Author/s:
Cafri, G;Graves, SE;Sedrakyan, A;Fan, J;Calhoun, P;de Steiger, RN;Cuthbert, A;Lorimer, M;Paxton, EW

Title:
Postmarket surveillance of arthroplasty device components using machine learning methods

Date:
2019-11-01

Citation:
Cafri, G., Graves, S. E., Sedrakyan, A., Fan, J., Calhoun, P., de Steiger, R. N., Cuthbert, A., Lorimer, M. & Paxton, E. W. (2019). Postmarket surveillance of arthroplasty device components using machine learning methods. Pharmacoepidemiology and Drug Safety, 28 (11), pp.1440-1447. https://doi.org/10.1002/pds.4882.

Cafri Guy (Orcid ID: 0000-0002-1743-429X)

Full Title: Post-Market Surveillance of Arthroplasty Device Components Using Machine Learning Methods

Running Title: Post-Market Surveillance of Arthroplasty Devices

Author Names: Guy Cafri, Stephen E. Graves, Art Sedrakyan, Juanjuan Fan, Peter Calhoun, Richard N. de Steiger, Alana Cuthbert, Michelle Lorimer, Elizabeth W. Paxton

Author affiliations: Surgical Outcomes and Analysis, Kaiser Permanente (Guy Cafri and Elizabeth W. Paxton), Australian Orthopaedic Association National Joint Replacement Registry (Stephen E. Graves, Alana Cuthbert, Michelle Lorimer), Healthcare Policy and Research, Weill Cornell Medical College (Art Sedrakyan), Department of Mathematics and Statistics, San Diego State University (Juanjuan Fan), Computational Science Research Center, San Diego State University (Peter Calhoun), Department of Surgery Epworth HealthCare, The University of Melbourne (Richard N. de Steiger)

Correspondence to: Guy Cafri, PhD, MStat, Kaiser Permanente National Implant Registries, Surgical Outcomes & Analysis, 8954 Rio San Diego Drive, Suite 406. San Diego, CA 92108; phone: 813-486-9875; email: guycafri@gmail.com

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pds.4882

Keywords: Arthroplasty, Elastic Net, Machine Learning, FDA, Random Forest, Post-Market Surveillance

Key Points:

- Early identification of total hip arthroplasty devices with increased risk of failure can be challenging because hundreds of distinct components are available for use and any estimated effect needs to address confounding due to device and patient factors
- Machine learning methods are promising for detection of components at increased risk of failure
- Machine learning methods were effective at detecting device components recalled by the U.S. Food and Drug Administration using data from a U.S. total joint replacement registry.

Word Count: 3348

Conflicts of Interest: None declared.

Related Presentations:

i) Food and Drug Administration, 21st Century Active Surveillance: Developing Methods for Evolving National System for Medical Devices; June 1, 2016; Silver Spring, MD.

ii) Seventh International Congress of Arthroplasty Registries; June 9, 2018; Reykjavik, Iceland

Funding: None

ABSTRACT

Purpose: While joint arthroplasty is generally a safe and effective procedure there are concerns that some devices are at increased risk of failure. Early identification of total hip arthroplasty devices with increased risk of failure can be challenging because devices consist of multiple components, hundreds of distinct components are currently used in surgery, and any estimated effect needs to address confounding due to device and patient factors. The purpose of this study was to assess the effectiveness of machine learning approaches at identifying recalled components listed by the U.S. Food and Drug Administration using data from a U.S. total joint arthroplasty registry. **Methods:** An open cohort study was conducted using data (January 1, 2001 to December 31, 2015) from 74,520 implantations and 348 unique components in the Kaiser Permanente Total Joint Replacement Registry. Exposures of interest were device components used in elective primary total hip arthroplasty. The outcome was time to first revision surgery, defined as exchange, removal or addition of any component. Machine learning methods included regularized/unregularized Cox models and random survival forest. **Results:** Among the recalled components detected were ASR acetabular shell/large femoral head, Durom acetabular shell/ Metasul large femoral head and Rejuvenate modular neck stem. The three components not identified were characterized by small numbers of devices recorded in the registry. **Conclusions**: The novel approaches to signal detection may improve post-market surveillance of frequently used arthroplasty devices which in turn will improve public health.

INTRODUCTION

The effectiveness and safety of medical devices is a major public health issue. Lack of pre- and post-market safety assurances for these devices is a well-recognized problem worldwide (1-4). Arthroplasty devices are among the most relevant given their widespread use and the presence of underperforming devices. Primary total knee and hip arthroplasty exceed one million per year in the U.S. alone (5) and the incidence of these procedures is expected to increase (6). It is not uncommon for newly introduced arthroplasty devices to place patients at increased risk of revision surgery compared to those already available on the market (7). Not only individual devices but entire classes of devices have been identified as having significantly higher risks of revision surgery (8-13). The most well-known of these is large head metal-on-metal (LHMoM) designs in total hip arthroplasty (THA) and hip resurfacing (8-12). Registries have played a critically important role in identifying these problems. It was data collected and reported by registries that identified the problem and led to the recall of many of these devices and the abandonment of the entire class of LHMoM devices. There is growing consensus that large scale multinational evaluations of devices using all-inclusive registries are essential for determining if a device is at increased risk for failure (14,15).

Arthroplasty devices are comprised of multiple components working together to ensure the success of the procedure. In the case of total hip replacements there are four major components: the femoral stem implanted in the femur, a femoral head that attaches to the stem, an acetabular

shell that is implanted in the hip socket and a liner which is inserted into the shell. Revision surgery may occur because one or more of these components fail. To identify specific components that are at increased risk of revision surgery is challenging as there are hundreds of individual components and they are used in many different combinations. Past signal detection efforts in this area use a multi-step process that includes empirical methods and clinical judgement in identifying device components at increased risk of revision surgery (16,17). The first stage of the process, which determines whether components are examined further, relies on calculating incidence density for individual components and comparing them to the density of other components in the same class. However, the method ignores the ordering of time and does not address the possibility of confounding due to device and patient factors. Ideally, a method can identify individual components with an increased risk of revision surgery using a time-to-event endpoint while also limiting the confounding effects of other components in the device and patient characteristics.

Machine learning methods are appealing for this type of problem because they can be used to address high dimensional data, which conventional methods generally cannot. Moreover, the methods address the added complexity introduced by having each component be both of substantive interest and potentially a confounder of other component effects. The principal objective of this paper is to evaluate use of machine learning methods for surveillance of total hip arthroplasty components. The effectiveness of the methods are determined based on their

ability to detect recalled components identified by the U.S. Food and Drug Administration (FDA) using data from a U.S. joint replacement registry on patients undergoing elective primary total hip arthroplasty. Use of recalled components in this study provides a gold standard by which to evaluate the proposed methods, and conditional on the presence of sufficient evidence of the method's validity, could be used to enhance post-market surveillance efforts using contemporary data from national and regional arthroplasty registries.

METHODS

An open cohort study was conducted in which patients were continuously enrolled over time and the design was conceived after the data were collected.

Data Source

The Kaiser Permanente Total Joint Replacement Registry (KPTJRR) (18,19) was used to identify a cohort of 74,520 implantations in patients with primary elective total hip arthroplasty from January 1, 2001 to December 31, 2015 (Figure 1). The registry includes data from 52 hospitals in 6 geographical regions of the U.S. (California, Colorado, Georgia, Hawaii, Northwest, Mid-Atlantic). The participation of surgeons is voluntary and was 95% among all surgeons performing elective hip replacements. Registry data is validated using a hospital utilization database and independent chart review. An institutional review board approved the study prior to data collection.

Figure 1 here

Device Components

Four major components of a total hip replacement prosthesis were considered: femoral stem, femoral head, acetabular shell, and acetabular liner. Each device component is defined using manufacturer product numbers that correspond to specific model names and clinical attributes. For the femoral stem the component was defined based on the model name, whether the design indicated use of cement, the surface finish, type of coating applied, material and neck connection. For femoral classification is based on model name, head size and material. The acetabular shell was defined based on model name, indicated use of cement, the surface finish, type of coating applied and material. Lastly the insert, if present, was classified according to the model name, articulation, material, hood and use of antioxidants. The information was obtained from the International Society for Arthroplasty Registries implant library, a validated source for standardizing the naming conventions of components and their clinical attributes.

There were 348 unique components (42 acetabular shells, 103 femoral heads, 122 stems, and 81 liners). Components were made by six manufacturers: DePuy Synthes (Warsaw, IN, USA), Zimmer Biomet (Warsaw, IN, USA), Smith & Nephew (Memphis, TN, USA), Stryker (Mahwah, NJ, USA), Exactech (Gainesville, FL, USA), and Wright Medical Technology (Memphis, TN, USA). In some cases, modifications to variable inputs were made to limit collinearity in the data. For example, >36 mm ASR femoral heads were always paired with an

ASR shell, and vice versa. Given perfect collinearity among these components a single predictor representing joint use of the shell and femoral head was used.

Patient Characteristics

Patient covariates included: age, gender, diagnosis associated with joint replacement (osteoarthritis vs. other), race (African American, Asian/Pacific Islander, Caucasian, Hispanic, Native American, Multi-Race), body mass index (BMI), diagnosis of diabetes, and American Society of Anesthesiologists (ASA) score (<3 vs. ≥ 3). All patient covariates were treated as potential confounders.

Outcome

The outcome was time to first revision surgery, defined as exchange, removal or addition of any device components. Revision surgeries were confirmed by chart review conducted by a trained clinical associate. Health insurance member terminations or death were treated as censored cases with survival time based on the time those cases exited the study sample. The end of the follow-up period for implantations in KPTJRR was December 31, 2015. Patients not experiencing a revision, lost to follow-up or death have survival times based on time elapsed between their initial implantation date and the end of the follow-up period.

Identifying Components with Known Design Defects Using FDA Listed Recalls

Reporting on recalled components with design defects was motivated by a desire to know with relative certainty that a component or combination of components were ineffective, providing a gold standard by which the statistical methods could be judged regarding accuracy of detection. FDA's medical device recall database (www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfRES/res.cfm) was searched for components used in total hip arthroplasty from November, 2002 (inception) to December 31, 2017. Problems with the design of the device (as the reason for the recall listed by FDA) are of exclusive interest in this study. Recalls related to design of a device component corresponded to a relatively small number of distinct implant components in KPTJRR, most associated with LHMoM designs or modular neck stems: DePuy Synthes' ASR femoral head and acetabular shell, Zimmer's Durom acetabular shell, Stryker Howmedica's ABG II and Rejuvenate modular neck stems, Smith & Nephew's Modular SMF stem and Zimmer's alumina ceramic head when paired with cobalt chromium hip stems. The FDA's reason for the recall of Durom acetabular shells was "Labeling False and Misleading", although a design defect appears to be a better explanation and was therefore grouped with implants described as having problems related to design.

Statistical Analyses

Device components and patient characteristics were predictor variables and time to first revision surgery was the outcome. Each device component distinctly identified based on its model name attributes was entered as an indicator variable. Patient-level covariates were categorical variables

with two levels (gender, diabetes, ASA score, diagnosis associated with joint replacement), more than two levels (race), or continuous variables (age, BMI) that were transformed into categorical variables based on quintiles of their distributions. For modeling of categorical variables with K levels we form K-1 indicator variables in the Cox modeling approaches and for tree construction implement an ordinal (age, BMI) or nominal (race) splitting approach. Missing data was only present on the patient covariates (0.01% gender, 0.17% race, 1.79% BMI, 2.32% diabetes and 2.35% ASA score), which were imputed using iterative adaptive tree imputation (20).

The first approach makes use of both regularized and unregularized Cox proportional hazards models. A regularized model with a mixture of L_1 (lasso) and L_2 (ridge) penalties was used to select a subset of device components most predictive of survival (21-23). The choice of elastic net over either the lasso or ridge regression was based on its superior performance with highly correlated predictors, including selection of more than one predictor in a group of correlated predictors and less erratic numerical behavior, while still resulting in a sparse selection of those predictors (21-22). The extent of the penalties was based on choosing a priori a value for a parameter in the model that is midway between lasso and ridge regression (α=0.5; α ranges from 0 to 1). The parameter that determines model complexity was chosen based on the value that minimizes the -2*log(partial likelihood) via 10-fold cross validation (21). No penalty is applied to patient covariates in the model given a desire to fully control for the effects of relatively few patient characteristics. While a sparse solution is desirable because it allows identification of a

more parsimonious set of device components, parameter estimates for selected components are shrunken toward zero with no accompanying estimate of the precision of the effect. Therefore, as a second step, components that are selected and patient covariates are included in an unregularized Cox proportional hazards model. Given a desire to conduct inference, but appreciating that a selection process was initially undertaken, *P*-values that maintain the false discovery rate (24) at 0.05 are calculated using the total number of device components that went into the regularized model (for unselected variables we set *P*-values ≈ 1, as implied by zero coefficient in the model). Control over the false discovery rate maintains the fraction of false discoveries among the rejected null hypotheses at the nominal value.

The second approach makes used of random survival forest (25). Forests in this application are a collection of 2000 trees in which each tree is constructed by repeatedly performing binary splits of the data using the logrank test until a stopping rule is reached (e.g., < 10 observations or 2 events in a parent node). The variables considered at each split in each tree were chosen randomly, but the number of variables was fixed at $\lceil P/4 \rceil$ (P is the number of variables). The number of variables considered at each split is larger than convention because bias in variable selection with correlated inputs can be limited by increasing the number of variables considered at each split (26). This is important to the extent that it can limit patient- and device-level confounding in the selection process. Intuitively, considering more variables at each split foster more competition, which limits selection of variables simply because they are associated with

other inputs that have a relationship with the response. Rankings of variables are based on minimal depth (25). Minimal depth for a variable in a tree is the distance from the root node of the tree to the node a variable is first split on. The distance is recorded for each variable in a tree and then for each variable an average is calculated over trees in the forest. Shorter distances denote variables with stronger effects. To determine whether a device component's minimal depth exceeds chance, a threshold *P*-value of 0.05 was used based on each variable's empirical null distribution (27). The null distribution is based on 1000 permutations of the response, growing a forest (200 trees) on each permutation of the response and calculating minimal depth rank for each variable (smallest attainable permutation *P*-value is 1/1001, or 0.000999). Given the small number of permutations performed due to large computational cost, *P*-values based on false discovery rate adjustment were not calculated. Therefore, a variable is considered significant if the permutation *P*-value is less than 0.05. See supplementary materials for sensitivity analyses related to surgeon effects, mixing parameter in Cox elastic net and the number of variables considered at each split in random forest. R statistical software was used for all analyses, randomForestSRC version 2.4.1 for imputation, MST (28) version 2.1 for random survival forest, glmnet (29,30) version 2.0-16 for Cox elastic net, and the survival package version 2.42-3 for unregularized Cox regression.

RESULTS

A majority of patients in the sample underwent arthroplasty for osteoarthritis (93.2%), had an ASA score <3 (66.1%), were female (57.4%), Caucasian (79.2%), had an average age of 65.5 and BMI of 29.3 (Table 1). Overall device survival is also provided over the study period (Figure 2).

Table 1 here

Figure 2 here

Results are displayed that illustrate the extent to which recalled components are identified using the regularized/unregularized Cox model and random forest methods (Table 2). The categories of devices investigated fall into three groups, LHMoM, modular necks and alumina femoral heads. With respect to recalled LHMoM devices, the ASR shell/ >36 mm femoral head was identified using both approaches, however only in the case of the regularized/unregularized Cox model can we ensure that the false discovery rate is maintained at 0.05. It is important to mention that in KPTJRR only >36 mm ASR femoral heads were implanted. Another set of LHMoM devices use the Durom acetabular shell and Metasul femoral head. While Metasul femoral heads $\leq$ 36 mm were implanted, these were never paired with a Durom shell. For implantations of the Durom shell paired with > 36 mm Metasul femoral head, the regularized/unregularized Cox models were able to identify the Durom shell, even while maintaining the false discovery rate at the nominal level. The shells were not detected using the random forest approach. With respect to modular

neck stems, the Rejuvenate was identified using both approaches, with the regularized/unregularized Cox models identifying the component while also maintaining the false discovery rate at 0.05. Neither method was able to identify the ABG II or the Modular SMF, but these components had only 1 revision surgery each, out of 3 and 47 implantations, respectively. For alumina femoral heads neither approach was able to identify this device. The recall in this case however was more specific, the femoral head paired with a cobalt chrome femoral stem. Among the 290 implantations that used this femoral head, 18 were paired with a cobalt chrome femoral stem.

Table 2 here

Although our primary interest is in detecting components that have been recalled because these represent a gold standard (i.e., known design defect) by which to judge the success of the methods, it is also possible to report on non-recalled components that were detected. Among the components screened, only two were detected based on statistical significance after controlling for the false discovery rate (Table 2). Both components were acetabular liners used in Pinnacle shells, with the difference in the two being the use or not of an elevated liner rim (which is normally used to prevent dislocation of the hip prosthesis). An important point is that for both of these components the effect suggests that the components are protective against revision surgery (Figure 3).

Figure 3 here

A central motivation for the use of machine learning methods is to control for component confounding. The extent of component confounding among the reported components is evaluated. For each component, we compare the hazard ratio for that component in two models: a) a Cox model with a variable indicating the use of that component and all patient covariates, and b) the unregularized Cox model, which includes all components selected using the elastic net and all patient covariates and represents the effect of each component after conditioning on the other components and patient characteristics. Therefore, the difference in hazard ratios between these two models represents the extent of component confounding. We can see that for most of the components there is at least modest evidence of component confounding, relative differences in model coefficients (i.e., natural log of the hazard ratio) range from 3 to 85 percent (Figure 3).

DISCUSSION

Effective post-market surveillance of arthroplasty devices is necessary to protect public health. Three of the six recalled components were identified using the machine learning approaches. Three implants were not detected; however, the number of implantations was either moderately or prohibitively small to detect two of the implant components, and in another case a recall applied to a combination of components that was infrequently utilized. This suggests that a small

number of implantations places a practical constraint on detection. Nevertheless, the results suggest that the machine learning methods explored in this paper can be effective at detection.

Some comparison of the two methods is in order. First, one of the recalled components was detected using regularized/unregularized Cox modelling but was not identified using random forest, providing greater support for the former approach. Second, one of the primary motivations for the current work was to address both device- and patient-level confounding when identifying harmful device components. The two approaches handle this differently, the regularized/ unregularized Cox modelling approach simultaneously conditions on device and patient characteristics while the random forest approach has a large number of variables compete in the data splitting process. To the extent that the Cox models rely on a more conventional way of addressing confounding they may be preferable but may be limited by extrapolation and model misspecification often cited with the use of regression models (31) as well as a pre-selection process that may omit some relevant confounders. There are some additional advantages that characterize the regularized/unregurlaized Cox modelling approach, including availability of direction (and magnitude) of component effects and improved computational efficiency. In fact, calculating a sufficient number of permutations in the random forest approach to obtain *P*-values that could be corrected to maintain the false discovery rate at a desired level was computationally too burdensome.

In future applications, machine learning methods could act as an initial screen of device components that would be followed-up with comparative effectiveness studies of detected components. An important consideration in this two-step approach is statistical modeling. Although prediction models were used in this paper to identify components, covariate balancing methods (e.g., based on the propensity score) may be more appropriate in a confirmatory study given advantages with respect to design and inference (32,33).

The current study is not without potential limitations. One important consideration is that the success of the screening process relies on identifying relevant component characteristics. That is, the process will be compromised if some attributes that contribute to survival of the device are not accounted for. However, in this study we included all known clinically relevant attributes. The contrary may be a concern as well, too many attributes were considered, resulting in a parsing of components that is too fine, which may have hindered detection. One possibility to address this issue is a post hoc pooling of results across device components. Another possible solution is to combine data from several registries throughout the world that have information on the same device components.

There is a need to use effective methods in early identification of hip arthroplasty device components with increased risk of revision. Future research may be directed at applying the proposed methods to different arthroplasty devices (e.g., knee, shoulder) as well as other multi-

component medical devices in which many different components can be interchanged (e.g., spinal fusion). Early identification of harmful devices using these methods can lead to fewer patients at risk of receiving these devices.

TABLES

Table 1. Patient Characteristics

Characteristic	Mean (SD)
Age	65.5 (11.2)
BMI	29.3 (5.7)
	n (%)
Race	
Caucasian	59005 (79.2)
African American	6132 (8.2)
Asian Pacific Islander	2795 (3.8)
Hispanic	5921 (7.9)
Native American	179 (0.2)
Multi-Race	488 (0.7)
Female	42782 (57.4)
Diabetes diagnosis	16468 (22.1)
Osteoarthritis diagnosis	69453 (93.2)
ASA score $\geq$ 3	25295 (33.9)

Note: BMI= Body Mass Index; ASA= American Society of Anesthesiologists

Table 2. Results for Recalled Components

	Descriptive Information			Regularized/ Unregularized Cox Model	Random Forest	
Component	Number of Implants	Number of Events	Average Follow-up Years	*P*-value	Rank	*P*-value
FDA Recalled Components						
ASR shell/head (>36 mm)	649	186	6.0	<0.001*	1	<0.001
Durom shell/ Metasul head (>36 mm)	159	23	6.3	<0.001*	4	0.192
Rejuvenate	71	36	3.3	<0.001*	2	<0.001
ABG II	3	1	3.3	-	23	0.347
Modular SMF	47	1	3.8	-	-	-
Alumina head (32 mm)	290	9	7.5	0.652	140	0.420
Non-Recalled Components[†]						
Pinnacle Liner (15 degree hood)[‡]	2745	54	5.8	<0.001*	44	0.389
Pinnacle Liner (0 degree hood)[‡]	27327	491	3.4	<0.001*	3	0.146

Note: Regularized Cox model selected 103 components. *P*-values reported for regularized/unregularized Cox model approach are based on a Wald test from an unregularized Cox model with selected components and patient covariates. A * indicates that the *P*-value < 0.05 after an adjustment that maintains the false discovery rate at 0.05. The rank column is based on a rank of minimal depth values. Ranks closer to zero denote smaller minimal depths that indicate stronger variable effects. In the case of the regularized/unregularized Cox model approach "-" denotes that the device feature was not selected therefore no *P*-value is provided. In the case of the random forest approach "-" denotes that the device feature was not included in any trees in the forest, therefore no rank or *P*-value is provided. [†]Reporting only components that are statistically significant after adjustment that maintains the false discovery rate at 0.05. [‡]The acetabular liner includes two models, Marathon and ALTRX.

FIGURES

Figure 1. Flow Chart Depicting Study Inclusion Criteria

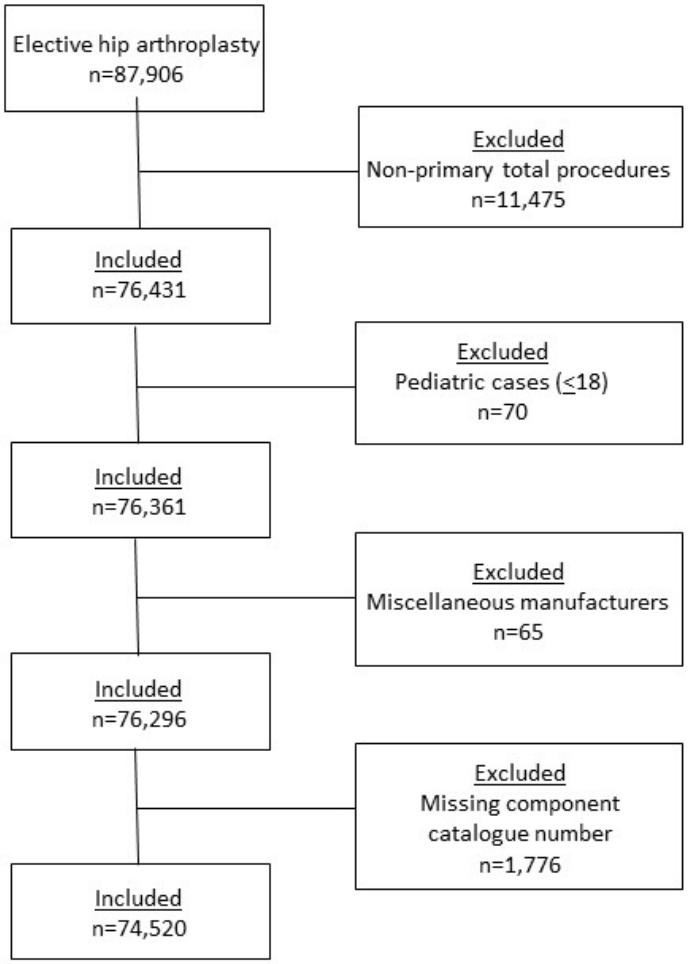


Figure 2. Time-to- First Revision Surgery for the Sample

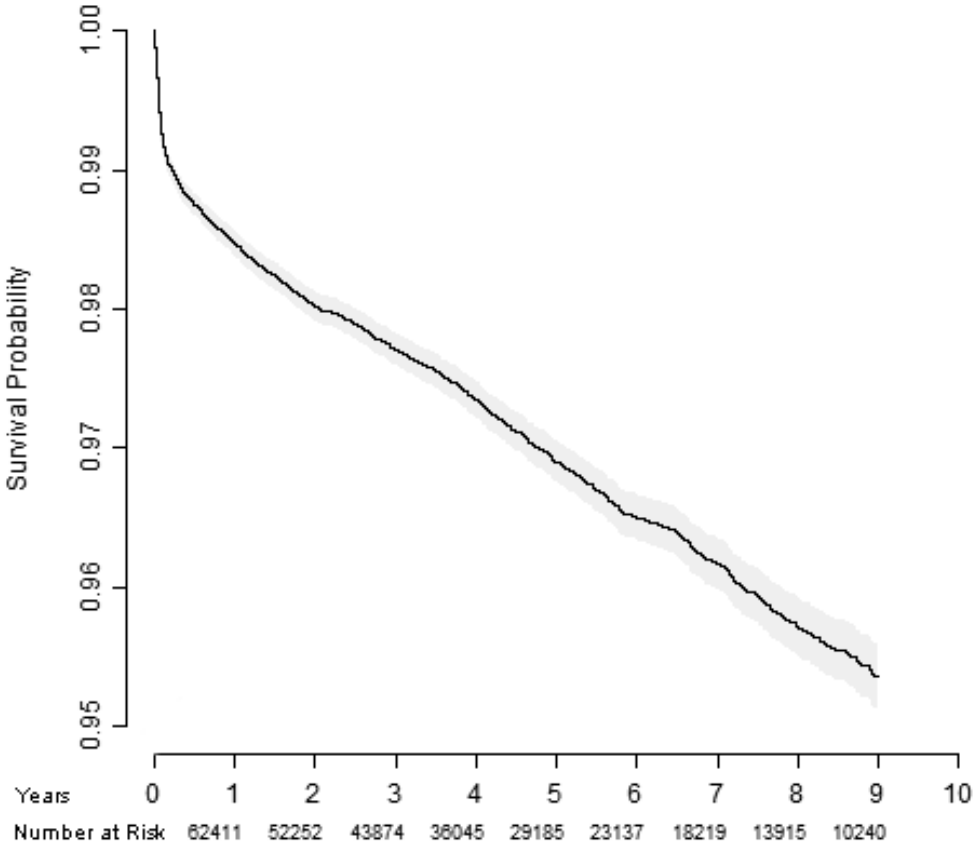
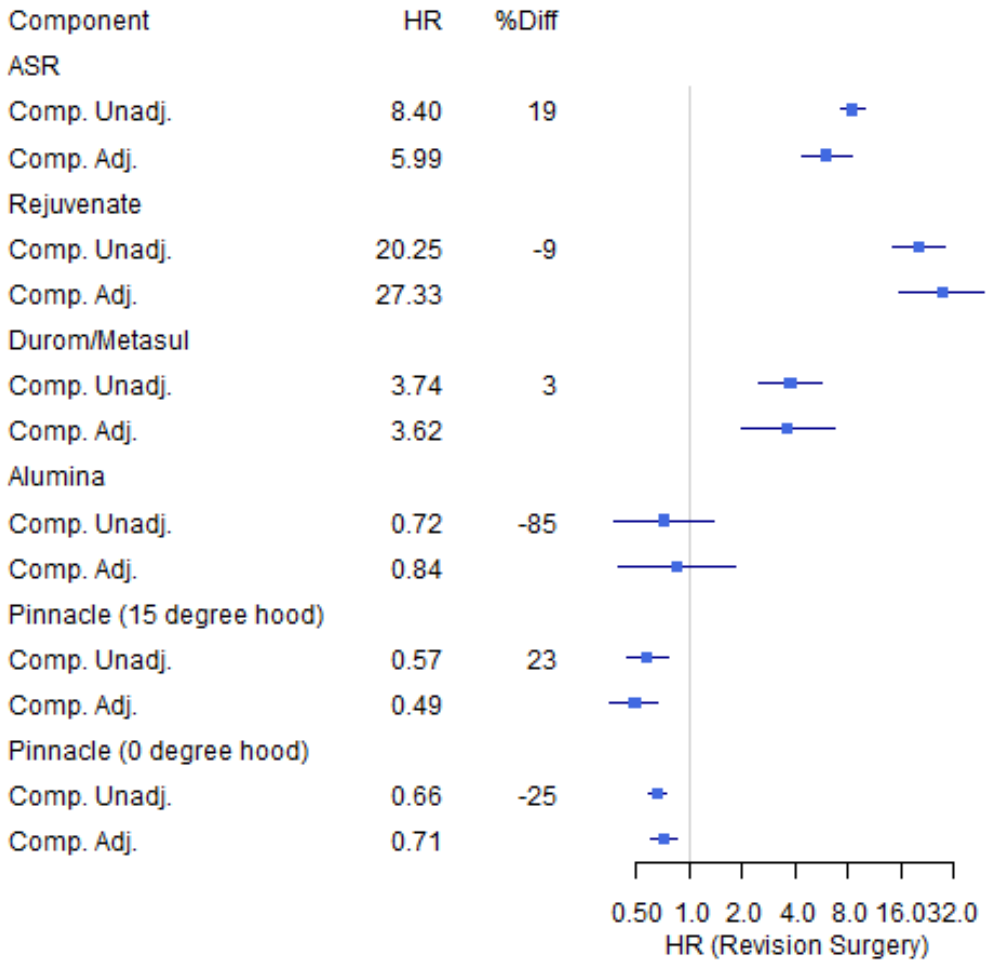


Figure 3. Hazard Ratio Comparison to Illustrate Component Confounding



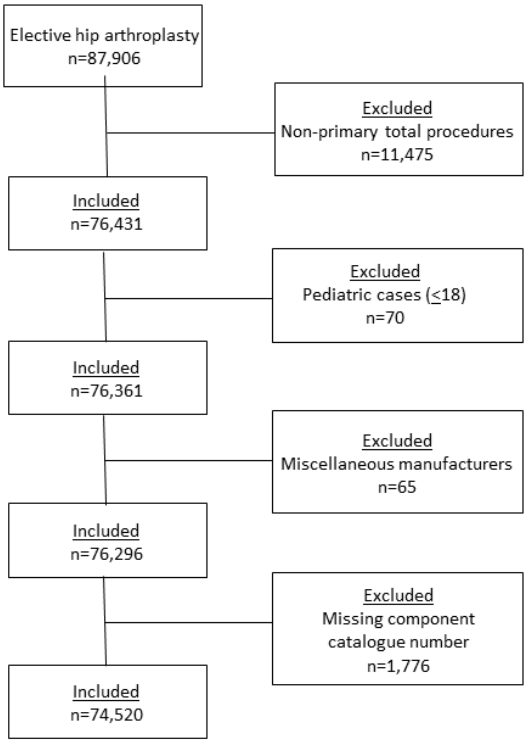
Note: HR= Hazard Ratio; Comp. Unadj.= Unadjusted for Component Confounding; Comp. Adj.=Adjusted for Component Confounding.

%Diff= $[\ln(HR_{\text{Comp. Unadj.}}) - \ln(HR_{\text{Comp. Adj.}})] / |\ln(HR_{\text{Comp. Adj.}})|$

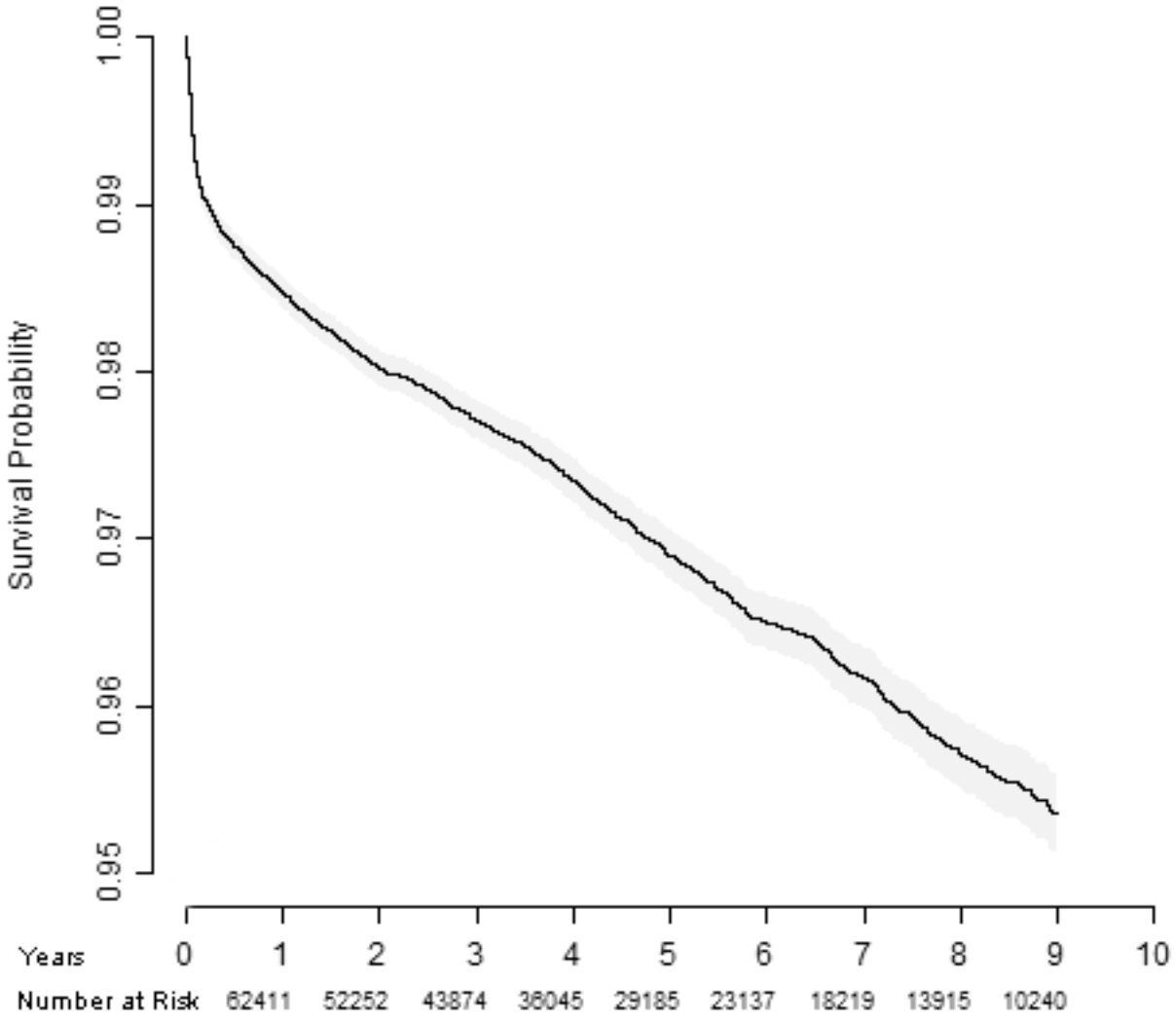
REFERENCES

1. Shah JS, Maisel WH. Recalls and safety alerts affecting automated external defibrillators. *JAMA.* 2006;296:655-60.
2. Maisel WH. Semper fidelis — consumer protection for patients with implanted medical devices. *N Engl J Med.* 2008;358:985-7.
3. Resnic FS, Normand SL. Postmarketing surveillance of medical devices — filling in the gaps. *N Engl J Med.* 2012;366:875-7.
4.Sedrakyan A. Metal-on-metal failures—in science, regulation, and policy. *Lancet.* 2012; 379:1174-6.
5. Centers for Disease Control and Prevention (CDC). *National Center for Health Statistics. National Hospital Discharge Survey: 2010 table, Procedures by selected patient characteristics - Number by procedure category and age*. CDC web site. http://www.cdc.gov/nchs/fastats/inpatient-surgery.htm. 2013.
6. Kurtz S, Ong K, Lau E, Mowat F, Halpern M. Projections of primary and revision hip and knee arthroplasty in the United States from 2005-2030. *J Bone Joint Surg Am.* 2007; 89: 780-5.
7. Anand R., Graves SE, De Steiger RN, Davidson DC, Ryan P, Miller LN et al. What is the benefit of introducing new hip and knee prostheses? *J Bone Joint Surg Am*. 2011; 93 Supple 3:51-4.
8. De Steiger RN, Hang JR, Miller LN, Graves SE, Davidson DC. Five-year results of the ASR XL acetabular system and the ASR hip resurfacing system: An analysis from the Australian Orthopaedic Association National Joint Replacement Registry. *J Bone Joint Surg Am*. 2011; 93:2287-93.
9. Graves SE, Rothwell A, Tucker K, Jacobs JJ, Sedrakyan A. A multinational assessment of metal-on-metal bearings in hip replacement. *J Bone Joint Surg Am.* 2011; 93 Suppl 3:43-7.
10. Smith AJ, Dieppe P, Vernon K, Porter M, Blom AW. Failure rates of stemmed metal-on-metal hip replacements: Analysis of data from the National Joint Registry of England and Wales. *Lancet.* 2012; 379: 1199-1204.
11. Smith AJ, Dieppe P, Howard PW, Blom AW; 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. *Lance*t. 2012;380:1759-66.
12. Ardaugh BM, Graves SE, Redberg RF. The 510(k) Ancestry of a metal-on-metal hip implant. *New Engl J Med.* 2013; 368: 97-100.
13. Graves SE, de Steiger R, Davidson D, Donnelly W, Rainbird S, Lorimer MF et al. The use of femoral stems with exchangeable necks in primary total hip arthroplasty increases the rate of revision. *Bone Joint J*. 2017; 99-B:766-73.
14. Krucoff MW, Sedrakyan A, Normand SL. Bridging unmet medical device ecosystem needs with strategically coordinated registries networks. *JAMA.* 2015;314:1691-92
15. Sedrakyan A, Campbell B, Graves S, Cronenwett JL. Surgical registries for advancing quality and device surveillance. *Lancet.* 2016;388:1358-60

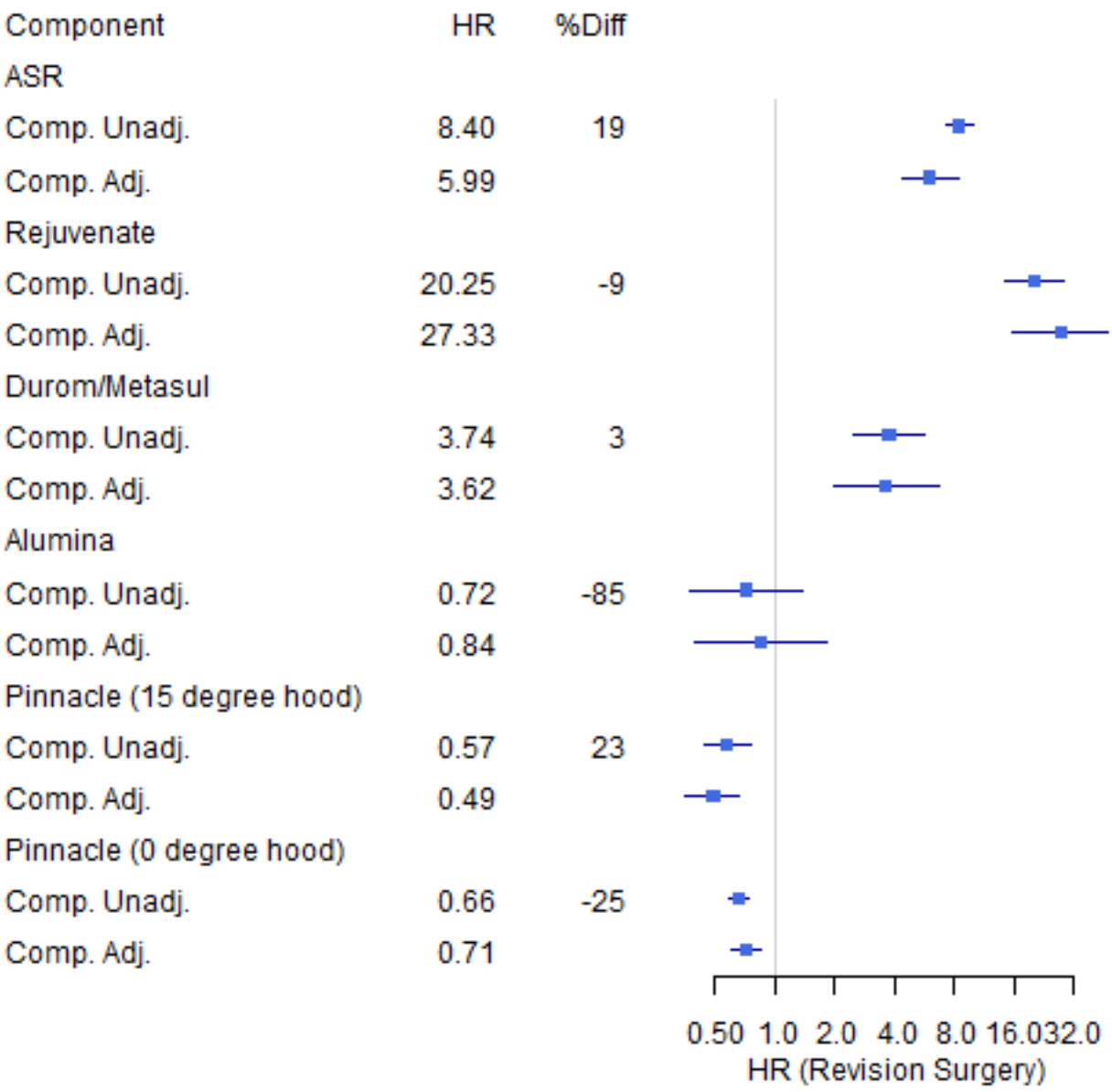
16. De Steiger RN, Miller LN, Davidson DC, Ryan P, Graves SE. Joint registry approach for identification of outlier prostheses. *Acta Orthop*. 2013;84:348-52.
17. Australian Orthopaedic Association National Joint Replacement Registry (AOANJRR). *Hip, knee & shoulder arthroplasty: 2017 annual report*. Adelaide: AOA, 2017.
18. Paxton E W, Inacio M C, Kiley M L. The Kaiser Permanente Implant Registries: effect on patient safety, quality improvement, cost effectiveness, and research opportunities. *Perm J*. 2012; 16: 36-44.
19. Paxton EW, Kiley ML, Love R, Barber TC, Funahashi TT, Inacio M C. Kaiser Permanente implant registries benefit patient safety, quality improvement, cost-effectiveness. *Jt Comm J Qual Patient Saf.* 2013; 39: 246-52.
20. Ishwaran H, Kogalur UB, Blackstone E, et al. Random survival forests. *Ann Appl Stat.* 2008; 2: 841-60.
21. Hastie T, Tibshirani R, Wainwright M. *Statistical learning with sparsity: The lasso and generalizations*. Boca Raton, FL: CRC Press; 2015.
22. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol.* 2005, 67: 301–20.
23. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw.* 2011; 39: 1-13.
24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol.* 1995; 57: 289–300.
25. Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High-dimensional variable selection for survival data. *J Am Stat Assoc* 2010; 105: 205-17.
26. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC Bioinformatics* 2008; 9: 307.
27. Altmann A, Tolosi L, Sander O, Lenqauer T. Permutation importance: A corrected feature importance measure. *Bioinformatics.* 2010; 26: 1340-47.
28. Calhoun P, Su X, Nunn M, Fan J. Constructing multivariate survival trees: The MST package for R. *J Stat Softw* 2018; *83*: 1 - 21.
29. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010; 33: 1-22.
30. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw*. 2011; 39: 1-13.
31. Stuart EA, Marcus SM, Horvitz-Lennon MV, Gibbons RD, Normand SLT. Using non-experimental data to estimate treatment effects. *Psychiatr Ann.* 2009; 39:414-51.
32. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med.* 2007; 26:20–36.
33. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Stat Med.* 2014; 33: 1242–58.



PDS_4882_F1.tif



PDS_4882_F2.tif



PDS_4882_F3.tif