

Electronic Thesis and Dissertation Repository

6-27-2022 4:45 PM

Large-scale Analysis and Automated Detection of Trunnion Corrosion on Hip Arthroplasty Devices

Anastasia M. Codirezzi, *The University of Western Ontario*

Supervisor: Teeter, Matthew G., *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Engineering Science degree in Biomedical Engineering

© Anastasia M. Codirezzi 2022

Follow this and additional works at: <https://ir.lib.uwo.ca/etd>



Part of the [Biomedical Engineering and Bioengineering Commons](#)

Recommended Citation

Codirezzi, Anastasia M., "Large-scale Analysis and Automated Detection of Trunnion Corrosion on Hip Arthroplasty Devices" (2022). *Electronic Thesis and Dissertation Repository*. 8614.
<https://ir.lib.uwo.ca/etd/8614>

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlsadmin@uwo.ca.

Abstract

Corrosion at the modular head-neck taper interface of total and hemiarthroplasty hip implants (trunnionosis) is a cause of implant failure and thus a clinical concern. Patient and device factors contributing to the occurrence of trunnionosis have been investigated in prior implant retrieval studies. The Goldberg corrosion scoring method is considered the gold standard for observing trunnionosis, but it is labour-intensive. As a result, previous studies have generally looked at under 250 implants for analysis. The purpose of this thesis was to do a large-scale analysis of trunnionosis and explore its relationship to device and patient factors and compare to previously known trends from more limited studies. Additionally, it was to develop a tool using machine learning for rapid screening of implants to identify for further study in order to reduce the labour burden associated with implant retrieval studies.

Keywords

Hip arthroplasty, trunnionosis, corrosion, machine learning, convolutional neural networks

Summary for Lay Audience

Hip replacements are an increasingly common procedure for Canadians. Despite their safety and efficacy, sometimes the devices fail, requiring the patient to undergo an additional surgery to remove the original device. This accounts for ~10% of hip replacement surgeries done each year. Corrosion of the device has become an increasing reason for failure, and it is thought that corrosion is underreported. It is known that wear and corrosion of implants in the body may affect the tissue in the immediate area in a negative way and some patient factors may contribute to a more corrosive environment in the body. There is a need to study these retrieved devices to better understand potential patient factors that may contribute to increased rate of failure.

Goldberg scoring is a method used to observe corrosion at the taper interface for these devices. This method is labour-intensive and as a result, studies have generally looked at under 250 implants when studying their corrosion and the patient and device factors that may contribute to it. This thesis has done a 664-device study of implants and their corrosion and determined relationships between corrosion severity and patient and device factors and compared them to previously identified relationships in smaller studies. Additionally, it has developed a tool to distinguish no/mild corrosion from moderate/severe corrosion to allow for rapid screening of implants for further study, reducing the labour barrier for implant retrieval studies.

This thesis has provided the first large-scale study of retrieved hip arthroplasty devices and created a tool to make large-scale studies more accessible by reducing the labour required for early identification of devices with significant corrosion. The ability to conduct more large-scale studies allows for refinement of device design and identification of patients who may be at increased risk for corrosion of the taper. As hip arthroplasty surgeries continue to become more frequent, it is important to attempt to minimize their possible failure.

Co-Authorship Statement

The following thesis contains manuscripts intended for publication within scientific journals. Chapter 2 is an original manuscript entitled “What patient and implant factors affect trunnionosis severity? An implant retrieval study of 664 femoral stems”. The manuscript is co-authored by Anastasia M. Codirezzi, Matthew G. Teeter and Brent A. Lanting. In my role as MESc candidate, I participated in designing the study, performed the data collection, did the statistical analysis and wrote the manuscript text. Matthew Teeter and Brent Lanting, as the candidate’s supervisor and mentor, designed the study, reviewed the results and gave editorial assistance and provided mentorship.

Chapter 2 is an original manuscript entitled “A convolutional neural network for high throughput screening of trunnion corrosion”. The manuscript is co-authored by Anastasia M. Codirezzi, Matthew G. Teeter and Brent A. Lanting. In my role as MESc candidate, I participated in designing the study, performed the data collection, created the convolutional neural network, computed the associated metrics, and wrote the manuscript text. Christopher Del Balso advised only as the secondary observer for implant scoring. Matthew Teeter and Brent Lanting, as the candidate’s supervisor and mentor, designed the study, reviewed the results and gave editorial assistance and provided mentorship.

Acknowledgments

I'd like to start off by thanking my supervisor, Dr. Matthew Teeter, for his supervision and mentorship during my graduate studies. Dr. Teeter was there to guide me and have a meeting to chat, whether it be about my project, my career and goals, or recommending a local auto body repair shop.

I'd also like to thank my mentors and advisors, interdisciplinary projects take a lot of effort and care to advise meaningfully, and you all did. Dr. Brent Lanting, for giving me a window into the world of arthroplasty decision making and the history and lore of implants. Our meetings always left me feeling engaged, interested, and inspired by my project. Dr. Aaron Ward, for always being available to discuss machine learning and my project, as well as the realities of being a first-generation university student. It was good to have someone to talk about that with. Dr. Yolanda Hedberg, for her insight on corrosion and meticulous eye that helped me hammer down details of this project, its goals, and definitions within orthopaedics. I am grateful to have had such a strong advisory committee.

Next, I'd like to thank my lab mates in the Teeter lab, especially Jennifer and Shahnaz. Our pact to go into Robarts every day and just do one thing helped to keep me on track. Group meetings were always a highlight to my week and they reminded me all projects have their bumps in the road.

Finally, I'd like to thank my friends and family for their support during my university education. To my parents, Mary-Ann and Dominic, for their continued support even when I decided to switch fields and for always answering my calls and texts. My partner, Nyell, for always making me celebrate the small wins. My friends, especially Peggy, McKenna, Sam, Theresa, Julia, Sarah, Amgad, and Alex. You all supported me in different ways, but I knew I always had you to rely on.

Even when school was difficult, knowing I had my community behind me every step of the way kept me moving forward and I knew I was never truly alone. My community, both academic and personal, is what made this thesis possible.

Table of Contents

Abstract.....	ii
Summary for Lay Audience.....	iii
Co-Authorship Statement.....	iv
Acknowledgments.....	v
Table of Contents.....	vi
List of Tables.....	ix
List of Figures.....	xi
List of Appendices.....	xiii
Chapter 1.....	1
1.1 Osteoarthritis.....	1
1.2 The Hip.....	1
1.3 Hip Arthroplasty.....	2
1.4 Device Design.....	3
1.4.1 History of Device Design.....	3
1.4.2 Modern Device Design.....	4
1.4.3 The Acetabular Component.....	5
1.4.4 The Femoral Component.....	5
1.5 Revision, Wear, and Damage.....	9
1.5.1 Revision.....	9
1.5.2 Wear and Corrosion.....	10
1.5.3 The Bearing Surface.....	11
1.6 Implant Retrieval Studies.....	12
1.6.1 Scoring of Implant Damage.....	12
1.7 Machine Learning.....	14

1.7.1	Convolutional Neural Networks	15
1.7.2	Previous Attempts to Automate Goldberg Scoring	17
1.8	Thesis Objectives and Hypothesis	17
1.9	References.....	19
Chapter 2.....		25
2	What patient and implant factors affect trunnionosis severity? An implant retrieval study of 664 femoral stems	25
2.1	Introduction.....	26
2.2	Methods.....	27
2.2.1	Study Population.....	27
2.2.2	Visual Scoring.....	28
2.2.3	Statistical Analysis.....	29
2.3	Results.....	30
2.4	Discussion.....	36
2.5	Conclusion	41
2.6	References.....	42
Chapter 3.....		47
3	A convolutional neural network for high throughput screening of trunnion corrosion.....	47
3.1	Introduction.....	48
3.2	Methods.....	49
3.2.1	Implant Imaging and Visual Scoring	49
3.2.2	Data Curation	51
3.2.3	Neural Network Architecture.....	52
3.2.4	Network Training and Testing	54
3.3	Results.....	54
3.3.1	Imaging and corrosion scoring.....	54

3.3.2 Neural Network Training and Evaluation.....	55
3.4 Discussion.....	60
3.5 Conclusions.....	62
3.6 Acknowledgements.....	63
3.7 References.....	64
Chapter 4.....	69
4 General Discussions and Conclusions	69
4.1 Overview of Objectives	69
4.2 Summary of Results.....	69
4.3 Limitations	70
4.4 Applications and Future Directions	71
4.5 Conclusions.....	71
Appendices.....	73
Curriculum Vitae	74

List of Tables

Table 1: Reported wear rates for bearing surfaces.....	11
Table 2: Goldberg Corrosion scoring criteria	12
Table 3: Goldberg Fretting Criteria	14
Table 4: Goldberg scoring criteria [18]	29
Table 5: Implant characteristics. B Type 1: Mallory (n = 9), Taperloc (n = 6), Integral (n = 2) from Zimmer-Biomet (Warsaw, IN). C-taper: ODC (n = 1), Omnifit (n = 43), Restoration (n = 2), Secur-fit (n = 16)) from Stryker (Mahwah, NJ). D 11/13: S-ROM (n = 9) from DePuy Synthes (Raynham, MA). D 12/14: AML (n = 16), Corail (n = 42), Endurance (n = 19), Prodigy (n = 2), Reclaim (n = 7), Response (n = 1), Solution (n = 12), Summit (n = 29) from DePuy Synthes (Raynham, MA). D 14/16: Solution (n = 5), AML (n = 2), CML (n = 1), from DePuy Synthes (Raynham, MA). PCA Taper: PCA (n = 40), Precision (n = 3), Strata (n = 1), from Stryker (Mahwah, NJ). S 10/12: Richards Modular (n = 1) from Smith & Nephew (Memphis, TN). S 12/14: Anthology (n = 4), Conquest (n = 13). CPCS (n = 7), Echelon (n = 24), Polarstem (n = 8), Redapt (n = 4), SL plus (n = 1), SMF (n = 2), Spectron (n = 32), Synergy (n = 70) from Smith & Nephew (Memphis, TN). S 14/16: Biofit (n = 1), TriWedge (n = 1) from Smith & Nephew (Memphis, TN). V40: ABG (n = 3), Accolade (n = 26), Definition PM (n = 1), Exeter (n = 47), GRMS (n = 1), Precision (n = 3), Rejuvenate (n = 32), Restoration (n = 14) from Stryker (Mahwah, NJ). W 12/14: Profemur (n = 7), Gladiator (n = 2) from Stryker (Mahwah, NJ). Z 12/14: Advocate (n = 1), Apollo (n = 1), CLS (n = 1), CPT (n = 1), M/L taper (n = 31), MS30 (n = 1), Versys (n = 44) from Zimmer-Biomet (Warsaw, IN). Z 6 degree: Harris (n = 8), Versys (n =2) from Zimmer-Biomet (Warsaw, IN). This table reflects modern ownership of the taper designs and companies.....	31
Table 6: Ordinal logistic regression results	32
Table 7: Previous implant retrieval studies.....	36
Table 8: Goldberg scoring criteria [5]	51

Table 9: Description of the different datasets	52
Table 10: Datasets and images included	55
Table 11: Computed error metrics for each dataset	59

List of Figures

Figure 1: a) Natural hip anatomy, b) hemiarthroplasty, c) total hip arthroplasty 2

Figure 2: Sketch of the Charnley implant 3

Figure 3: Modern Total Hip Arthroplasty Design, showing a modular stem and a metal-on-polyethylene bearing surface. 4

Figure 4: Modular stem (left) beside a dual modular stem (right) showing modularity at the base of the neck..... 5

Figure 5: A variety of available stems, showcasing major design differences in the distal stem component of the device. 6

Figure 6: a) Zimmer Biomet Type-1, b) DePuy 12/14, c) Stryker PCA, d) Stryker C-taper, e) Stryker V40, f) Zimmer Biomet 12/14 7

Figure 7: Exeter stems showcasing two different offsets. 8

Figure 8: Different head sizes, from left to right: 22mm, 26mm, 28mm, 32mm, 36mm. 8

Figure 9: (left) Hemiarthroplasty device utilising a large head, (right) hemiarthroplasty device utilizing a head within a head with a polyethylene liner between..... 9

Figure 10: Example of a CNN for classifying a trunnion. The first two blocks extract features that can be used to identify different features (shapes, colour, brightness, etc.) and how they can be associated with the different classes. The knowledge from the last block moves forward to the next, then all the information is gathered together, and the system makes a statistical guess of the class. Image adapted from [40]..... 16

Figure 11: Study design 27

Figure 12: Representative images of each class (a) class 1, (b) class 2, (c) class 3, (d) class 4 28

Figure 13: Study design for implant inclusion..... 49

Figure 14: Representative images of each class (a) class 1, (b) class 2, (c) class 3, (d) class 4 50

Figure 15: Convolutional neural network architecture. The neural network is comprised of an input of 900x900 images with RGB colouring. It then has a convolutional layer with a filter size of 3x3x32. Batch normalization was then employed. Then had a rectified linear unit activation function. That led to a max pooling layer with a filter size of 2x2 and a stride of 1,1. Then it had the fully connected layer, SoftMax function, and the output layer. 53

Figure 16: Confusion matrix for dataset 1 (class 1 versus class 2 versus class 3 versus class 4). Left is all images, right is with glare removed. The blue diagonal shows correct classifications (predicted class matches the true class) while the orange off-diagonal shows misclassifications. Intensity of colour is based off count in each category. 56

Figure 17: Confusion matrix for dataset 2, class 1 versus class 2,3,4. Right is all images, left is with glare removed. The blue diagonal shows correct classifications (predicted class matches the true class) while the orange off-diagonal shows misclassifications. Intensity of colour is based off count in each category. 56

Figure 18: Confusion matrix for dataset 3, C12 versus C34. Right is all images, left is images with glare removed. The blue diagonal shows correct classifications (predicted class matches the true class) while the orange off-diagonal shows misclassifications. Intensity of colour is based off count in each category. 57

Figure 19: ROC for all datasets, with all images included 58

List of Appendices

Appendix A: Study Approvals.....	73
----------------------------------	----

Chapter 1

1.1 Osteoarthritis

Osteoarthritis is a degenerative disease that leads to the breakdown of joint cartilage and the underlying bone and may have associated synovial inflammation. [1]. Risk factors for osteoarthritis include prior injury, occupation, joint misalignment/deformity, muscle weakness, lifestyle, sex, obesity, and genetics [2]. Although all risk factors contribute, genetic factors have been found to be strong determinants of the disease [3]. Osteoarthritis is possible in any joint but is most commonly in weight bearing joints such as the knees, hips, big toes, and spine, as well as hands.

Osteoarthritis is a debilitating disease. Approximately 25% of patients with osteoarthritis cannot perform their usually daily activities and 80% have limitations in movement [4]. It is the most common form of arthritis, currently affecting 4.6 million people in Canada and is projected to affect 10 million people in Canada within the next 30 years [5]. Joint-specific symptoms include joint pain, stiffness, swelling, crepitus (creaking/grinding noise) and instability [2].

Early detection of osteoarthritis allows for conservative management of the disease, but the disease cannot be reversed and will continue to progress. Osteoarthritis is confirmed through physical examination and x-ray imaging is used to grade the disease. The most common grading system used is Kellgren and Lawrence, where grade 0 is normal/non-diseased, and grade 4 is end-stage. Grade 4 shows a complete loss of joint space and a “bone on bone” appearance [6].

1.2 The Hip

The hip joint (acetabulofemoral joint) is one of the most common sites of osteoarthritis. It is a synovial ball and socket joint, with a large articulating surface created by the head of the femur and the acetabulum of the pelvis and lined with hyaline articular cartilage. It is a weight bearing joint and its primary function is to support the weight of the body in static and dynamic postures [7].

Symptoms of osteoarthritis in the hip includes pain in the groin, pain that flares up with vigorous activity, locking of the joint with a grinding noise, and decreased range of motion. In imaging, osteoarthritis of the hip is indicated by a narrowing joint space (thinning of the cartilage), the presence of bone spurs, and damaged cartilage. By end-stage osteoarthritis, conservative management methods no longer relieve pain, and the only available treatment is joint replacement [2].

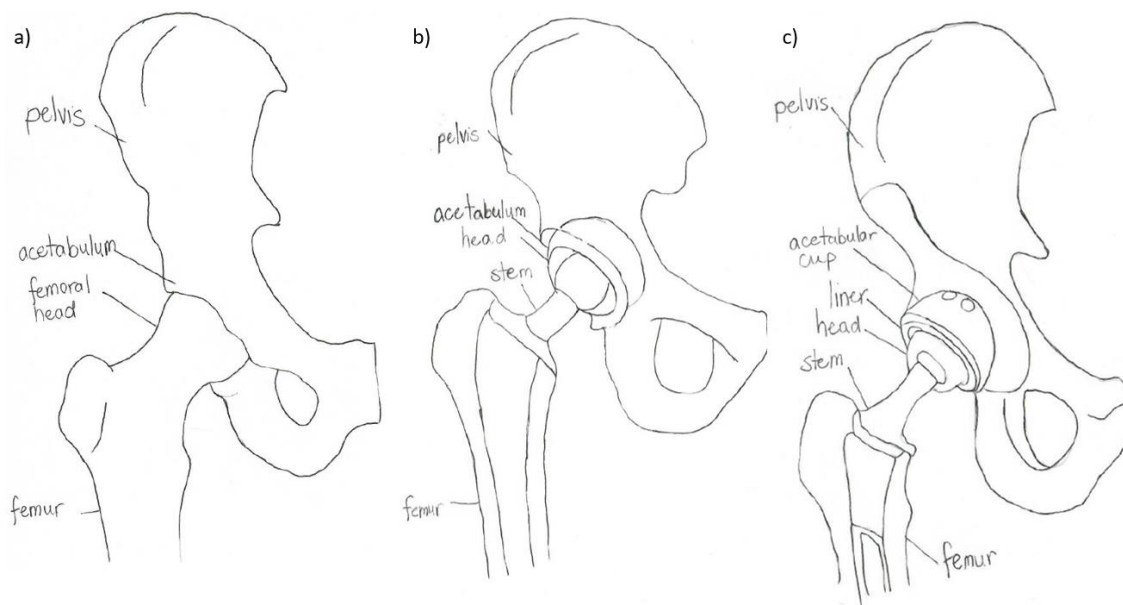


Figure 1: a) Natural hip anatomy, b) hemiarthroplasty, c) total hip arthroplasty

1.3 Hip Arthroplasty

Hip arthroplasty refers to surgery where part of or all of the hip joint is replaced by an artificial implant. Hemiarthroplasty refers to the replacement of half of the hip joint, where the femoral component is replaced with an artificial implant while natural hip socket remains. Total hip arthroplasty (THA) refers to the replacement of both the femoral and acetabular articulating surfaces with an artificial implant (Figure 1). This is most used as a treatment for end-stage degenerative arthritis (primarily osteoarthritis), but other reasons include trauma and hip dysplasia [8]. The goals of hip arthroplasty are to remove the diseased or defective joint and replace it to be as anatomically similar to the patient's original joint as possible [7].

In Canada, hip arthroplasty procedures have seen an upward trend- increasing 20.1% between 2015 and 2019, with a total of 62,016 surgeries reported in 2018 [9]. 81.3% of primary hip replacements performed in 2017 – 2018 were performed because of degenerative arthritis. It is expected that hip arthroplasty procedures will continue to increase, especially as osteoarthritis cases increase in Canada.

1.4 Device Design

1.4.1 History of Device Design

The earliest reported attempts at hip replacements occurred in Germany in 1891, but modern total hip arthroplasty as we know it today is associated with orthopaedic surgeon Sir John Charnley of the United Kingdom, with the first surgery completed in 1962. It consisted of three parts: a metal femoral stem, a polyethylene acetabular component, and an acrylic bone cement, a sketch of the device is shown in Figure 2 [10].

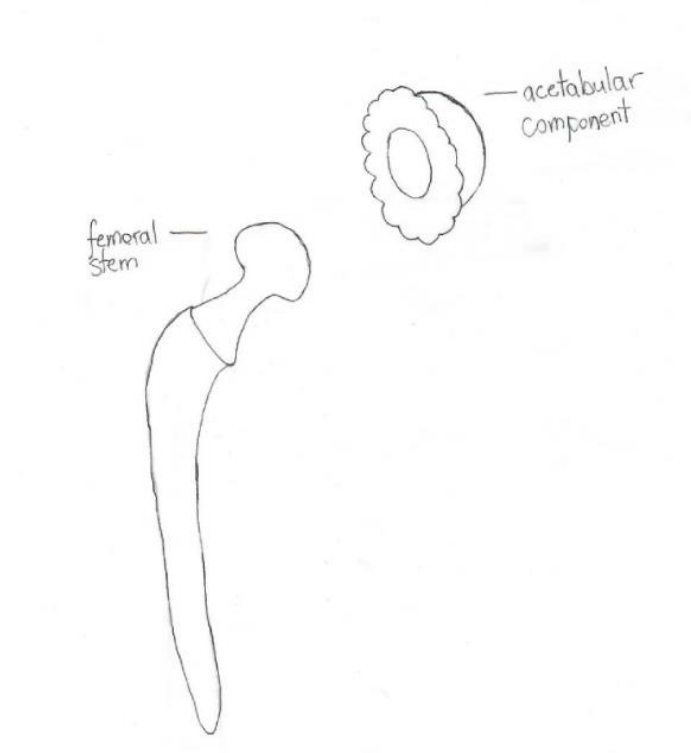


Figure 2: Sketch of the Charnley implant

This design is largely similar in concept to the total hip arthroplasty devices used today. Major design changes include advances in materials design and usage, geometry of the implants, segmentation of the stem into modular components, and the inclusion of the acetabular cup in the pelvis.

1.4.2 Modern Device Design

Today, most devices are now modular (stem and ball are separate pieces), with a metal-on-polyethylene design (Figure 3) being the most common in Canada [9]. According to the Canadian Agency for Drugs and Technologies in Health, these implants are considered the “gold standard” in total hip prosthesis [11].

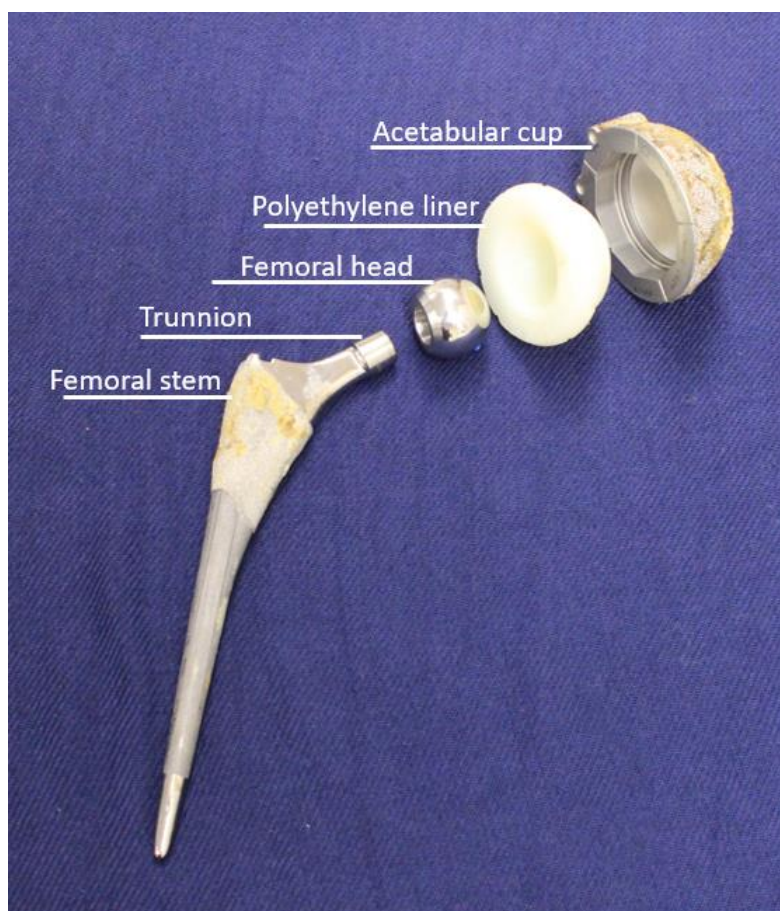


Figure 3: Modern Total Hip Arthroplasty Design, showing a modular stem and a metal-on-polyethylene bearing surface.

In THA, matching the patient anatomy as closely as possible to the original configuration allows for the best outcome in surgery. Different sizes and variations to match patient anatomy are available.

1.4.3 The Acetabular Component

Present in total hip arthroplasty, the acetabular component is comprised of the acetabular cup and the liner. Most modern cups are made from titanium or tantalum metals [12]. Liners may be made of polyurethane, ceramic, or metal, however polyethylene is the most commonly used material. The liners come in varying thicknesses to accommodate the size of the head used and there are different design types to help accommodate patient anatomy and to prevent dislocation [13].

1.4.4 The Femoral Component

In modular implants, the femoral component consists of the stem and the head. Stems are typically manufactured from stainless steel, cobalt-chromium alloys, or titanium alloys. There are two main versions of the stem, modular and dual modular. Modular refers to the stem and ball being separate components whereas dual modular typically refers to the ball, neck, and stem being separate components (Figure 4). Modularity can exist in other locations, but their use is typically reserved for implants in revision cases.



Figure 4: Modular stem (left) beside a dual modular stem (right) showing modularity at the base of the neck.

There is no one perfect stem design, as patient factors such as anatomy, bone quality and structure, and stature, all widely vary. There are two major places for variation: the distal stem design and the trunnion/taper design. Figure 5 shows just a few of the widely available stem models from different major manufacturers.



Figure 5: A variety of available stems, showcasing major design differences in the distal stem component of the device.

Taper geometry and design can widely vary even within manufactures, although most companies try to limit the taper geometries produced and sold simultaneously. Figure 6 shows variation in the taper design with popular implants. Taper design may change when the distal stem design has not, most notably when a company acquires a device from another company. An example of this is the Exeter stem which has had the same distal stem design since 1988 but saw a change in its taper geometry in 2000 after the stem was acquired by Stryker [14]. Taper design has been considered a significant factor for implant wear [15]. Machining lines (also called microgroove finish), parallel lines on the trunnion surface, are present on some taper designs, but this specific design factor has not been

found to be a significant factor for implant wear [16]. Figure 6 shows variation in taper designs, including machining lines.

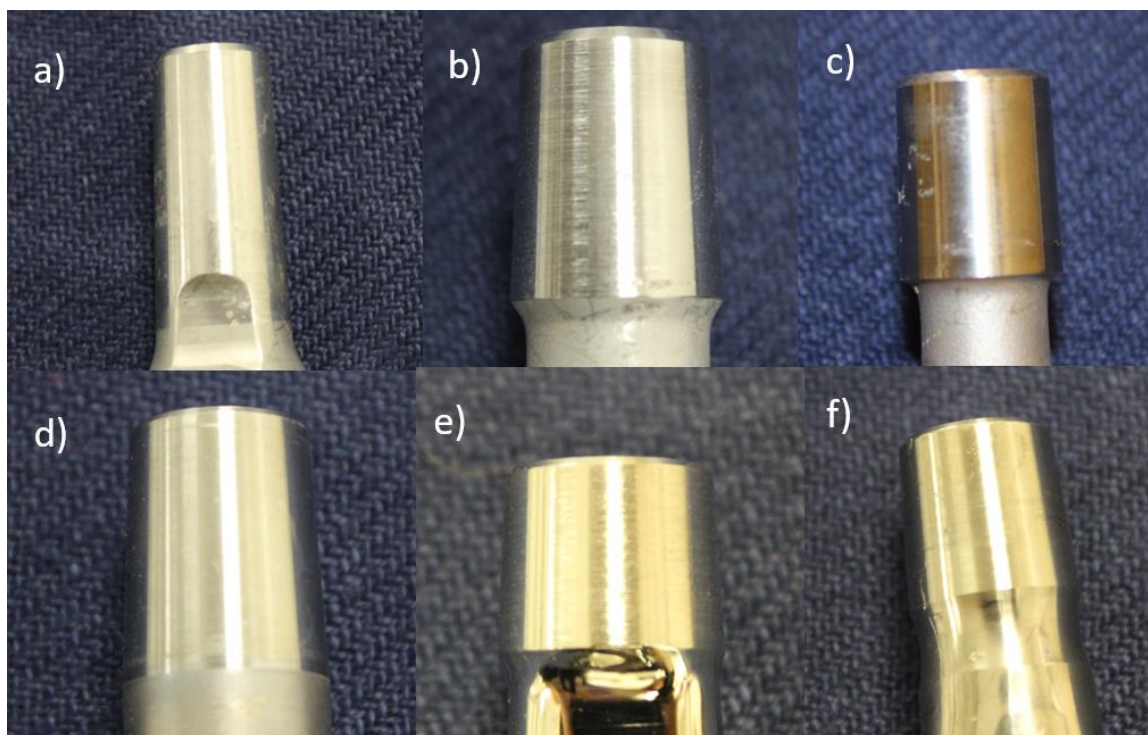


Figure 6: a) Zimmer Biomet Type-1, b) DePuy 12/14, c) Stryker PCA, d) Stryker C-taper, e) Stryker V40, f) Zimmer Biomet 12/14

Beyond the specific stem and taper design, there is a wide variety of different versions of the same stem with different sizes and offsets. Figure 7 shows the same stem model and two of its possible offset (distance from the center of the rotation of the femoral head to a line bisecting the long axis of the femur) offerings. Generally, the same stems are used for both hemi- and total hip arthroplasty, with a select number of stems being more common for usage in hemiarthroplasty.

The head is typically made of chromium cobalt or ceramic and all designs are spherical, but similar to the stem, there is variance in ball height and diameter, as well as trunnion designs. Similarly, they are available in different outer diameter sizes (Figure 8) and have an offset to match the stem selection.



Figure 7: Exeter stems showcasing two different offsets.



Figure 8: Different head sizes, from left to right: 22mm, 26mm, 28mm, 32mm, 36mm.

Total hip arthroplasty devices have the head fit into the acetabular component, while hemiarthroplasty devices make use of a larger head or a head within a larger head (Figure 9) with a liner between them and placed into the natural acetabulum.



Figure 9: (left) Hemiarthroplasty device utilising a large head, (right) hemiarthroplasty device utilizing a head within a head with a polyethylene liner between.

1.5 Revision, Wear, and Damage

1.5.1 Revision

Revision surgeries refer to an additional surgery that is done to correct the primary implant, often resulting in removal of part, or the entire implant. They account for 7.3% of all joint replacement surgeries done in Canada in 2018. Typically, these surgeries are 80% more costly than primary surgery due to the increased complexity as well as extended hospital stays for the patient [9]. Common reasons for revision include joint infection, aseptic loosening, and instability. The retrieved implant can hold insight into the *in vivo* implant behaviour, particularly when damage to the implant is observed.

1.5.2 Wear and Corrosion

Wear may be defined as a cumulative surface damage phenomena in which material is removed from a body in the form of small particles, primarily by mechanical processes [17]. Previously, mechanical wear of the polyethylene component was a major cause for implant revision, but with the introduction of cross-lined polyethylene (XLPE) this is widely considered a “solved problem.” Fretting is a specific type of wear that occurs as a result of small oscillatory motions between two surfaces. The modular head and trunnion create a space for this type of wear to occur. Another type of damage, corrosion, has also been of concern. Corrosion of the trunnion, trunnionosis, has been identified as a growing cause of THA failure [18]. Corrosion also creates wear debris that can trigger adverse soft-tissue reactions. Adverse tissue reactions are well documented in metal-on-metal hip replacements and had led to their mass revision and phasing out of their usage. There is an argument that corrosion-related soft-tissue reactions may be overlooked and misdiagnosed as recurrent instability or infection, which may lead to inadequate treatment of the issue [19].

Some examples of the damages commonly found on retrieved implants are: pitting and crevice corrosion, tribocorrosion, intergranular corrosion, and inflammatory cell-induced corrosion. Pitting and crevice corrosion refer to corrosion where the surface oxide is locally damaged leading to either pits or crevices on the material surface [20]. Tribocorrosion is when both corrosion and wear occur simultaneously; which has been reported for the trunnion [21]. Intergranular corrosion refers to corrosion that occurs at grain sites of the material. This is commonly seen in alloys, where materials have been combined [22]. Inflammatory cell-induced corrosion refers to a biological corrosion that is caused by inflammatory cells adhered to the metal surface [23]. Corrosion is observed visually as surface discolouration and more advanced forms are able to be visually observed, but not necessarily able to be distinguished from each other (with the exception of crevice and pitting)[24]. Knowing the mechanism of corrosion is important to identify the cause of the corrosion and employ appropriate methods to prevent it, however distinguishing between the mechanisms is difficult to do visually and often requires advanced analysis techniques.

These techniques are time consuming and cost-prohibitive, thus there is a need to screen for the presence of corrosion before employing these methods.

1.5.3 The Bearing Surface

The surface between the acetabular and femoral component, made up of the liner and the head, is the bearing surface. In total hip arthroplasty and hemiarthroplasty that uses a liner, this surface is susceptible to mechanical wear. There are four types of bearings that are studied and applied in THA: metal-on-polyethylene (MoP), metal-on-metal (MoM), ceramic-on-ceramic, and ceramic-on-polyethylene [25],[26]. Conventionally, the first material listed is with regards to the head and the second is the liner. Reported *in vitro* wear rates for popular materials of each bearing combination are included in Table 1.

Table 1: Reported wear rates for bearing surfaces.

Bearing Combination	Specific Material	Reported wear rate (mm ³ /Mc)
Metal-on-metal	CoCr-CoCr	0.60±0.18[27]
Metal-on-polyethylene	CoCr-XLPE	6.71±1.03[28]
Ceramic-on-polyethylene	CoCr-XLPE	4.09±0.64[29]
	Alumina-XLPE	3.35±0.29[30]
Ceramic-on-ceramic	Alumina-Alumina	0.74±1.73[31], 0.03[32]
	Zirconia-Zirconia	0.024[32], 0.06±0.004[31]

Although ceramic-on-ceramic and ceramic-on-polyethylene have the lowest wear rates, these bearing combinations have not been shown to be meaningfully better performing when looking at mid-term results of patients. The lack of evidence that they perform significantly better in patient context, and the increased cost of these combinations have prevented their widespread usage [33,34]. Metal-on-metal, a once popular bearing combination due to its cost effectiveness and low wear rate, has been associated with

adverse tissue reactions and metal ion release into the bloodstream [35]. Metal-on-polyethylene is the most widely used and it is considered the gold-standard in Canada. This bearing combination is both cost effective and has a low wear rate [11].

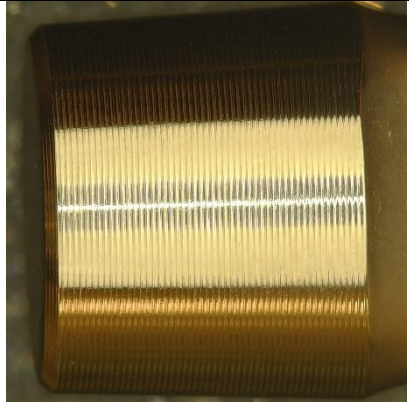
1.6 Implant Retrieval Studies

Analyzing implants retrieved gives insight into the behaviour of the implants *in vivo* that may have not been considered during clinical trial testing and can inform design considerations for improved implants and surgical techniques [19,25]. Implant retrieval analysis can also help identify models that are experiencing catastrophic failure *in vivo* despite medical device approval [36]. Furthermore, connection of implant damage with patient records allows for an understanding of conditions that may contribute to a more damaging environment and higher rates of implant failure.

1.6.1 Scoring of Implant Damage

The Goldberg damage scoring method is the industry standard for analysis of retrieved hip replacements focusing on the trunnion. The damage scoring process is separated into two categories, fretting and corrosion, and they are each given a score between 1-4. The criteria is summarized in Table 2 and Table 3 [37].

Table 2: Goldberg Corrosion scoring criteria

Severity of Corrosion	Score	Criteria	Example
None	1	No visible corrosion observed	

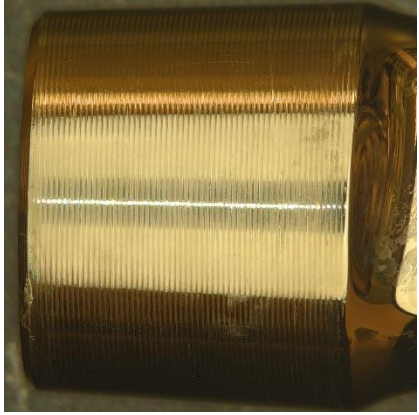
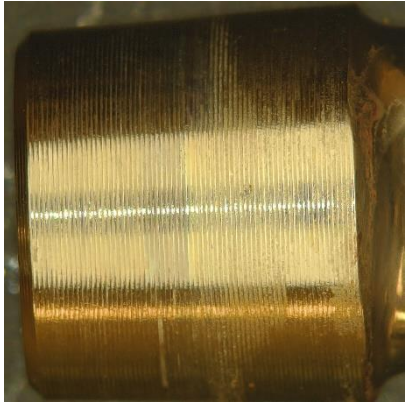
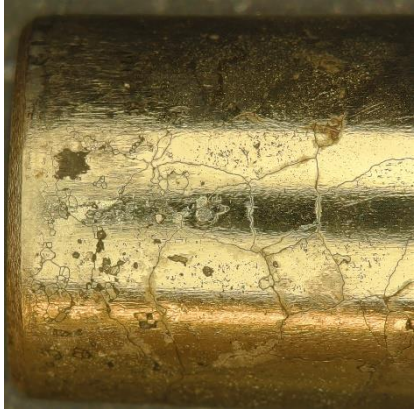
Mild	2	<30% of taper surface discoloured or dull	
Moderate	3	>30% of surface discoloured or dull, or, <10% of taper surface containing black debris, pits, or etch marks	
Severe	4	>10% of taper surface containing black debris, pits, or etch marks	

Table 3: Goldberg Fretting Criteria

Severity of Fretting	Score	Criteria
None	1	No visible sign of fretting observes
Mild	2	Single band or bands of fretting scars involving three or fewer machine lines on taper surface
Moderate	3	Several bands of fretting scars or a single band involving more than three machining lines
Severe	4	Several bands of fretting scars involving several adjacent machine lines, or flattened areas with nearby fretting scars

Typically, the scoring is done visually unaided or by observation through a low-power microscope with the observer deciding which threshold the implant meets in the criteria. This leads to an unintended qualitative nature of the scoring and possible interobserver variation of score. As a result, for reliable damage scoring, multiple parties with expertise are required to score the implants, leading to a large labour requirement. When conducting studies on retrieved implants that include damage scoring, the sample size is often <100 to account for this increased labour. This is considered underpowered, as the original criteria recommended at least 200 implants be scored to achieve appropriate power to detect small differences between groups. To try to mitigate this, single models of implants will often be selected. Goldberg scoring is a key method to observe and determine implant corrosion and fretting, however the labour requirement has prevented large-scale (>500 implants) studies from being undertaken.

1.7 Machine Learning

Artificial intelligence as a field emerged in the 1950's. Machine learning is arguably the most popular subset of this field and it refers to algorithms that can update themselves using statistics to self-optimize [38]. The topic of machine learning is broad but its

applications can generally be sorted into two categories, supervised and unsupervised learning. Unsupervised is when an unlabelled collection of data is given to the algorithm, and it finds patterns and makes assumptions from the given data. Supervised is when a set of labelled data is given to train the algorithm, after which it is able to perform a task based on how it has been trained from the labelled images [39]. Both unsupervised and supervised machine learning have been used for biomedical applications and the decision regarding which one is implemented is based on the research question being investigated. Image classification, such as classifying the damage score of an object using a predetermined criterion, is considered a supervised machine learning problem.

1.7.1 Convolutional Neural Networks

Convolutional neural networks (CNN) are a member of a subset of machine learning called deep learning. The theory for them was meaningfully developed in the 1980's by Kunihiro Fukushima, but they are extremely computationally intensive and could not reasonably be used in application until the development of graphics processing units in the 2000s.

Convolutional neural networks for image classification contain both a feature extractor and a classifier. In a supervised problem, the network is fed labelled information (typically images) that belong to a series of classes. The model then trains itself by moving filters over the images to create feature maps. The values of the maps are then associated with different features of the images, then the results are summarized and passed to the next layer. The final stage of the network is the classification, where it takes all the information gathered from the feature maps associated with different classes and statistically calculates the probability of it being in each class. It then labels the image with the class that had the highest probability, and it checks its answer against the true label. If the label is correct, the network accepts its summaries of the feature maps as accurate and does not update. If the label is incorrect, the network updates its summaries of the feature maps to correct the network classification [40]. This is done through calculating the loss function and penalizing the network for incorrect classifications and feeding this back through the network layers (a concept called “back-propagation”). A visual example of a CNN for classifying an object is shown in Figure 10.

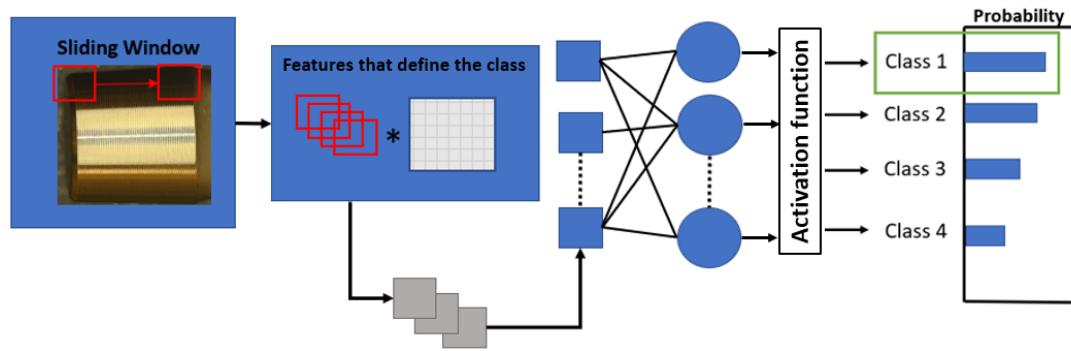


Figure 10: Example of a CNN for classifying a trunnion. The first two blocks extract features that can be used to identify different features (shapes, colour, brightness, etc.) and how they can be associated with the different classes. The knowledge from the last block moves forward to the next, then all the information is gathered together, and the system makes a statistical guess of the class. Image adapted from [40].

The data used to train the convolutional neural network is termed the training data. The data held and used during training to spot check the network and tune hyperparameters in the network is called validation data. The testing data refers to a set of data that the network does not see during training or validation. This is used to test the network and determine its accuracy and effectiveness.

CNNs are common in a variety of biomedical applications, and their use in classification tasks has been second only to segmentation [38]. An example in arthroplasty includes identifying hip arthroplasty designs from an x-ray [41]. Obstacles to using CNNs can include sufficient data (typically >1000 images required for training). Cases where data is not sufficient may be remedied through data augmentation which includes varying your images (rotation, brightness, colour, etc.) to create what the system will perceive as unique images. However, care must be taken to ensure the augmentation still represents the images input and over-augmentation can affect generalizability of the network [42,43]. Ronneberger et al. demonstrated with U-Net that training a convolutional neural network for segmentation with limited data was possible if the architecture was well designed [44]. It can reasonably be inferred that for a classification task this should also be possible.

1.7.2 Previous Attempts to Automate Goldberg Scoring

A previous attempt to automate Goldberg corrosion damage scoring of the trunnion was using a different machine learning method called support vector machine learning [45]. The key difference between support vector machine learning and convolutional neural networks is that a support vector machine maps the inputted data while a convolutional neural network extracts features before mapping the data. This means that support vector machines require significant image preprocessing as the image features must be extracted prior to using the algorithm. They also do not provide class probabilities [46].

Milionfared et al. achieved Goldberg scoring of the trunnions with a cross-validated accuracy of 85% [45]. However, they only observed 138 modular stems, and it is unclear how many were of each class and which models of stems were used. Without this information and effort to ensure presence of the lesser common classes in each validation run, it is difficult to ensure that the 85% accuracy is not a result of the accidental exclusion of less common classes (such as class 3 and 4). Aside from accuracy, there is little discussion on other evaluation metrics that better distinguish a methods performance (such as sensitivity, specificity). Furthermore, knowing that there are greatly varying taper designs available, it is unclear if the modular stems they selected include all models currently used in practice, both in Australia where the study was conducted, and in Canada and the US, where there is interest to apply it. Notably, none of their example images included machining lines and instead all had smooth finishes. In Canada, machining lines are present on many of the taper designs currently used and their affect on the trunnion surface no longer appearing homogenous in colouring without corrosion may affect the feature extraction methods proposed by Milimonfared et al.. Lastly, this method has failed to be used in any implant retrieval studies since its publication, calling into question its ease of use and effectiveness in a laboratory setting.

1.8 Thesis Objectives and Hypothesis

Retrieved implants offer a wealth of information to understand the *in vivo* behaviour of implant devices and to identify potentially problematic implants before widespread catastrophic failure occurs. Machine learning, specifically convolutional neural networks,

offer a possibility to automate the detection of corrosion from high-quality photos. As a result, the objective of this thesis is to 1) perform a large-scale survey of all stems in the possession of the IRL laboratory, including mass imaging and Goldberg scoring of the implants, and look for possible trends in device and patient characteristics, 2) create a convolutional neural network able to discern corrosion severity using the Goldberg scoring method.

1.9 References

- [1] Osteoarthritis (OA) | Arthritis | CDC n.d.
<https://www.cdc.gov/arthritis/basics/osteoarthritis.htm> (accessed March 1, 2022).
- [2] Osteoarthritis - Symptoms, Causes, Diagnosis & Treatments | Arthritis Society n.d.
[https://arthritis.ca/about-arthritis/arthritis-types-\(a-z\)/types/osteoarthritis](https://arthritis.ca/about-arthritis/arthritis-types-(a-z)/types/osteoarthritis) (accessed March 1, 2022).
- [3] Spector TD, MacGregor AJ. Risk factors for osteoarthritis: genetics. *Osteoarthritis and Cartilage* 2004;12:39–44. <https://doi.org/10.1016/J.JOCA.2003.09.005>.
- [4] Tarride JE, Haq M, O'Reilly DJ, Bowen JM, Xie F, Dolovich L, et al. The excess burden of osteoarthritis in the province of Ontario, Canada. *Arthritis & Rheumatism* 2012;64:1153–61. <https://doi.org/10.1002/ART.33467>.
- [5] Vos T, Flaxman AD, Naghavi M, Lozano R, Michaud C, Ezzati M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* 2012;380:2163–96. [https://doi.org/10.1016/S0140-6736\(12\)61729-2](https://doi.org/10.1016/S0140-6736(12)61729-2).
- [6] Hayashi D, Roemer FW, Guermazi A. Imaging of osteoarthritis—recent research developments and future perspective. *The British Journal of Radiology* 2018;91. <https://doi.org/10.1259/BJR.20170349>.
- [7] Rivière C, Vendittoli PA. Personalized hip and knee joint replacement. *Personalized Hip and Knee Joint Replacement* 2020:1–350. <https://doi.org/10.1007/978-3-030-24243-5>.
- [8] Hip Replacement Surgery and arthritis | Arthritis Society n.d.
<https://arthritis.ca/treatment/surgery/hip-replacement-surgery> (accessed March 1, 2022).
- [9] CJRR annual report: Hip and knee replacements in Canada | CIHI n.d.
<https://www.cihi.ca/en/cjrr-annual-report-hip-and-knee-replacements-in-canada> (accessed March 1, 2022).

- [10] Sarmiento A. Sir John Charnley and his legacy to total hip arthroplasty, 1970-1993. *Current Orthopaedic Practice* 2014;25:115–8. <https://doi.org/10.1097/BCO.0000000000000084>.
- [11] TITLE: Components and Materials used for Total Hip Replacement: A Review of the Comparative Clinical Effectiveness. 2013.
- [12] Beckmann NA, Bitsch RG, Schonhoff M, Siebenrock KA, Schwarze M, Jaeger S. Comparison of the Primary Stability of Porous Tantalum and Titanium Acetabular Revision Constructs. *Materials* 2020;13. <https://doi.org/10.3390/MA13071783>.
- [13] Kunze KN, Premkumar A, Bovonratwet P, Sculco PK. Acetabular Component and Liner Selection for the Prevention of Dislocation After Primary Total Hip Arthroplasty. *JBJS Rev* 2021;9. <https://doi.org/10.2106/JBJS.RVW.21.00148>.
- [14] Mahon J, McCarthy CJ, Sheridan GA, Cashman JP, O’Byrne JM, Kenny P. Outcomes of the Exeter V40 cemented femoral stem at a minimum of ten years in a non-designer centre. <https://doi.org/10.1302/2633-1462.112.BJO-2020-0163R1> 2020;1:743–8. <https://doi.org/10.1302/2633-1462.112.BJO-2020-0163.R1>.
- [15] Tan SC, Teeter MG, del Balso C, Howard JL, Lanting BA. Effect of Taper Design on Trunnionosis in Metal on Polyethylene Total Hip Arthroplasty. *The Journal of Arthroplasty* 2015;30:1269–72. <https://doi.org/10.1016/J.ARTH.2015.02.031>.
- [16] Arnholt CM, MacDonald DW, Underwood RJ, Guyer EP, Rimnac CM, Kurtz SM, et al. Do Stem Taper Microgrooves Influence Taper Corrosion in Total Hip Arthroplasty? A Matched Cohort Retrieval Study. *The Journal of Arthroplasty* 2017;32:1363–73. <https://doi.org/10.1016/J.ARTH.2016.11.018>.
- [17] Rémond G, Nockolds C, Phillips M, Roques-Carmes C. Implications of Polishing Techniques in Quantitative X-Ray Microanalysis. *Journal of Research of the National Institute of Standards and Technology* 2002;107:639. <https://doi.org/10.6028/JRES.107.052>.

- [18] Mistry JB, Chughtai M, Elmallah RK, Diedrich A, Le S, Thomas M, et al. Trunnionosis in total hip arthroplasty: a review. *Journal of Orthopaedics and Traumatology : Official Journal of the Italian Society of Orthopaedics and Traumatology* 2016;17:1. <https://doi.org/10.1007/S10195-016-0391-1>.
- [19] Whitehouse MR, Endo M, Zachara S, Nielsen TO, Greidanus N v., Masri BA, et al. Adverse local tissue reactions in metal-onpolyethylene total hip arthroplasty due to trunnion corrosion: The risk of misdiagnosis. *Bone and Joint Journal* 2015;97-B:1024–30. <https://doi.org/10.1302/0301-620X.97B8.34682/ASSET/IMAGES/LARGE/34682-GALLEYFIG4B.JPEG>.
- [20] Gilbert JL, Mali S, Urban RM, Silverton CD, Jacobs JJ. In vivo oxide-induced stress corrosion cracking of Ti-6Al-4V in a neck–stem modular taper: Emergent behavior in a new mechanism of in vivo corrosion. *Journal of Biomedical Materials Research Part B: Applied Biomaterials* 2012;100B:584–94. <https://doi.org/10.1002/JBM.B.31943>.
- [21] Cudjoe E. Tribocorrosion Behavior of Metallic Implants: A Comparative Study of CoCrMo and Ti6Al4V in Simulated Synovial Environments 2019.
- [22] Gilbert JL. Corrosion in the Human Body: Metallic Implants in the Complex Body Environment. *Corrosion* 2017;73:1478–95. <https://doi.org/10.5006/2563>.
- [23] Gilbert JL, Sivan S, Liu Y, Kocagöz SB, Arnholt CM, Kurtz SM. Direct in vivo inflammatory cell-induced corrosion of CoCrMo alloy orthopedic implant surfaces. *Journal of Biomedical Materials Research Part A* 2015;103:211–23. <https://doi.org/10.1002/JBM.A.35165>.
- [24] Forms of Corrosion - AMPP n.d. <https://www.ampp.org/resources/what-is-corrosion/forms-of-corrosion#uniform> (accessed March 1, 2022).
- [25] Merola M, Affatato S. Materials for Hip Prostheses: A Review of Wear and Loading Considerations. *Materials* 2019;12. <https://doi.org/10.3390/MA12030495>.

- [26] Hu CY, Yoon TR. Recent updates for biomaterials used in total hip arthroplasty. *Biomaterials Research* 2018 22:1 2018;22:1–12. <https://doi.org/10.1186/S40824-018-0144-8>.
- [27] Halma JJ, Señaris J, Delfosse D, Lerf R, Oberbach T, van Gaalen SM, et al. Edge loading does not increase wear rates of ceramic-on-ceramic and metal-on-polyethylene articulations. *Journal of Biomedical Materials Research Part B: Applied Biomaterials* 2014;102:1627–38. <https://doi.org/10.1002/JBM.B.33147>.
- [28] Brandt JM, Vecherya A, Guenther LE, Koval SF, Petrak MJ, Bohm ER, et al. Wear testing of crosslinked polyethylene: Wear rate variability and microbial contamination. *Journal of the Mechanical Behavior of Biomedical Materials* 2014;34:208–16. <https://doi.org/10.1016/J.JMBBM.2014.02.016>.
- [29] Moro T, Takatori Y, Kyomoto M, Ishihara K, Kawaguchi H, Hashimoto M, et al. Wear resistance of the biocompatible phospholipid polymer-grafted highly cross-linked polyethylene liner against larger femoral head. *Journal of Orthopaedic Research* 2015;33:1103–10. <https://doi.org/10.1002/JOR.22868>.
- [30] Zietz C, Fabry C, Baum F, Bader R, Kluess D. The Divergence of Wear Propagation and Stress at Steep Acetabular Cup Positions Using Ceramic Heads and Sequentially Cross-Linked Polyethylene Liners. *The Journal of Arthroplasty* 2015;30:1458–63. <https://doi.org/10.1016/J.ARTH.2015.02.025>.
- [31] Al-Hajjar M, Carbone S, Jennings LM, Begand S, Oberbach T, Delfosse D, et al. Wear of composite ceramics in mixed-material combinations in total hip replacement under adverse edge loading conditions. *Journal of Biomedical Materials Research Part B: Applied Biomaterials* 2017;105:1361–8. <https://doi.org/10.1002/JBM.B.33671>.
- [32] Al-Hajjar M, Jennings LM, Begand S, Oberbach T, Delfosse D, Fisher J. Wear of novel ceramic-on-ceramic bearings under adverse and clinically relevant hip simulator conditions. *Journal of Biomedical Materials Research Part B: Applied Biomaterials* 2013;101:1456–62. <https://doi.org/10.1002/JBM.B.32965>.

- [33] Carnes KJ, Odum SM, Troyer JL, Fehring TK. Cost analysis of ceramic heads in primary total hip arthroplasty. *Journal of Bone and Joint Surgery - American Volume* 2016;98:1794–800. <https://doi.org/10.2106/JBJS.15.00831>.
- [34] The Effect of Bearing Surface on Early Revisions Following Total Hip Arthroplasty | CIHI n.d. <https://secure.cihi.ca/estore/productFamily.htm?pf=PFC2288> (accessed May 27, 2022).
- [35] Bijukumar DR, Segu A, Souza JCM, Li XJ, Barba M, Mercuri LG, et al. Systemic and local toxicity of metal debris released from hip prostheses: A review of experimental approaches. *Nanomedicine* 2018;14:951–63. <https://doi.org/10.1016/J.NANO.2018.01.001>.
- [36] Wylde CW, Jenkins E, Pabbruwe M, Bucher T. Catastrophic failure of the Accolade I hip arthroplasty stem: a retrieval analysis study. *HIP International* 2020;30:481–7. <https://doi.org/10.1177/1120700020919665>.
- [37] Goldberg JR, Gilbert JL, Jacobs JJ, Bauer TW, Paprosky W, Leurgans S. A multicenter retrieval study of the taper interfaces of modular hip prostheses. *Clinical Orthopaedics and Related Research* 2002;401:149–61. <https://doi.org/10.1097/00003086-200208000-00018>.
- [38] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Medical Image Analysis* 2017;42:60–88. <https://doi.org/10.1016/J.MEDIA.2017.07.005>.
- [39] [PDF] Understanding Machine Learning - From Theory to Algorithms | Semantic Scholar n.d. <https://www.semanticscholar.org/paper/Understanding-Machine-Learning-From-Theory-to-Shalev-Shwartz-Ben-David/ce615ae61d67db8537e981a0a08da7f0f2ff1cee> (accessed March 1, 2022).
- [40] What is a Convolutional Neural Network? - MATLAB & Simulink n.d. <https://www.mathworks.com/discovery/convolutional-neural-network-matlab.html> (accessed March 1, 2022).

- [41] Kang YJ, Yoo J il, Cha YH, Park CH, Kim JT. Machine learning–based identification of hip arthroplasty designs. *Journal of Orthopaedic Translation* 2020;21:13–7. <https://doi.org/10.1016/J.JOT.2019.11.004>.
- [42] Abraham GK, Jayanthi VS, Bhaskaran P. Convolutional neural network for biomedical applications. *Computational Intelligence and Its Applications in Healthcare* 2020:145–56. <https://doi.org/10.1016/B978-0-12-820604-1.00010-8>.
- [43] Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift Für Medizinische Physik* 2019;29:102–27. <https://doi.org/10.1016/J.ZEMEDI.2018.11.002>.
- [44] Ronneberger O. Invited Talk: U-Net Convolutional Networks for Biomedical Image Segmentation 2017:3–3. https://doi.org/10.1007/978-3-662-54345-0_3.
- [45] Milimonfared R, Oskouei RH, Taylor M, Solomon LB. An intelligent system for image-based rating of corrosion severity at stem taper of retrieved hip replacement implants. *Medical Engineering & Physics* 2018;61:13–24. <https://doi.org/10.1016/J.MEDENGPHY.2018.08.002>.
- [46] Shalev-Shwartz S, Ben-David S. UNDERSTANDING MACHINE LEARNING From Theory to Algorithms n.d.

Chapter 2

2 What patient and implant factors affect trunnionosis severity? An implant retrieval study of 664 femoral stems

Anastasia M. Codireni¹, Brent A. Lanting², Matthew G. Teeter^{1,3}

1. School of Biomedical Engineering, Western University, London, Ontario, Canada
2. Division of Orthopaedic Surgery, Department of Surgery, Schulich School of Medicine & Dentistry, Western University, London, Ontario, Canada
3. Department of Medical Biophysics, Schulich School of Medicine & Dentistry, Western University, London, Ontario, Canada

Background: Corrosion at the modular head-neck taper interface of total and hemiarthroplasty hip implants (trunnionosis) is a cause of implant failure and thus a clinical concern. Patient and device factors contributing to the occurrence of trunnionosis have been investigated in prior implant retrieval studies, but generally with limited sample sizes and a narrow range of models. The purpose of the present investigation was to determine which patient and device factors were associated with corrosion damage on the femoral stem taper across a large collection of different implant models retrieved following revision hip arthroplasty.

Methods: A retrieval study of 664 hip arthroplasty modular stem components was performed. Patient and device information was collected. Trunnions were imaged under digital microscopy and scored for corrosion damage using the Goldberg scale. Damage was related to patient and device factors using regression analysis.

Results: Greater duration of implantation ($p = 0.005$) and larger head size ($p < 0.001$) were associated with an elevated corrosion class. Older age at index surgery ($p = 0.035$), stainless steel stem material ($p = 0.022$), indication for revision as bone or periprosthetic fracture ($p = 0.017$) and infection ($p = 0.018$), and certain larger taper geometries were associated with a decreased corrosion class.

Conclusions: Factors identified as contributing to a higher or lower risk of more severe corrosion are consistent with most prior smaller retrieval studies. Surgeons should be aware

of these risk factors when selecting implants for their patients, and when diagnosing trunnionosis in symptomatic hip replacement patients.

Keyword (max of 6 keywords): Corrosion, trunnionosis, hip arthroplasty

2.1 Introduction

Modularity in modern hip arthroplasty implant designs enable surgeons to closely match the original anatomy of the patient [1]. Most devices provide modularity at the head-neck interface using a Morse taper, where the head and trunnion are attached using an interface fit [2]. The disadvantage of modularity is the potential for fretting and corrosion to occur due to the biomechanical forces acting at the head-neck taper junction [3]. This can produce debris in the form of metal ions, particles, and other corrosion products [4]. The presence of this debris can cause adverse local tissue reactions [5]. The corrosion process, termed trunnionosis, has been identified as a cause of hip arthroplasty failure and thus a clinical concern [6–8]. Corrosion-related soft-tissue reactions may potentially be overlooked and misdiagnosed as recurrent instability or infection [9]. Therefore, understanding potential risk factors for the development of trunnionosis can assist surgeons in making a proper diagnosis.

Implant retrieval studies have had an important role in identifying a variety of implant and patient factors that contribute to trunnionosis, including head material, taper design, implantation time, femoral offset, body mass index (BMI), and taper rigidity [10–16]. However, such studies have made these observations from a limited number of devices sampled from select manufacturers. At most, these studies have included 252 femoral heads with 148 femoral tapers, with some studies including as few as 46 implants [13,16]. In contrast, large-scale implant retrieval studies examining other device failure modes such as polyethylene wear have provided a more rigorous assessment of the variables contributing to implant damage [17].

The purpose of the present investigation was to determine which patient and device factors were associated with corrosion damage on the femoral stem taper across different implant models retrieved following revision hip arthroplasty.

2.2 Methods

2.2.1 Study Population

Institutional research ethics board approval was obtained for review of patient charts and implant retrieval analysis. All hip implants in our institutional implant retrieval laboratory were reviewed for inclusion (Figure 11). Implants included for analysis were designs with a modular head-neck taper where the femoral head and stem were retrieved at the time of revision surgery. Excluded were implants that were non-modular, cases where the femoral stem was not retrieved or had gross taper failure, and when there was fewer than five instances of a particular femoral stem model. Patient information including sex, age at implantation, hip joint laterality (left or right), reason for revision, and length of stem implantation were obtained from chart review. Implant information including taper design, stem material, head material, head size, stem model, and manufacturer was collected from the institutional implant retrieval laboratory catalogue and analysis of the device.

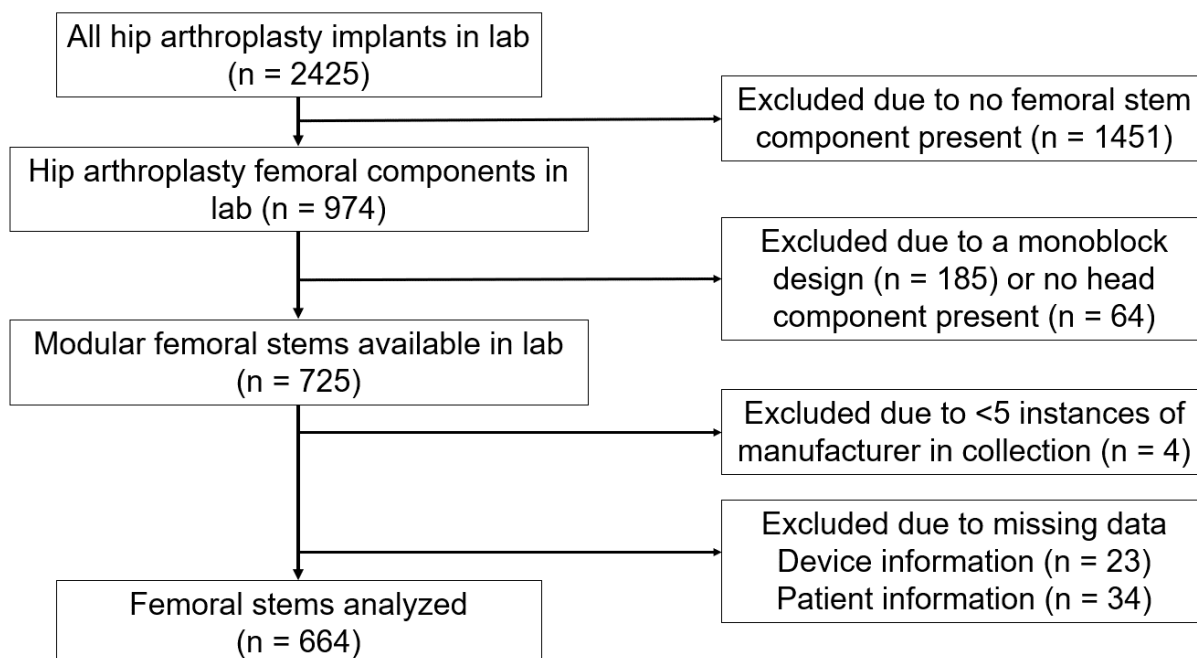


Figure 11: Study design

2.2.2 Visual Scoring

Each stem was imaged using a digital microscope at 20x magnification (DSX1000, Olympus Corporation, Tokyo, Japan). The surface was divided into four areas (medial, lateral, posterior and anterior) each represented by one image. The images (Figure 12) were then examined and assigned a corrosion score using the method of Goldberg et al., described in Table 4 [18]. A single score was assigned to each image and recorded; the maximum score observed on the implant from the four sides of the taper was used as the score for statistical analysis. All scoring was done by a single observer (A. Codirenze).

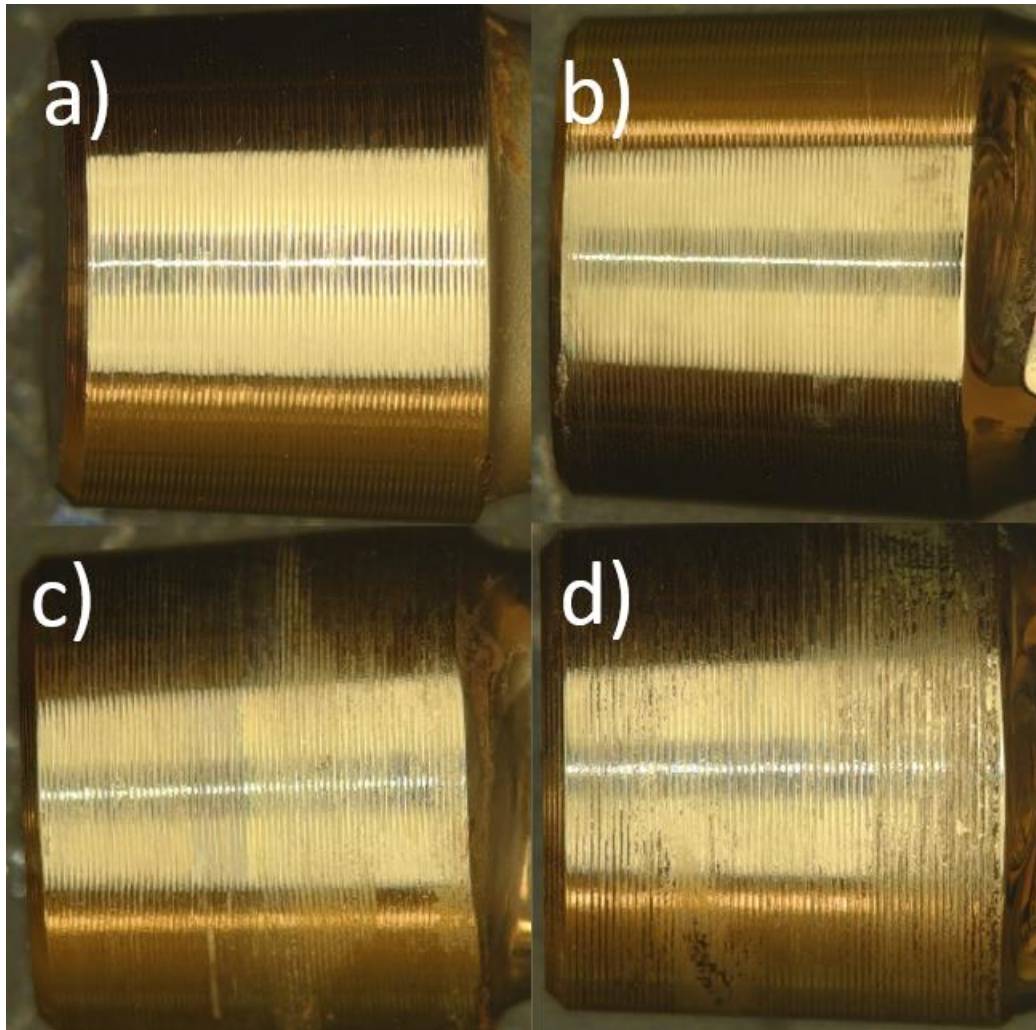


Figure 12: Representative images of each class (a) class 1, (b) class 2, (c) class 3, (d) class 4

Table 4: Goldberg scoring criteria [18]

	Class 1	Class 2	Class 3	Class 4
Severity	None	Mild	Moderate	Severe
Criteria	No visible corrosion observed	<30% of taper surface discoloured or dull	>30% of surface discoloured or dull, or, <10% of taper surface containing black debris, pits, or etch marks	>10% of taper surface containing black debris, pits, or etch marks

2.2.3 Statistical Analysis

Ordinal logistic regression was used to compare the maximum observed trunnion corrosion score with the gathered device and patient factors. Odds ratios (OR) were calculated to by exponentiating the parameter estimates of the final ordinal logistic regression model.

Possible interactions between patient and device factors identified in the ordinal logistic regression were further investigated. A Shapiro-Wilk test was used to determine normality. Accordingly, the Kruskal-Wallis one-way ANOVA was used to delineate differences in the factors with an alpha-value of 0.05. Data analysis was performed using IBM SPSS for Windows OS, version 28.0.1.1 and Prism 9 for Windows, version 9.3.1.

2.3 Results

In total, 664 implants were included in the analysis. Mean patient age at original implantation was 66 years (range, 23 to 97 years). There was slightly more female patients (51.3%, n = 339) and most were right hips (57%, n = 376). The mean length of implantation was 5 years (range, <1 to 24 years). Reasons for revision included bone and periprosthetic fracture (19.3%, n = 127), implant fracture (2.0%, n = 13), infection (37.9%, n = 249), instability (3.7%, n = 24), metal debris reaction (3.0%, n = 20), polyethylene wear, osteolysis and aseptic loosening (32.4%, n = 213), and undifferentiated pain (1.7%, n = 11). Details on the implants included in the analysis are shown in Table 5.

Stem material	Titanium	47% (n = 312)
	Cobalt-chromium	45.8% (n = 304)
	Stainless steel	7.2% (n = 48)
Head material	Cobalt-chromium	93.0% (n = 618)
	Stainless steel	2.5% (n = 17)
	Ceramic	2.8% (n = 19)
	Oxinium	1.2% (n = 8)
	Zirconia	0.3% (n = 2)
Head size	22 mm	0.8% (n = 5)
	26 mm	2.9% (n = 19)
	28 mm	39.9% (n = 265)
	32 mm	26.1% (n = 173)
	36 mm	25.9% (n = 172)

	38 mm	0.2% (n = 1)
	40 mm	2.1% (n = 14)
	>40 mm	2.2% (n = 15)
Taper design	B Type 1	2.6% (n = 17)
	C-taper	9.3% (n = 62)
	D 11/13	1.4% (n = 9)
	D 12/14	19.3% (n = 128)
	D 14/16	1.2% (n = 8)
	PCA taper	6.6% (n = 44)
	S 10/12	0.2% (n = 1)
	S 12/14	24.7% (n = 164)
	S 14/16	0.3% (n = 2)
	V40	19.4% (n = 129)
	W 12/14	1.4% (n = 9)
	Z 12/14	12.2% (n = 81)
	Z 6 degree	1.5% (n = 10)

Table 5: Implant characteristics. B Type 1: Mallory (n = 9), Taperloc (n = 6), Integral (n = 2) from Zimmer-Biomet (Warsaw, IN). C-taper: ODC (n = 1), Omnifit (n = 43), Restoration (n = 2), Secur-fit (n = 16)) from Stryker (Mahwah, NJ). D 11/13: S-ROM (n = 9) from DePuy Synthes (Raynham, MA). D 12/14: AML (n = 16), Corail (n = 42), Endurance (n = 19), Prodigy (n = 2), Reclaim (n = 7), Response (n = 1), Solution (n = 12), Summit (n = 29) from DePuy Synthes (Raynham, MA). D

14/16: Solution (n = 5), AML (n = 2), CML (n = 1), from DePuy Synthes (Raynham, MA). PCA Taper: PCA (n = 40), Precision (n = 3), Strata (n = 1), from Stryker (Mahwah, NJ). S 10/12: Richards Modular (n = 1) from Smith & Nephew (Memphis, TN). S 12/14: Anthology (n = 4), Conquest (n = 13). CPCS (n = 7), Echelon (n = 24), Polarstem (n = 8), Redapt (n = 4), SL plus (n = 1), SMF (n = 2), Spectron (n = 32), Synergy (n = 70) from Smith & Nephew (Memphis, TN). S 14/16: Biofit (n = 1), TriWedge (n = 1) from Smith & Nephew (Memphis, TN). V40: ABG (n = 3), Accolade (n = 26), Definition PM (n = 1), Exeter (n = 47), GRMS (n = 1), Precision (n = 3), Rejuvenate (n = 32), Restoration (n = 14) from Stryker (Mahwah, NJ). W 12/14: Profemur (n = 7), Gladiator (n = 2) from Stryker (Mahwah, NJ). Z 12/14: Advocate (n = 1), Apollo (n = 1), CLS (n = 1), CPT (n = 1), M/L taper (n = 31), MS30 (n = 1), Versys (n = 44) from Zimmer-Biomet (Warsaw, IN). Z 6 degree: Harris (n = 8), Versys (n = 2) from Zimmer-Biomet (Warsaw, IN). This table reflects modern ownership of the taper designs and companies.

The ordinal logistic regression revealed length of implantation, age at implantation, reason for revision, head size, stem material, and taper design to be significant factors associated with the severity of corrosion class summarized in Table 6.

Table 6: Ordinal logistic regression results

Variable	Estimate	95% CI	OR	P-value
Implantation time	0.048	[0.015, 0.081]	1.049	0.005
Age at implant	-0.015	[-0.029, -0.001]	0.985	0.035
Head size	0.092	[0.050, 0.134]	1.096	<0.001
Sex	-0.133	[-0.448, 0.181]	0.875	0.406
Side	-0.095	[-0.401, 0.212]	0.910	0.545

Primary or repeat revision		0.167	[-0.216, 0.551]	1.182	0.392
Reason for revision	Bone and periprosthetic fracture	-1.470	[-2.682, -0.257]	0.230	0.017
	Implant fracture	-0.800	[-2.416, 0.816]	0.449	0.332
	Infection	-1.437	[-2.631, -0.243]	0.238	0.018
	Instability	-1.326	[-2.759, 0.108]	0.266	0.070
	Metal debris reaction	0.162	[-1.233, 1.557]	1.176	0.820
	Polyethylene wear, osteolysis, and aseptic loosening	-1.115	[-2.313, 0.083]	0.328	0.068
	Undifferentiated pain	0			
Taper geometry	B Type 1	-0.835	[-2.341, 0.671]	0.434	0.277
	C-taper	-0.738	[-2.025, 0.550]	0.478	0.261

	D11/13	-1.108	[-2.947, -0.731]	0.330	0.238
	D12/14	-2.214	[-3.387, -0.861]	0.120	0.001
	D14/16	-2.792	[-4.641, -0.943]	0.061	0.003
	PCA	-3.144	[-4.497, -1.790]	0.043	<0.001
	S10/12	-2.259	[-6.446, 1.928]	0.104	0.290
	S12/14	-1.371	[2.612, -0.129]	0.254	0.030
	S14/16	-0.648	[-3.524, 2.227]	0.523	0.659
	V40	-0.698	[-1.990, 0.593]	0.497	0.289
	W 12/14	-0.585	[-2.363, 1.193]	0.557	0.519
	Z 12/14	-1.299	[-2.570, -0.029]	0.273	0.045
	Z 6 Degree	0			

Stem Material	Chromium Cobalt	-0.325	[-0.704, 0.055]	0.723	0.094
	Stainless Steel	-1.016	[-1.885, -0.146]	0.362	0.022
	Titanium	0			
Head Material	Ceramic	0.422	[-2.561, 3.405]	1.525	0.781
	Chromium Cobalt	-0.268	[-3.122, 2.587]	0.765	0.854
	Oxinium	-0.017	[-3.159, 3.125]	0.983	0.992
	Stainless Steel	0.548	[-2.547, 3.643]	1.730	0.729
	Zirconia	0			

For patient factors, increasing length of implantation was associated with a greater probability of elevated corrosion class (OR 1.049, $p = 0.005$), while increasing patient age at implantation was associated with a lower probability of elevated corrosion class (OR 0.985, $p = 0.035$). Two reasons for revision were associated with a lower probability of elevated corrosion class: bone or periprosthetic fracture (OR 0.230, $p = 0.017$) and infection (OR = 0.238, $p = 0.018$).

For device factors, increasing head size was associated with a greater probability of elevated corrosion class (OR = 1.096, $p < 0.001$). Stainless steel stem material was associated with a lower probability of elevated corrosion class (OR = 0.362, $p = 0.022$). Five taper geometries were associated with a lower probability of elevated corrosion class:

D12/14 (OR = 0.120, $p = 0.001$), D14/16 (OR = 0.061, $p = 0.003$), PCA (OR = 0.043, $p < 0.001$), S12/14 (OR = 0.254, $p = 0.030$), and Z12/14 (OR = 0.273, $p = 0.045$).

There were two interactions between patient and device factors. Length of implantation varied between reasons for revision ($p < 0.001$) with median implantation times of 1.1 years for infection, 1.1 years for instability, 3.2 years for bone and periprosthetic fracture, 5.2 years for undifferentiated pain, 5.3 years for metal debris reactions, 6.1 years for implant fracture, and 7.1 years for wear/osteolysis/loosening. Head size varied between stem material ($p < 0.001$) with median head size of 28 mm for the cobalt-chromium and stainless-steel stems, and 32 mm for the titanium stems.

2.4 Discussion

Trunnionosis and the potential for adverse local tissue reactions remains an area of concern for modular hip arthroplasty devices. Although prior implant retrieval studies have identified a number of risk factors associated with trunnionosis (Table 7), these studies have generally examined only specific models of devices with low population numbers (typically below 150 implants). The purpose of this study was to determine which patient and device factors were associated with corrosion damage on the femoral stem taper across different implant models retrieved following revision hip arthroplasty. With 664 femoral stems examined for corrosion severity, this is the largest implant retrieval study of its kind.

Table 7: Previous implant retrieval studies

Study	Study Population	Factors associated with corrosion	Factors not associated with corrosion
El Zein et al. (2021) [1]	Eight cohorts defined from 157 retrieved THA based on femoral head composition	Head material, taper geometry	Head size

	and taper geometry		
Hampton et al. (2019)[2]	Cohort matched: 28mm Oxinium heads with 28mm CoCr heads	Head material	N/A
Silijander et al. (2018)[3]	92 femoral stems and heads	Head size, age at implant	Gender, alloy combination
Lange et al. (2018) [4]	56 CoCr designs from a single manufacturer, two trunnion designs, mated with ceramic or CoCr heads	Length of implantation	Head size, neck length
Del Balso et al. (2018)[5]	Single taper design, single manufacturer, matched Bipolar and THA	Implantation time	Bipolar hemiarthroplasty versus total hip arthroplasty
Triantafyllopoulos et al. (2016)[6]	154 femoral stem and heads, single modular neck	Taper design, alloy combinations, implantation time	Head size
Tan et al. (2016)[7]	Cohort matched, two cohorts: 52 Ceramic, Cobalt	Head material, taper design, implantation time	Age, gender, body mass index

	heads and 8 oxinimum, CoCr		
Higgs et al. (2016) [8]	252 CoCr femoral heads, 148 femoral tapers, C- taper and V-40	Head offset, implantation time, weight, flexural rigidity	Taper design
Del Balso et al. (2016) [9]	Cohort matched, 23 femoral heads, 32 mm and 28 mm, single manufacture and 23 femoral stems, 10 28mm and 13 32mm	Implantation time	Stem offset, stem type, stem fixation method
Tan et al. (2015) [10]	44 implants with 6 taper designs, from four manufacturers, 28+0mm heads	Taper design	N/A
Brock et al. (2015)[11]	104 heads and 11 stem trunnions from a single manufacturer, grounded by stem model (Corail and S-ROM)	Taper design, threading/machining lines	N/A

Kurtz et al. (2013) [12]	100 femoral head-stem pairs, 50 ceramics matched with 50 CoCr based on implantation, lateral offset, stem design, flexural rigidity	Head material, stem alloy, stem flexural rigidity	N/A
Dyrkacz et al. (2013)[13]	74 implants with 28mm and 36mm heads, 12/14 mm taper, CoCr heads and stems, two manufactures	Head size	N/A

Longer Implantation Time

Longer implantation duration (time in vivo) was associated with higher corrosion scores. This finding is consistent with prior literature [11,13–16,27]. With increased time implanted, there is increased time for corrosion to occur. This is furthered by shear stresses and mechanically assisted wear, such as crevice corrosion, that has been associated with repeated load cycling [28].

Age at Time of Implantation and Stem Material

Greater age at the time of implantation was associated with lower corrosion scores. This finding is inconsistent with prior studies by Tan et al. and Hothi et al [11,14]. Younger patients would potentially have had an implant for a longer time before revision and may be more active than older patients, increasing stress on the modular head-neck taper junction. Younger patients would also be more likely to receive a cementless cobalt-

chromium or titanium femoral stem than a cemented stainless-steel stem, where stainless steel material was also associated with lower corrosion scores in the present study.

Indication for Revision

Infection and bone and periprosthetic fracture as the reason for revision were associated with lower corrosion scores. Prior studies have not reported on reason for revision as being associated with corrosion. Stems within the infection and bone and periprosthetic fracture groups had some of the lowest durations of implantation in the study, and length of implantation was associated with corrosion severity. However, instability also had a low implantation time, but was not significantly associated with lower corrosion scores. Instability may affect the stresses at the head-neck taper junction causing increased mechanical wear and contributing to a higher corrosion score even with less time implanted. Conversely, corrosion and local debris could cause the soft tissue damage and subsequent dislocation.

Head Size, Trunnion Design, and Head Material

Larger head sizes were associated with higher corrosion scores. This finding is consistent with prior literature [12]. This has previously been attributed to larger heads having a greater torque acting along the taper junction during daily activity, leading to more micromotion and increased deterioration of the passive oxide film.

Certain taper designs were associated with lower corrosion scores. Prior studies have found a difference in corrosion severity between taper designs, with smaller tapers having greater corrosion than larger tapers (e.g. 11/13 versus 12/14 tapers, where the numbers refer to the proximal and distal diameter of the trunnion, respectively) but this finding has been inconsistent, with Brock et al. claiming the opposite on a single-manufacture study [10,23,25,27,29]. In the current study, 12/14 tapers from several manufacturers were associated with lower corrosion scores. Although all are labelled as 12/14 tapers, it is important to note that the cone angle and trunnion length differ between them. Two additional taper designs, both larger than the 12/14 tapers, were also associated with lower

corrosion scores. Having a smaller taper geometry with a smaller contact surface may increase the stress concentrations within the head-neck taper junction.

Previous studies have identified head material as a significant contributing factor to corrosion, but that factor was not found to be significant in this study [11,19,26]. 93% of heads in this study were cobalt-chromium, which may have affected the model as there was little representation of other head materials. Weight has previously been reported as a significant factor in increased corrosion by Higgs et al., but Tan et al. has reported that BMI is not [11,13]. As height, weight, and BMI data was missing for a large portion of the retrieved stems, these variables were not examined in the present study.

A limitation of this study is that we were unable to include head offset due to lack of information in implant database systems and inconsistent presence of offset printed on the physical components. This factor has been identified as contributing to greater corrosion scores [13]. Similarly, we did not examine flexural rigidity [30]. We scored only corrosion and not fretting, but the two scores are related. Although with 664 femoral stems examined this is a large study including many different implant models and manufacturers, it reflects a single institution and does not include every device available on the market. A subset of the implants examined were also included in previously published retrieval studies by our group, although we have included a discussion of literature from multiple institutions. We assessed damage visually with the common Goldberg score and did not directly measure volumetric changes to the stem tapers, which might yield different results [31]. Finally, all implant retrieval studies examine failed implants and may not be entirely representative of well-performing devices that have not been revised.

2.5 Conclusion

In this large-scale study, the first of its size for retrieved hip arthroplasty devices, length of implantation time, age at implantation, reason for revision, head size, and taper design were found to have a significant effect on corrosion score. These findings are generally supportive of prior implant retrieval studies that examined fewer implants. Surgeons should be aware of these risk factors when choosing femoral stems and heads for their patients, and in diagnosing trunnionosis in patients presenting with complaints after surgery.

2.6 References

- [1] Vierra BM, Blumenthal SR, Amanatullah DF. Modularity in total hip arthroplasty: Benefits, risks, mechanisms, diagnosis, and management. *Orthopedics* 2017;40:355–66. <https://doi.org/10.3928/01477447-20170606-01>.
- [2] Hernigou P, Queinnec S, Flouzat Lachaniette CH. One hundred and fifty years of history of the Morse taper: from Stephen A. Morse in 1864 to complications related to modularity in hip arthroplasty. *International Orthopaedics* 2013;37:2081. <https://doi.org/10.1007/S00264-013-1927-0>.
- [3] Lavernia CJ, Iacobelli DA, Villa JM, Jones K, Gonzalez JL, Jones WK. Trunnion–Head Stresses in THA: Are Big Heads Trouble? *The Journal of Arthroplasty* 2015;30:1085–8. <https://doi.org/10.1016/J.ARTH.2015.01.021>.
- [4] Shulman RM, Zywiell MG, Gandhi R, Davey JR, Salonen DC. Trunnionosis: the latest culprit in adverse reactions to metal debris following hip arthroplasty. *Skeletal Radiology* 2015;44:433–40. <https://doi.org/10.1007/S00256-014-1978-3/FIGURES/10>.
- [5] Cooper HJ, della Valle CJ, Berger RA, Tetreault M, Paprosky WG, Sporer SM, et al. Corrosion at the Head-Neck Taper as a Cause for Adverse Local Tissue Reactions After Total Hip Arthroplasty. *The Journal of Bone and Joint Surgery American Volume* 2012;94:1655. <https://doi.org/10.2106/JBJS.K.01352>.
- [6] Berstock JR, Whitehouse MR, Duncan CP. Trunnion corrosion: What surgeons need to know in 2018. *Bone and Joint Journal* 2018;100B:44–9. <https://doi.org/10.1302/0301-620X.100B1.BJJ-2017-0569.R1/ASSET/IMAGES/LARGE/BJJ-2017-0569.R1-GALLEYFIG1.JPEG>.
- [7] Banerjee S, Cherian JJ, Bono J v., Kurtz SM, Geesink R, Meneghini RM, et al. Gross Trunnion Failure After Primary Total Hip Arthroplasty. *The Journal of Arthroplasty* 2015;30:641–8. <https://doi.org/10.1016/J.ARTH.2014.11.023>.

- [8] Hussenbocus S, Kosuge D, Solomon LB, Howie DW, Oskouei RH. Head-Neck Taper Corrosion in Hip Arthroplasty. *BioMed Research International* 2015;2015. <https://doi.org/10.1155/2015/758123>.
- [9] Bijukumar DR, Segu A, Souza JCM, Li XJ, Barba M, Mercuri LG, et al. Systemic and local toxicity of metal debris released from hip prostheses: A review of experimental approaches. *Nanomedicine* 2018;14:951–63. <https://doi.org/10.1016/J.NANO.2018.01.001>.
- [10] Tan SC, Teeter MG, del Balso C, Howard JL, Lanting BA. Effect of Taper Design on Trunnionosis in Metal on Polyethylene Total Hip Arthroplasty. *The Journal of Arthroplasty* 2015;30:1269–72. <https://doi.org/10.1016/J.ARTH.2015.02.031>.
- [11] Tan SC, Lau ACK, del Balso C, Howard JL, Lanting BA, Teeter MG. Tribocorrosion: Ceramic and Oxidized Zirconium vs Cobalt-Chromium Heads in Total Hip Arthroplasty. *The Journal of Arthroplasty* 2016;31:2064–71. <https://doi.org/10.1016/J.ARTH.2016.02.027>.
- [12] Dyrkacz RMR, Brandt JM, Ojo OA, Turgeon TR, Wyss UP. The influence of head size on corrosion and fretting behaviour at the head-neck interface of artificial hip joints. *J Arthroplasty* 2013;28:1036–40. <https://doi.org/10.1016/J.ARTH.2012.10.017>.
- [13] Higgs GB, Macdonald DW, Gilbert JL, Rimnac CM, Kurtz SM. Basic Science Does Taper Size Have an Effect on Taper Damage in Retrieved Metal-on-Polyethylene Total Hip Devices? *The Journal of Arthroplasty* 2016;31:277–81. <https://doi.org/10.1016/j.arth.2016.06.053>.
- [14] Hothi HS, Eskelinen AP, Berber R, Lainiala OS, Moilanen TPS, Skinner JA, et al. Factors Associated With Trunnionosis in the Metal-on-Metal Pinnacle Hip. *J Arthroplasty* 2017;32:286–90. <https://doi.org/10.1016/J.ARTH.2016.06.038>.
- [15] del Balso C, Teeter MG, Tan SC, Lanting BA, Howard JL. Does the Additional Articulation in Retrieved Bipolar Hemiarthroplasty Implants Decrease Trunnionosis

Compared to Total Hip Arthroplasty? *The Journal of Arthroplasty* 2018;33:268–72.
<https://doi.org/10.1016/J.ARTH.2017.08.027>.

[16] del Balso C, Teeter MG, Tan SC, Howard JL, Lanting BA. Trunnionosis: Does Head Size Affect Fretting and Corrosion in Total Hip Arthroplasty? *The Journal of Arthroplasty* 2016;31:2332–6. <https://doi.org/10.1016/J.ARTH.2016.03.009>.

[17] Currier JH, Currier BH, Abdel MP, Berry DJ, Titus AJ, van Citters DW. What factors drive polyethylene wear in total knee arthroplasty? <https://doi.org/10.1302/0301-620X.103B11.BJJ-2020-2334R1> 2021;103-B:1695–701. <https://doi.org/10.1302/0301-620X.103B11.BJJ-2020-2334.R1>.

[18] Goldberg JR, Gilbert JL, Jacobs JJ, Bauer TW, Paprosky W, Leurgans S. A multicenter retrieval study of the taper interfaces of modular hip prostheses. *Clinical Orthopaedics and Related Research* 2002;401:149–61. <https://doi.org/10.1097/00003086-200208000-00018>.

[19] El-Zein ZS, Gehrke CK, Croley JS, Siljander MP, Mallow MA, Flierl MA, et al. Assessing Taper Geometry, Head Size, Head Material, and Their Interactions in Taper Fretting Corrosion of Retrieved Total Hip Arthroplasty Implants. *The Journal of Arthroplasty* 2021;36:S386-S394.e4. <https://doi.org/10.1016/J.ARTH.2021.02.041>.

[20] Hampton C, Weitzler L, Baral E, Wright TM, Bostrom MPG. Do oxidized zirconium heads decrease tribocorrosion in total hip arthroplasty? <https://doi.org/10.1302/0301-620X.101B4.BJJ-2018-1316R1> 2019;101 B:386–9. <https://doi.org/10.1302/0301-620X.101B4.BJJ-2018-1316.R1>.

[21] Siljander MP, Baker EA, Baker KC, Salisbury MR, Thor CC, Verner JJ. Fretting and Corrosion Damage in Retrieved Metal-on-Polyethylene Modular Total Hip Arthroplasty Systems: What Is the Importance of Femoral Head Size? *The Journal of Arthroplasty* 2018;33:931–8. <https://doi.org/10.1016/J.ARTH.2017.10.010>.

[22] Lange J, Wach A, Koch CN, Hopper RH, Ho H, Engh CA, et al. Do Well-functioning THAs Retrieved at Autopsy Exhibit Evidence of Fretting and Corrosion?

Clinical Orthopaedics and Related Research 2018;476:2017.

<https://doi.org/10.1097/CORR.0000000000000369>.

[23] Triantafyllopoulos GK, Elpers ME, Burket JC, Esposito CI, Padgett DE, Wright TM. Otto Aufranc Award: Large Heads Do Not Increase Damage at the Head-neck Taper of Metal-on-polyethylene Total Hip Arthroplasties. *Clinical Orthopaedics and Related Research* 2016;474:330. <https://doi.org/10.1007/S11999-015-4468-6>.

[24] Higgs GB, MacDonald DW, Gilbert JL, Rimnac CM, Kurtz SM, Chen AF, et al. Does Taper Size Have an Effect on Taper Damage in Retrieved Metal-on-Polyethylene Total Hip Devices? *The Journal of Arthroplasty* 2016;31:277–81. <https://doi.org/10.1016/J.ARTH.2016.06.053>.

[25] Brock TM, Sidaginamale R, Rushton S, Nargol AVF, Bowsher JG, Savisaar C, et al. Shorter, rough trunnion surfaces are associated with higher taper wear rates than longer, smooth trunnion surfaces in a contemporary large head metal-on-metal total hip arthroplasty system. *Journal of Orthopaedic Research* 2015;33:1868–74. <https://doi.org/10.1002/JOR.22970>.

[26] Kurtz SM, Kocagöz SB, Hanzlik JA, Underwood RJ, Gilbert JL, MacDonald DW, et al. Do Ceramic Femoral Heads Reduce Taper Fretting Corrosion in Hip Arthroplasty? A Retrieval Study. *Clinical Orthopaedics and Related Research* 2013;471:3270. <https://doi.org/10.1007/S11999-013-3096-2>.

[27] Tan SC, Teeter MG, del Balso C, Howard JL, Lanting BA. Effect of Taper Design on Trunnionosis in Metal on Polyethylene Total Hip Arthroplasty. *The Journal of Arthroplasty* 2015;30:1269–72. <https://doi.org/10.1016/J.ARTH.2015.02.031>.

[28] Panagiotidou A, Meswania J, Hua J, Muirhead-Allwood S, Hart A, Blunn G. Enhanced Wear and Corrosion in Modular Tapers in Total Hip Replacement Is Associated With the Contact Area and Surface Topography 2103. <https://doi.org/10.1002/jor.22461>.

[29] Mueller U, Braun S, Schroeder S, Sonntag R, Kretzer JP. Same Same but Different? 12/14 Stem and Head Tapers in Total Hip Arthroplasty. *The Journal of Arthroplasty* 2017;32:3191–9. <https://doi.org/10.1016/J.ARTH.2017.04.027>.

[30] Kao YYJ, Koch CN, Wright TM, Padgett DE. Flexural Rigidity, Taper Angle, and Contact Length Affect Fretting of the Femoral Stem Trunnion in Total Hip Arthroplasty. *The Journal of Arthroplasty* 2016;31:254–8. <https://doi.org/10.1016/J.ARTH.2016.02.079>.

[31] McCarthy SM, Hall DJ, Mathew MT, Jacobs JJ, Lundberg HJ, Pourzal R. Are Damage Modes Related to Microstructure and Material Loss in Severely Damaged CoCrMo Femoral Heads? *Clin Orthop Relat Res* 2021;479:2083–96. <https://doi.org/10.1097/CORR.0000000000001819>.

Chapter 3

3 A convolutional neural network for high throughput screening of trunnion corrosion

Anastasia M. Codirezzi¹, Brent A. Lanting², Matthew G. Teeter^{1,3}

1. School of Biomedical Engineering, Western University, London, Ontario, Canada
2. Division of Orthopaedic Surgery, Department of Surgery, Schulich School of Medicine & Dentistry, Western University, London, Ontario, Canada
3. Department of Medical Biophysics, Schulich School of Medicine & Dentistry, Western University, London, Ontario, Canada

Background: Corrosion at the modular head-neck taper interface of total and hemiarthroplasty hip implants (trunnionosis) is a cause of implant failure and clinical concern. The Goldberg corrosion scoring method is considered the gold standard for observing trunnionosis, but it is labour-intensive and often requires multiple observers. This limit the quantity of implants trunnionosis studies typically study. Machine learning, particularly convolutional neural networks, have been used in various medical imaging applications and corrosion detection applications to help reduce repetitive and tedious image identification tasks.

Methods: 725 modular femoral stem arthroplasty devices had their trunnion imaged in four positions and scored by an observer. A convolutional neural network was designed and trained from scratch using the images

Results: The convolutional neural network was able to distinguish no and mild corrosion from moderate and severe corrosion with an accuracy of 98.32%, a class 1 and 2 sensitivity of 0.9881, a class 3 and 4 sensitivity of 0.9556 and an area under the curve of 0.9740.

Conclusions: This convolutional neural network may be used as a screening tool to identify retrieved modular hip arthroplasty device trunnions for further study and the presence of moderate and severe corrosion with high reliability, reducing the burden on skilled observers in early stages of a study.

Keywords: machine learning, arthroplasty, trunnionosis, corrosion

3.1 Introduction

Trunnionosis refers to the fretting and corrosion of modular hip arthroplasty devices at the head-neck taper junction. This process can cause debris that has been shown to cause adverse tissue reactions and it has been of clinical concern because of its identification as a cause of hip arthroplasty failure [1–4]. Trunnionosis is believed to be underreported [1]. The presence of corrosion and fretting on explanted hip arthroplasty devices can be quantified using the Goldberg scoring method during implant retrieval studies [5]. Implant retrieval studies have helped identify areas for advancement in implant design, manufacturing, and installation, and they are of importance to identify specific device issues that may have been previously unknown [3,4,6–17]. Large-scale retrieved studies in knee arthroplasty have given new insight to drivers of wear, but similar studies for hip arthroplasty have yet to be completed on the same scale [12].

The Goldberg scoring method requires a trained observer to classify the corrosion and fretting class on the head and the taper typically under low-power microscopy. This method is time consuming and prohibits large scale study of retrieved arthroplasty devices. Studies that have looked at trunnionosis generally observe less than 150 implants, and often less than 100 [3,4,6–8,14,15,18,19]. Centers that do not have a research space must ship their implants to centers that do. This is logistically intensive and leads to infrequent collaboration between centers and limitations in the ability of centers to participate in implant orthopaedic research without an already established program.

Automation of corrosion detection would allow centers to collect data on explanted implants and identify implants that may require further analysis. It would reduce the labour-intensive aspect of Goldberg scoring and allow for more careful use of shipping implants and logistics. Machine learning has been successfully used in several imaging and corrosion detection classifications, and this may be extended to use for wear detection on medical devices [20–25]. Milimonfared et al. described an automated corrosion scoring approach with 85% accuracy, however, their method required substantial image pre-processing, and was trained using few trunnions of a specific design [25].

The purpose of the present investigation is to create a novel machine learning pipeline using a convolutional neural network that can rapidly identify implants that should be selected for further study using images of the implant trunnion.

3.2 Methods

3.2.1 Implant Imaging and Visual Scoring

All hip arthroplasty implants in our institutional implant retrieval laboratory were reviewed for inclusion (Figure 14). Implants included for imaging were designs with a modular head-neck taper where the femoral stem was retrieved at the time of revision surgery. Excluded were implants that were non-modular, and cases that had gross taper failure (“bird-beaking”). Each stem was imaged using a digital microscope at 20x magnification (DSX1000, Olympus Cooperation, Tokyo, Japan) at 1200x1200 and in RGB colour. The surface was divided into four areas, with each area represented by one image. Images were taken with the aid of an image diffuser when possible, as this minimized the amount of metallic glare. The images (Figure 2) were then examined and assigned a corrosion score using the method of Goldberg et al, described in Table 9. A single score was assigned to each image and recorded. Images were scored independently of other images in the trunnion set. Images were excluded if they were not of sufficient quality (unfocused, glare that obstructed view of 30% or more of the surface). A subset of 100 images from the test set was provided to a secondary observer to check for reliability of scoring between observers and calculated using the interclass correlation coefficient with a 95% confidence level.

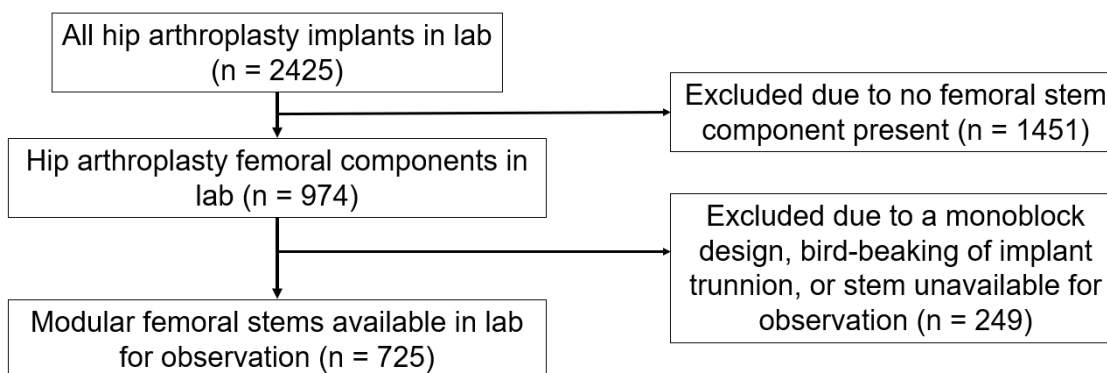


Figure 13: Study design for implant inclusion

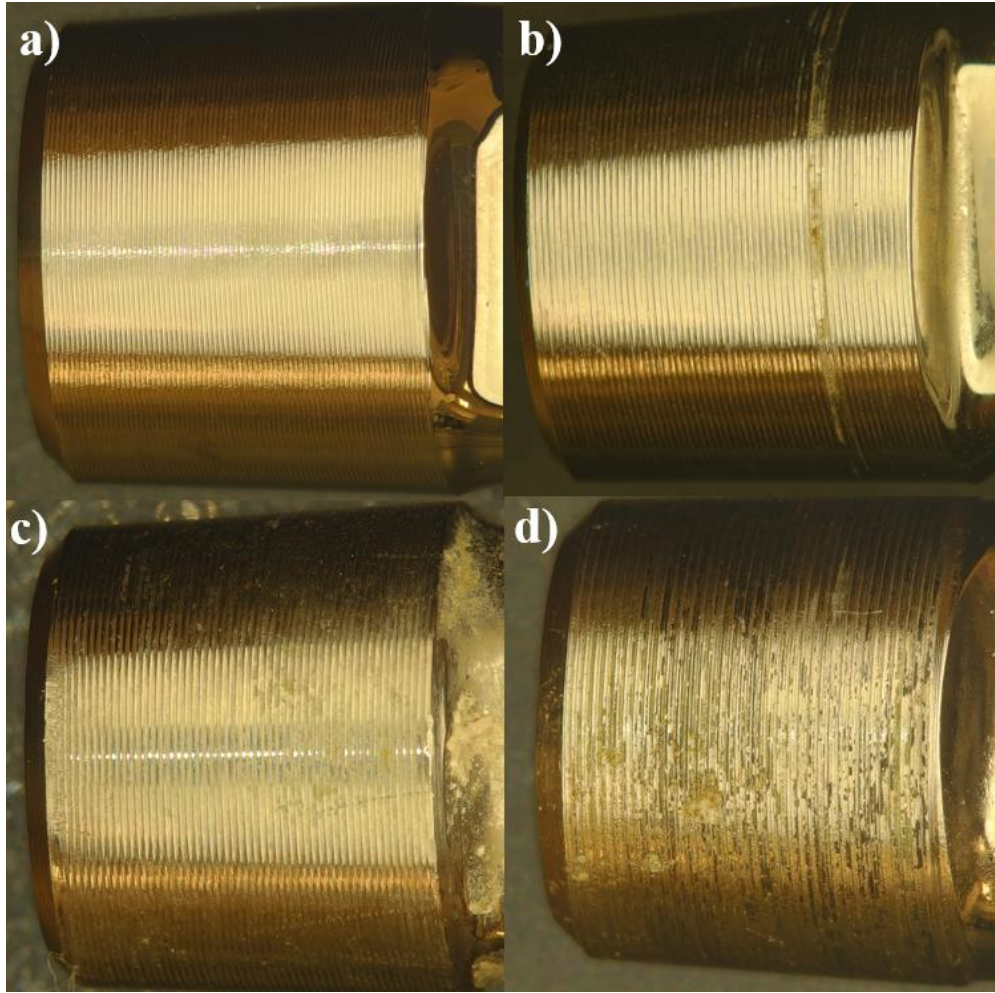


Figure 14: Representative images of each class (a) class 1, (b) class 2, (c) class 3, (d) class 4

Table 8: Goldberg scoring criteria [5]

	Class 1	Class 2	Class 3	Class 4
Severity	None	Mild	Moderate	Severe
Criteria	No visible corrosion observed	<30% of taper surface discoloured or dull	>30% of surface discoloured or dull, or, <10% of taper surface containing black debris, pits, or etch marks	>10% of taper surface containing black debris, pits, or etch marks

3.2.2 Data Curation

Images were sorted into classes based on their Goldberg score by a single observer (A. Codirezzi). 10% of the full dataset was separated out to create a testing set, keeping the proportion to their representation within the class. A subset of 100 images from the testing set were provided to a secondary observer to check the interclass correlation coefficient. Keeping in line with the best practice guidelines for machine learning for medical devices, this testing set was maintained separately from the training/validation set [26]. The data was organized into three datasets each with a training/validation set and an associated test set (Table 9). Each set included the same images in their training/validation and testing sets, but organized in a different manner.

Table 9: Description of the different datasets

Dataset	Dataset description
1. All classes (C1 vs C2 vs C3 vs C4)	All images included in their respective Goldberg corrosion class (class 1, class 2, class 3, class 4)
2. No corrosion/Corrosion (C1 vs C2, C3, C4)	All images included, separated into two classes, no corrosion and corrosion. No corrosion is comprised of class 1 images and corrosion is comprised of class 2, 3, and 4 images.
3. No corrosion and mild corrosion/moderate and severe corrosion (C1, C2 vs C3, C4)	No/mild corrosion class comprised of class 1 and 2 images and moderate/severe corrosion class comprised of class 3 and 4 images.

3.2.3 Neural Network Architecture

A convolutional neural network was designed and trained from scratch using MATLAB's DeepNetworkDesigner application (MATLAB for Windows, version 2021b). The network architecture was based off the concept of starting with a small network and expanding outward using a trial-and-error method, first using extreme cases (ie. Class 1 versus Class 4) and then including intermediate cases. A network diagram for the network used in this study is shown in Figure 15. It utilizes a single convolutional layer for feature learning. The same network architecture and training parameters were used for all datasets.

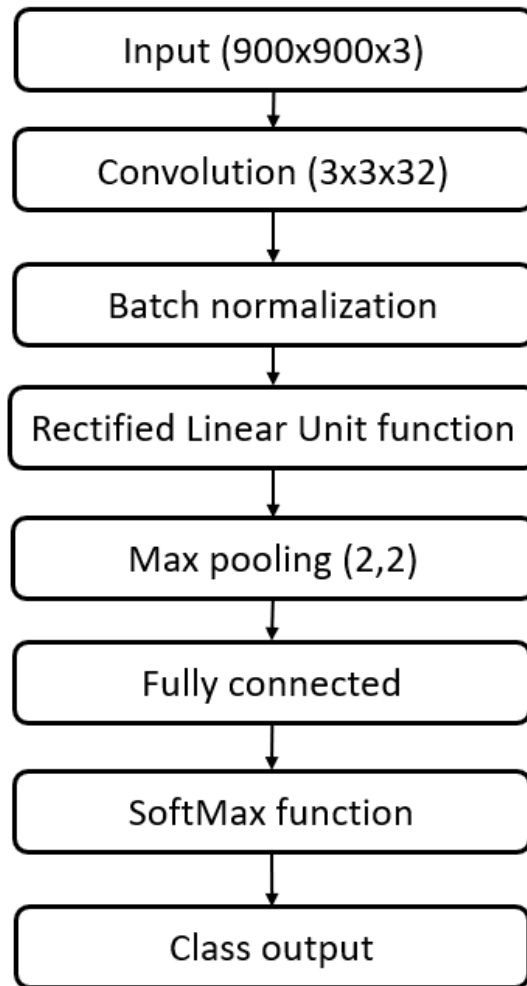


Figure 15: Convolutional neural network architecture. The neural network is comprised of an input of 900x900 images with RGB colouring. It then has a convolutional layer with a filter size of 3x3x32. Batch normalization was then employed. Then had a rectified linear unit activation function. That led to a max pooling layer with a filter size of 2x2 and a stride of 1,1. Then it had the fully connected layer, SoftMax function, and the output layer.

Three regularization techniques were used: batch normalization, L2 regularization, and early-stopping. Batch normalization normalizes along a mini-batch of the data across all observations for each colour channel independently. L2 regularization works by adding a term to the error function which prevents overfitting. Early stopping refers to ending the training before the determined number of epochs to prevent overfitting and generally an early stop is used when stagnation in the loss is observed during training [27].

3.2.4 Network Training and Testing

The same network architecture was trained separately using each of the three curated datasets. 10% of all images were split to create the holdout testing set. The training/validation set had 15% randomly split in order to create the validation set, with the remainder being used as the training set. The training set was used during training to which was used during training to spot-check the network and to aid in hyperparameter tuning of the network. The test set was used to evaluate the fully trained network. The convolutional neural network was trained separately using each dataset using the version with all images. The training parameters included using an ADAM optimizer with a learning rate of 0.003. All networks were given 14 epochs to train and the images were shuffled every epoch. The images were read in with a mini-batch size of 15 images and validation was done every 25 iterations. An L2 regularization of 0.001 was used.

Each trained network was tested using both versions of its associated test dataset, one with all images and one with images only taken with an image diffuser. Accuracy and sensitivity were computed. Accuracy refers to the overall proportion of the images that were correctly classified. Sensitivity refers to the proportion of images the network classified correctly for each class [27]. Confusion matrices were used to show the classifications made in each class, both correctly and incorrectly. The reliability of the neural network was evaluated by plotting the receiver operating characteristic curve (ROC) and determining the area under the curve (AUC). An area under the curve of 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding [28].

3.3 Results

3.3.1 Imaging and corrosion scoring

In total, 725 stems were imaged in four positions for a total of $n=2890$ images, with $n=10$ excluded due to poor image quality. The images were assigned a Goldberg corrosion score of class 1 ($n=1228$), class 2 ($n=1225$), class 3 ($n=335$), and class 4 ($n=102$). The test set was split off, comprising of 10% of the images ($n=298$), with $n=2592$ remaining in the training/validation set. The interclass correlation coefficient was reported to be 0.60 (+/- 0.13), rating as moderately reliable.

The testing images were further organized into two versions of the three datasets, one with all images and one with images that were taken without the diffuser adapter removed. Table 10 summarizes the content of each dataset.

Table 10: Datasets and images included

Dataset	All images		Images taken with an image diffuser
	Training/Validation	Testing	Testing
1	Class 1 (n=1100) Class 2 (n=1101) Class 3 (n=301) Class 4 (n=90)	Class 1 (n=128) Class 2 (n=125) Class 3 (n=34) Class 4 (n=12)	Class 1 (n=93) Class 2 (n=95) Class 3 (n=27) Class 4 (n=9)
2	Class 1 (n=1100) Class 2, 3, 4 (n=1492)	Class 1 (n=128) Class 2, 3, 4 (n=170)	Class 1 (n=93) Class 2, 3, 4 (n=131)
3	Class 1, 2 (n=2201) Class 3,4 (n= 391)	Class 1, 2 (n=252) Class 3,4 (n= 46)	Class 1,2 (n=188) Class 3,4 (n=36)

3.3.2 Neural Network Training and Evaluation

The confusion matrices are shown in Figures 16-18. Figure 16 shows the confusion matrices for dataset 1 (C1 vs C2 vs C3 vs C4). In both cases, the network failed to classify any test images as class 4 and the most common misclassification was classifying as class 2 when the true image class was class 1. Class 1 never misclassified an image on the other end of the spectrum from it (ie class 4 images were never misclassified as class 1, and class 1 images were never misclassified as class 3). Figure 17 shows the confusion matrices for dataset 2 (class 1 versus class 2, 3, 4). There were similar amounts of misclassifications for both classes.

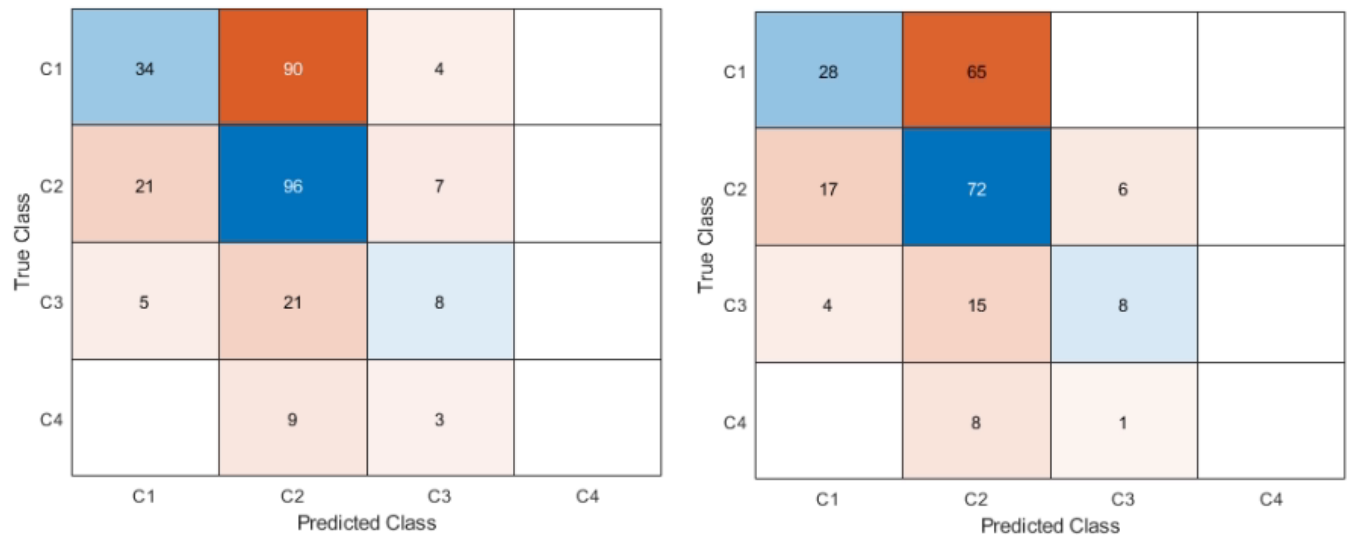


Figure 16: Confusion matrix for dataset 1 (class 1 versus class 2 versus class 3 versus class 4). Left is all images, right is with glare removed. The blue diagonal shows correct classifications (predicted class matches the true class) while the orange off-diagonal shows misclassifications. Intensity of colour is based off count in each category.

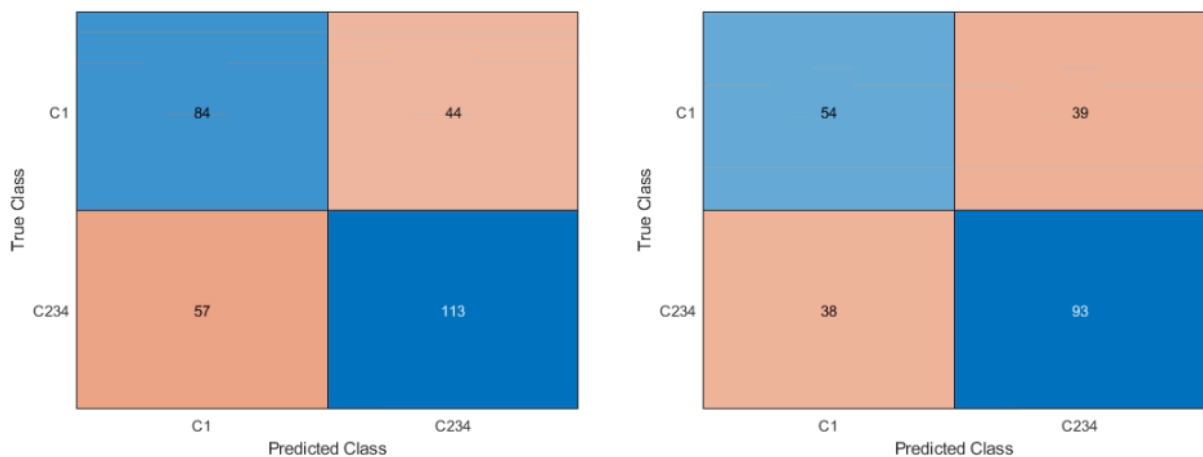


Figure 17: Confusion matrix for dataset 2, class 1 versus class 2,3,4. Right is all images, left is with glare removed. The blue diagonal shows correct classifications (predicted class matches the true class) while the orange off-diagonal shows misclassifications. Intensity of colour is based off count in each category.

Figure 18 shows the confusion matrices for dataset 3 (class 1 and 2 versus 3 and 4). This dataset had the fewest number of misclassifications, with little difference between if images with glare were included or not.

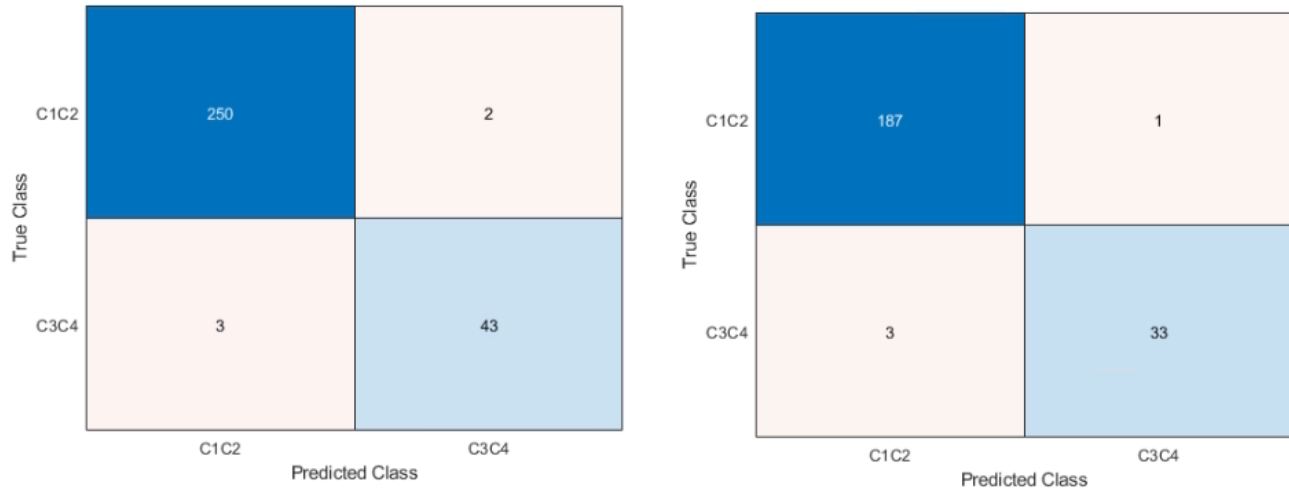


Figure 18: Confusion matrix for dataset 3, C12 versus C34. Right is all images, left is images with glare removed. The blue diagonal shows correct classifications (predicted class matches the true class) while the orange off-diagonal shows misclassifications. Intensity of colour is based off count in each category.

The receiver operating characteristic was plotted and the plots for all datasets with all images are shown in Figure 19. Classes 1 and 2 versus 3 and 4 (dataset 3) show the highest area under the curve and has the fewest individual points, showing that the network made guesses with similar probabilities for many of the images. Class 1 versus 2, 3 and 4 (dataset 2) had a much lower area under the curve, but showed many more points, showing a range in the probabilities for different guesses. Class 1 versus 2 versus 3 versus 4 (dataset 1), had the lowest area under the curve and showed a number or probabilities guessed, but significantly fewer outside the center of the graph.

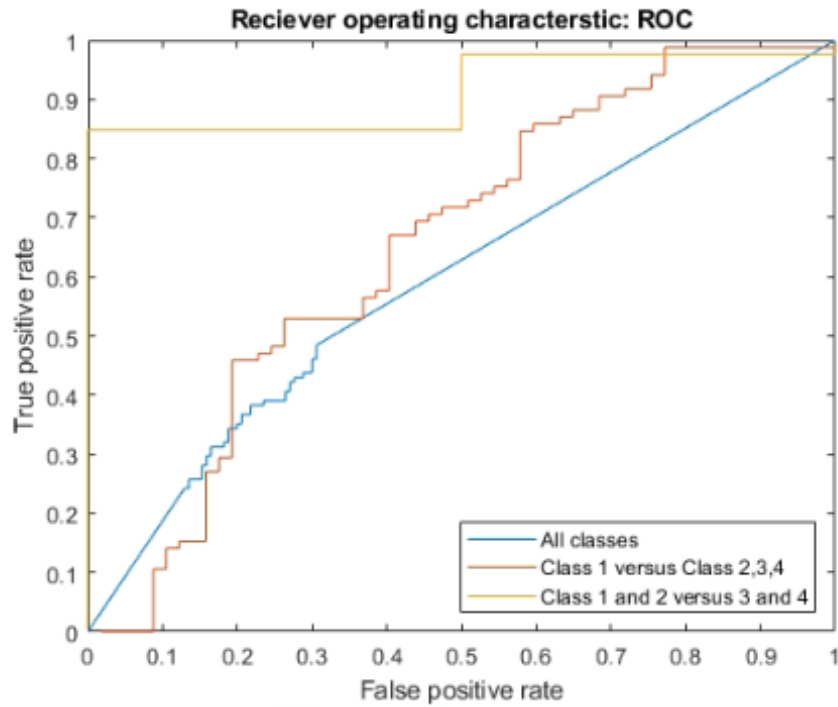


Figure 19: ROC for all datasets, with all images included

The accuracy, sensitivity for each class, and area under the curve were computed for each dataset and for both versions of the testing set. Table 11 shows a summary of these metrics.

Table 11: Computed error metrics for each dataset

Dataset	Sensitivity		Accuracy	Area under the curve (AUC)
1. All classes				
All images	C1	0.5667	48.21%	0.5941
	C2	0.4444		
	C3	0.3636		
	C4	0		
Glare removed	C1	0.5714	48.21%	0.6116
	C2	0.4500		
	C3	0.5333		
	C4	0		
2. C1 versus C234				
All images	C1	0.5957	66.11%	0.6875
	C234	0.7197		
Glare removed	C1	0.5869	65.63%	0.6661
	C234	0.7045		
3. C12 versus C34				
All images	C12	0.9881	98.32%	0.9740
	C34	0.9556		
Glare removed	C12	0.9842	98.31%	0.9693
	C34	0.9706		

Dataset 3, class 1 and 2 versus class 3 and 4, showed the highest sensitivity, accuracy and area under the curve for each class. The inability of the network to classify any images as

class 4 for dataset one is reflected in the 0 sensitivity for class 4 in both versions of the testing set.

3.4 Discussion

The ability to rapidly identify trunnions for further study is an important aspect in being able to achieve large-scale trunnion implant studies with reduced labour. In this study, a convolutional neural network was developed that could discriminate no and mild corrosion (Goldberg corrosion class 1 and 2) from moderate and severe corrosion (Goldberg corrosion class 3 and 4) with 98.32% accuracy.

Dataset one, which has the network classify the images into each Goldberg corrosion class separately, had the poorest performance. This dataset gives insight into the behaviour of the network. Of note, the network failed to classify any images (correctly or incorrectly) as class 4 during testing. There was significantly fewer class 4 images available than any other class. This is unsurprising, as class 4 corrosion is the rare occurrence of severe corrosion that has >10% of taper surface containing black debris, pits, or etch marks [5]. Although the network is trained with all available images, there was less opportunity for the network to learn from the class 4 images because they were so few in frequency. The images that were available for training may not have generalized well to the test set or there was so few that the network was never able to distinguish a high probability of class 4 for images. The confusion matrix also showed that there was a number of images that were classified as class 2 when their true class was class 1. When class 1 and class 2 images were combined in dataset 3, the network was able to perform with excellent accuracy, specificity, and a high area under the curve that showed excellent discrimination between the two classes.

Class 1 versus classes 2, 3, and 4 showed a marginal improvement in its sensitivity for detecting class 1 images, but the area under the curve remained similar to dataset 1 (class 1 versus class 2 versus class 3 versus class 4). The accuracies here are less meaningful as we compare between a multiclass and binary case, although the accuracy is significantly higher, the area under the curve still is poor.

The difficulty in discriminating between class 1 and class 2 Goldberg corrosion scoring could be due to the semi-quantitative nature of the Goldberg scale. Class 2 has the largest range of corrosion information- with any discolouration up to 30% of the surface being considered. Small patches of discolouration or corrosion may be difficult to distinguish on the network and may point to segmentation being necessary to distinguish these. The results of dataset three supports this, as when we combined class 1 and 2 versus 3 and 4, the network was able to distinguish with a high accuracy and very well. Segmentation is commonly used in other corrosion detection applications and in other applications before classification is done, but it was excluded here due to its intensity of requiring an observer to manually segment the training images [29,30].

It was anticipated that the glare on images taken without the diffusion adapter would likely affect the network's ability to discriminate the corrosion score. However, for all cases, there was little difference in the error metrics between having all the images and only images without glare. This is believed to be because the image scoring was done from the same photos as the network was given and the network was trained using images with glare as well. It was noted during the secondary observing of the images that some were difficult to score because of the glare. Best practice would be to take all images with a diffusion adapter; however incorporation of less-than-perfect images improves the generalizability of the results. There was no trend seen amongst different trunnion designs and misclassification of images, thus the network generalizes well across the difference designs present in the data. The interclass correlation coefficient of the test set that was scored was considered moderately reliable, which is consistent with a previous study determining the reliability of scoring [31].

A previous attempt to automate damage scoring of trunnions was done by Milimonfared et al with a reported accuracy of 85% and the ability to distinguish across the four classes using support vector machine learning [25]. They imaged 138 stems, with a total of eight images per stem. A description of the image population in each class was not shared. Accuracy was reported but additional metrics such as sensitivity, confusion matrices, and area under the curve were not reported. Furthermore, it does not appear they ensured every class was present in their testing set. Although their method had high reported accuracy,

without knowing information such as the sensitivity per class, it is difficult to determine if this pipeline could be used to reliably score implants. Polce et al. have pointed out that underreporting of models and network evaluation is a common theme in machine learning studies in total joint arthroplasty and has called for more reliable reporting, including adequate reporting of results beyond accuracy [32]. In contrast, our network can distinguish between class 1 and 2 versus class 3 and 4 with a higher accuracy than Milimonfared et al but it is unable to distinguish between the four classes effectively. We also had an increased number of trunnions available- including 725 trunnions in our study as opposed to Milimonfared's 138. The increased reporting, including class population descriptions, confusion matrices and class sensitivity characterize the network performance to better understand the reliability of the network and where it may fail. This network can be reliably used as a screening tool to select implants for further study but in its current state cannot be used to classify across the full scale. This tool is effective as a screening tool to identify implants for further research, especially implants that show a class 3 or 4 corrosion score.

Limitations of this study include an unbalanced dataset for testing and training. There was significantly more class 1 and 2 images available than class 3 and 4, ideally a balanced dataset is best practice for neural network training. This network also does not represent a full automation of damage scoring across the Goldberg scores, which is sought after to further reduce labour barriers to large-scale studies and the need for skilled observers. The images were largely scored by a single observer and intrareliability was not quantized. Furthermore, the images that this network was trained and tested on were taken using a high-quality digital microscope. This microscope is unlikely to be found in a center that does not have a strong research focus and limits the accessibility of this network to be used at smaller centers. Further studies should look to incorporate images acquired using various acquisition systems, such as a smartphone, to improve accessibility of this network. They should also involve a measure of intrareliability and a secondary scorer for the entire training dataset to ensure the network is being trained with the best possible quality of data.

3.5 Conclusions

In this study, a convolutional neural network was developed that could discriminate no and mild corrosion (Goldberg class 1 and 2) from moderate and severe corrosion (Goldberg

class 3 and 4) with 98.32% accuracy, and a class 1 and 2 sensitivity of 0.9881 and class 3 and 4 sensitivity of 0.9556. This network is suitable for use as a screening tool to discriminate class 1 and 2 implants from class 3 and 4, to help rapidly identify implants that should be considered for further study. Future work should include an expansion of the network to do full corrosion scoring and for images from a smartphone to be used in the network.

3.6 Acknowledgements

Implant retrievals were performed by author Dr Brent Lanting and 6 other fellowship trained arthroplasty surgeons. The authors would like to thank Dr Christopher Del Balso, Dr James Howard, Dr Robert Bourne, Dr Steve MacDonald, Dr Richard McCalden, Dr Douglas Naudie, Dr Cecil Rorabeck, and Dr Edward Vasarhelyi for their contribution. Anastasia M. Codireni is supported by a Transdisciplinary Bone and Joint Training Award. Dr Matthew G. Teeter is supported by an NSERC Discovery Grant. This study was conducted at a center that received institutional research support provided by Stryker, Smith and Nephew, and DePuy.

3.7 References

- [1] Shulman RM, Zywiell MG, Gandhi R, Davey JR, Salonen DC. Trunnionosis: the latest culprit in adverse reactions to metal debris following hip arthroplasty. *Skeletal Radiology* 2015;44:433–40. <https://doi.org/10.1007/S00256-014-1978-3/FIGURES/10>.
- [2] Cooper HJ, della Valle CJ, Berger RA, Tetreault M, Paprosky WG, Sporer SM, et al. Corrosion at the Head-Neck Taper as a Cause for Adverse Local Tissue Reactions After Total Hip Arthroplasty. *The Journal of Bone and Joint Surgery American Volume* 2012;94:1655. <https://doi.org/10.2106/JBJS.K.01352>.
- [3] Berstock JR, Whitehouse MR, Duncan CP. Trunnion corrosion: What surgeons need to know in 2018. *Bone and Joint Journal* 2018;100B:44–9. <https://doi.org/10.1302/0301-620X.100B1.BJJ-2017-0569.R1/ASSET/IMAGES/LARGE/BJJ-2017-0569.R1-GALLEYFIG1.JPEG>.
- [4] Banerjee S, Cherian JJ, Bono J v., Kurtz SM, Geesink R, Meneghini RM, et al. Gross Trunnion Failure After Primary Total Hip Arthroplasty. *The Journal of Arthroplasty* 2015;30:641–8. <https://doi.org/10.1016/J.ARTH.2014.11.023>.
- [5] Goldberg JR, Gilbert JL, Jacobs JJ, Bauer TW, Paprosky W, Leurgans S. A multicenter retrieval study of the taper interfaces of modular hip prostheses. *Clinical Orthopaedics and Related Research* 2002;401:149–61. <https://doi.org/10.1097/00003086-200208000-00018>.
- [6] Lavernia CJ, Iacobelli DA, Villa JM, Jones K, Gonzalez JL, Jones WK. Trunnion–Head Stresses in THA: Are Big Heads Trouble? *The Journal of Arthroplasty* 2015;30:1085–8. <https://doi.org/10.1016/J.ARTH.2015.01.021>.
- [7] Hussencocus S, Kosuge D, Solomon LB, Howie DW, Oskouei RH. Head-Neck Taper Corrosion in Hip Arthroplasty. *BioMed Research International* 2015;2015. <https://doi.org/10.1155/2015/758123>.

- [8] Tan SC, Teeter MG, del Balso C, Howard JL, Lanting BA. Effect of Taper Design on Trunnionosis in Metal on Polyethylene Total Hip Arthroplasty. *The Journal of Arthroplasty* 2015;30:1269–72. <https://doi.org/10.1016/J.ARTH.2015.02.031>.
- [9] Tan SC, Lau ACK, del Balso C, Howard JL, Lanting BA, Teeter MG. Tribocorrosion: Ceramic and Oxidized Zirconium vs Cobalt-Chromium Heads in Total Hip Arthroplasty. *The Journal of Arthroplasty* 2016;31:2064–71. <https://doi.org/10.1016/J.ARTH.2016.02.027>.
- [10] Hampton C, Weitzler L, Baral E, Wright TM, Bostrom MPG. Do oxidized zirconium heads decrease tribocorrosion in total hip arthroplasty? <https://doi.org/10.1302/0301-620X.101B4.BJJ-2018-1316R1> 2019;101 B:386–9. <https://doi.org/10.1302/0301-620X.101B4.BJJ-2018-1316.R1>.
- [11] Dyrkacz RMR, Brandt JM, Ojo OA, Turgeon TR, Wyss UP. The influence of head size on corrosion and fretting behaviour at the head-neck interface of artificial hip joints. *J Arthroplasty* 2013;28:1036–40. <https://doi.org/10.1016/J.ARTH.2012.10.017>.
- [12] Higgs GB, Macdonald DW, Gilbert JL, Rimnac CM, Kurtz SM. Basic Science Does Taper Size Have an Effect on Taper Damage in Retrieved Metal-on-Polyethylene Total Hip Devices? *The Journal of Arthroplasty* 2016;31:277–81. <https://doi.org/10.1016/j.arth.2016.06.053>.
- [13] Hothi HS, Eskelinen AP, Berber R, Lainiala OS, Moilanen TPS, Skinner JA, et al. Factors Associated With Trunnionosis in the Metal-on-Metal Pinnacle Hip. *J Arthroplasty* 2017;32:286–90. <https://doi.org/10.1016/J.ARTH.2016.06.038>.
- [14] del Balso C, Teeter MG, Tan SC, Lanting BA, Howard JL. Does the Additional Articulation in Retrieved Bipolar Hemiarthroplasty Implants Decrease Trunnionosis Compared to Total Hip Arthroplasty? *The Journal of Arthroplasty* 2018;33:268–72. <https://doi.org/10.1016/J.ARTH.2017.08.027>.

- [15] del Balso C, Teeter MG, Tan SC, Howard JL, Lanting BA. Trunnionosis: Does Head Size Affect Fretting and Corrosion in Total Hip Arthroplasty? *The Journal of Arthroplasty* 2016;31:2332–6. <https://doi.org/10.1016/J.ARTH.2016.03.009>.
- [16] El-Zein ZS, Gehrke CK, Croley JS, Siljander MP, Mallow MA, Flierl MA, et al. Assessing Taper Geometry, Head Size, Head Material, and Their Interactions in Taper Fretting Corrosion of Retrieved Total Hip Arthroplasty Implants. *The Journal of Arthroplasty* 2021;36:S386-S394.e4. <https://doi.org/10.1016/J.ARTH.2021.02.041>.
- [17] Lange J, Wach A, Koch CN, Hopper RH, Ho H, Engh CA, et al. Do Well-functioning THAs Retrieved at Autopsy Exhibit Evidence of Fretting and Corrosion? *Clinical Orthopaedics and Related Research* 2018;476:2017. <https://doi.org/10.1097/CORR.0000000000000369>.
- [18] Tan SC, Teeter MG, del Balso C, Howard JL, Lanting BA. Effect of Taper Design on Trunnionosis in Metal on Polyethylene Total Hip Arthroplasty. *The Journal of Arthroplasty* 2015;30:1269–72. <https://doi.org/10.1016/J.ARTH.2015.02.031>.
- [19] del Balso C, Teeter MG, Tan SC, Lanting BA, Howard JL. Taperosis: Does head length affect fretting and corrosion in total hip arthroplasty? *Bone and Joint Journal* 2015;97-B:911–6. <https://doi.org/10.1302/0301-620X.97B7.35149/ASSET/IMAGES/LARGE/35149-GALLEYFIG3.JPEG>.
- [20] Niculescu B, Faur CI, Tataru T, Diaconu BM, Cruceru M. Investigation of biomechanical characteristics of orthopedic implants for tibial plateau fractures by means of deep learning and support vector machine classification. *Applied Sciences (Switzerland)* 2020;10. <https://doi.org/10.3390/APP10144697>.
- [21] Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift Für Medizinische Physik* 2019;29:102–27. <https://doi.org/10.1016/J.ZEMEDI.2018.11.002>.

- [22] Abraham GK, Jayanthi VS, Bhaskaran P. Convolutional neural network for biomedical applications. *Computational Intelligence and Its Applications in Healthcare* 2020:145–56. <https://doi.org/10.1016/B978-0-12-820604-1.00010-8>.
- [23] Ronneberger O. Invited Talk: U-Net Convolutional Networks for Biomedical Image Segmentation 2017:3–3. https://doi.org/10.1007/978-3-662-54345-0_3.
- [24] Kang YJ, Yoo J il, Cha YH, Park CH, Kim JT. Machine learning–based identification of hip arthroplasty designs. *Journal of Orthopaedic Translation* 2020;21:13–7. <https://doi.org/10.1016/J.JOT.2019.11.004>.
- [25] Milimonfared R, Oskouei RH, Taylor M, Solomon LB. An intelligent system for image-based rating of corrosion severity at stem taper of retrieved hip replacement implants. *Medical Engineering & Physics* 2018;61:13–24. <https://doi.org/10.1016/J.MEDENGPHY.2018.08.002>.
- [26] Good machine learning practice for medical device development: Guiding principles - Canada.ca n.d. <https://www.canada.ca/en/health-canada/services/drugs-health-products/medical-devices/good-machine-learning-practice-medical-device-development.html> (accessed May 16, 2022).
- [27] Shalev-Shwartz S, Ben-David S. UNDERSTANDING MACHINE LEARNING From Theory to Algorithms n.d.
- [28] Andersen PK. 3. Applied Logistic Regression. 2nd edn. David W. Hosmer and Stanley Lemeshow. Wiley, New York, 2000. No. of pages: xii+373. Price: £60.95. ISBN 0-471-35632-8. *Statistics in Medicine* 2002;21:1963–4. <https://doi.org/10.1002/SIM.1236>.
- [29] Nash W, Zheng L, Birbilis N. Deep learning corrosion detection with confidence. *Npj Materials Degradation* 2022 6:1 2022;6:1–13. <https://doi.org/10.1038/s41529-022-00232-6>.
- [30] Audebert N, le Saux B, Lefèvre S. Segment-before-Detect: Vehicle Detection and Classification through Semantic Segmentation of Aerial Images. *Remote Sensing* 2017, Vol 9, Page 368 2017;9:368. <https://doi.org/10.3390/RS9040368>.

- [31] Hothi HS, Matthies AK, Berber R, Whittaker RK, Skinner JA, Hart AJ. The Reliability of a Scoring System for Corrosion and Fretting, and Its Relationship to Material Loss of Tapered, Modular Junctions of Retrieved Hip Implants. *The Journal of Arthroplasty* 2014;29:1313–7. <https://doi.org/10.1016/J.ARTH.2013.12.003>.
- [32] Kunze KN, Polce EM, Sadauskas AJ, Levine BR. Primary Knee Development of Machine Learning Algorithms to Predict Patient Dissatisfaction After Primary Total Knee Arthroplasty 2020. <https://doi.org/10.1016/j.arth.2020.05.061>.

Chapter 4

4 General Discussions and Conclusions

4.1 Overview of Objectives

This thesis sought to explore large-scale studies within retrieved hip arthroplasty device analysis. First, a large-scale study of hip arthroplasty devices and patient factors and their relationship to trunnionosis was explored through a statistical analysis of the factors. Secondly, an attempt to automate the corrosion detection of devices to identify implants more rapidly for corrosion using a convolutional neural network was explored to help reduce the labour requirement for large-scale trunnionosis studies. Together, these objectives illustrate the value of large-scale studies and showing that “big data” is relevant to hip arthroplasty, as well as the need to simplify data collection of these devices.

4.2 Summary of Results

Chapter 2 included the large-scale analysis of hip arthroplasty devices saw 664 modular femoral stem components analysed, collecting both patient and device characteristics and relating them to the presence and severity of trunnionosis. It was found that greater duration of implantation and larger head size were associated with elevated corrosion class. Older age at surgery, a stainless-steel stem material, and indication for revision a bone or periprosthetic fracture and infection, and certain larger taper geometries were associated with a decreased corrosion class. This large-scale study of retrieved hip arthroplasty trunnions is the first of its kind.

Chapter 3 explored automation of damage scoring on the Goldberg scale using a convolutional neural network. 725 implants were imaged, the largest imaging undertaking of trunnions to date, and a convolutional neural network was designed and trained from scratch to be able to classify the trunnion images. The study was successful in creating a convolutional neural network with the ability to distinguish between no to mild corrosion versus moderate to severe corrosion with high accuracy and reliability. This tool is suitable for use in a research environment to screen for moderate/severely corroded implants to identify for further study.

4.3 Limitations

Both the studies in this thesis were limited to the implants present in the implant retrieval laboratory, which reflects the implants used in Southwestern Ontario and may not be reflective of overall implant usage in Canada or globally. Similarly, the implants are explanted when they have failed, well-functioning implants are absent from both analysis because of the lack of a cadaver retrieval program as is present in some other regions.

The large-scale study relating trunnion corrosion to device and patient factors was further limited by the availability of device information. The registration of orthopaedic devices as they are installed through databases such as Ortech allow for comprehensive device information to be available and quickly accessed for large-scale studies. Retrieved implants can be assessed to determine some device factors, but this can be unreliable as information such as material and offset are not printed on the physical device by every manufacturer. Furthermore, it adds significant labour and in the case of this study, led to the exclusion of implants and device information (such as offset). Similarly, the retrieved implants were overwhelmingly with one head material, chromium cobalt. Ceramic and oxinium heads have previously been identified as having lower trunnion corrosion than chromium cobalt heads, but there was not a significant enough presence of these heads in the study and a relationship was not identified.

The study that created a convolutional neural network to distinguish between no/mild corrosion and moderate/severe corrosion was limited by the availability of images in Goldberg corrosion classes 1-4. Class 1 and 2 had about equal representation of instances, but class 3 and 4 had significantly less images available (class 3 had about 1/3rd the available images when compared to class 1 and 2, while class 4 had 1/10th). The few instances of class 4 specifically led to having to combine class 3 and 4 images to train the network. This study also did not explore the use of aggressive augmentation in images to try to limit the class imbalances seen in the images. Transfer learning was also not considered due to the image size but may be useful in a segmentation application to create masks that show corrosion before attempting to place them into their Goldberg class.

4.4 Applications and Future Directions

The results from the large-scale study show that there are patient and device factors that generalize across different models of stems. Previously studies have focused strongly on specific models and manufacturers which has limited the ability to generalize factors that influence implant failure. This study shows the importance of large-scale data collection and the need for more centers to utilize systems like Ortech to allow for mass studies to occur with reduced labour. Moving forward, additional patient and device factors may be studied using these methods to allow for a more general understanding of what factors affect trunnionosis and early device failure.

The convolutional neural network for rapid identification of no/mild and moderate/severe corrosion may be used as a screening tool in centers that perform arthroplasty retrieval. In its current state, it requires a similar imaging setup, but future work should include the ability to use images from various sources (ex. iPhone, point-and-shoot camera), and even live identification. By including more images, there also is an ability to further train the network, especially with more class 3 and 4 images to eventually achieve the goal of full Goldberg classification of the images. The implementation of a segmentation pipeline before the classification pipeline may also help distinguish better between class 1 and 2, which did have equal images present, but the broadness of the Goldberg scale for class 2 made it difficult for the network to identify and distinguish class 1 and 2 effectively.

Achieving this would contribute a strong research tool that could even be used to observed *in situ* stems during revision surgeries, allowing us to have a better understanding of the true prevalence of trunnionosis and the factors that contribute to it.

4.5 Conclusions

In conclusion, this Master's thesis has shown the importance of large-scale studies in hip arthroplasty as it related to trunnionosis, and it has developed a tool to help aid in rapid identification of moderate/severe corrosion to help identify implants for further study. As the subject of 'big-data' grows within healthcare, it is important that orthopaedics and arthroplasty engage in big-data research. This thesis offers a path forward for both rapid

identification of implants required for further research and shows the value of collecting patient and device information gradually and in a format that allows for big-data methods to be explored.

Appendices

Appendix A: Study Approvals



LAWSON FINAL APPROVAL NOTICE

LAWSON APPROVAL NUMBER: R-22-116

PROJECT TITLE: Retrospective retrieved total hip arthroplasty femoral stem review

PRINCIPAL INVESTIGATOR: Dr. Matthew Teeter

LAWSON APPROVAL DATE: 4/03/2022

ReDA ID: 12108

Overall Study Status: Active

Please be advised that the above project was reviewed by Lawson Administration and the project was approved.

“COVID-19: Please note that Lawson is continuing to review and approve research studies. However, this does not mean the study can be implemented during the COVID-19 pandemic. Principal Investigators, in consultation with their program leader or Chair/Chief, should use their judgment and consult [Lawson’s research directive and guidelines](#) to determine the appropriateness of starting the study. Compliance with hospital, Lawson, and government public health directives and participant and research team safety supersede Lawson Approval.”

Please provide your Lawson Approval Number (R#) to the appropriate contact(s) in supporting departments (eg. Lab Services, Diagnostic Imaging, etc.) to inform them that your study is starting. The Lawson Approval Number must be provided each time services are requested.

**Dr. David Hill
V.P. Research
Lawson Health Research Institute**

Curriculum Vitae

Name: Anastasia M. Codirezzi

Post-secondary Education and Degrees:

MESc in Biomedical Engineering with Collaborative Specialization in Musculoskeletal Health
2020 – 2022, Western University, London, Ontario

BSc Honours Specialization in Integrated Sciences with Physics

2016 – 2020, Western University, London, Ontario

Thesis title: Luminescence and defect engineering, silicon quantum dots in $\alpha - \text{Al}_2\text{O}_3$

Honours and Awards

2021-2022

- Queen Elizabeth II Graduate Scholarship in Science and Technology (QEII), Government of Ontario
- Collaborative Specialization in Musculoskeletal Health Research Transdisciplinary Training Award, Bone & Joint Institute, Western University

2020

- Donald R. Hay Prize, Department of Physics & Astronomy, Western University
- Physics Undergraduate Research Award of Distinction, Department of Physics & Astronomy, Western University

2018

- Ontario Baden-Württemberg Scholar, Ontario Universities International

2016

- Western Scholarship of Excellence, Western University

Related Work Experience

- Graduate Student Assistantship, Implant Retrieval Laboratory
2020 – 2022, Department of Medical Biophysics, Schulich School of Medicine and Dentistry, Western University
- Teaching Assistant, Foundations of Engineering Practice (ES1050)
2021 – 2022, Faculty of Engineering, Western University

- Teaching Assistant, Electromagnetism for Electrical Engineers (ECE3336) 2021, Faculty of Engineering, Western University
- Content Creator, Introductory Astronomy (ASTRO1021) 2018 – 2019, Faculty of Science, Western University

Publications

1. Y. Zhao, M. Amirmaleki, Q. Sun, C. Zhao, **A. Codirenze**, L.V.Goncharova, C. Wang, K. Adair, X. Li, X. Yang, F. Zhao, R. Li, T. Filleter, M. Cai, X. Sun, “Natural SEI-Inspired Dual-Protective Layers via Atomic/Molecular Layer Deposition for Long-Life Metallic Lithium Anode”, *Matter* Vol 1(5), 2019.
2. **A. M. Codirenze**, M. G. Teeter, B. A. Lanting, “What patient and implant factors affect trunnionosis severity? An implant retrieval study of 664 femoral stems”. *Submitted to Journal of Arthroplasty*.
3. E. Jin, K. Tantratian, C. Zhao, **A. Codirenze**, Lyudmila V. Goncharova, C. Wang, F. Yang, P. Pirayesh, J. Guo, L. Chen, X. Sun, Y. Zhao, “Ionic Conductive and Highly-Stable Interface for Alkali Metal Anodes,” *Submitted to Advanced Energy Materials*.
4. **A. M. Codirenze**, M. G. Teeter, B. A. Lanting, “A convolutional neural network for high throughput screening of trunnion corrosion”. *Manuscript in preparation*.

Abstracts and Presentations

1. **Anastasia Codirenze**, Brent A. Lanting, Matthew G. Teeter, “What patient and implant factors affect trunnionosis severity? An implant retrieval study of 664 femoral stems”. *2022 International Society for Technology in Arthroplasty*, Maui, Hawaii, USA, August 31st – September 3rd 2022. Accepted for presentation.
2. **Anastasia Codirenze**, Brent A. Lanting, Matthew G. Teeter, “Automated detection of trunnion corrosion on retrieved hip arthroplasty,” *2022 COA/CORS/CORA Annual Meeting*, Quebec City, Quebec, June 8th – 11th, 2022. Poster presentation.
3. **Anastasia Codirenze**, Brent A. Lanting, Matthew G. Teeter, “A convolutional neural network for detection of corrosion on retrieved hip arthroplasty systems” *Imaging Network Ontario 2022*, Virtual Conference, March 22-24th, 2022. Pitch presentation.
4. **Anastasia Codirenze**, Brent A. Lanting, Matthew G. Teeter, “Convolutional neural networks for detection of trunnion corrosion on retrieved total hip arthroplasty,” *Orthopaedic Research Society Annual Meeting*, Tampa, Florida, February 4th-8th, 2022. Poster displayed *in absentia*.
5. **Anastasia Codirenze**, Lyudmila V. Goncharova, Yang Zhou, Xeuling Sun, “Proof of concept for in-situ ion beam observation of a Li-ion cell,” *National Association*

of Corrosion Engineers: Southern Ontario Student Section 10th Annual Symposium, McMaster University, Hamilton, Ontario (Online Format), July 23rd-24th, 2020. Poster presentation.

6. **Anastasia Codirezzi**, Lyudmila V. Goncharova, Peter J. Simpson, “Luminescence and defect engineering: silicon quantum dots in $\alpha\text{-Al}_2\text{O}_3$,” *Physics Undergraduate Conference (PhUnC)*, Western University, London Ontario, March 13th, 2020. Poster Presentation.
7. **Anastasia Codirezzi**, Lyudmila V. Goncharova, “Li-ion cell stack design for in-situ observation,” *Physics Undergraduate Conference (PhUnC)*, Western University, London Ontario, March 13th, 2020. Poster Presentation.
8. Lyudmila V. Goncharova, **Anastasia Codirezzi**, Yang Zhou, Xueling Sun, “In-situ and ex-situ IBA for Atomic/Molecular Layer Deposition of Protective Layers for Li electrodes,” *24th International Conference on Ion Beam Analysis*, Antibes, French Riviera, France, October 13th-18th, 2019. Poster presentation.

Professional Development

- Graduate Scholars Innovation Program, Winter 2022 Cohort
- Fundamentals of OCAP, First Nations Information Governance Centre, November 2021
- Understanding How Medical Devices are Regulated in Canada, Health Canada, October 2021
- Siemens Healthineers Innovation Think Tank, Western University, October 2021
- Ivey International Centre for Health Innovation Workshops
 - ❖ Establishing and Sustaining High Functioning Teams, March 2022
 - ❖ Systems Thinking, October 2021
 - ❖ Operations & Supply Chain Management, February 2021
 - ❖ Value Based Healthcare & Payment Systems, January 2021
 - ❖ Managerial Accounting, November 2020