Observational Studies 6 (2020) 86-102 Submitted 2/19; Published 5/20

Survival of Ceramic vs. Metal Femoral Heads in Total Primary Hip Arthroplasty; An Observational Registry Study

Guy Cafri
Medical Device Epidemiology and Real World Data Sciences, J&J Medical Devices and Office of the Chief Medical Officer
Email: gcafri@its.jnj.com

Yuexin Chen, Elizabeth W. Paxton, Priscilla Chan
Surgical Outcomes & Analysis, Kaiser Permanente

Matthew Kelly
South Bay Medical Center, Kaiser Permanente

Brian Hallstrom
Department of Orthopedic Surgery, University of Michigan

Steven Kurtz
Biomedical Engineering and Health Sciences, Drexel University

Keywords: Protocol, Survival, Propensity Score, Covariate Balancing, Total Hip Arthroplasty

Observational Study Protocol

The following study protocol is based on the Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) guidelines (Chan et al., 2013). The structure of the protocol is amended for an observational study. Given the availability of the data during the writing of the protocol, we detail which methods are used to attain covariate balance (Cafri and Paxton, 2018). The outcome analysis is undertaken only after finalizing the protocol.

1. Administrative information

1.1 Title

Survival of Ceramic vs. Metal Femoral Heads in Total Primary Hip Arthroplasty; An Observational Registry Study

1.2 Study registration

None

1.3 Protocol date and version

Issue Date: 2020-March-30 (Revision). Protocol amendments: 02. Authors: *GC*
Revision chronology
2019-Feb-12 (Origainl)
2019-Sep-24 (Amendment 01)
2020-March-30 (Amendment 02)

1.4 Funding

The study is supported exclusively by Kaiser Permanente

1.5 Contributors

GC [Guy Cafri, Johnson & Johnson]
YC [Yuexin Chen, Surgical Outcomes and Analysis, Kaiser Permanente]
LP [Liz Paxton, Surgical Outcomes and Analysis, Kaiser Permanente]
PC [Priscilla Chan, Surgical Outcomes and Analysis, Kaiser Permanente]
MK [Matthew Kelly, South Bay Medical Center, Kaiser Permanente]
BR [Brian Hallstrom, Department of Orthopedic Surgery, University of Michigan]
SK [Steven Kurtz, Biomedical Engineering and Health Sciences, Drexel University]

GC conceived and designed the study. *YC* prepared the data. *GC* conducted the covariate balancing and *PC* will conduct the outcome analysis, sensitivity analyses and other outputs described in the protocol. *GC*, *LP*, *MK*, *BR*, *SK* contributed to refinement of the study protocol and approved the final version.

2. Introduction

2.1 Background

Hip arthroplasty exceeded half a million per year in the U.S. in 2014 (McDermott et al., 2017) and the incidence of this procedure is expected to increase (Kurtz et al., 2007). Cost of hip arthroplasty exceeded 8 billion U.S. dollars in the U.S. in 2014 (McDermott et al., 2017) and as such represents an important procedure to target for future healthcare savings (Lam et al., 2018). An underappreciated feature of components used in hip arthroplasty devices is that they are characterized by continual changes in design, often with a lack of evidence supporting their ability to prolong survival of the device and almost always with increased cost. A prime example is ceramic femoral heads, which despite their common use (Cafri et al., 2016), have notable added cost (Carnes et al., 2016; Wyles et al., 2016). Therefore, it is critical to determine the extent of the clinical benefit associated with these newer device components.

Limited research has evaluated the clinical effectiveness of ceramic femoral heads. Relative to metal femoral heads, ceramic femoral heads are thought to reduce implant wear, debris,

corrosion and metal toxicity (Semlitsch et al., 1977; Pivec et al., 2014; Gilbert et al., 1993). Ceramic heads may be at increased risk of femoral head fracturing, but the risk appears to be reduced in the most common applications (e.g., pairing with a polyethylene acetabular liner) (Traina et al., 2013; Amanatullah et al., 2011). One registry reported that among patients with osteoarthritis, ceramic heads had a slightly elevated but nonsignificant risk of any-component revision surgery as compared to metal heads (Stephen et al., 2014). In one study ceramic heads had a lower nonsignificant risk of revision surgery for any-component revision surgery as compared to metal heads (Cafri et al., 2016). Both results were limited insofar as they did not: sufficiently address device-level confounding (thereby unable to isolate the effect to the femoral head) (Cafri et al., 2014), include the most clinically relevant comparison group, include all available ceramic femoral heads and had limited follow-up. Device-level confounding is particularly noteworthy and denotes confounding due to the effects of other component characteristics in the device apart from the characteristic of interest. In this study the characteristic of interest is the material of the femoral head, but the effect of this characteristic may be confounded with one or more characteristics of other components in the device (i.e., femoral stem, acetabular shell and the acetabular liner).

In this study we will evaluate the comparative effectiveness of ceramic femoral heads in elective primary total hip arthroplasty. The outcome of interest is time-to-event, time to first replacement of any component in the device. Given the observational nature of the data and the possibility of confounding due to patient, procedure and device characteristics, propensity scores are used to balance the treatment groups on these characteristics. Additional complexity is introduced by possible variation in the effect across different manufacturers. To address this concern, treatment effects are first estimated within manufacturer and then a meta-analysis is performed to calculate the effect averaged across manufacturers.

2.2 Choice of treatments and comparators

The treatment of interest consists of devices in which a ceramic femoral head is used. Ceramic heads are either alumina, zirconia-toughened alumina matrix composite, or oxidized zirconium. These included the following model names recorded in the International Society for Arthroplasty Registries implant library: Articul/EZE Biolox Delta, Articul/EZE Ceramic, S-Rom Biolox Delta (DePuy Synthes), Femoral Head (Zimmer) (Zimmer Biomet), and Femoral Head (S & N) (Smith and Nephew). For ceramic heads manufactured by Smith and Nephew only those manufactured from oxidized zirconium were included because other ceramic material were rarely utilized.

The comparison group of interest consists of devices in which a metal femoral head is used. Metal heads are either stainless steel or cobalt chrome. These included model names of: Articul/EZE, Asphere, Bantam, Elite Modular, Femoral Head (DePuy), PFC, S-Rom (DePuy Synthes), Metasul, Total Head, Versys (Zimmer Biomet) and Femoral Head (S & N) (Smith and Nephew).

Device selection is at the discretion of the surgeon and a choice of which components to use in a device for a surgeon is often from a single manufacturer. Our choice of a comparator reflects this practice by comparing ceramic and metal femoral heads in total hip devices within manufacturer. In addition to being clinically relevant, our choice of comparator can be used to address the confounding effects of auxiliary components in the device (femoral stem, acetabular shell and liner) (Amanatullah et al., 2011) that would otherwise not be possible if comparisons are not within manufacturer.

2.3 Objectives

The objective is to determine to what extent ceramic femoral heads are superior to metal femoral heads with respect to risk of any-component revision surgery.

2.4 Study design

Observational Study. Patients in this study were continuously enrolled from January 1, 2003-December 31, 2017. Patients were followed-up prospectively. The study was designed retrospectively, after all data from patients were collected into the registry. Superiority hypotheses are tested.

3. Methods

3.1 Study setting

Data are collected from Kaiser Permanente health plan members from 52 hospitals in 6 geographical regions of the U.S. (California, Colorado, Georgia, Hawaii, Northwest, Mid-Atlantic).

3.2 Eligibility criteria

- Adults (age ≥ 18)
  Rationale: Pediatric cases are a relatively small subgroup that could have a distinctly different response to elective primary total hip arthroplasty than adult patients.

- Diagnoses: osteoarthritis, rheumatoid arthritis, inflammatory arthritis, hip dysplasia, osteo/avascular necrosis
  Rationale: These represent the most common diagnoses in the Kaiser Permanente Total Joint Replacement Registry (KPTJRR). Other diagnoses are sufficiently infrequent that their inclusion can produce problems with respect to balancing the data.

- Implantations in operative years ranging from 2003-2017
  Rationale: Ceramic femoral heads began to be implanted in 2003 for all manufacturers in Kaiser Permanente, therefore the inclusion of metal head implantations will be limited to operative years beginning in 2003. The last day of 2017 represents the last record of an implantation in the registry.

- Highly cross-linked polyethene liners (unconstrained)
  Rationale: Highly crosslinked ultrahigh-molecular-weight polyethylene acetabular liners are used quite often. While other liner materials have been used in patients in KPTJRR (conventional polyethylene, ceramic, metal) their use has been discontinued. Constrained liners are excluded given their special indications for use.

- Femoral head sizes: 28mm, 32mm, 36mm, 40mm
  Rationale: These represent most head sizes in KPTJRR. Although occasionally smaller (<28mm) and larger (>40mm) head sizes are used, these are sufficiently infrequent that their inclusion can produce problems with respect to balancing the data.

- Patients implanted with femoral heads, acetabular shells, liners and femoral stems from the same manufacturer
  Rationale: Mixing of components from different manufacturers represents off-label use.

- Excluding patients that have either extremely rare covariate values (< 5 observations within a treatment group) or covariate values that are not sufficiently present in the alternative treatment (< 5 observations). The exclusion is applied to all covariates: implant components, surgical approaches, patient characteristics and missing values for the covariates (i.e., the treatment of missing values in the analysis is to create a separate level for a nominal variable with missing values and for a continuous variable with missing data a missing indicator is created).
  Rationale: This exclusion provides a reasonable chance at attaining good balance on the covariates.

3.3 Outcome Rationale

The primary outcome is time to first revision surgery, defined as any exchange or removal of any component of the device for any reason. It is among the most common and clinically relevant endpoints in comparative effectiveness studies of total joint arthroplasty. The endpoint reflects cost to the patient (pain, burden, and risk of additional surgery) and institution (financial). The relative effect, hazard ratio, will be used to quantify the treatment effect. This will be supplemented by descriptive information related to the absolute treatment effect, the difference in survival at fixed times. Additional descriptive information will also be provided pertaining to the reason for revision (e.g., septic vs. aseptic).

3.4 Enrollment and Follow-Up

For as long as patients remain members of the Kaiser Permanente health plan, the date of revision surgery and other information pertaining to revision surgery are captured on an operative form, which is linked back to the index procedure using the patient's medical record number. Revision surgeries are chart reviewed by a clinical associate to confirm that a revision surgery has taken place and the reason for revision.

There are no efforts to follow-up patients who terminate their health membership and those who terminate their Kaiser Permanente membership will be censored at that time.

3.5 Sample Size

Study design was informed by sample size considerations, which were based on statistical power. Power was calculated using Monte Carlo simulation. The individual performing covariate balancing also performed the power analysis (GC).

Power was calculated for each manufacturer-specific analysis and the pooled analysis. The minimum effect of clinical interest was deemed to be a 30% reduction in the hazard for ceramic femoral heads relative to metal femoral heads. If survival of devices with metal heads is assumed to be 0.950 at 10 years, this minimum effect translates to survival of 0.965 for devices with ceramic heads.

Parameters for the simulation, except for the magnitude of the treatment effect and between-manufacturer variability in the treatment effect, were based on available data. The sample size used to calculate power for each manufacturer analysis was based on the number of observations available from the covariate balancing datasets (CBD). The proportion treated with ceramic heads was specific to each of the three manufacturers from the CBD and simulated as $X \sim Bernoulli$ (p). A time-to-event outcome for each subject was generated using the method of Bender et al. (Bender et al., 2005). A linear predictor (LP) is first calculated: $LP_{ki} = \beta X_{ki} + \gamma_k$, with K surgeon clusters ($k = 1, \ldots, K$), n_k observations per cluster ($i = 1, \ldots, n_k$) and $N = \sum_{k=1}^{K} n_k$. In this model the log hazard ratio β takes on the value of $exp(\beta) = 0.70$ and γ_k denotes a cluster or surgeon-specific random effect generated as $\gamma_k \sim N(0, \sigma_\gamma^2)$ with $\sigma_\gamma^2 = 0.18$. The cluster size and variance of the random effect were based on an auxiliary dataset (AD) not used for covariate balancing, only to inform parameter selection for the simulated data. The median cluster size in this dataset was 100 and the variance of the surgeon random effect was estimated from a survival model with a normal random effect. The event time was calculated by generating a random number from a standard uniform distribution, $u_{ki} \sim U(0, 1)$, and using this value to generate event times from a Weibull distribution: $(\frac{-log(u_{ki})}{\lambda exp(LP_{ki})})^{1/\eta}$, with λ and η fixed at 0.16 and 1.16, respectively. The parameters for generating event times and the (uniform) censoring rate of 0.974 were based on AD. Additional censoring was undertaken for observations with survival times past 15 years to mimic the administrative censoring in the real data.

For the simulated data to reflect covariate balancing, we used the weights that were created for each manufacturer-specific analysis after covariate balancing. A weighted Cox proportional hazard model with robust standard errors was fit to each simulated dataset and the proportion of null rejections using alpha= 0.05 (two-sided) from 10,000 simulations were the estimates of statistical power. Initially, power was only considered for significance tests of each manufacturer-specific analysis without consideration of any adjustment for multiplicity. As can be seen from the first column in the table below, power is good for data involving DePuy Synthes implants, and low for Zimmer Biomet and Smith and Nephew. Based on these initial results significance testing for Zimmer Biomet and Smith and Nephew implants were not planned, although information from those manufacturers was used in the pooled analysis. We note that power for the pooled analysis is less than for DePuy Synthes in the absence of multiplicity adjustment and no between-manufacturer heterogeneity because the

pooling method forces between-manufacturer heterogeneity to be greater than zero, which in turn reduces power.

Next, we considered power based on a multiplicity adjustment using Holm's sequential Bonferroni procedure (Holm, 1979). Three scenarios were considered that make different assumptions about the magnitude of the between-manufacturer heterogeneity (τ^2=0.000,0.025,0.050). As can be seen from columns 2-4, as the between-studies heterogeneity increases power decreases. Despite less than ideal power in simulated conditions with non-zero between-studies variability, the tests have been retained under the presumption of little or no variability across manufacturers.

Table 1. Power Under Varying Levels of Heterogeneity and Multiplicity Adjustment

Manufacturer	No Multiplicity Adjustment ($\tau^2 = 0.000$)	Multiplicity Adjustment ($\tau^2 = 0.000$)	Multiplicity Adjustment ($\tau^2 = 0.025$)	Multiplicity Adjustment ($\tau^2 = 0.050$)
Zimmer Biomet	0.57	-	-	-
DePuy Synthes	0.88	0.85	0.75	0.71
Smith and Nephew	0.31	-	-	-
Pooled	0.79	0.75	0.60	0.51

3.6 Data Collection

A detailed description of the data collected in the registry has been previously published (Paxton et al., 2010). Briefly, core data elements of the hip replacement registry consist of standardized operative data collected from the surgeon by paper or electronically at the time of surgery. Electronic health record database tables, claims databases and health plan membership databases are used for validation of information provided by the surgeon, identify patients who die or are lost to follow up, and to provide additional data elements that populate the registry. Catalogue numbers for implant components are linked to the International Society of Arthroplasty Registries implant library, a validate source of model names and component attributes. A current version of the operative form is available upon request.

3.7 Data Entry, Storage and Management

Information from the KP hip arthroplasty registry is accessible through a SAS dataset, the primary source for conducting research studies. The dataset is updated annually. The dataset is stored on KP servers with access limited to individuals in the surgical outcomes and analysis department of the institution. The dataset is based on information stored in an SQL database with front-end Microsoft® Access®. Any data entry has strict validation rules limited by predetermined characters, dates, and integers. In addition to the predefined validation rules, quality control queries are constantly applied to the data to identify out of range values, duplicate entries, missing information and inconsistent values. Any values

suspected of being inaccurate are followed up with chart review. Further information can be found in the following publication (Paxton et al., 2010).

3.8 Statistical Methods

3.8.1 Treatment.

The treatment of interest is a primary total hip replacement device in which the surface of the femoral head is made of a ceramic material. The reference or control group for all analyses is a hip implant with a femoral head whose surface material is metal.

3.8.2 Treatment Effect Heterogeneity.

There are several potential sources of heterogeneity in the treatment effect. One source of heterogeneity may be differences across manufacturers. Differences among manufacturers can be attributed to differences in device components (femoral head, stem, acetabular shell, liner). Estimating a treatment effect separately by manufacturer allows us to evaluate these potential differences.

3.8.3 Endpoint Creation and Utilization.

There is a single endpoint used in this study to test the study hypotheses, time to first revision surgery, defined as any exchange or removal of any component of the device for any reason. The endpoint is validated by a trained clinical associate. Validations are conducted independent of the study objectives described in the protocol. Patients who terminate their health insurance membership or experience a death prior to experiencing revision surgery are treated as censored cases, with survival time based on the time those cases exit the study sample. The end of the follow-up period for implantations is December 31, 2017. Because patients in this study are enrolled January 1, 2003-December 31, 2017, this provides all patients follow-up at least through the initial surgery and immediately thereafter. Additional endpoints are used to provide clinical insights into the results without hypothesis testing. Specifically, time to revision surgery of a specific component (for each of the four major components) is used. The specific component being exchanged is based on surgeon report.

3.8.4 Covariates.

Covariates included in calculation of the propensity score are body mass index (BMI), age, gender, race, American Society of Anesthesiologists (ASA) score, operative year, surgical approach, diagnosis, femoral head size, whether the femoral stem was designed to be used with cement (all acetabular shells were cementless designs), the model name of the shell, liner, and femoral stem (which capture clinical attributes of these components, such as coating applied to the fixation surface). Apart from the model names of the acetabular shells and liners, which are from product catalogues, model names of other components and their attributes are obtained from the International Society of Arthroplasty Registries Implant Library. The covariates considered prior to balancing the data was the same as the covariates used to ultimately balance the data, however some modifications were made for the

balancing process to be effective. For balancing involving Zimmer Biomet devices the liner model used is perfectly predicted by the shell type, apart from whether the liner was infused with Vitamin E. Therefore, a variable indicating the use of a liner with Vitamin E was the only liner-related variable used. The same issue arose with Smith and Nephew, except in that case no variable was used to indicate Vitamin E infusion because liners are not manufactured with this characteristic. In order to improve balance for both Zimmer Biomet and Smith and Nephew analyses, operative year was categorized into the following time periods: 2003-2006, 2007-2010, 2011-2014, 2015-2017.

A brief rationale is provided for the inclusion of the aforementioned covariates as potential confounders. Generally, younger and healthier/more active patients are treated with ceramic femoral heads because they are thought to be more active and live longer and would likely benefit most from any reductions in implant wear arising from the use of ceramic heads. Moreover, the age and health of the patients are risk factors for revision surgery (Khatod et al., 2014; Paxton et al., 2008). This is broadly the rationale for including the covariates of BMI, age, and ASA score. Race may be important to the extent that it is a proxy for socioeconomic status, which may be used in implant selection and can impact risk for revision surgery (Khatod et al., 2014). Gender may influence implant selection and has been shown to be related to risk of revision surgery (Inacio et al., 2013). Changes to surgical practice can occur over time and those changes can affect the risks for revision surgery (e.g., improved infection prophylaxis). Moreover, the use of ceramic heads has been used with increasing frequency over time. Some diagnoses may be indicative of a subgroup of patients at increased risk of revision surgery because they involve more complicated index procedures. While there is no specific reason to suspect that ceramic heads are used in more complex procedures it may nevertheless be important to provide some assurance that any observed difference is not due to case complexity. Differences in surgical approaches and auxiliary components (i.e., acetabular shell, liner and femoral stem) may arise between the treatment and control conditions because use of more novel approaches and components in an arthroplasty device tend to co-occur (Cafri et al., 2014), and these newer approaches and components may impact risk of revision surgery.

3.9 Statistical Analysis

3.9.1 Covariate Balancing

In this observational study covariate balancing methods are used, therefore the act of balancing the data is kept separate from the analysis (Rubin, 2006). A physical separation between the design and analysis is also put in place (Cafri and Paxton, 2018), such that the person responsible for covariate balancing (GC) does not have access to the outcome data until after balancing the data, at which time access is granted to the outcome data by an intermediary (YC) that links data for each observation (treatment indicator, observation weight, surgeon ID variable, value for physical activity at three times, loss to follow up indicator and time to loss) to its event time and event indicator using a unique implant identifier. A separate individual (PC) will conduct the outcome analysis.

While an ideal situation is one in which the individual who balances the data is from another institution or company than the individual who performs the outcome analysis, given the resources available for this project such a separation was not feasible. Therefore, we have chosen an approach that creates maximal separation between these individuals given the available resources. An additional approach that can be used to ensure fidelity to the proposed separation is that the individual performing the covariate balancing signs a document attesting to not having access to the outcome data while balancing the data, and similarly, the individual performing the outcome analysis signs a document stating that the covariate balancing was not altered in any way upon receipt of the balanced data.

Average treatment effect (ATE) propensity score weights are calculated using a multivariable logistic regression model that includes all covariates as predictors of treatment assignment. Missing data were only present on the covariates. We create separate levels for nominal variables with missing values and a missing indicator variable for continuous variables with missing data (i.e., BMI) while also imputing the mean (Rosenbaum, 2009). Estimating the ATE as opposed to the average treatment effect on the threated (ATT) or controls (ATC) is based on the ease with which the surgeon can transition from metal to ceramic femoral heads, or vice versa. We used stabilized weights, $w_i = \frac{Z_i Pr(Z=1)}{e_i} + \frac{(1-Z_i)Pr(Z=0)}{1-e_i}$, where $Pr(Z=1)$ and $Pr(Z=0)$ correspond to the marginal probability of treated and control individuals in the sample (Robins et al., 2000). We also considered weight trimming (Lee et al., 2011) at the $0.1/99.9, 0.25/99.75, 0.5/99.5, 1^{st}/99^{th}, 2^{nd}/98^{th}$, or $3^{rd}/97^{th}$ percentiles of the stabilized weight distribution. When used with logistic regression to estimate the propensity score, trimming can reduce bias and increase precision of the estimate (Lee et al., 2011). However, since balancing is done independent of the outcome data, trimming is only undertaken if it does not worsen balance relative to the untrimmed stabilized weights. For covariate balancing using DePuy Synthes data, no weight trimming was done, while for Zimmer Biomet and Smith and Nephew trimming was performed at the $0.25/99.75$ percentiles of the stabilized weight distribution.

3.9.2 Outcome Models

We test for the possibility that primary hip arthroplasties with ceramic femoral heads have reduced risk of time to first revision surgery for any component relative to those with metal femoral heads. The hypothesis is tested within one of the three manufacturers and the effect is also averaged across manufacturers and tested for significance. Hypothesis testing is based on a relative measure of effect (hazard ratio), although absolute measures of effect (risk difference at fixed times) will be reported descriptively for improved interpretability.

Each manufacturer-specific analysis is based on a weighted Cox regression model. Specifically, fitting a model with a single variable, indicator for treatment, is used to obtain a time-averaged estimate of the treatment effect. Variance estimation of the treatment effect should incorporate the nonindependence of observations being nested within surgeon as well as impact of weights. Elsewhere, variance estimation has been shown to be conservative when using cluster robust standard errors (an independence working correlation structure)

(Lee et al., 1992) with propensity score weights in the absence (Austin, 2016) and presence of clustered data (Cafri et al., 2019). When estimating the ATT with clustered data a cluster bootstrap was shown to be too liberal with larger cluster sizes (Cafri et al., 2019). Cluster robust standard errors are used given a preference for a conservative vs. liberal result.

To obtain an overall effect (hazard ratio) averaged across manufacturers we utilize a Bayes modal estimate of the random-effects variance combined with a normal approximation approach for confidence interval construction and hypothesis testing (Chung et al., 2013). The differences across manufacturers that are expected are small given that the treatment and comparison groups are similar, covariates are the same and the treatment effects are estimated over the same time in the same institution. Provided that the amount of between-manufacturer variation is small, the method will be effective at estimation (e.g., maintaining nominal coverage) (Friede et al., 2016). Reported P-values are based on alpha=0.05 (two-tailed). Given a concern about multiplicity (2 tests), Holm's sequential Bonferroni approach (Holm, 1979) is applied to arrive at conclusions about statistical significance.
Several statistics are reported for descriptive purposes to gain better clinical insight. For each manufacturer we report the relative hazard for specific time intervals in order to evaluate the possibility of a time-dependent effect. The issue of time dependency is relevant because it may take considerable time for some of the benefits of ceramic heads to be realized (e.g. reduced acetabular liner wear). We estimate the hazard ratio in three distinct time intervals: (0-4], (4-8] and (8-12] years. We also calculate the risk difference at the midpoint of these intervals: 2, 6 and 10 years. Additionally, we report hazard ratios for time to revision surgery of specific components to gain insights into what types of revisions ceramic heads may be most effective at preventing.

3.10 Sensitivity Analyses

Sensitivity analyses will be used to examine how the main study results change as a function of making different assumptions in our statistical analyses. A common form of sensitivity analysis in observational studies is to examine how the study results change because of unmeasured confounding. In this study we take a more targeted approach to the problem. One of the key potential confounding variables is physical activity levels of the patient. This information is partially available in our registry (since 2008 for patients choosing to respond). Measures of physical activity were obtained for patients at three time points, just prior to the operative date, one year prior to operative date and two years prior to the operative date. Reporting physical activity up to two years prior to the index procedure is considered because impairment associated with the indication for hip replacement could attenuate the amount of physical activity a person engages in. Each measure of physical activity is the average number of minutes per week the patient reports engaging in physical activity. Using the method of Lin et al. (Lin et al., 1998), a sensitivity analysis can be performed by specifying the imbalance in the two groups on a covariate, as well as that covariate's relationship to the response. Our approach is to take the largest estimated standardized difference that compares the two treatments on physical activity and the largest estimated effect of physical activity on time to revision surgery, among the three different measurements of physical activity. By choosing the maximum values for group imbalance and the outcome relationship,

the maximum impact of physical activity on the study results is being considered based on the point estimates of the partially measured variables.

Additional sensitivity analyses will be based on making different assumption about loss to follow-up. In the registry a non-trivial portion of patients (10.1%) are lost to follow-up due to terminations of their health insurance membership. The main analysis assumes that the censoring mechanism is noninformative, but this may not be plausible as those that chose to end their membership at some set of times may not be representative of patients who did not end their membership at those times. While there is no specific reason to think that those individuals who terminate membership are at increased (or decreased) risk of failure, a sensitivity analysis can be used to determine how robust the reported analyses are to the presence of informative censoring. A method of imputing the response (event and event times) can be undertaken in which those lost to follow-up are assumed to have increased or decreased risk of the event of interest (Jackson et al., 2014). Our implementation considers increasing or decreasing the hazard ratio of device failure among those lost to follow-up by a factor of 2.0 or 0.5, respectively. These values are selected based on their subjective plausibility. Another approach we also consider, in the event of a statistically significant result, is to identify the value for the parameter (hazard ratio of device failure among those lost to follow-up) that makes the finding no longer significant. The imputation model consists of a Cox proportional hazards model that is stratified on treatment and includes patient age, with the maximum failure time based on the end of the follow-up period (December 31, 2017). Ten imputations of the response will be performed, with the weighted model described in the main text applied to each imputed data set and the results aggregated using Rubin's rules (Rubin, 1987).

R statistical software will be used for all analyses.

3.11 Table Shells

Table X.1 Main Study Results, Hazard Ratios by Manufacturer

Manufacturer	HR	CI	P
DePuy Synthes			
Zimmer Biomet			Not Reported
Smith and Nephew			Not Reported
Pooled			

*Indicates p-value$<$0.05 after multiplicity adjustment

Table X.2 Hazard Ratios by Manufacturer and Time

	(0-4] yrs.		(4-8] yrs.		(8-12] yrs.	
Manufacturer	HR	CI	HR	CI	HR	CI
DePuy Synthes						
Zimmer Biomet						
Smith and Nephew						

Table X.3 Hazard Ratios by Component-Specific Endpoints

	Femoral Head		Femoral Stem		Acetabular Liner		Acetabular Shell	
Manufacturer	HR	CI	HR	CI	HR	CI	HR	CI
DePuy Synthes								
Zimmer Biomet								
Smith and Nephew								

Note: Component -specific endpoints are defined as time to first revision surgery of the component specified in the table with or without revision to any other component

Table X.4 Sensitivity Analyses

	Physical Activity Confounding		Loss to Follow-Up Dec. Risk Revision		Loss to Follow-Up Inc. Risk Revision	
Manufacturer	HR	CI	HR	CI	HR	CI
DePuy Synthes						
Zimmer Biomet						
Smith and Nephew						

Note: Component -specific endpoints are defined as time to first revision surgery of the component specified in the table with or without revision to any other component

4. Ethics and Dissemination

4.1 Research Ethics Approval

The study has been approved by the Kaiser Permanente Southern California IRB (#5488)

4.2 Protocol Amendments

Any modifications to the protocol that may impact on the study design or analysis will require an amendment to the protocol. The amendment will be agreed to by the study authors. Any other changes to the protocol that do not alter the design or analysis (e.g., administrative changes) will be documented in a memorandum.

4.3 Informed Consent

The study has been approved with a waiver of requirement to document and obtain informed consent

4.4 Confidentiality

The storage of data has been previously described. Paper-based forms are locked in file cabinets in an area with limited access. Records that contain personal identifiers are limited with database linkage utilizing patient medical record numbers. Restricted access is provided to databases by system administrators.

4.5 Declaration of Interests

GC is an employee of Johnson & Johnson. The work conducted by *GC* is in his personal capacity and not in his capacity as a Johnson & Johnson employee.

SK is and officer and shareholder of Exponent. Exponent has been paid fees by companies and suppliers for consulting services of *SK* on behalf of such companies and suppliers, including: Stryker, Zimmer Biomet, Invibio, Stelkast, Wright Medical Technology, Ceramtec, Celanese, Simplify Medical, Formae, and Ferring Pharmaceuticals
No conflicts of interests reported by any other contributors.

4.6 Access to data

Only one person has complete access to the dataset [*YC*]. The person responsible for analysis of the data [*GC*] has limited access, as previously described.

4.7 Public Access

Public access is available to the protocol. The statistical code is available upon request. The patient-level dataset is not available due to policies set forth by Kaiser Permanente.

References

Amanatullah, D. F., Landa, J., Strauss, E. J., Garino, J. P., Kim, S. H., and Di Cesare, P. E. (2011). Comparison of surgical outcomes and implant wear between ceramic-ceramic and ceramic-polyethylene articulations in total hip arthroplasty. *The Journal of Arthroplasty*, 26(6):72–77. Available from: https://doi.org/10.1016%2Fj.arth.2011.04.032.

Austin, P. C. (2016). Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in Medicine*, 35(30):5642–5655. Available from: https://doi.org/10.1002%2Fsim.7084.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723. Available from: https://doi.org/10.1002%2Fsim.2059.

Cafri, Graves, Sedrakyan, Fan, Calhoun, de Steiger, Cuthbert, Lorimer, and Paxton (2014). Confounding by exogenous treatment in observational comparative effectiveness studies of medical devices: An illustration using porous tantalum acetabular shells in primary total hip arthroplasty. *Manuscript Submitted for Publication*.

Cafri, G. and Paxton, E. W. (2018). Mitigating reporting bias in observational studies using covariate balancing methods. *Observational Studies*, 4:292–296.

Cafri, G., Paxton, E. W., Love, R., Bini, S. A., and Kurtz, S. M. (2016). Is there a difference in revision risk between metal and ceramic heads on highly crosslinked polyethylene liners? *Clinical Orthopaedics and Related Research®*, 475(5):1349–1355. Available from: https://doi.org/10.1007%2Fs11999-016-4966-1.

Cafri, G., Wang, W., Chan, P. H., and Austin, P. C. (2019). A review of causal inference methods for clustered observational data with application to the evaluation of the effectiveness of medical devices. *Statistical Methods in Medical Research*, 28:3142–3162. Available from: https://doi.org/10.1177%2F0962280218799540.

Carnes, K. J., Odum, S. M., Troyer, J. L., and Fehring, T. K. (2016). Cost analysis of ceramic heads in primary total hip arthroplasty. *The Journal of Bone and Joint Surgery*, 98(21):1794–1800. Available from: https://doi.org/10.2106%2Fjbjs.15.00831.

Chan, A.-W., Tetzlaff, J. M., Gøtzsche, P. C., Altman, D. G., Mann, H., Berlin, J. A., Dickersin, K., Hróbjartsson, A., Schulz, K. F., Parulekar, W. R., et al. (2013). Spirit 2013 explanation and elaboration: guidance for protocols of clinical trials. *Bmj*, 346:e7586.

Chung, Y., Rabe-Hesketh, S., and Choi, I.-H. (2013). Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in Medicine*, 32(23):4071–4089. Available from: https://doi.org/10.1002%2Fsim.5821.

Friede, T., Röver, C., Wandel, S., and Neuenschwander, B. (2016). Meta-analysis of few small studies in orphan diseases. *Research Synthesis Methods*, 8(1):79–91. Available from: https://doi.org/10.1002%2Fjrsm.1217.

Gilbert, J. L., Buckley, C. A., and Jacobs, J. J. (1993). In vivo corrosion of modular hip prosthesis components in mixed and similar metal combinations. the effect of crevice, stress, motion, and alloy coupling. *Journal of Biomedical Materials Research*, 27(12):1533–1544. Available from: https://doi.org/10.1002%2Fjbm.820271210.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Inacio, M. C., Ake, C. F., Paxton, E. W., Khatod, M., Wang, C., Gross, T. P., Kaczmarek, R. G., Marinac-Dabic, D., and Sedrakyan, A. (2013). Sex and risk of hip implant failure: Assessing total hip arthroplasty outcomes in the united states. *JAMA Intern Med*, 173(6):435–441.

Jackson, D., White, I. R., Seaman, S., Evans, H., Baisley, K., and Carpenter, J. (2014). Relaxing the independent censoring assumption in the cox proportional hazards model using multiple imputation. *Statistics in Medicine*, 33(27):4681–4694. Available from: https://doi.org/10.1002%2Fsim.6274.

Khatod, M., Cafri, G., Namba, R. S., Inacio, M. C., and Paxton, E. W. (2014). Risk factors for total hip arthroplasty aseptic revision. *The Journal of Arthroplasty*, 29(7):1412–1417. Available from: https://doi.org/10.1016%2Fj.arth.2014.01.023.

Kurtz, S., Ong, K., Lau, E., Mowat, F., and Halpern, M. (2007). Projections of primary and revision hip and knee arthroplasty in the united states from 2005 to 2030. *The Journal of Bone and Joint Surgery-American Volume*, 89(4):780–785. Available from: https://doi.org/10.2106%2Fjbjs.f.00222.

Lam, V., Teutsch, S., and Fielding, J. (2018). Hip and knee replacements. *JAMA*, 319(10):977. Available from: https://doi.org/10.1001%2Fjama.2018.2310.

Lee, B. K., Lessler, J., and Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PLoS ONE*, 6(3):e18174. Available from: https://doi.org/10.1371%2Fjournal.pone.0018174.

Lee, E. W., Wei, L., Amato, D. A., and Leurgans, S. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival analysis: state of the art*, pages 237–247. Springer.

Lin, D. Y., Psaty, B. M., and Kronmal, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, 54(3):948. Available from: https://doi.org/10.2307%2F2533848.

McDermott, K. W., Freeman, W. J., and Elixhauser, A. (2017). Overview of operating room procedures during inpatient stays in us hospitals, 2014: statistical brief# 233. *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Rockville: Agency for Healthcare Research and Quality.*

Paxton, Cafri, Lorimer, Kärrholm, Graves, Malchau, Namba, and Rolfson (2008). Patient risk factors of total hip arthroplasty revision in patients with osteoarthritis. *Manuscript*

Submitted for Publication, pages 1412–1417. Available from: https://doi.org/10.1016%2Fj.arth.2014.01.023.

Paxton, E. W., Inacio, M. C., Khatod, M., Yue, E. J., and Namba, R. S. (2010). Kaiser permanente national total joint replacement registry: Aligning operations with information technology. *Clinical Orthopaedics and Related Research®*, 468(10):2646–2663. Available from: https://doi.org/10.1007%2Fs11999-010-1463-9.

Pivec, R., Meneghini, R. M., Hozack, W. J., Westrich, G. H., and Mont, M. A. (2014). Modular taper junction corrosion and failure: How to approach a recalled total hip arthroplasty implant. *The Journal of Arthroplasty*, 29(1):1–6. Available from: https://doi.org/10.1016%2Fj.arth.2013.08.026.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560. Available from: https://doi.org/10.1097%2F00001648-200009000-00011.

Rosenbaum, P. R. (2009). *Design of Observational Studies. Springer: New York, NY.*

Rubin, D. (2006). The designversus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):20–36. Available from: https://doi.org/10.1002%2Fsim.2739.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc. Available from: https://doi.org/10.1002%2F9780470316696.

Semlitsch, M., Lehmann, M., Weber, H., Doerre, E., and Willert, H. (1977). New prospects for a prolonged functional life-span of artificial hip joints by using the material combination polyethylene/aluminium oxide ceramic/metal. *Journal of Biomedical Materials Research*, 11(4):537–552. Available from: https://doi.org/10.1002%2Fjbm.820110409.

Stephen, G., David, D., Philip, R., Lisa, I., and et al (2014). Australian orthopaedic association national joint replacement registry (aoanjrr). *Hip, knee & shoulder arthroplasty: 2014 annual report*, Adelaide: AOA(6):72–77.

Traina, F., De Fine, M., Di Martino, A., and Faldini, C. (2013). Fracture of ceramic bearing surfaces following total hip replacement: A systematic review. *BioMed Research International*, 2013:1–8. Available from: https://doi.org/10.1155%2F2013%2F157247.

Wyles, C. C., McArthur, B. A., Wagner, E. R., Houdek, M. T., Jimenez-Almonte, J. H., and Trousdale, R. T. (2016). Ceramic femoral heads for all patients? an argument for cost containment in hip surgery. *Am J Orthop*, 45:E362–E366.