The Journal of Arthroplasty xxx (2019) 1–7

Contents lists available at ScienceDirect

The Journal of Arthroplasty

journal homepage: www.arthroplastyjournal.org

What Can We Expect Following Anterior Total Hip Arthroplasty on a Regular Operating Table? A Validation Study of an Artificial Intelligence Algorithm to Monitor Adverse Events in a High-Volume, Nonacademic Setting

Casper Van de Meulebroucke, MD, Joris Beckers, MD, Kristoff Corten, MD, PhD *

Orthopedic Department, Hip Unit, Ziekenhuis Oost-Limburg, Genk, Belgium

ARTICLE INFO

Article history:
Received 12 June 2019
Received in revised form
17 July 2019
Accepted 30 July 2019
Available online xxx

Keywords:
total hip arthroplasty
clinical research informatics
search algorithm
complication reporting
artificial intelligence

ABSTRACT

Background: Quality monitoring is increasingly important to support and assure sustainability of the orthopedic practice. Surgeons in nonacademic settings often lack resources to accurately monitor quality of care. Widespread use of electronic medical records (EMR) provides easier access to medical information, facilitating its analysis. However, manual review of EMRs is highly inefficient. Artificial intelligence (AI) software allows for the development of algorithms for extracting relevant complications from EMRs. We hypothesized that an AI-supported algorithm for complication data extraction would have an accuracy level equal to or higher than manual review after total hip arthroplasty (THA).
Methods: A total of 532 consecutive patients underwent 613 THA between January 1 and December 31, 2017. A random derivation cohort (100 patients, 115 hips) was used to determine accuracy. After generation of a gold standard, the algorithm was compared to manual extraction to validate performance in raw data extraction. The full cohort (532 patients, 613 hips) was used to determine recall, precision, and F-value.
Results: AI accuracy was 95.0%, compared to 94.5% for manual review ($P = .69$). Recall of 96.0% (84.0%-100%), precision of 88.0% (33%-100%) and F-measure of 0.85 (0.5-1) was achieved for all adverse events. No adverse events were recorded in 80.6%, 1.3% required reintervention and 18.1% had "transient" events.
Conclusion: The use of an automated, AI-supported search algorithm for EMRs provided continuous feedback on the quality of care with a performance level comparable to manual data extraction, but with greater speed. New clinical information surfaced, as 18.1% of patients can be expected to have "transient" problems.

Total hip arthroplasty (THA) has often been called one of, if not the most, successful orthopedic procedures developed in the twentieth century [1]. The number of THA is expected to grow along with its impact on healthcare budgets [2,3]. Greater government spending, however, warrants proper quality control and accountability. The importance of quality monitoring is expected to grow in the near future, increasing its influence on the orthopedic practice and on medicine in general. However, many surgeons in a nonacademic setting often lack the manpower and resources to accurately monitor their quality of care. The widespread use of electronic medical records (EMR) provides easier access to vast quantities of medical information and facilitates its analysis. Manual review of these medical records remains the golden standard, despite its inefficient and time-consuming character.

Although THA offers low complication rates, research has shown that reporting of postoperative surgical adverse events is rarely performed in a structured manner [4–7]. Reviewing sufficiently large numbers of records within a reasonable time frame remains a major obstacle monitoring the quality of care following THA. Natural language processing technologies using artificial intelligence (AI) with machine learning has allowed for the development of an automated, electronic search algorithm capable of extracting relevant postoperative complications from "unstructured" EMR. We questioned whether such an AI-supported algorithm could be used to provide accurate, continuous feedback on the quality of care following THA in a high-volume, nonacademic clinical setting. We

One or more of the authors of this paper have disclosed potential or pertinent conflicts of interest, which may include receipt of payment, either direct or indirect, institutional support, or association with an entity in the biomedical field which may be perceived to have potential conflict of interest with this work. For full disclosure statements refer to https://doi.org/10.1016/j.arth.2019.07.039.

* Reprint requests: Kristoff Corten, MD, PhD, Hip Unit, Orthopedic Department, Ziekenhuis Oost-Limburg, Schiepse Bos 6, 3600 Genk, Belgium.

https://doi.org/10.1016/j.arth.2019.07.039
0883-5403/

hypothesized that an automated algorithm would achieve a performance level similar to or superior to that of a human reviewer (ie, the gold standard).

The primary goal of this study is to develop and train an automated search algorithm for extracting key clinical concepts from unstructured data in outpatient, hospitalization, and surgery reports from a high-volume, nonacademic center. The secondary goal was to evaluate and compare the accuracy of complication detection and extraction to the gold standard of "manual extraction."

Methods

The clinical, surgery, and hospitalization notes of 532 consecutive patients undergoing 613 primary THA between January 1 and December 31, 2017, were prospectively followed by a human reviewer (C.V.D.M.) and by an automated AI algorithm (LynxCare, Brussels, Belgium). All patients were operated and followed by the senior surgeon (K.C.). The surgery was conducted through the efficient direct anterior approach on a regular operating table. The same cementless implant, Corail stem, and Pinnacle socket (DePuy Synthes, Warsaw, IN) were used in all cases. Exclusion criteria were patients younger than 18 years, failure to provide informed consent to prospective follow-up, and revision surgery. Approval by the local ethical committee was obtained. All patients were prospectively followed preoperatively and postoperatively at 6 weeks, 3 months, and 1 year. Patients who encountered adverse events were seen more frequently depending on the course of the event. The entire study group was seen at our outpatient clinic by the surgeon who created all clinical notes and reported every adverse event or complication. All patients were requested to provide consent for the automated monitoring algorithm. They were followed throughout an automated patient care path in which the AI algorithm was embedded and which provided detailed information regarding the procedure (SmartHip; DEO, Brussels, Belgium). The algorithm is General Data Protection Regulation approved which is in accordance with the Belgian laws of privacy protection. The algorithm is connected to the EMR system of the hospital. It is connected to the personal files of the patient and is automatically updated in case new reports are being generated. The algorithm screens the clinical, surgery, and hospitalization notes for key sentences, key words, and clinical definitions that are compiled by the surgical team. A list of adverse events based on the study of the American Hip Society [8] was compiled. All 19 complications were retained, although 2 modifications were made, as displayed in Table 1. First, a separate category was added for lower back pain and/or sciatica. Second, other soft tissue complications such as groin pain, psoas tendinitis, adductor tendinitis, and tensor fascia latae pain were added next to abductor muscle weakness. Five separate subtypes of muscular complications were defined concerning pain or weakness related to the abductors, psoas, adductors, tensor fascia latae, and quadriceps. Treatment was categorized as either conservative (physiotherapy or outpatient steroid injection) or interventional (new surgery).

Development and validation of the algorithm was done in 3 phases. There is no readily available gold standard for the use of natural language processing in orthopedics. As such, it was necessary to create a training set for initial retraining of the algorithm. By using the validation/annotation interface, an experienced orthopedic surgeon manually annotated 60 EMRs (on average 3 documents per patient) to generate relevant input for the algorithm to train on. Sample size was based on previous experience with the AI in cardiology research, showing that 50 EMRs are sufficient to achieve accuracy over 90%. In total, 180 documents were annotated with an equal distribution of operative reports, 6-week, 3-month, and 1-year consultation reports. No further direct manual verification of extracted data was performed afterward.

Afterward, an independent gold standard was generated. Two independent reviewers (J.B. and C.M.) manually reviewed a derivation cohort of 100 EMRs (115 hips). The extraction by these 2 reviewers allowed to compose an exhaustive list of complications present in the derivation cohort. These 2 reviewers were not involved in the treatment of patients or in the development of the algorithm.

In the third phase, the entire dataset of 532 patients (613 hips) was manually reviewed by 2 reviewers (C.V.D.M. and J.B.) in order to calculate the program's recall, precision, and F-measure.

Patient records were analyzed in a "classical manner" by a manual reviewer using the hospital's EMR system. Every hospitalization report, surgery report, and postoperative outpatient clinical report was manually reviewed. Radiographic, laboratory, and pathology reports were not reviewed. Surgery and implant characteristics such as implant size and implant articulation were extracted. Any reported adverse events and their respective treatment were recorded.

The algorithm consists of a text mining engine based on natural language processing technology and machine learning to extract key concepts from EMRs. A validation interface was provided to manually verify the output of the system and provide user feedback for automatic retraining of the AI. This interface was only used during development, not during the validation process. Adverse events and their respective treatment, the operated side, as well as implant characteristics were collected into a database. The extracted clinical concepts were matched with standard terminologies (Systematized Nomenclature of Medicine—Clinical Terms [SNOMED-CT]). These concepts have been encoded into the natural language processing algorithm and allow for unique concept identification and secondary usage of these data for information and knowledge processing as well as language independent benchmarking on a large scale. Validation of the algorithm in a clinical research context was achieved by applying the algorithm to patient records and cross-validating the extracted data with data obtained via the gold standard, that is, manual extraction. First, a random derivation cohort of 100 patients (115 hips) was used to determine the accuracy of the AI algorithm in raw data extraction. Accuracy is a measure of the percentage of relevant concepts that are extracted by the algorithm. Accuracy is defined as the total number of true-positive and false-negative concepts extracted by the algorithm or the human reviewer, divided by the total number of relevant concepts. A scoring system was used in which one point was attributed to each correct clinical concept. Points were distributed based on extraction by the reviewer or the algorithm as illustrated in Figure 1. A correct clinical concept was defined as a match with the gold standard. The system was applied to every listed clinical concept. Using the total scores, accuracy was calculated as displayed in Figure 2 and Table 2.

Secondly, the full cohort of 532 patients and 613 hips was used to determine the precision and recall values of the algorithm in classifying and interpreting free text. R-statistical programming was used to calculate accuracy, F measures, precision, and recall of the algorithm. The program was configured to recognize whether an extracted concept was relevant and to match it with SNOMED-CT terminology. Recall, precision, and F-value were used as statistical parameters (Fig. 1). Precision was chosen over specificity because it better illustrates a program's ability to correctly classify a concept, whereas specificity is more useful for a diagnostic test. Relevance can be determined by person, date, or context (eg, hypothetical phrasing or a negation). The algorithm was trained to tag a concept as relevant, negated, or hypothetical. Recognition of temporality and subject (patient, physician, or other) were also trained. The generated database was queried on the relevant clinical concepts using SNOMED-CT coding. A gold standard was generated through manual review of the entire database (532 patients, 613 hips) by 2 independent reviewers (J.B. and C.V.D.M.), allowing the calculation of recall, precision, and F-value.

Table 1
Complication and Intervention Categories.

Complication	Subgroup	Definition	Treatment: Conservative (C) Interventional (I)
Soft tissue complications			
Wound complications	Bleeding	Symptomatic and/or significant postoperative bleeding.	C: local treatment, transfusion I: redo wound closure, surgical debridement
	Hematoma	Symptomatic and/or significant postoperative hematoma.	C: local treatment I: aspiration, surgical debridement
	Prolonged wound drainage	Clinically significant postoperative wound drainage.	C: local treatment I: redo wound closure, surgical debridement
Infection	Superficial wound infection	Wound infection requiring antibiotic treatment or surgical debridement.	C: antibiotic and local treatment I: readmission, surgical debridement
	Deep PJI	PJI was graded using a modified MSIS criteria [9].	C: suppressive antibiotic treatment I: debridement and implant retention, 1- or 2-stage revision
Neural deficit	Motor nerve deficit	Postoperative motor nerve dysfunction.	C: expectation I: surgical exploration
	Sensory deficit	Postoperative sensory nerve dysfunction.	C: expectation I: surgical exploration
Vascular injury		Intraoperative vascular injury requiring surgical intervention.	C: N/A I: endovascular or surgical intervention
HTO		Symptomatic HTO associated with stiffness and reduced ROM.	C: medical treatment, physiotherapy I: surgical excision of HTO, irradiation
Muscle pain or weakness	Abductor muscles or greater trochanter	Pain or gait disturbance due to abductor muscle deficiency. Greater trochanter pain.	C: expectation, physiotherapy modification, steroid injection I: surgical intervention
	Psoas muscle and tendon	Pain or loss of muscle strength related to psoas tendinitis or dysfunction.	C: expectation, modified physiotherapy, steroid injection I: surgical intervention
	Adductor	Clinically significant symptoms related to adductor muscles.	C: expectation, modified physiotherapy, steroid injection I: surgical intervention
	TFL	TFL-related pain or weakness.	C: expectation, modified physiotherapy, steroid injection I: surgical intervention
	Quadriceps	Loss of strength or pain related to quadriceps muscles.	C: expectation, modified physiotherapy, steroid injection I: surgical intervention
Lower back pain	Lumbalgia	Postoperative lower back pain without leg pain.	C: analgesia, physiotherapy, pain-therapeutic treatment I: spinal surgery
	Sciatica	Postoperative radicular pain, motor weakness, or sensory deficit.	C: analgesia, physiotherapy, pain-therapeutic treatment I: spinal surgery
Thromboembolic disease		Postoperative thromboembolic event.	C: anticoagulant therapy I: endovascular intervention
Mechanical complications			
Instability		Prosthetic joint dislocation or subluxation requiring treatment.	C: N/A I: closed reduction, revision
Periprosthetic fracture	Intraoperative	Intraoperative fracture	C: weight-bearing restrictions, intraoperative cerclage I: osteosynthesis, intraoperative implant modification
	Postoperative	Postoperative femur or pelvic fracture	C: weight-bearing restrictions I: osteosynthesis, revision
Implant fracture			C: N/A I: revision
Loosening		Radiographic or intraoperatively identified implant loosening.	C: expectation I: revision
Cup-liner dislocation		Dislocation of the cup liner from the acetabular shell.	C: N/A I: revision
Bearing surface wear			C: expectation I: revision
Osteolysis		Expansile lytic lesion adjacent to implant.	C: expectation I: revision
Intervention			
Intraoperative	Osteosynthesis	Intraoperative fracture treatment.	
Outpatient	Infiltration	Local corticosteroid or anesthetic injection.	
	Aspiration	Diagnostic or evacuating joint aspiration.	
Reoperation	Wound debridement	Return to operating room for wound-related complication.	
	Closed reduction	Closed reduction of postoperative dislocation.	
	Osteosynthesis/cerclage	Return to operating room for treatment of postoperative fracture.	
Revision		Removal or exchange of THA components.	
Readmission		Readmission for adverse event related to THA.	
Death		Death attributed to adverse event after THA.	

PJI, prosthetic joint infection; MSIS, Musculoskeletal Infection Society; HTO, heterotopic ossification; ROM, range of motion; TFL, tensor fascia latae; THA, total hip arthroplasty; N/A, not applicable.

A

Measurement	Definition	Calculation
Accuracy	% of relevant concepts that were extracted by method A/B	Number of concepts/total number of concepts that should have been extracted
Recall/sensitivity	Ratio of all identified concepts and all existing concepts in a given text (0-1)	True positives/(true positives + false negatives)
Precision	Precision is the ratio between all adequately identified concepts and all identified concepts (0-1)	True positives/(true positives + false positives)
F measure	Harmonic mean of recall and precision (0-1)	2*(Recall*precision)/(recall + precision)

B

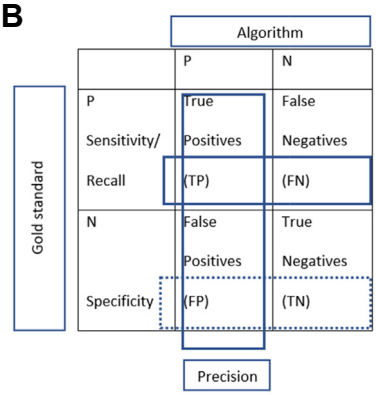

Fig. 1. Statistical parameters. Definitions (A) and graphical representation (B).

Results

Overall, statistically significant differences in accuracy could not be demonstrated ($P = .69$) between the algorithm and the human reviewer (95.0% vs 94.5%, respectively; Table 2). For implant characteristics, the algorithm showed a significantly higher accuracy than the reviewer (94.8% vs 93.4%; $P = .01$). This demonstrates that for data points with a clear pattern such as implant sizes, the algorithm outperforms the manual process even with minimal training. For data points concerning intraoperative events and interventions noted in the OR reports, no significant difference in accuracy could be demonstrated ($P = .2$ and $P = .1$, respectively). The accuracy of the algorithm for postoperative complications was slightly superior compared to the human reviewer (94.5% vs 90.0%) but not to a significant level ($P = .13$), indicating that the human reviewer was more prone to missing some data when going through large datasets. The algorithm had a lower accuracy for extracting postoperative interventions (0.88 vs 0.96, respectively), although this result was not significant ($P = .5$).

Overall, the algorithm achieved a recall of 96.0%, a precision of 88.0%, and an F-value of 0.89 for all adverse events (Table 3). Precision and recall for soft tissue complications (94% and 85%, respectively) was lower than for mechanical complications (100%) due to high grade of variability in description of soft tissue–related concepts, which creates difficulty for training of the algorithm. In addition, precision of rare complications such as bleeding and heterotopic ossification (HTO; 2 and 1 cases, respectively) was only 33%.

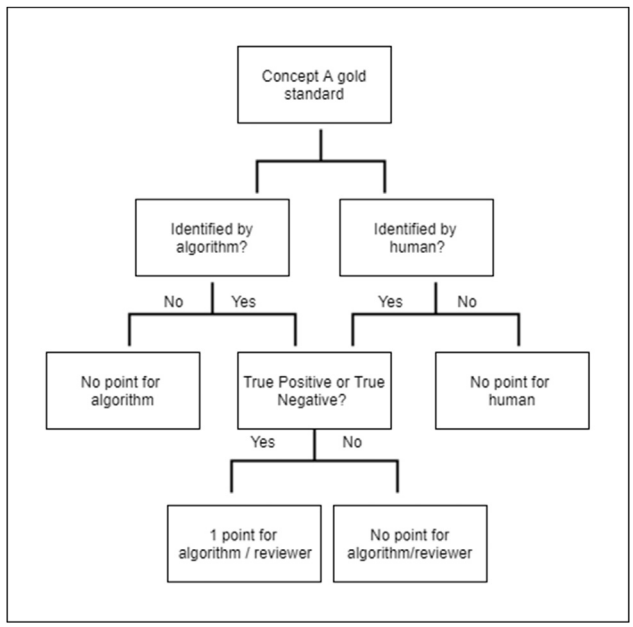

Fig. 2. Scoring system for accuracy.

Table 2
Accuracy.

Clinical Concept	Human Reviewer Accuracy	Algorithm Accuracy	*P* Value Groups
Stem size	0.940	0.948	N/A
Head size	0.931	0.957	N/A
Cup size	0.939	0.948	N/A
Cup brand	0.948	0.957	N/A
Insert size	0.913	0.930	N/A
Implant subset	0.934	0.948	**.01**[a]
Intraoperative adverse event	0.975	0.983	.19[a]
Intraoperative intervention	0.956	0.967	.10[a]
Postoperative adverse event	0.900	0.945	.13[a]
Postoperative intervention	0.959	0.878	.50[a]
Side	0.991	0.982	N/A
Overall average	0.945	0.950	.69[b]

Statistically significant results are marked in bold. N/A, not applicable.
[a] Paired *t*-test 1 tail.
[b] Paired *t*-test 2 tails.

Table 3
Recall, Precision, and F-Value.

Clinical Concept	Recall	Precision	F-Value
Bleeding	1.00	0.33	0.50
Hematoma	0.88	0.88	0.88
Prolonged wound drainage	1.00	1.00	1.00
Superficial wound infection	1.00	1.00	1.00
Lateral femoral cutaneous nerve	0.92	0.85	0.64
Motor nerve	N/A	N/A	N/A
Heterotopic ossifications	1.00	0.33	0.50
Abductor/greater trochanter	0.92	0.92	0.92
Psoas	0.86	0.90	0.88
Adductor	0.86	1.00	0.92
Tensor fascia latae	1.00	1.00	1.00
Quadriceps	1.00	1.00	1.00
Lumbago	0.84	0.84	0.84
Sciatica	0.90	0.82	0.86
Thromboembolic disease	N/A	N/A	N/A
Deep periprosthetic joint infection	1.00	1.00	1.00
Soft tissue overview	0.94	0.85	0.85
Dislocation/instability	1	1	1
Periprosthetic fracture intraoperatively	1	1	1
Periprosthetic fracture postoperatively	1	1	1
Implant fracture	N/A	N/A	N/A
Implant loosening	N/A	N/A	N/A
Cup-liner dislocation	N/A	N/A	N/A
Bearing surface wear	N/A	N/A	N/A
Osteolysis	N/A	N/A	N/A
Mechanical/prosthesis	1.00	1.00	1.00
Infiltration	0.89	0.84	0.86
Wound debridement	1.00	1.00	1.00
Punction, aspiration	1.00	0.83	0.91
Draining of hematoma	N/A	N/A	N/A
Psoas release	N/A	N/A	N/A
Osteosynthesis	N/A	N/A	N/A
Cerclage	1.00	1.00	1.00
Revision	1.00	0.83	0.91
Reintervention	0.98	0.90	0.94
Readmission	0.92	0.85	0.88
Overall results	0.96	0.88	0.89
Standard deviation	0.06	0.18	0.14

N/A, not applicable.

Clinical results are presented in Table 4. In total, 119 patients (19.4%) presented with some form of perioperative problem during the first postoperative year. Most (77%) adverse events were transient and soft tissue related. Lower back pain or sciatica were the most prevalent adverse soft tissue events with an incidence of 4.2% and 1.6%. They were all managed conservatively and generally resolved within 3 months. Psoas and abductor/greater trochanter–related problems were seen in 19 and 11 patients (3.1% and 1.8%), requiring corticosteroid infiltration in 9 and 8 cases. One further infiltration was required for adductor pain, resulting in a total of 18 infiltrations (2.9%). Cessation of physiotherapy was the most frequent "intervention" for lower back pain and muscle-related problems. Sensory problems (1.8%) were all related to the lateral femoral cutaneous nerve and were all managed conservatively. Bleeding, hematoma, and HTO were seen in 0.3%, 1.3%, and 0.2%, respectively, requiring no further intervention. One superficial wound infection (0.2%) required oral antibiotics and local wound care. Ten patients (1.6%) suffered an intraoperative calcar fracture, requiring intraoperative cerclage wiring without any further consequence (1.3%).

Nine patients (1.5%) were readmitted within the first postoperative year. Four patients (0.7%) were readmitted for wound-related problems in which debridement was performed in 3 patients (0.5%). Two patients were readmitted for closed reduction following a dislocation of which one was traumatic due to a fall (0.3%). There were 3 major reinterventions (0.5%). One patient developed a deep periprosthetic infection [9] and underwent a 2-stage revision. One patient was revised for a traumatic rectus femoris tear. One late postoperative periprosthetic fracture required stem revision. Overall, a surgical reintervention was required in 6 of 613 implants (0.9%).

Discussion

The algorithm showed to be noninferior to a human reviewer for the detection and extraction of relevant data from a large number of EMRs following the anterior THA procedure on a regular operating table. However, the data extraction is done at a much greater speed and could be optimized following machine training sequences. The program analyzes every single word in a document and is not prone to lapses in concentration like a human reviewer. Therefore, concept recognition and extraction are done at an accuracy level at least as good and even slightly better than a human reviewer. For data points with limited variability, such as mechanical events and implant data noted in the surgery report, the algorithm achieved a statistically significant higher accuracy. The time required for the program to process a dataset is negligible. Depending on system power, it takes mere minutes to mine a dataset of several hundreds of EMRs. Precision was considerably harder to optimize. The program performed well in categorizing frequent concepts. However, it struggled to retrain on rare concepts such as bleeding or HTO due to a lack of source material. This could be improved by increasing the volume of the dataset. Some subgroups (eg, motor nerve dysfunction) were not found within the investigated patient population, making it impossible to train the algorithm and calculate their recall, precision, and F-value. Other studies have developed algorithms to extract complications [10]. To our knowledge, however, a comprehensive outcome analysis tool in the context of THA has not been described yet. Having the algorithm continuously mine all of our data allows us to keep track of our clinical key point indicators without having to review each individual EMR to see whether an adverse event has occurred. Therefore, a manual validation of extracted adverse events and interventions is still required, currently representing 15% of the EMR dataset (79 unique patients). A follow-up study will be executed with a larger dataset to further optimize the program's performance for accuracy, recall, and precision. We expect a manual review limited to 5% of the dataset would be sufficient. This facilitates clinical research at a greater efficiency level with short-term feedback on outcomes and clinical performance, allowing faster optimization of our department, products, and clinical pathways.

Table 4
Complications and Interventions.

Complication	Subgroup	(N) %	Treatment: Conservative (N) % Interventional (N) %
Soft tissue complications			
Wound complications	Bleeding	(2) 0.3%	C: (2) 0.3%; I: (0) 0%
	Hematoma	(8) 1.3%	C: (8) 1.3%; I: (0) 0%
	Prolonged wound drainage	(4) 0.7%	C: (1) 0.2%; I: (3) 0.5%
Infection	Superficial wound infection	(1) 0.2%	C: (1) 0.2%; I: (0) 0%
	Deep prosthetic joint infection	(1) 0.2%	C: (0) 0%; I: (1) 0.2%
Neural deficit	Motor nerve deficit	(0) 0%	C: (0) 0%; I: (0) 0%
	Sensory deficit	(11) 1.8%	C: (11) 1.8%; I: (0) 0%
Vascular injury		(0) 0%	C: (0) 0%; I: (0) 0%
Heterotopic ossification		(1) 0.2%	C: (1) 0.2%; I: (0) 0%
Muscle pain or weakness	Abductor muscles	(11) 1.8%	C: (11) 1.8%; I: (0) 0%
	Greater trochanter		
	Psoas	(19) 3.1%	C: (19) 3.1%; I: (0) 0%
	Adductor	(6) 1%	C: (6) 1%; I: (0) 0%
	Tensor fascia latae	(2) 0.3%	C: (2) 0.3%; I: (0) 0%
	Quadriceps	(4) 0.7%	C: (3) 0.5%; I: (1) 0.2%
Lower back pain	Lumbago	(26) 4.2%	C: (26) 4.2%; I: (0) 0%
	Sciatica	(10) 1.6%	C: (10) 1.6%; I: (0) 0%
Thromboembolic disease		(0) 0%	C: (0) 0%; I: (0) 0%
Mechanical complications			
Instability		(2) 0.3%	C: N/A; I: (2) 0.3%
Periprosthetic fracture	Intraoperative	(10) 1.6%	C: (10) 1.6%; I: (0) 0%
	Postoperative	(1) 0.2%	C: (0) 0%; I: (1) 0.2%
Implant fracture		(0) 0%	C: N/A; I: (0) 0%
Loosening		(0) 0%	C: (0) 0%; I: (0) 0%
Cup-liner dislocation		(0) 0%	C: N/A; I: (0) 0%
Bearing surface wear		(0) 0%	C: (0) 0%; I: (0) 0%
Osteolysis		(0) 0%	C: (0) 0%; I: (0) 0%
Intervention			
Intraoperative	Cerclage wiring	(8) 1.3%	
Outpatient	Infiltration	(18) 2.9%	
	Aspiration	(2) 1.3%	
Reoperation	Wound debridement	(3) 0.5%	
	Closed reduction	(2) 0.3%	
	Osteosynthesis/cerclage	(1) 0.2%	
Revision		(3) 0.5%	
Readmission		(8) 1.3%	
Death		(0) 0%	

New and important insights in the postoperative course following the efficient direct anterior THA have surfaced due to the use of the AI algorithm. Based upon these data, patients can now be consented to have an expected postoperative uneventful course in approximately 81% of the cases. Eighteen percent of patients will encounter a transient adverse event, of which soft tissue problems such as low back pain or sciatica are the most prevalent. Periarticular muscle soreness was found to be the second most prevalent. These problems could be managed conservatively in 93% of cases within the first 3 postoperative months. In approximately 1%, a reintervention can be expected. Using an automated algorithm reduces the amount of individual judgment involved in extracting adverse events, potentially generating a more realistic complication profile [4]. Due to the automated algorithm, it is possible to provide a continuously updated feedback loop to the surgeon which allows for swift recognition of emerging systematic problems and adjustment of daily practice.

This study is not without limitations. First, the algorithm was validated in the context of THA in Dutch for a single high-volume center. Extrapolation of these results to other clinical settings or languages requires further validation, as different surgical teams use different terminologies. We believe this to be a minor limitation, as the program is not a simple search engine. The AI recognizes a large array of terminology and context and subsequently categorizes a concept rather than simply detecting a specific combination of words. It may however be possible that a new clinical context requires some specific training to get the AI up to optimal performance. The program extracts concepts from outpatient, surgery, and hospitalization reports. It is therefore directly dependent on the quality of records kept by the surgeon. The performance of the program is also dependent on its training. Some issues were encountered with recognition of concepts related to activity moderation, demonstrating that the program is only able to extract and interpret concepts on which it has been trained. The performance of manual review could have been greater if radiographs were reviewed. However, as the AI is currently unable to process radiographic data, this was not included in manual review. As discussed above, it is still required to review 15% of the dataset manually. The amount of EMRs to be reviewed is proportional to the size of the dataset. In very large cohorts, this might still be a significant obstacle. However, precision will increase as more records are mined by the program, reducing the number of false positives.

Conclusion

An automated, AI-supported search algorithm can analyze and interpret large quantities of EMRs at a much greater speed but with a performance level comparable to or slightly higher than a human reviewer. Due to the use of the AI search engine, new clinical information surfaced regarding the postoperative course because 18% of patients can be expected to have a “transient” problem following a THA procedure, which has not been reported before. The algorithm provides continuously updated feedback on quality of care, facilitating swift adjustment of clinical practice if required, especially in a high-volume, nonacademic setting.

Acknowledgments

We would like to thank D. Hens, MD, W. Daelemans, PhD, and C. Motmans for their assistance in development of the algorithm and statistical analysis.

References

[1] Learmonth ID, Young C, Rorabeck C. The operation of the century: total hip replacement. Lancet 2007;370:1508–19.
[2] Lee GC, Marconi D. Complications following direct anterior hip procedures: costs to both patients and surgeons. J Arthroplasty 2015;30:98–101. https://doi.org/10.1016/j.arth.2015.03.043.
[3] Woolf AD, Pfleger B. Burden of major musculoskeletal conditions. Bull World Health Organ 2003;81:646–56.
[4] Goldhahn S, Sawaguchi T, Audigé L, Mundi R, Hanson B, Bhandari M, et al. Complication reporting in orthopaedic trials. J Bone Joint Surg Am 2009;91:1847–53.
[5] Chen BP, Garland K, Roffey DM, Poitras S, Dervin G, Lapner P, et al. Can surgeons adequately capture adverse events using the spinal Adverse Events severity System (SAVES) and OrthoSAVES? Clin Orthop Relat Res 2017;475:253–60.
[6] Mercer J, Penner M, Wing K, Younger ASE. Inconsistency in the reporting of adverse events in total ankle arthroplasty: a systematic review of the literature. Foot Ankle Int 2016;37:124–36.
[7] Bozic KJ, Grosso LM, Lin Z, Parzynski CS, Suter LG, Krumholz HM, et al. Variation in hospital-level risk-standardized complication rates following elective primary total hip and knee arthroplasty. J Bone Joint Surg Am 2014;96:640–7.
[8] Healy WL, Iorio R, Clair AJ, Pellegrini VD, Della Valle CJ, Berend KR. Complications of total hip arthroplasty: standardized list, definitions, and stratification developed by the Hip society. Clin Orthop Relat Res 2016;474:357–64.
[9] Parvizi J, Tan TL, Goswami K, Higuera C, Della Valle C, Chen AF, et al. The 2018 definition of periprosthetic hip and knee infection: an evidence-based and validated criteria. J Arthroplasty 2018;33:1309–1314.e2. https://doi.org/10.1016/j.arth.2018.02.078.
[10] Tien M, Kashyap R, Wilson GA. Retrospective derivation and validation of an automated electronic search algorithm to identify postoperative cardiovascular and thromboembolic complications. Appl Clin Inform 2015;6:565–76.