Registry; OPC, objective performance criteria; OSHPD, Office of Statewide Health Planning and Development; SPARCS, Statewide Planning and Research Cooperative System.

Methods

Overview of the study design

This study synthesized evidence for the most widely used, well-studied arthroplasties – THR and TKR – for the construction of OPC for safety and effectiveness. Data from three sources were combined: a systematic review of published data, a direct analysis of registry data, and a direct analysis of claims databases (Fig. 1). Meta-analyses were then performed to synthesize evidence from these three data sources and construct OPC for safety and effectiveness endpoints.

Safety

The primary OPC considered for safety was the 2-year all-cause revision rate. This was defined as the removal, replacement, or addition (or alteration) of an implant. All-cause revision is a recognized metric used globally for device benchmarking. Revision due to infection was included as a secondary safety OPC to aid the evaluation of innovations aimed at reducing septic revision surgery.

Effectiveness

OPC considered for effectiveness measures were patient-reported outcome measures (PROMs) at 2-year follow-ups. Disease-specific PROMs used for the construction of OPC were Hip disability and Osteoarthritis Outcome Score (HOOS)[14], Oxford Hip Score (OHS), Knee disability and Osteoarthritis Outcome Score (KOOS)[15], and Oxford Knee Score (OKS)[16]. For general Health-Related Quality of Life (HRQoL), EuroQol-5D (EQ-5D)[17], Short Form 12 (SF-12)[18], and Short-Form 36 (SF-36) were used[19]. Global (population normalized) Physical Health is measured using the SF-36 Physical Component Score (PCS) and global (population normalized) Emotional Health using the SF-36 Mental Component Score (MCS). We also developed OPC for the physician-evaluated Harris Hip Score (HHS)[20] and mixed (physician and patient) outcome measure Knee Society Scores (KSS)[21].

(1) Systematic review of available literature

A systematic review was carried out according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines[22]. PubMed, MEDLINE, EMBASE, Web of Science, Cochrane Library, CINAHL, and Academic Search Premier were searched from January 2010 through January 2020 (Appendix 1, Supplemental Digital Content 3, http://links.lww.com/JS9/A255). Bibliographies of trials and reviews were cross-referenced for additional studies. We also reviewed annual reports from major institutional registries in the U.S., like Mayo Clinic, Michigan Arthroplasty Registry Collaborative Quality Initiative (MARCQI), and American Joint Replacement Registries (AJRR) that have prospective cohort series for THR and TKR survival estimates.

The following selection criteria were used to identify eligible studies:

(a) Cohort studies reporting on prospectively acquired data (including registry studies) on THR and TKR irrespective of design, excluding THR with metal-on-metal articulation.
(b) Study population aged at least 18 years, at least 90% of whom had a diagnosis of symptomatic primary OA as an indication for surgery.
(c) Reporting of 2-year PROMs, HRQoL, or physician-evaluated function measures in at least 100 subjects and/or implant survival for at least 250 implants.

Evaluations were limited to the U.S. population to construct representative and contemporary OPC, and study results had to be published in the last decade (2010 to January 2020). Search results were independently evaluated by at least two assessors (M.J.N., P.-H.R., A.C., J.H., S.A., A.L.). In case of disagreement, the consensus was reached by the referee (A.S., M.J.N.). After inclusion, information was independently extracted by two abstractors and assessed for agreement. Information on study design, study quality, setting, time period, number of implants, demographics, follow-up length and completion, and manufacturer were extracted.

(2) Registries

The Functional Outcomes Research for Comparative Effectiveness in Total Joint Replacement and Quality Improvement Registry (Force-TJR) and Kaiser Permanente Implant Registries (KPIR) were utilized to provide RWD for the construction of THR and TKR OPC (Supplementary File 1, Supplemental Digital Content 1, http://links.lww.com/JS9/A253). Patients at least 18 years who underwent primary elective THR or TKR for OA from 2011 to 2017 were identified in the KPIR and included in our analysis.

Table 1

Revision estimates stratified by age and gender in hip and knee arthroplasty: Kaiser Permanente registry.

	Baseline *N*	Two-year overall revision rate	Two-year septic revision rate
Total hip arthroplasty			
Total	88 126	2.1% (2.0–2.2%)	0.6% (0.6–0.7%)
Age < 55	11 028		
Male, age <55	5635	1.8% (1.5–2.2%)	0.6% (0.4–0.8%)
Female, age <55	5393	2.0% (1.7–2.4%)	0.5% (0.3–0.7%)
Age 55–64	26 152		
Male, age 55–64	11 947	1.8% (1.5–2.0%)	0.7% (0.5–0.8%)
Female, age 55–64	14 205	2.1% (1.9–2.3%)	0.6% (0.5–0.7%)
Age 65 +	50 946		
Male, age 65 +	18 969	2.2% (2.0–2.4%)	0.8% (0.7–1.0%)
Female, age 65 +	31 977	2.1% (1.9–2.3%)	0.5% (0.4–0.6%)
Total knee arthroplasty			
Total	17 3918	1.7% (1.7–1.8%)	0.8% (0.8–0.8%)
Age <55	13 076		
Male, age <55	5054	4.2% (3.6–4.7%)	1.6% (1.2–1.9%)
Female, age <55	8022	2.7% (2.4–3.1%)	0.7% (0.5–0.9%)
Age 55–64	52 205		
Male, age 55–64	20 455	2.3% (2.1–2.5%)	1.1% (0.9–1.2%)
Female, age 55–64	31 750	1.6% (1.5–1.8%)	0.6% (0.5–0.7%)
Age 65 +	108 637		
Male, age 65 +	41 352	1.8% (1.7–2.0%)	1.1% (1.0–1.2%)
Female, age 65 +	67 285	1.2% (1.1–1.3%)	0.6% (0.5–0.6%)

Table 2

Two-year patient-reported outcomes in hip and knee arthroplasty: FORCE-TJR registry.

	Baseline *N*	Mean (SD)	Two-year *N*	Mean (SD)	Two-year change *N*	Mean (SD)
Total hip arthroplasty						
SF-36 MCS	3875	51.92 (11.87)	3875	54.03 (9.18)	3875	2.11 (10.63)
SF-36 PCS	3874	32.12 (8.56)	3874	46.41 (10.2)	3874	14.29 (10.57)
HOOS Score	3845	41.55 (16.71)	3845	87.09 (14.55)	3845	45.54 (19.02)
HOOS Pain Score	3980	44.22 (17.93)	3980	90.94 (14.28)	3980	46.72 (20.45)
HOOS ADL Score	3910	46.94 (18.91)	3910	88.89 (15.22)	3910	41.95 (20.61)
HOOS QoL Score	3897	26.64 (18.35)	3897	80.3 (20.85)	3897	53.67 (24.73)
Total knee arthroplasty						
SF-36 MCS	5393	53.23 (11.45)	5393	54.03 (9.42)	5393	0.8 (10.17)
SF-36 PCS	5393	33.67 (8.29)	5393	44.2 (10.03)	5393	10.53 (9.84)
KOOS Score	5378	45.58 (15.18)	5378	80.63 (16.03)	5378	35.05 (18.25)
KOOS Pain Score	5623	48.14 (17.19)	5623	86.8 (16.35)	5623	38.66 (20.41)
KOOS ADL Score	5515	54.74 (18.1)	5515	86.04 (16.25)	5515	31.29 (19.69)
KOOS QoL Score	5481	27.41 (17.87)	5481	71.9 (23.57)	5481	44.49 (25.81)

ADL, activities of daily living; FORCE-TJR, The Functional Outcomes Research for Comparative Effectiveness in Total Joint Replacement and Quality Improvement Registry; HOOS, Hip disability and Osteoarthritis Outcome Score; KOOS, Knee disability and Osteoarthritis Outcome Score; MCS, Mental Component Score; PCS, Physical Component Score; QoL, quality of life; SF-36, Short-Form 36.

(2) Claims data

The New York State Department of Health Statewide Planning and Research Cooperative System (SPARCS) and the California Office of Statewide Health Planning and Development (OSHPD) provided claims data (Supplementary File 1, Supplemental Digital Content 1, http://links.lww.com/JS9/A253). Patients at least 18 years who underwent THR or TKR in New York (January 2016 to December 2018) or California (October 2015 to December 2017) with primary OA were included.

Statistical analyses

Meta-analyses of the literature

Study heterogeneity was evaluated by extracting data on study design, demographics, relevant devices, and outcome definitions. The number of patients, mean and standard deviation (SD) of PROM scores, number of revision events, and estimated implant survivals and 95% confidence intervals (CIs) were extracted. When not reported, the SD of the performance measure was imputed using a pooled estimate obtained from studies that

Table 3

Two-year revision estimates stratified by age and gender in hip and knee arthroplasty: New York and California statewide claims data.

	New York			California		
	Total *N*	Two-year revision rate (95% CI)	Septic revision rate (95% CI)	Total *N*	Two-year revision rate (95% CI)	Septic revision rate (95% CI)
Total hip arthroplasty						
Total	65 109	2.1% (1.9–2.2%)	0.5% (0.4–0.5%)	67 717	2.6% (2.4–2.7%)	0.5% (0.5–0.6%)
< 55 age						
< 55 Males	5412	1.6% (1.3–2.0%)	0.4% (0.3–0.7%)	4675	2.3% (1.8–2.9%)	0.7% (0.5–1.0%)
< 55 Females	4836	2.4% (1.9–2.9%)	0.6% (0.4–0.9%)	3888	3.4% (2.8–4.2%)	0.7% (0.4–1.0%)
55–64 age						
55–64 Males	10 248	1.7% (1.4–2.0%)	0.5% (0.4–0.7%)	9729	2.4% (2.1–2.9%)	0.6% (0.5–0.9%)
55–64 Females	9923	2.0% (1.7–2.3%)	0.5% (0.3–0.6%)	9941	2.5% (2.2–2.9%)	0.4% (0.2–0.5%)
≥ 65 age						
≥ 65 Males	13 347	2.1% (1.9–2.4%)	0.6% (0.4–0.7%)	15 413	2.3% (2.0–2.5%)	0.6% (0.5–0.8%)
≥ 65 Females	21 343	2.3% (2.1–2.5%)	0.4% (0.3–0.5%)	24 071	2.7% (2.5–3.0%)	0.4% (0.3–0.5%)
Total knee arthroplasty						
Total	96 453	1.8% (1.7–1.9%)	0.5% (0.5%-0.6%)	116 559	2.1% (1.9–2.2%)	0.6% (0.5–0.6%)
< 55 age						
< 55 Males	3836	3.6% (2.9–4.4%)	0.9% (0.6–1.3%)	3707	4.4% (3.6–5.4%)	1.3% (1.0–1.8%)
< 55 Females	6547	2.9% (2.4–3.4%)	0.6% (0.4–0.8%)	5320	3.7% (3.0–4.5%)	0.8% (0.5–1.2%)
55–64 age						
55–64 Males	11 342	2.5% (2.2–2.9%)	0.9% (0.7–1.2%)	13 353	2.2% (1.9–2.6%)	0.6% (0.5–0.8%)
55–64 Females	17 927	2.0% (1.8–2.3%)	0.4% (0.3–0.5%)	18 910	2.2% (1.9–2.5%)	0.5% (0.4–0.6%)
≥ 65 age						
≥ 65 Males	19 664	1.6% (1.4–1.8%)	0.6% (0.5–0.7%)	28 319	2.0% (1.8–2.2%)	0.7% (0.6–0.9%)
≥ 65 Females	37 137	1.2% (1.1–1.3%)	0.3% (0.3–0.4%)	46 950	1.6% (1.4–1.8%)	0.4% (0.3–0.5%)

reported SD. The 2-year all-cause revision was reported either as a proportion or as a Kaplan–Meier estimate. Normal approximation to binomial proportions was used to calculate the SD.

Registry and claims data analysis

Kaplan–Meier statistics were used to determine 2-year all-cause and septic revision. Patients were censored at death or the end of follow-up, whichever occurred first. For the septic revision endpoint, patients were additionally censored at revisions, not due to infection.

OPC construction

Point estimates of OPC were derived from performance estimates pooled from all three data sources (literature, registries, and claims databases) using a random-effects meta-analysis model. The model was constructed of fixed effects in a linear mixed effects model that included a random effect to capture between-study heterogeneity. Limits of 95% CI and sample SD were presented alongside the synthesized point estimate for revision rate and PROMs, respectively. Statistical heterogeneity was estimated using the I^2 statistic. All meta-analysis models were stratified by study type (literature, registry, or claims) to explore the source of heterogeneity. For revision outcomes, two additional stratified random-effects models were constructed to examine the effect of censoring (operationalized as censoring rate unknown, $>5\%$ or $\leq 5\%$) and type of revision probability estimates (Kaplan–Meier estimate or naïve proportions) on between-study heterogeneity. Sensitivity analyses assessed the influence of each individual study or data source on the pooled estimate using leave-one-out diagnostics. Publication bias was examined using funnel plots and the Begg–Mazumdar rank correlation test.

Finally, based on the constructed OPC, a suggested analysis margin to show if the investigational results meet OPC was provided. Note that for valid use of an OPC, analysis margins must be established by both clinical and statistical consensus. For safety, a candidate for a margin could be based on the approach used by Cardiac Valve OPC methods[22,23]: in this case, the analysis margin is represented by the upper limit that is two times the (OPC) point estimate. Thus, in case of a 2% revision rate OPC, the upper bound of the 95% CI of the estimate should be below 4%, which for this to happen means the estimate itself has to be close to 2%, that is, the OPC. For effectiveness, an effect size approach could be chosen, and a candidate margin could be 0.2 × SD: a difference in the observed results smaller in magnitude than 0.2 SD corresponds to a Cohen's effect size measure less than 0.2, which is small. We did not calculate the analysis margin recommended by the prosthesis benchmarking international working group[23], but these margins can be easily calculated for international reference.

Direct analyses of registry and claims data were performed using SAS 9.4 (Cary, North Carolina). Meta-analyses were performed using R (R package 'metaphor', Viechtbauer, 2010).

Results

Systematic review

Of the 7979 screened abstracts, 294 studies underwent full-text review. Ultimately, 31 publications, 29 from literature, and 2 institutional registry reports (Mayo and MARCQI) provided data for 333 995 cases that were eligible for the construction of OPC (Supplementary Fig. 1, Supplemental Digital Content 2, http://links.lww.com/JS9/A254).

For THR, eight studies with a total of 12 211 cases and two institutional registry reports with a total of 103 451 cases were included in the final analysis. Seven studies included PROM data, three reported implant survival, and two reported both PROMs and implant survival. The most commonly reported PROMs were SF-36 and SF-12 scores (see Appendix 2, supplemental Digital Content 2, http://links.lww.com/JS9/A254).

For TKR, 23 studies with a total of 65 235 cases and the 2 institutional registry reports with 153 098 cases were included. Nine studies reported implant survival data and 17 studies reported PROM data, where 3 reported both PROMs and implant survival. The most commonly reported PROM was the KSS function (see Appendix 2, Supplemental Digital Content 2, http://links.lww.com/JS9/A254).

Pooled estimates for 2-year revision rates of THR and TKR are shown in Figures 3 and 4 (details in Appendix 3, Supplemental Digital Content 2, http://links.lww.com/JS9/A254). There was no evidence of substantial heterogeneity, and the leave-one-out sensitivity analysis indicated robust findings.

Registries

Kaiser Permanente

A total of 88 126 THR patients were included (Supplementary Fig. 2, Supplemental Digital Content 2, http://links.lww.com/JS9/A254). The overall rate of revision was 2.1% (95% CI: 2.0–2.2) at 2 years; the 2-year rate of septic revision was 0.6% (95% CI: 0.6–0.7; Table 1). For TKR, 173 918 patients were included. The overall rate of revision in TKR was 1.7% (95% CI: 1.7–1.8) at 2 years; the 2-year septic revision rate was 0.8% (95% CI: 0.8–0.8; Table 1).

FORCE-TJR

In all, 3845 THR patients completed the HOOS and 5378 TKR patients completed the KOOS score at 2 years (Table 2). The overall mean (SD) score for HOOS and KOOS were 87.1 (14.6) and KOOS was 80.6 (16.0). There were 3875 THR and 5393 TKR patients who completed the SF-36 in 2 years. For THR, SF-36 PCS and MCS scores were 46.1 (10.2) and 54.0 (9.2); for TKR, SF-36 PCS and MCS scores were 44.2 (10.3) and 54.0 (9.4).

Claims database analyses

A total of 132 826 THR patients were included (Supplementary Fig. 3, Supplemental Digital Content 2, http://links.lww.com/JS9/A254). The 2-year all-cause rate of revision was 2.0% (95% CI:1.9–2.3) in New York State and 2.6% (95% CI:2.4–3.0) in California. The 2-year septic revision rate was 0.5% (95% CI: 0.4–0.6) and 0.5% (95% CI:0.3–0.7) in New York State and California, respectively (Table 3).

For TKR, 213 012 patients met the inclusion criteria (Supplementary Fig. 4, Supplemental Digital Content 2, http://links.lww.com/JS9/A254). The 2-year all-cause revision rate was 1.8% (95% CI:1.7-1.9) in New York State and 2.1% (95% CI:1.9-2.3) in California, whereas the 2-year septic revision rates were 0.5% (95% CI:0.4-0.6) and 0.6% (95%

Table 4

Safety and effectiveness objective performance criteria (OPC) and proposed benchmarks for (A) hip and (B) knee arthroplasty at 2 years.

	Revision rate (95% CI)	Combined revision rate (95% CI)	OPC (with example analysis margin)
(A) Hip arthroplasty			
THA			
Safety			
Revision – overall			
Aggregate literature/report	1.8 (1.3–2.4)		
Registry (KP)	2.1 (2.0–2.2)		
Claims – NY	2.0 (1.9–2.3)	2.0 (1.8–2.2)	2.0% (upper limit 4.0%)
Claims – CA	2.6 (2.4–3.0)		
Revision – septic			
Aggregate literature/report	0.7 (0.0–1.4)	0.6 (0.5–0.6)	0.6% (upper limit 1.2%)
Registry (KP)	0.6 (0.6–0.7)		
Claims – NY	0.5 (0.4–0.6)		
Claims – CA	0.5 (0.3–0.7)		
	Mean (SD)	Combined mean scores (SD)	
Effectiveness			
HOOS			
Aggregate studies	Symptoms: 92.0 (12.0) Pain: 93.0 (12.0) ADL: 90.0 (14.0) QoL: 83.0 (20.0)	Symptoms: 92 (12) Pain: 92.1 (9.2) ADL: 89.5 (10.3) QoL: 81.7 (14.4) Overall: 87.1 (14.5)	Symptoms: 92.0 (margin 2.4; lower limit 89.6) Pain: 92.1 (margin 1.8; lower limit 90.3) ADL: 89.5 (margin 2.1; lower limit 87.5) QoL: 81.7 (margin 2.6; lower limit 87.5) Overall: 87.1 (margin 2.9; lower limit 84.2)
Registry (FORCE-TJR)	Overall: 87.1 (14.5) Pain: 90.9 (14.3) ADL: 88.9 (15.2) QoL: 80.3 (20.9)		
HHS			
Aggregate literature	94.4 (5.0)	94.4 (5.0)	HHS: 94.4 (margin 1.0; lower limit 93.4)
Aggregate registries	–		
OHS			
Aggregate literature	–	–	
Aggregate registries	–		
EQ-5D			
Aggregate literature	EQ-5D 0.88 (0.1) EQ-VAS 82.7 (14.1)	EQ-5D 0.88 (0.1) EQ-VAS 82.7 (14.1)	EQ-5D: 0.9 (margin 0.02; lower limit 0.86) EQ-VAS: 82.7 (margin 2.8; lower limit 79.9)
Aggregate registries	–		
SF-12/SF-36			
Aggregate literature	SF-12-PCS: 46.5 (1.1) SF-12-MCS: 54.2 (0.7) SF-36-PCS: 47.0 (7.6) SF-36-MCS: 55.9 (5.7)	SF-12-PCS: 46.5 (1.1) SF-12-MCS: 54.2 (0.7) SF-36-PCS: 46.9 (5.9) SF-36-MCS: 55.3 (5.0)	SF-12-PCS: 46.5 (margin 0.22; lower limit 46.3) SF-12-MCS: 54.2 (margin 0.14; lower limit 54.1) SF-36-PCS: 46.9 (margin 0.18; lower limit 45.7) SF-36-MCS: 55.3 (margin 1.0; lower limit 54.3)
Registry (FORCE-TJR)	SF-36-PCS: 46.4 (10.1) SF-36-MCS: 54.0 (9.1)		
(B) Knee arthroplasty			
TKA			
Safety			
Revision – overall			
Aggregate literature/report	1.6 (1.2–2.0)		
Registry (KP)	1.7 (1.7–1.8)		
Claims – NY	1.8 (1.7–1.9)	1.6 (1.3–1.9)	

Table 4

(Continued)

	Revision rate (95% CI)	Combined revision rate (95% CI)	OPC (with example analysis margin)
			1.6% (upper limit 3.2%)
Claims – CA	2.1 (1.9–2.3)		
Revision – septic			
Aggregate literature/report	0.7 (0.3–1.1)		
Registry (KP)	0.8 (0.8–0.8)		
Claims – NY	0.5 (0.4–0.6)	0.7 (0.5–0.9)	0.7% (upper limit 1.4%)
Claims – CA	0.6 (0.5–0.7)		
	Mean (SD)	Combined mean scores (SD)	
Effectiveness			
KOOS			
Aggregate studies	Symptoms: 83.0 (15.0) Sport: 80.0 (14.0) Pain: 88.0 (15.0) ADL: 85.0 (16.0) QoL: 72.0 (24.0)	Symptoms: 83 (15.0) Sport: 80 (14.0) Pain: 87.5 (11.1) ADL: 85.5 (11.4) QoL: 71.9 (16.8) Overall: 80.6 (16.0)	Symptoms: 83.0 (margin 3.0; lower limit 80.0) Sport: 80.0 (margin 2.8; lower limit 77.2) Pain: 87.5 (margin 2.2; lower limit 85.3) ADL: 85.5 (margin 2.3; lower limit 83.2) QoL: 71.9 (margin 3.4; lower limit 68.5) Overall: 80.6 (margin 3.2; lower limit 77.4)
Registry (FORCE-TJR)	Overall: 80.6 (16.0) Pain: 86.8 (16.4) ADL: 86.0 (16.3) QoL: 71.9 (23.6)		
KSS			
Aggregate literature	KSS Pain: 93.0 (4.2) KSS Function: 90.6 (4.6)	KSS Pain: 93.0 (4.2) KSS Function: 90.6 (4.6)	KSS Pain: 93.0 (margin 0.84; lower limit 92.6) KSS Function: 90.6 (margin 0.92; lower limit 89.7)
Aggregate registries	–		
OKS			
Aggregate literature	28.4 (7.6)	28.4 (7.6)	OKS: 28.4 (margin 1.52; lower limit 26.8)
Aggregate registries	–		
EQ-5D			
Aggregate literature	EQ-5D 0.84(0.1) EQ-VAS 80.1 (14.9)	EQ-5D 0.84 (0.1) EQ-VAS 80.1 (14.9)	EQ-5D: 0.84 (margin 0.02; lower limit 0.82) EQ-VAS: 80.1 (margin 3.0; lower limit 77.1)
Aggregate registries	–		
SF-12/SF-36			
Aggregate literature	SF-12-PCS: 41.9 (0.9) SF-12-MCS: 51.6 (0.8) SF-36-PCS: 46.9 (3.5) SF-36-MCS: 55.5 (3.2)	SF-12-PCS: 41.9 (0.9) SF-12-MCS: 51.6 (0.8) SF-36-PCS: 46.7 (3.4) SF-36-MCS: 55.4 (3.1)	SF-12-PCS: 41.9 (margin 0.18; lower limit 41.7) SF-12-MCS: 51.6 (margin 0.16; lower limit 52.4) SF-36-PCS: 46.7 (margin 0.68; lower limit 46.0) SF-36-MCS: 55.4 (margin 0.62; lower limit 54.8)
Registry (FORCE-TJR)	SF-36-PCS: 44.2 (10.0) SF-36-MCS: 54.0 (9.4)		

(A) HOOS Symptoms; EQ-VAS; SF-12 MCS were derived from only one study with a generalizable cohort from large volume hospital.

Safety (revision rate): Upper limit = OPC + margin; and Effectiveness (PROMs): Lower limit = OPC – 0.2*SD margin.

(B) KOOS Symptoms, Sport; EQ-VAS were derived from only one study with a generalizable cohort from a large volume hospital.

Safety (revision rate): Upper limit = OPC + margin; and Effectiveness (PROMs): Lower limit = PC – 0.2*SD margin.

ADL, activities of daily living; FORCE-TJR, The Functional Outcomes Research for Comparative Effectiveness in Total Joint Replacement and Quality Improvement Registry; HHS, Harris Hip Score; HOOS, Hip disability and Osteoarthritis Outcome Score; KOOS, Knee disability and Osteoarthritis Outcome Score; KSS, Knee Society Scores; MCS, Mental Component Score; OHS, Oxford Hip Score; OKS, Oxford Knee Score; PCS, Physical Component Score; QoL, quality of life; SF-12, Short Form 12; SF-36, Short-Form 36; THA, total hip arthroplasty; TKA, total knee arthroplasty.

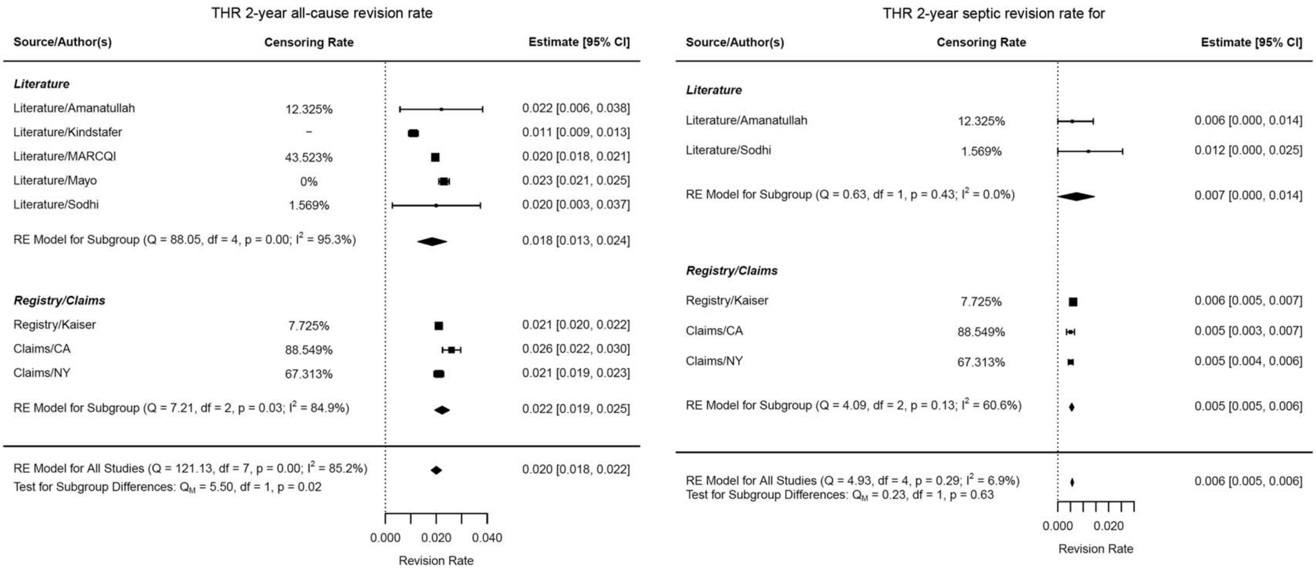

Figure 2. Forest plot revision rate THA (total hip replacement) (all-cause and septic).

CI:0.5-0.7) in New York State and California, respectively (Table 3).

Final estimates with or without data synthesis to inform the construction of OPC:

Based on these estimates, OPC for safety and effectiveness outcome measures were constructed (Table 4A, B; Figs 2, 3).

All-cause cumulative revision rate was 2.0% and 1.6% for THA and TKA, and septic revision was 0.6% and 0.7%, respectively.

For effectiveness, the estimates for HOOS and KOOS were 87.1 and 80.6; for HSS and KSS function were 94.4 and 90.6. For HrQoL, THR SF-12 PCS and MCS were 46.5 and 54.2, and SF-36 PCS and MCS were 46.9 and 55.3, respectively. For TKR, scores for SF-12 PCS and MCS were 41.9 and 51.6 and SF-36 PCS and MCS were 46.7 and 55.4 (Table 4A, B). See Appendix 3, Supplemental Digital Content 2, http://links.lww.com/JS9/A254 for details.

Example analysis margins derived from the constructed OPC are provided in Table 4A, B.

Discussion

Based on extensive assessment and synthesis of various U.S. data sources, we were able to include data from over 950 000 THR and TKR patients and construct representative and

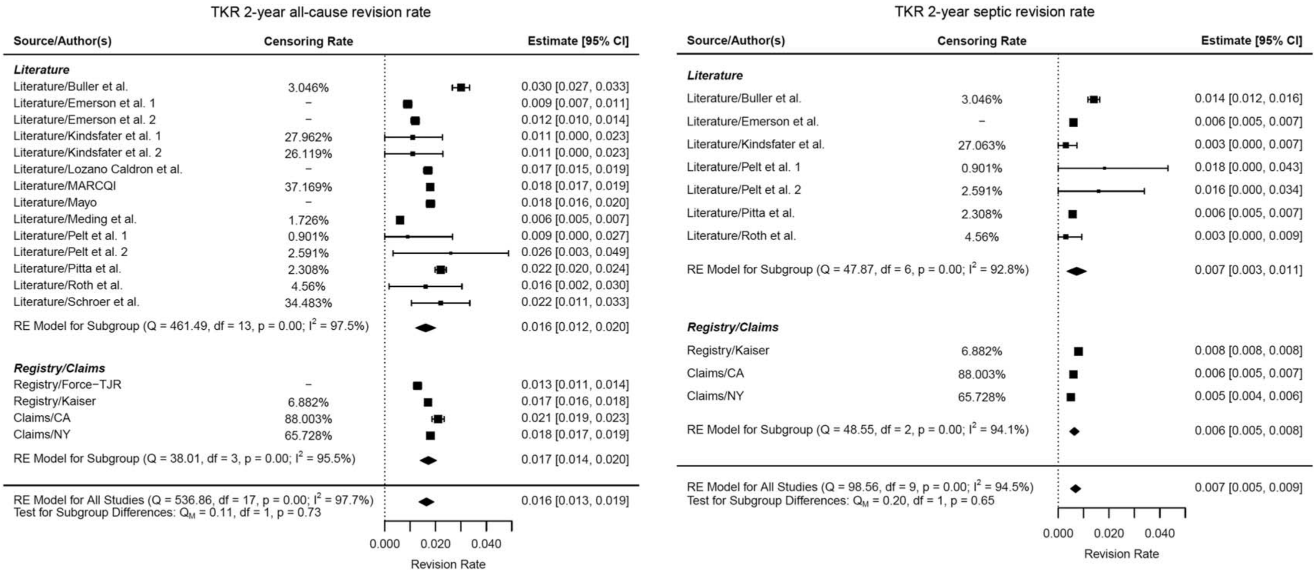

Figure 3. Forest plot revision rate TKA (total knee replacement) (all-cause and septic).

contemporary U.S. OPC with appropriate narrow confidence intervals for evaluation of 2-year safety and effectiveness of implants used in THR and TKR. These estimates were until now not available and can be used by stakeholders to conduct single-arm investigations throughout the TPLC of THR and TKR implants in order to ensure that original or incremental innovations seeking clearance or approval do not lead to harm. These single-arm clinical studies are more feasible, less time consuming, and less expensive than traditional clinical studies and can generate fit-for-purpose evidence for decision making by all stakeholders. Importantly, most of the evidence to construct OPC is based on RWE, and hence RWD can be used to conduct these single-arm studies. Furthermore, these performance metrics are invaluable for postmarket evaluations of device technologies using registries and, in some instances, also administrative data.

RWD sources such as registries and claims databases are gaining major attention for conducting device safety assessments[7,24,25]. Several attempts have been made internationally to establish safety benchmarks for joint replacements, including organizations such as the National Institute for Health and Care Excellence (NICE), UK Orthopaedic Data Evaluation Panel (ODEP), and International Society of Arthroplasty Registries (ISAR)[23,26]. Our study takes advantage of this internationally accumulated knowledge, advances the methodological approach, and focuses on evidence in the U.S. with its unique healthcare and regulatory system. The inclusion of literature, registry, and claims data enabled a comprehensive evaluation of early device performance. All stakeholders agree that 2-year all-cause revision is a patient-centered measure of implant survival and derives an important and meaningful benchmark.

Among orthopedic devices, effectiveness is best estimated using PROMs since the major value of the joint replacement is the patient's pain relief and functional improvement. Therefore, PROMS are important in benchmarking acceptable outcomes of joint replacements similar to other device products that aim to improve patient HRQoL[27]. However, normative values and benchmarks for PROMs have not been established for THR and TKR and are currently unavailable for reference. Hence, the effectiveness of new implants may have insufficiently been addressed. This study intends to fill this gap. We were able to construct 2-year performance metrics for the effectiveness of THR and TKR within the U.S. population. There is a general agreement that little functional and PROM improvement is expected beyond 2 years[28], making this an appropriate time point for OPC creation and the implementation for performance assessment of new implants.

Implementation and future perspective

OPC are evolving and dynamic estimates, which require periodic updating as new information becomes available, thereby becoming more reliable and robust over time. As our knowledge base evolves, OPC can be refined for the population in which the new implant is to be used. Furthermore, OPC should be established for mid-term and long-term safety and effectiveness.

Ultimately, predetermined device performance OPC can be employed by regulators, payors, and other stakeholders to propose benchmarks for TPLC evaluations, ranging from premarket clinical trials to postmarket clinical studies and/or surveillance. We believe that it is necessary for stakeholders to reach an agreement on analysis margins for OPC (at the 2-year follow-up to begin with) for a given purpose, such as premarket evaluation and review. This consensus should at least incorporate both clinical and statistical considerations, but further consideration of, for example, cost-effectiveness may further enhance their applicability. In order for OPC to be validly used, an OPC margin should be well established. The FDA has a history of well-established analysis margins for cardiovascular and urological devices. We believe that the evidence for the OPC presented in this study is sufficiently robust to initiate these discussions for orthopedic devices.

Study limitations

The main limitation of the literature review was a lack of standardized data reporting by researchers, which significantly reduced the number of comparable endpoints among included studies. Registry data may be incomplete or have high attrition rates, which can introduce a potential bias. However, completeness from registries in this study is high, and loss to follow-up was limited. The limitations of using stand-alone registry data and secondary research (e.g. systematic and narrative reviews) are well recognized[29,30].

When using OPC, it is important to ensure that the clinical cohorts generating outcome data are sufficiently comparable to the cohorts used for developing OPC. This will ensure that confounding is as low as possible. Our study included major robust data sources in the U.S., providing a wealth of information for properly designing single-arm investigation or registry-based (device) evaluations. We confirmed that both FORCE-TJR and KPIR results have been subject to extensive audits. We also validated the generalizability of registry data using all-inclusive New York and California state discharge data sources. Administrative databases have their own limitations related to coding errors and lack of device data. However, the consistency of results across different data sources is reassuring.

Data reporting was insufficient to construct OPC based on certain subgroups of device designs such as cemented and cementless THR and TKR, cruciate-retaining and posterior-stabilized TKR designs, or safety risk class II and class III THR and TKR. However, it could be argued that OPC should be independent of these factors since they provide standards of care by presenting benchmarks of safety and effectiveness for all implants and thus should be applied irrespective of design, material, or technique. For fundamentally different but widely used designs without much evidence, one can construct different performance goals to assist the evaluation of these devices[31].

Also, it is recognized that this study did not include any radiographic data for 2-year outcomes, which will be considered in future work. Finally, our study focused on a population of at least 18 years for all evaluations, while FDA defines adults as those at least 22 and older. However, this is unlikely to change the OPC presented in our study since joint replacement is extremely rare in these very young patients.

In conclusion, this study is the first to provide OPC for the safety and effectiveness of THR and TKR in the U.S., based on which new implant designs can be evaluated at 2 years of follow-up. Subsequent benchmarking based on established OPC for both safety and effectiveness offers a unique contribution to the regulated (commercial) introduction of new implants and, by its increasing reliability over time, has the potential to play a pivotal role in this process.

Ethics approval

This study was conducted with approval from the Institutional Review Board (IRB) of Weill Cornell Medical College (#1209013064) for the usage of SPARCS and OSHPD data. Approval was also obtained from Kaiser Permanente IRB (#5488) for the use of registry data.

Sources of funding

The study was funded by National Evaluation System for health Technology (NEST) grant U01FD006292 and AHRQ funding for the FORCE-TJR cohort (P50 HS018910-04).

Author contribution

M.J.N., P.-H.R., D.M.D., E.P., and A.S.: study design; M.J.N., P.-H.R., J.H.H., P.F., J.M., S.A., P.C., A.C., A.L., P.B., P.V., L.E.G., C.D., R.P., and E.P.: data collection; W.X., L.S., X.Z., and S.B.: data analysis; M.J.N., P.-H.R., J.H.H., S.A., D.M.D., E.P., and A.S.: writing.

Conflicts of interest disclosure

There are no conflicts of interest.

Research registration unique identifying number (UIN)

1. Name of the registry: not applicable.
2. Unique identifying number or registration ID: not applicable.
3. Hyperlink to your specific registration (must be publicly accessible and will be checked): not applicable.

Guarantor

Marc J. Nieuwenhuijse and Art Sedrakyan.

Data availability

Due to sensitivities related to the Data Use Agreement (DUA) between Weill Cornell and each of the data owners like New York State (SPARCS), California State (OSHPD), Kaiser Permanente, and FORCE-TJR, the data used in this study is not readily available for sharing. The data may be available upon request if appropriate DUA is to be executed between the interested parties. Systematic Literature Review data has been provided in the Appendix of this manuscript.

References

[1] Zhao X, Shah D, Gandhi K, *et al.* Clinical, humanistic, and economic burden of osteoarthritis among noninstitutionalized adults in the United States. Osteoarthr Cartil 2019;27:1618–26.

[2] Hunter DJ, Bierma-Zeinstra S. Osteoarthritis. Lancet 2019;393: 1745–59.

[3] Hunter DJ, Schofield D, Callander E. The individual and socioeconomic impact of osteoarthritis. Nat Rev Rheumatol 2014;10:437–41.

[4] Singh JA, Yu S, Chen L, *et al.* Rates of total joint replacement in the United States: future projections to 2020–2040 using the National Inpatient Sample. J Rheumatol 2019;46:1134–40.

[5] Nelissen RG, Pijls BG, Kärrholm J, *et al.* RSA and registries: the quest for phased introduction of new implants. J Bone Joint Surg Am 2011;93(suppl 3):62–5.

[6] Cohen D. Faulty hip implant shows up failings of EU regulation. BMJ 2012;345:e7163.

[7] Furnes O, Paxton E, Cafri G, *et al.* Distributed analysis of hip implants using six national and regional registries: comparing metal-on-metal with metal-on-highly cross-linked polyethylene bearings in cementless total hip arthroplasty in young patients. J Bone Joint Surg Am 2014;96(suppl 1):25–33.

[8] Smith AJ, Dieppe P, Howard PW, *et al.* National Joint Registry for England and Wales. Failure rates of metal-on-metal hip resurfacings: analysis of data from the National Joint Registry for England and Wales. Lancet 2012;380:1759–66.

[9] U.S. FDA. The 510(k) Program: Evaluating Substantial Equivalence in Premarket Notifications [510(k)]: Guidance for Industry and Food and Drug Administration Staff. Accessed 5 May 2021. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/510k-program-evaluating-substantial-equivalence-premarket-notifications-510k

[10] Ardaugh BM, Graves SE, Redberg RF. The 510(k) ancestry of a metal-on-metal hip implant. N Engl J Med 2013;368:97–100.

[11] Nieuwenhuijse MJ, Nelissen RGHH, Schoones JW, *et al.* Appraisal of evidence base for introduction of new implants in hip and knee replacement: a systematic review of five widely used device technologies. BMJ: Br Med J 2014;349:g5133.

[12] U.S. FDA. FDA Guidance: Design Considerations for Pivotal Clinical Investigations for Medical Devices. Accessed 28 May 2020. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/design-considerations-pivotal-clinical-investigations-medical-devices

[13] U.S. FDA. Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices. Guidance for Industry and Food and Drug Administration, 2017. https://www.fda.gov/media/99447/download

[14] Nilsdotter AK, Lohmander LS, Klassbo M, *et al.* Hip disability and osteoarthritis outcome score (HOOS) – validity and responsiveness in total hip replacement. BMC Musculoskelet Disord 2003;4:10.

[15] Roos EM, Roos HP, Lohmander LS, *et al.* Knee Injury and Osteoarthritis Outcome Score (KOOS) – development of a self-administered outcome measure. J Orthop Sports Phys Ther 1998;28:88–96.

[16] Dawson J, Fitzpatrick R, Murray D, *et al.* Questionnaire on the perceptions of patients about total knee replacement. J Bone Joint Surg Br 1998;80:63–9.

[17] EuroQol Group. EuroQol Group – a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199–208.

[18] Gandek B, Ware JE, Aaronson NK, *et al.* Cross-validation of item selection and scoring for the SF-12 Health Survey in nine countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol 1998;51:1171–8.

[19] Brazier JE, Harper R, Jones NM, *et al.* Validating the SF-36 health survey questionnaire: new outcome measure for primary care. BMJ 1992;305: 160–4.

[20] Harris WH. Traumatic arthritis of the hip after dislocation and acetabular fractures: treatment by mold arthroplasty. An end-result study using a new method of result evaluation. J Bone Joint Surg Am 1969;51: 737–55.

[21] Insall JN, Dorr LD, Scott RD, *et al.* Rationale of the Knee Society clinical rating system. Clin Orthop Relat Res 1989;248:13–4.

[22] Moher D, Liberati A, Tetzlaff J, *et al.* Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009;339:b2535.

[23] ISAR. International Prosthesis Benchmarking Working Group Guidance Document. ISAR; 2018, www.isarhome.com.

[24] Cram P, Lu X, Kates SL, *et al.* Total knee arthroplasty volume, utilization, and outcomes among Medicare beneficiaries, 1991–2010. JAMA 2012;308:1227–36.

[25] Evans JT, Evans JP, Walker RW, *et al.* How long does a hip replacement last? A systematic review and meta-analysis of case series and national registry reports with more than 15 years of follow-up. Lancet 2019;393: 647–54.

[26] Kandala NB, Connock M, Pulikottil-Jacob R, *et al.* Setting benchmark revision rates for total hip replacement: analysis of registry evidence. BMJ 2015;350:h756.

[27] Guidance for Industry: Patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance by U.S. Department of Health Human Services, and F.D.A. Center for Drug Evaluation Research, Center for Biologics Evaluation Research, Center

for Devices Radiological Health. Health and Quality of Life Outcomes 2006;4:79.

[28] Ramkumar PN, Harris JD, Noble PC. Patient-reported outcome measures after total knee arthroplasty: a systematic review. Bone Joint Res 2015;4:120–7.

[29] American Academy of Orthopaedic Surgeons: Surgical Management of Osteoarthritis of the Knee: Evidence-Based Clinical Practice Guideline, 2015. Accessed 1 June 2021. http://www.aaos.org/uploadedFiles/PreProduction/Quality/Guidelines_and_Reviews/SMOAK%20CPG__12.4.15.pdf

[30] McGrory BJ, Weber KL, Jevsevar DS, *et al*. Surgical management of osteoarthritis of the knee: evidence-based guideline. J Am Acad Orthop Surg 2016;24:e87–93.

[31] Head SJ, Mylotte D, Mack MJ, *et al*. Considerations and recommendations for the introduction of objective performance criteria for transcatheter aortic heart valve device approval. Circulation 2016;133:2086–93.