Application of a Causal Discovery Algorithm to the Analysis of Arthroplasty Registry Data

Biomedical Engineering and Computational Biology
Volume 9: 1–9
© The Author(s) 2018
Reprints and permissions: sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1179597218756896
SAGE

Camden Cheek[1], Huiyong Zheng[2], Brian R Hallstrom[2] and Richard E Hughes[1,2,3]

[1]Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI, USA.
[2]Department of Orthopaedic Surgery, University of Michigan, Ann Arbor, MI, USA.
[3]Department of Industrial & Operations Engineering, University of Michigan, Ann Arbor, MI, USA.

ABSTRACT: Improving the quality of care for hip arthroplasty (replacement) patients requires the systematic evaluation of clinical performance of implants and the identification of "outlier" devices that have an especially high risk of reoperation ("revision"). Postmarket surveillance of arthroplasty implants, which rests on the analysis of large patient registries, has been effective in identifying outlier implants such as the ASR metal-on-metal hip resurfacing device that was recalled. Although identifying an implant as an outlier implies a causal relationship between the implant and revision risk, traditional signal detection methods use classical biostatistical methods. The field of probabilistic graphical modeling of causal relationships has developed tools for rigorous analysis of causal relationships in observational data. The purpose of this study was to evaluate one causal discovery algorithm (PC) to determine its suitability for hip arthroplasty implant signal detection. Simulated data were generated using distributions of patient and implant characteristics, and causal discovery was performed using the TETRAD software package. Two sizes of registries were simulated: (1) a statewide registry in Michigan and (2) a nationwide registry in the United Kingdom. The results showed that the algorithm performed better for the simulation of a large national registry. The conclusion is that the causal discovery algorithm used in this study may be a useful tool for implant signal detection for large arthroplasty registries; regional registries may only be able to only detect implants that perform especially poorly.

KEYWORDS: Causal discovery, probabilistic graphical models, arthroplasty, hip

RECEIVED: September 22, 2017. **ACCEPTED:** December 1, 2017.

TYPE: Review: Special Collection: Big Data Methods in Medicine

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Support for the Michigan Arthroplasty Registry Collaborative Quality Initiative is provided by Blue Cross and Blue Shield of Michigan and Blue Care Network as part of the BCBSM Value Partnerships program.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Richard E Hughes, Department of Orthopaedic Surgery, University of Michigan, 2003 BSRB, 109 Zina Pitcher Pl., Ann Arbor, MI 48104, USA. Email: rehughes@umich.edu

Introduction

Arthroplasty (joint replacement) is a very common and effective treatment for end-stage osteoarthritis of the hip. There were more than 370 000 primary total hip arthroplasty procedures performed during 2015 in the United States, and the number is projected to increase to more than 510 000 annually by 2020.[1] Unfortunately, there is wide variation in the risk of reoperation to replace prosthetic components ("revision") across implants: for uncemented implant designs the risk of revision at 10 years postoperatively ranges from 2.6% to 66.5%.[2] Because industry-sponsored studies are known to have bias,[3,4] regional and national patient registries are the best source of revision risk data. These registries are used to conduct postmarketing surveillance of implants.[5] For example, the ASR metal-on-metal hip implant was withdrawn from the market after it was shown to have an increased risk of revision by national arthroplasty registries.[6,7]

Arthroplasty registries seek to properly identify poorly performing implants ("signal detection") using classical biostatistical techniques.[5] Registry analyses of implant revision risk typically use revision as an end point and time to first revision (TTR) as the measure and then Kaplan-Meier estimates are computed.[6] To account for confounding that could occur by patient-level variables, Cox proportional hazards modeling is employed, including sex, age, and body mass index (BMI).

However, arthroplasty registries consist of observational data, and it is well known that analyses of observational data using classical statistical methods can produce misleading results.[7] This is a central problem because postmarketing surveillance implicitly seeks to identify a causal relationship between the choice of implant and revision risk. Although novel statistical methods[8–10] and multiregistry data sharing models[11] have recently been developed for orthopedic implant device postmarket surveillance, they do not address the causal inference problem that arises from the existence of confounding variables.

Statistical techniques for using directed acyclic graphs (DAGs) to model causal relationships have been developed in fields ranging from epidemiology[12–14] to artificial intelligence.[15] One thread of DAG modeling of causality has been casual discovery, which consists of algorithms to construct DAGs from empirical data that indicate causal relationships.[16] The TETRAD software package has been developed at Carnegie Mellon University for casual discovery[17,18] and it is freely available in open source format.

The purpose of this project was to test the ability of a causal discovery tool to properly identify causal relationships in arthroplasty registry data. The approach was to generate simulated data sets having known domain knowledge-based causal structure and apply the TETRAD software package to them. Two registry sizes were simulated to evaluate the effectiveness of TETRAD in national and regional registry settings.

Methods

To evaluate the performance of TETRAD, simulated data sets with a known causal structure were first generated. Each simulation consists of a number of cases, which represent a single patient along with its randomly generated sex, implant type, and TTR.

The PC causal discovery algorithm of Spirtes et al[16] was used to construct DAGs, which are graphs indicating causal structure, from the simulated data. Each node is a variable, and for each pair of nodes (A, B) there are 5 options for connection: (1) no edge indicating that there is no direct causal relationship, (2) a directed edge from A to B indicating that A directly causes B, (3) a directed edge from B to A indicating that B directly causes A, (4) an undirected edge indicating that there is a causal relationship between A and B but the direction cannot be determined from the data available, and (5) there may be a latent variable—or variables—confounding the relationship between A and B. Note that if there is a directed path from A to B, then A can be an indirect cause of B. Figure 1 illustrates the PC algorithm (a causal discovery algorithm named after its authors, Peter and Clark[16]) on a 4-node hypothetical DAG. Conceptually, the PC algorithm evaluates patterns in the data to determine whether they are consistent with hypothetical DAGs representing possible causal structures.

The causal structure used for the simulations (Figure 2) contained 3 variables: implant, sex, and TTR. The model represents the clinical situation where there is one implant of interest being analyzed (compared with cases performed with all implants other than the one of interest), and that implant affects TTR. In these simulations, the causal model also includes sex, which affects the selection of implant and TTR. This model was chosen as a simple example of a possible causal relationship between implant and TTR with a measured confounding variable (sex) that is a direct causal factor for both implant and TTR.

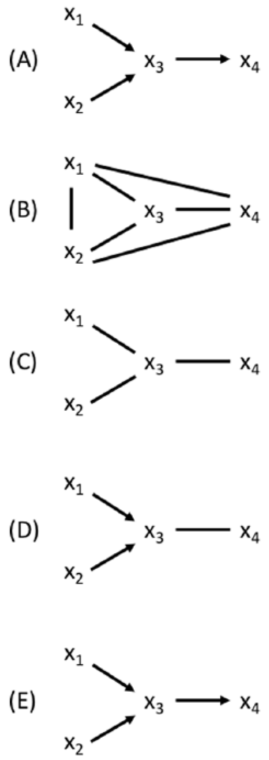

Figure 1. Example of causal discovery algorithm PC using 4 variables, x_1, x_3, x_3, and x_4, assuming that x_1 and x_2 are independent, x_1 and x_4 are conditionally independent given x_3, x_2, and x_4 are conditionally independent given x_3, x_1 is a cause of x_3, x_2 is a cause of x_3, and x_3 is a cause of x_4. This is illustrated as a DAG in (A). The algorithm has 3 steps. Step 1 consists of constructing a fully connected undirected graph (B). Step 2 consists of removing all edges where the data do not support a direct cause between 2 nodes. The edge between x_1 and x_2 is removed because they are independent. The edge between x_1 and x_4 is removed because x_1 and x_4 are conditionally independent given x_3. This means that if x_3 is fixed, any change in x_1 does not cause a change in x_4. Similarly, the edge between x_2 and x_4 is removed, resulting in the DAG shown in (C). Step 3 determines the direction of the edges remaining at the end of step 2. Directed edges from x_1 to x_3 and from x_2 to x_3 are determined because x_3 must be a collider node due to x_1 and x_2 being independent (the "collider test" for edge direction), resulting in (D). The "from collider test" is used to determine that the edge from x_3 to x_4 is directed from x_3 to x_4 (E). Observe that the PC algorithm reconstructed (E) from the statistical independence assumptions implied by the DAG in (A). DAG indicates directed acyclic graph.

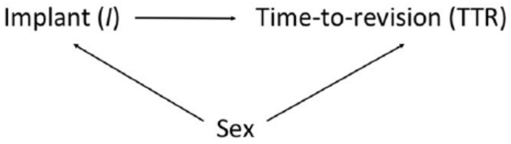

Figure 2. Causal diagram for simulated data. The simulation included 3 variables: (1) implant, (2) sex, and (3) time to first revision (TTR). Causal relationships are indicated by directed edges between nodes.

Simulated data sets were constructed using distributions from the Apache Commons Mathematics Library version 3.6.1,[19] using the included Mersenne Twister algorithm[20] seeded with the system time as a source of pseudo-randomness. The variables were generated in sequence according to the topologic ordering of the DAG that represents the relationship between the variables. In this situation, the order was sex, implant type, and then TTR. The sex variable was a Boolean variable where 0 represented female and 1 represented male. Sex was generated from a simple Bernoulli distribution with parameter, $\hat{p}_S$, which was the proportion of cases which were male. Implant type was a Boolean variable where 1 was used if the case used the implant to be analyzed as a potential outlier device (the "implant of interest"), and 0 was used to represent a case that used any implant other than the implant of interest. Implant type was generated from a Bernoulli distribution dependent on sex, where the parameter $\hat{p}_{I,F}$ was used if the case was female and $p_{I,M}$ was used if the case was male. For the sake of simplicity, in this initial exploration, it was assumed that every case had a revision surgery, so for every case, TTR was generated with a Weibull distribution $\text{Weibull}(\alpha,\beta)$ with shape parameter α and scale parameter β. Both distribution parameters were dependent on sex and implant type where $\alpha_{F,I}$ and $\beta_{F,I}$ were used when the case was female and the implant of interest was used, $\alpha_{F,\sim I}$ and $\beta_{F,\sim I}$ were used when the case was female and an implant *other* than *I* was used (~*I* representing "not the implant of interest"), $\alpha_{M,I}$ and $\beta_{M,I}$ were used when the case was male and implant of interest was used, and $\alpha_{M,\sim I}$ and $\beta_{M,\sim I}$ were used when the case was male with an implant other than *I*. Each simulation generated a set of N_R cases from the same set of parameters, and each simulation was repeated 1000 times. To simplify the concept of effect size, the variable E_I represented the effect size of the implant on TTR. It was used to calculate $\beta_{M,I}$ and $\beta_{F,I}$ as shown in the following equations:

$$\beta_{M,I} = \frac{\beta_{M,\sim I}}{E_I}$$

$$\beta_{F,I} = \frac{\beta_{F,\sim I}}{E_I}$$

The simulations were run on a machine with a 3.3-GHz AMD Ryzen 5 1400 Quad-Core Processor at a rate of approximately 10.6 simulations per second. We verified that the numerical simulations converge over multiple runs. The generated simulations were analyzed using the open source TETRAD (version 6.4.0) software package implemented in Java. We used the PC algorithm and the χ^2 independence test with a significance level of .05 to reconstruct the causal graphs from the simulation data. For each set of simulations, we calculated the proportion of graphs that contained each edge. We also calculated the proportion of those edges that were undirected, directed correctly, and directed incorrectly.

Data from the Michigan Arthroplasty Registry Collaborative Quality Initiative (MARCQI) were used to estimate baseline model parameters. The MARCQI is a consortium of 61 hospitals working to improve the quality of care for hip and knee arthroplasty patients in Michigan.[21] It collects data on devices implanted in patients as well as the dates of primary and revision cases. Therefore, TTR data were available. A total of 47 664 total hip arthroplasty cases were used to estimate parameters using SAS version 9.4 (SAS Institute, Cary, NC, USA).

The choice of parameters used in the simulation was based on data from MARCQI. The proportion of males (p_s) was fixed at the MARCQI average of 0.45. The proportions of cases with the implant of interest ($p_{I,F}$ and $p_{I,M}$) were varied between 0.0 and 0.3, a range that encompasses the implants found in the MARCQI data set. The TTR parameters $\alpha_{F,\sim I}$, $\alpha_{M,\sim I}$, $\beta_{F,\sim I}$, and $\beta_{M,\sim I}$ were fixed at the MARCQI average of 0.71, 0.71, 182.0, and 171.0, respectively. Because of the lack of variation seen in the Weibull shape parameter between implants in the MARCQI data, $\alpha_{F,I}$ and $\alpha_{M,I}$ were also fixed at 0.71. The effect size E_I, which was defined by the ratio $\beta_{M,\sim I} : \beta_{M,I}$, was varied around a baseline of 2.0, the size commonly used by registries for outlier detection.[5]

The simulations were run with the number of cases N_R set at 799 and 20 863, representative of the number of revised cases in a regional and national arthroplasty registry, respectively. The size of a regional registry was based on the MARCQI; the national registry was based on the National Joint Registry of England, Wales, Northern Ireland, and the Isle of Man (NJR).[22]

TETRAD and traditional Cox proportional hazards modeling were applied to real patient-level data from MARCQI. An implant was selected for analysis that appeared—based on inspection of the raw Kaplan-Meier curve—to have an increased risk of revision at 3 years following surgery.[23] There were 1452 primary cases out of 47 599 using this implant combination in the MARCQI database. Cox modeling was performed using TTR as an outcome, and implant (of interest or not of interest), sex, age, and BMI were included as predictors.

Results

With a data set size representative of smaller regional registries such as MARCQI (~800 revised cases), TETRAD was only able to consistently identify an edge between implant and TTR (TTR) when the effect size was large, and even with a large effect size of 4.0 and a $p_{I,M}$ and $p_{I,F}$ of 0.04, only 75% of reconstructed graphs identified a causal relationship between implant and TTR. However, with a data set representative of a large national registry such as NJR (~20 300 revised cases), an edge was identified even with very small effect sizes, exceeding 95% identification with an effect size of 2.0 for both $p_{I,M}$ and $p_{I,F}$ of 0.02 and $p_{I,M}$ and $p_{I,F}$ of 0.04 (Figure 3). False positives (detection of implant to TTR edge at an effect size of 1.0) occurred at a rate of 2.9%, 4.4%, 5.7%, and 4.7% in Figure 3A to D, respectively. These were consistent with the significance level of .05 set for the independence tests. To confirm the results with no effect of sex on implant, the false-positive rate and true-positive rate were calculated with various effect sizes and rates of implant of interest use. The results confirm a false-positive rate consistent with a significance level of .05 (Table 1).

To detect a causal relationship in 95% of the reconstructed graphs, 19 500 revised cases were needed for an effect size of 2.0 with $p_{I,M}$ and $p_{I,F}$ fixed at 0.02. When $p_{I,M}$ and $p_{I,F}$ were increased to 0.04, only 9000 cases were needed (Figure 4). This difference in number of revised cases needed to detect the proper relationship indicates that edge detection is sensitive to changes in $p_{I,M}$ and $p_{I,F}$.

The algorithm was sensitive to changes in $p_{I,M}$ and $p_{I,F}$ (Figures 5 and 6). With the regional-sized data set, detection was low at effect sizes of 1.5 and 2.0 (Figure 5A and B). Even at a large effect size of 4.0, the algorithm failed when $p_{I,M}$ and $p_{I,F}$ were both very small (Figure 5C). With the

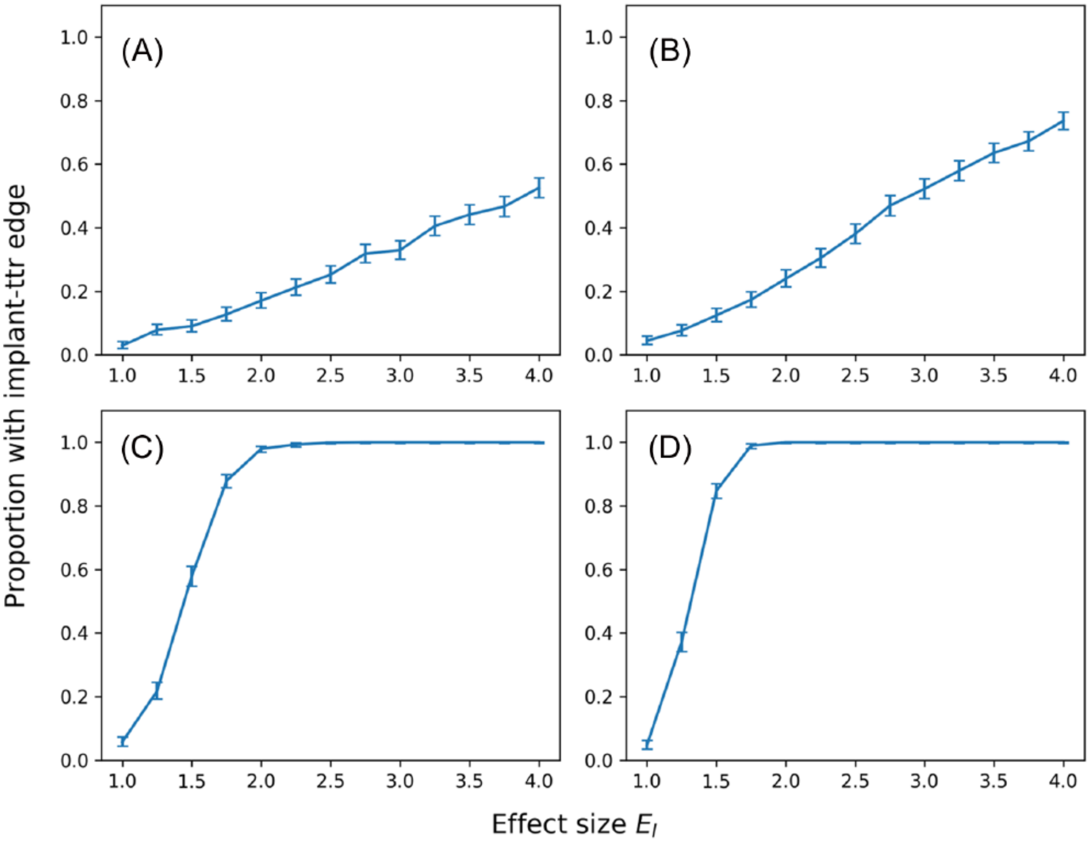

Figure 3. Proportion of reconstructed graphs with an edge between implant and TTR for a given effect size. Figure panels 3A and B use a registry size of 799 revised cases, and figure panels 3C and D use a registry size of 20 800 revised cases. Figure panels 3A and 3C use a $p_{I,F}$ and $p_{I,M}$ of 0.02, and figure panels 3B and 3D use a $p_{I,F}$ and $p_{I,M}$ of 0.04. Error bars represent 95% confidence interval. TTR indicates time to first revision.

national-sized data set, the algorithm performs extremely well at an effect size of 2.0 and performs sufficiently at an effect size of 1.5 when $\hat{p}_{I,M}$ and $\hat{p}_{I,F}$ were not small (Figure 6A to C).

The ability of the algorithm to detect direction was inconsistent and highly sensitive to errors in connectivity. When the reconstructed graph was fully connected, no directionality was inferred. In the simulations where one of the edges was undetected, the algorithm attributed a causal direction to the other edges, but often in the incorrect orientation. Specifically, when attempting to establish a causal relationship between implant and TTR, the performance is highly variable, and even a large effect size and a high usage of the implant of interest do little to improve the performance (Table 2).

Analysis of real patient data from MARCQI showed that the PC algorithm found an edge between implant of interest and TTR, whereas the Cox model did not indicate a statistically significant association between these 2 variables.

Discussion

This study aimed to provide a preliminary evaluation of the viability of using the TETRAD software package to aid in identifying causal relationships in arthroplasty registry data. The software performed well with data sets on the scale of large, national registries, but its ability to identify relationships was notably reduced when using data set sizes representative of regional registries such as MARCQI. The ability of PC algorithm implemented in TETRAD to determine direction of causation was unreliable and highly sensitive to any errors in the identified graph structure for simulations of both regional- and national-sized registries. Consequently, domain knowledge is required during postprocessing to produce the correct direction of the edge connecting implant to TTR.

The primary objective of study for this article was the identification of a properly directed edge between implant and TTR because that is necessary for postmarket surveillance outlier detection. The other edges were explored, but they performed as expected and were not included in the results for 3 reasons: (1) they did not display interesting variation between simulations, (2) the parameters that would cause interesting variation if changed do not vary considerably in our data set, and (3) they are not of primary interest to arthroplasty registry signal detection activities. The simulation parameters varied were chosen because they have some effect on the ability of the algorithm to detect the implant to TTR edge, and we wished to determine at what ranges of parameters this algorithm perform satisfactorily. The analysis was done based on an effect size of 2.0 based on the outlier detection methods used by the Australian Orthopaedic Association National Joint Replacement Registry.[5] The performance of the algorithm was dependent on the size of each group of cases being analyzed. If there are too few cases that use the implant of interest, the independence tests will be unable to detect an effect on TTR even if the effect size is large.

Table 1. Rates of detection of the arc between implant and time to first revision for a range of both effect size and proportion of cases that use the implant of interest.

EFFECT SIZE	PROPORTION WITH IMPLANT OF INTEREST	FALSE-POSITIVE RATE (FPR)[a]	TRUE-POSITIVE RATE (TPR)[b]
1.0	0.01	0.0490	NA
1.0	0.05	0.0474	NA
1.0	0.10	0.0486	NA
1.5	0.01	NA	0.0801
1.5	0.05	NA	0.1338
1.5	0.10	NA	0.1908
2.0	0.01	NA	0.1198
2.0	0.05	NA	0.2830
2.0	0.10	NA	0.4187
2.5	0.01	NA	0.1891
2.5	0.05	NA	0.4346
2.5	0.10	NA	0.6528
3.0	0.01	NA	0.2442
3.0	0.05	NA	0.5839
3.0	0.10	NA	0.8278

Abbreviation: NA, not applicable.

[a] FPR is only applicable when the effect size is 1.

[b] TPR is only applicable when the effect size is greater than 1.

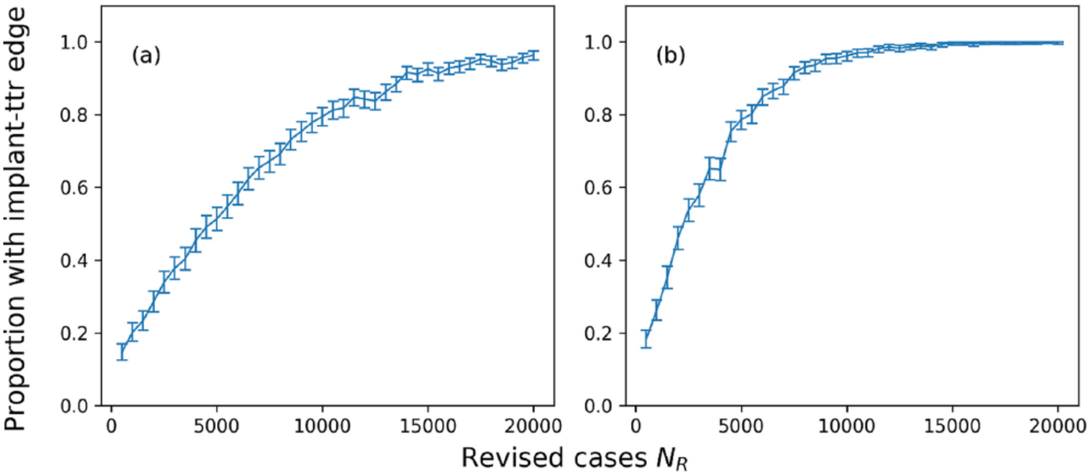

Figure 4. Proportion of reconstructed graphs with an edge between implant and TTR for a given number of revised cases. Both panels use an effect size (E_I) of 2.0. Figure panel 4A uses a $p_{I,F}$ and $p_{I,M}$ of 0.02 and figure panel 4B uses a $p_{I,F}$ and $p_{I,M}$ of 0.04. Error bars represent 95% confidence interval. TTR indicates time to first revision.

The number of cases that use the implant of interest depends on $\hat{p}_{I,M}$, $\hat{p}_{I,F}$, and N_R.

Although the inability to consistently determine the direction of the edge between implant and TTR is concerning, it does not disqualify TETRAD as a possible tool for causal discovery in arthroplasty because the directionality of the edge can be determined in postprocessing using logic. The PC algorithm prunes the fully connected DAG using statistical independence tests, and the result for a large registry is that the resulting undirected graph very likely has the implant-to-TTR edge correctly identified. The absence of direction is a result of the later steps in the PC algorithm that rely on analysis of 3-node chains as colliders or noncolliders (Figure 1). The very small DAG used for these simulations, which only had 3 nodes,

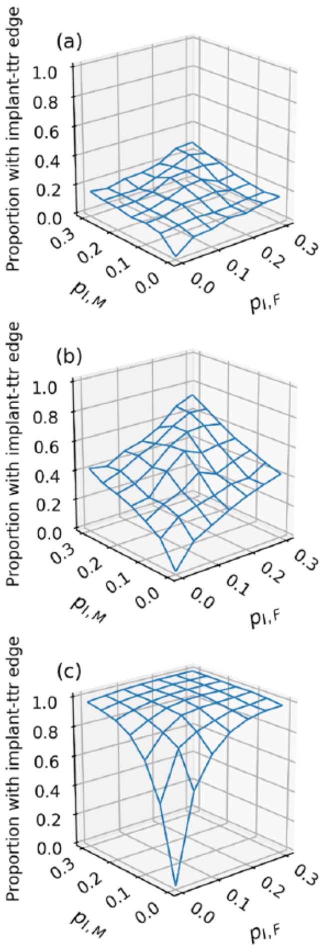

Figure 5. Proportion of reconstructed graphs with an edge between implant and TTR for a given $p_{I,F}$ and $p_{I,M}$. Number of revised cases is set at 799 for all panels. Effect size (E_I) is set at 1.5, 2.0, and 4.0 in figure panels 5A to C, respectively. TTR indicates time to first revision.

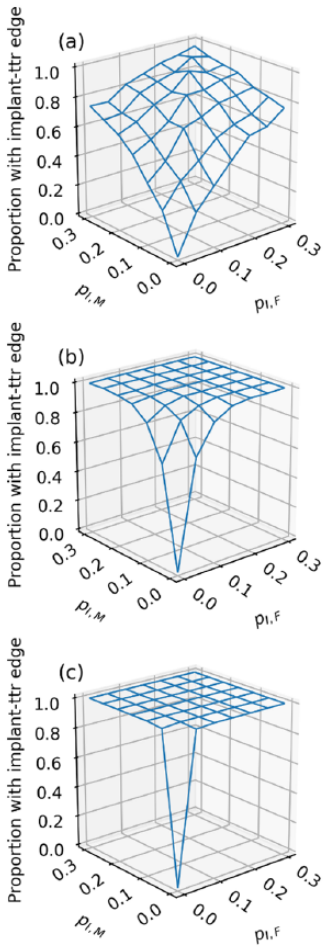

Figure 6. Proportion of reconstructed graphs with an edge between implant and TTR for a given $p_{I,F}$ and $p_{I,M}$. Number of revised cases is set at 20 863 for all panels. Effect size (E_I is set at 1.25, 1.5, and 2.0 in figure panels 6A to C, respectively). TTR indicates time to first revision.

will not provide the chains necessary for the edge directing portions of the PC algorithm. Larger DAGs that include age, BMI, surgeon volume, etc, are likely to provide DAGs that provide better directionality of edges. This should be an area of further investigation. However, even for the small DAGs analyzed here, the final step of causal inference can be done manually, making the method useful in practical settings. This is done by recognizing that TTR cannot cause implant because the choice of implant is made prior to the primary surgery, which is by definition prior to the revision event represented by the TTR variable. If the PC algorithm implemented in TETRAD returns an edge between implant and TTR, then the only logical direction can be from implant to TTR. Unfortunately, this domain knowledge cannot be implemented as a constraint in the PC algorithm; specifying the direction of this edge must be done as postprocessing following application of the PC algorithm.

With a large data set, the algorithm has the potential to be a useful tool in identifying a causal relationship between implant and TTR, but for smaller data sets such as MARCQI, there are too few revised cases to reliably detect an effect. For an N_R representative of a large, national registry such as NJR, the algorithm performed satisfactorily with $\hat{p}_{I,M}$ and $\hat{p}_{I,F}$ set at both 0.02 and 0.04 (Figure 3C and D). For the simulation of the smaller registry, however, the algorithm did not perform well for small values of $\hat{p}_{I,M}$ and $\hat{p}_{I,F}$ (Figure 3A and B) but it did achieve high detection ability when these parameters were larger (Figure 5C).

However, with a large data set, the algorithm reliably detects an edge down to an effect size of 2.0, even with a small $\hat{p}_{I,M}$ and $\hat{p}_{I,F}$ (Figure 6). Based on these results, for this algorithm to be useful for identifying a relationship between implant and TTR, a sufficiently large number of revised cases using the implant of interest are necessary for determining a causal relationship between implant and TTR. In our simulations, that number was approximately 300 cases but that number is expected to be larger with data sets where not every case has a TTR.

The ASR metal-on-metal implant experience was used to motivate this study, so it is reasonable to ask whether the PC algorithm would be able to identify an edge for an implant having the same clinical outcomes as the ASR at the time it was recalled from the market. That is, could the PC algorithm have identified it as an outlier device in 2010? That year was when data from the NJR prompted the voluntary recall of the device

Table 2. Rates of detection and direction for an arc between implant and TTR for a range of both effect size and proportion of cases that use the implant of interest.

EFFECT SIZE	PROPORTION WITH IMPLANT OF INTEREST	PROPORTION WITH ARC[a]	PROPORTION DIRECTED[b]	PROPORTION ORIENTED CORRECTLY[c]
1.0	0.01	0.0490	0.0022	0.7727
1.0	0.05	0.0474	0.0006	0.3333
1.0	0.10	0.0486	0.0020	0.6000
1.5	0.01	0.0801	0.0022	0.5000
1.5	0.05	0.1338	0.0048	0.5000
1.5	0.10	0.1908	0.0076	0.4342
2.0	0.01	0.1198	0.0045	0.8222
2.0	0.05	0.2830	0.0132	0.5909
2.0	0.10	0.4187	0.0193	0.5233
2.5	0.01	0.1891	0.0085	0.6353
2.5	0.05	0.4346	0.0194	0.5825
2.5	0.10	0.6528	0.0276	0.5797
3.0	0.01	0.2442	0.0105	0.6095
3.0	0.05	0.5839	0.0243	0.5556
3.0	0.10	0.8278	0.0423	0.5816

Abbreviation: TTR, time to first revision.
[a]Proportion of simulations that identified an arc between implant and TTR.
[b]Simulations that identified an arc, proportion that identified a direction to the arc.
[c]Simulations that identified a direction, proportion that identified the direction in the correct orientation.

after several years of data from Australia suggesting the same.[24] At that time, the 5-year revision risk of the ASR was 13% according to the NJR. For comparison, the overall revision risk at 5 years reported in the NJR 2010 annual report was 2.9%.[25] This results in a ratio of ASR revision risk to overall revision risk of 4.5 (because the 2.9% value includes ASRs, the actual ratio of ASR to non-ASR devices is likely larger than 4.5). Our simulation results indicate that both large national registry and smaller regional registry would have a high likelihood of detecting an edge between implant and TTR, provided the device was highly used in the regional registry. Using the knowledge that implant choice precedes revision can direct the edge; therefore, the causal connection between the ASR and revision could likely have been determined using the PC algorithm. This supports the possible utility of the PC algorithm in implant postmarket surveillance.

The most obvious comparison is with methods conventionally used for postmarketing surveillance of medical devices. More specifically, this method should be evaluated in the context of methods used for postmarketing surveillance of hip replacement implants. The Australian Orthopaedic Association National Joint Replacement Registry, which first identified the ASR resurfacing hip as an outlier,[24] uses a rigorous process of postmarketing surveillance and outlier detection.[5] That process begins with computing revision risks. A device that has twice the risk compared with all others in the class is selected for additional Cox proportional hazards modeling to adjust for age and sex. Ranstam et al[6,26,27] have made recommendations on statistical methods for analyzing implant data in arthroplasty registries, and they also focus on Cox models. Although these existing methods have been very useful, they lack the theoretical power of graphical causal model analysis for assessing causal relationship between implant and revision risk. Published analyses of revision risks associated with implant design features typically use multivariate modeling without reference to graphical methods for rigorously analyzing the problem causal inference, even though the ultimate purpose is to draw causal inference conclusions about the relationship between implant design and revision risk.[28–39] Many of these studies use extensive and careful modeling and propensity score matching to mitigate bias,[31,34,40] but they differ from this study because they do not describe formally employing concepts of d-separation and casual inference.[15] This is concerning because it has been shown by analysis of empirical data[7] and computer simulation[16] that regression modeling methods can produce results inconsistent with known causality.

The primary strength of this study was that it used a combination of real-world arthroplasty registry data and computer simulation. The simulation paradigm allowed for the causal discovery method to be evaluated against known causal relationships and effect magnitudes. Having known relationships enabled an analysis of whether the method could accurately determine the true causal relationship. Moreover, the statistical distributions and parameters used for the simulations were obtained by fitting models to data obtained from the MARCQI. Use of actual distributions strengthens the simulation results. Finally, a strength of this study was that it grew organically out of the work of the MARCQI device committee, which conducts statistical evaluations of revision risk by implant. Therefore, the selection of variables and the range to simulate over were based on actual postmarketing surveillance activities of hip replacement implants in the United States.

This study had 4 primary limitations: (1) it focuses on TTR outcome, (2) it used a simplified model of TTR, (3) it examined a limited number of causal relationships, and (4) no censoring was included. Joint replacement patients have many outcomes, including pain, range of motion, and avoidance of negative adverse events (death, revision, venous thromboembolism, etc). Revision is only one end point, and this is a limitation of the analysis. However, revision is the clinical end point that is virtually universally used in hip and knee arthroplasty registries. Because of the ubiquity of this outcome in the analysis of large arthroplasty registry quality improvement data sets, it makes sense to focus on revision. This study simplified revision to a random number assigned to each case. In actual registry data, not every patient has an associated TTR due to study length and censorship. This allowed us to run the simulation data directly through the causal discovery algorithm without interpolating missing values, but it also made the simulation data more poorly represent the registry data. Because implant selection is a comparatively easy factor to modify, quality improvement in hip and knee arthroplasty needs methods for determining causal relationships between implant and TTR. Conceptually, there are many factors that may confound statistical associations found between implant and TTR, including age, BMI, surgeon volume, hospital volume, etc. Although the focus on sex is well justified in the literature, it is only a first step in understanding the utility of causal discovery in hip arthroplasty implant analysis. The no censoring limitation likely has the biggest impact on registries that have not been around for a very long time because revisions often occur years after the primary surgery. Thus, it is more likely to affect a new registry such as MARCQI than an established registry such as the Australian Orthopaedic Association National Joint Replacement Registry.

We applied the PC algorithm to MARCQI data for a single hip stem/cup combination and found that the PC algorithm and Cox model produced different results regarding an association between implant and TTR. When the MARCQI data including age and BMI was used, the PC algorithm identified a number of relationships, including a directed arc from TTR to BMI. When we applied the PC algorithm to MARCQI data with only the 3 variables used in the simulations, the unused variables act as unmeasured noise variables as we previously found that there are relationships between them. As expected, the PC algorithm still identified an arc between implant and TTR. However, the Cox model did not indicate a statistically significant association between implant and TTR. Because we do not know the true causal relationship for this implant combination and TTR, we cannot conclude that the PC algorithm led to a false positive or the Cox model produced a false negative. However, this result does suggest that it is possible that the PC algorithm may detect outlier devices with higher sensitivity than Cox modeling.

This work prompts additional study into using causal discovery algorithms for arthroplasty registries. One potential area of further research includes improved modeling of TTR data. One of the limitations of this study is its simplified model of revision. Future work may implement a more complex model of TTR, such as semiparametric Bayesian survival analysis.[41] Additional work needs to be conducted to investigate other relevant clinical variables and the sensitivity of the results to unmeasured or noise variables. A potential future investigation may include the analysis of the MARCQI data set with the FCI (Fast Causal Inference) algorithm, which does not operate under the assumption that all variables are measured, which would be useful in a case as described above with MARCQI.

In conclusion, this study introduced and evaluated a novel tool for use in postmarketing surveillance of hip replacement implants conducted by large patient registries. The innovation lies in the fact that causal discovery algorithms have not—to our knowledge—been employed to analyze orthopedic implant data. The results indicated that it may be useful to identify edges in a DAG representing causality, but it is not nearly as good at identifying the causal direction of the edge. The method should continue to be investigated to determine how it performs in the presence of additional clinical variables. Finally, predictive modeling of implant revision risk is a long-term goal of this research. The identification of causal relationships is an important part of such modeling. Therefore, this work serves as a first step toward predictive analytics in arthroplasty.

Author Contributions

CC conducted computer simulations and wrote draft of manuscript; HZ analyzed MARCQI data and provided statistical insight to reporting error rates; BH gave clinical input on causal model to simulate; RH conceptualized, designed, and interpreted the study. All co-authors contributed to the writing of the manuscript and approved it.

Acknowledgements

Although Blue Cross Blue Shield of Michigan and the Michigan Arthroplasty Registry Collaborative Quality Initiative work collaboratively, the opinions, beliefs, and viewpoints expressed by the author do not necessarily reflect the opinions, beliefs, and viewpoints of BCBSM or any of its employees.

REFERENCES

1. Kurtz SM, Ong KL, Lau E, Bozic KJ. Impact of the economic downturn on total joint replacement demand in the United States: updated projections to 2021. *J Bone Joint Surg Am*. 2014;96:624–630.
2. Hughes RE, Batra A, Hallstrom BR. Arthroplasty registries around the world: valuable sources of hip implant revision risk data. *Curr Rev Musculoskel Med*. 2017;10:240–252.

3. Labek G, Neumann D, Agreiter M, Schuh R, Böhler N. Impact of implant developers on published outcome and reproducibility of cohort-based clinical studies in arthroplasty. *J Bone Joint Surg Am*. 2011;93:55–61.
4. Labek G, Sekyra K, Pawelka W, et al. Outcome and reproducibility of data concerning the oxford unicompartmental knee arthroplasty: a structured literature review including arthroplasty registry data. *Acta Orthop*. 2011;82:131–135.
5. de Steiger RN, Miller LN, Davidson DC, Ryan P, Graves SE. Joint registry approach for identification of outlier prostheses. *Acta Orthop*. 2013;84:348–352.
6. Ranstam J, Robertsson O. Statistical analysis of arthroplasty register data. *Acta Orthop*. 2010;81:10–14.
7. Ejima K, Li P, Smith DL Jr, et al. Observational research rigour alone does not justify causal inference. *Eur J Clin Invest*. 2016;46:985–993.
8. Normand S-L, Marinac-Dabic D, Sedrakyan A, et al. Rethinking analytical strategies for surveillance of medical devices: the case of hip arthroplasty. *Med Care*. 2010;48:S58–S67.
9. Cafri G, Banerjee S, Sedrakyan A, et al. Meta-analysis of survival curve data using distributed health data networks: application to hip arthroplasty studies of the international consortium of orthopaedic registries. *Res Synt Met*. 2015;6:347–356.
10. Cafri G, Fan J. Between-within effects in survival models with cross-classified clustering: application to the evaluation of the effectiveness of medical devices. *Stat Met Med Res*. 2016;27:312–319.
11. Banerjee S, Cafri G, Isaacs AJ, et al. A distributed health data network analysis of survival outcomes: the International Consortium of Orthopaedic Registries perspective. *J Bone Joint Surg Am*. 2014;96:7–11.
12. Robins JM. Data, design, and background knowledge in etiologic inference. *Epidemiology*. 2001;11:313–320.
13. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10:37–48.
14. Greenland S, Brumback B. An overview of relations among causal modelling methods. *Int J Epidemiol*. 2002;31:1030–1037.
15. Pearl J. *Causality: Models, Reasoning, and Inference*. New York, NY: Cambridge University Press; 2009.
16. Spirtes PC, Glymour C, Scheines R. *Causation, Prediction, and Search*. London, England: MIT; 2000.
17. Glymour C, Scheines R, Spirtes P, et al. TETRAD: discovering causal structure. *Multivar Behav Res*. 1988;23:279–280.
18. Scheines R, Spirtes P, Glymour C, Meek C, Richardson T. The TETRAD project: constraint based aids to causal model specification. *Multivar Behav Res*. 1988;33:65–117.
19. Apache Commons. Commons math: the apache commons mathematics library; 2017. https://commons.apache.org/proper/commons-math/.
20. Matsumoto M, Nishimura T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans Model Comp Simulat*. 1998;8:3–30.
21. Hughes RE, Hallstrom BR, Cowen ME, et al. Michigan arthroplasty registry collaborative quality initiative (MARCQI) as a model for regional registries in the United States. *Orthopaed Res Rev*. 2015;7:47–56.
22. National Joint Registry for England, Wales, Northern Ireland, and the Isle of Man 13th annual report, 2016. http://www.njrcentre.org.uk/njrcentre/Portals/0/Documents/England/Reports/13th%20Annual%20Report/07950%20NJR%20Annual%20Report%202016%20ONLINE%20REPORT.pdf.
23. Hughes RE, Hallstrom BR, Zheng H, et al. Five-year report of the Michigan arthroplasty registry collaborative quality initiative (MARCQI). Ann Arbor, MI: Michigan Arthroplasty Registry Collaborative Quality Initiative; 2017. http://marcqi.org/dev/wp-content/uploads/2017/11/MARCQI_Five-Year_Report_Nov2017.pdf. Accessed November 4, 2017.
24. de Steiger RN, Hang JR, Miller LN, et al. 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;93:2287–2293.
25. National Joint Registry for England and Wales, 7th annual report, 2010. http://www.njrcentre.org.uk/NjrCentre/Portals/0/NJR%207th%20Annual%20Report%202010.pdf.
26. Ranstam J, Karrholm J, Pulkkinen P, et al; NARA Study Group. Statistical analysis of arthroplasty data, I: introduction and background. *Acta Orthopaed*. 2011;82:253–257.
27. Ranstam J, Karrholm J, Pulkkinen P, et al; NARA Study Group. Statistical analysis of arthroplasty data, II: guidelines. *Acta Orthopaed*. 2011;82:258–267.
28. Namba R, Graves S, Robertsson O, et al. International comparative evaluation of knee replacement with fixed or mobile non-posterior-stabilized implants. *J Bone Joint Surg Am*. 2014;96:52–58.
29. Comfort T, Baste V, Froufe MA, et al. International comparative evaluation of fixed-bearing non-posterior-stabilized and posterior-stabilized total knee replacements. *J Bone Joint Surg Am*. 2014;96:65–72.
30. Graves S, Sedrakyan A, Baste V, et al. International comparative evaluation of knee replacement with fixed or mobile-bearing posterior-stabilized prostheses. *J Bone Joint Surg Am*. 2014;96:59–64.
31. Inacio MC, Cafri G, Paxton EW, Kurtz SM, Namba RS. Alternative bearings in total knee arthroplasty: risk of early revision compared to traditional bearings, an analysis of 62,177 primary cases. *Acta Orthopaed*. 2013;84:145–152.
32. Namba RS, Inacio MC, Cafri G. Increased risk of revision for high flexion total knee replacement with thicker tibial liners. *Bone Joint J*. 2014;96-B:217–223.
33. Namba RS, Inacio MC, Paxton EW, et al. Risk of revision for fixed versus mobile-bearing primary total knee replacements. *J Bone Joint Surg Am*. 2012;94:1929–1935.
34. Paxton EW, Inacio MC, Kurtz S, Love R, Cafri G, Namba RS. Is there a difference in total knee arthroplasty risk of revision in highly crosslinked versus conventional polyethylene. *Clin Orthopaed Rel Res*. 2015;473:999–1008.
35. Allepuz A, Havelin L, Barber T, et al. Effect of femoral head size on metal-on-HXLPE hip arthroplasty outcome in a combined analysis of six national and regional registries. *J Bone Joint Surg Am*. 2014;96:12–18.
36. 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;96:25–33.
37. Hooper GJ, Rothwell AG, Stringer M, Frampton C. Revision following cemented and uncemented primary total hip replacement: a seven-year analysis from the New Zealand Joint Registry. *J Bone Joint Surg Br*. 2009;91:451–458.
38. Jameson SS, Baker PN, Mason J, et al. The design of the acetabular component and size of the femoral head influence the risk of revision following 34 721 single-brand cemented hip replacements: a retrospective cohort study of medium-term data from a National Joint Registry. *J Bone Joint Surg Br*. 2012;94:1611–1617.
39. Kostensalo I, Junnila M, Virolainen P, et al. Effect of femoral head size on risk of revision for dislocation after total hip arthroplasty: a population-based analysis of 42,379 primary procedures from the Finnish Arthroplasty Register. *Acta Orthopaed*. 2013;84:342–347.
40. Cafri G, Paxton EW, Love R, Bini SA, Kurtz SM. Is there a difference in revision risk between metal and ceramic heads on highly crosslinked polyethylene liners. *Clin Orthopaed Relat Res*. 2016;475:1349–1355.
41. Ibrahim JG, Chen M-H, Sinha D. *Bayesian Survival Analysis*. New York, NY: Springer; 2001.

Appendix 1

Notation

E_I effect size of the implant on TTR, which is defined by the ratio of $\beta_{M,\sim I} : \beta_{M,I}$.

N_R number of cases in each of 1000 simulated data sets. It can vary according to scale of study, eg, state-level or national-level sizes.

p_s the proportion of cases which are male

$p_{I,F}$ the proportion of cases which are female using the implant of interest

$p_{I,M}$ the proportion of cases which are male using the implant of interest

TTR ~ Weibull(α, β) TTR follows Weibull distribution with shape parameter α and scale parameter β. Furthermore, for different groups,

$\alpha_{F,I}$ and $\beta_{F,I}$ are for females using the implant of interest;

$\alpha_{M,I}$ and $\beta_{M,I}$ are for males using the implant of interest;

$\alpha_{F,\sim I}$ and $\beta_{F,\sim I}$ are for females using an implant other than the one of interest;

$\alpha_{M,\sim I}$ and $\beta_{M,\sim I}$ are for males using an implant other than the one of interest.