Article published online: 2021-12-22

2020 SCHOT Research Award: Development and Validation of a Multivariable Prediction Model of Hospital Stay in Elderly Chilean Patients Undergoing Elective Total Hip Arthroplasty Using Machine Learning

Premio de Investigación SCHOT 2020: desarrollo y validación de un modelo multivariables de predicción de estadía hospitalaria en pacientes mayores de 65 años sometidos artroplastia total de cadera electiva en Chile utilizando aprendizaje de máquinas

Claudio Díaz-Ledezma[1] David Díaz-Solís[2] Raúl Muñoz-Reyes[3] Jonathan Torres Castro[4]

[1] Orthopedic Surgery and Traumatology Unit, Hospital El Carmen Dr. Luis Valentin Ferrada, Santiago, Chile
[2] Department of Orthopedics and Traumatology, Clínica Las Condes, Santiago, Chile
[3] Departamento de Administracion, Facultad de Economia y Negocios, Universidad de Chile, Santiago, Chile
[4] Data scientist, independent researcher, Santiago, Chile
[5] Hip Surgery Team, Clínica RedSalud Santiago, Santiago, Chile
[6] Hip Surgery Team, Instituto Traumatológico de Santiago, Santiago, Chile

Address for correspondence Claudio Díaz Ledezma, Av. Rinconada 1.201, Oficina 28, 5to piso, Maipú, Santiago, Chile (e-mail: claudioadiaz@gmail.com).

Rev Chil Ortop Traumatol 2021;62(3):e180–e192.

Abstract

Introduction The prediction of the length of hospital stay after elective total hip arthroplasty (THA) is crucial in the perioperative evaluation of the patients, and it plays a decisive role from the operational and economic point of view. Internationally, big data and artificial intelligence have been used to perform prognostic evaluations of this type. The present study aims to develop and validate, through the use of artificial intelligence (machine learning), a tool capable of predicting the hospital stay of patients over 65 years of age undergoing THA for osteoarthritis.
Material and Methods Using the electronic records of hospital discharges de-identified from the Department of Health Statistics and Information (Departamento

Keywords
- hospital stay
- machine learning
- total hip arthroplasty

received
March 18, 2021
accepted
August 6, 2021

DOI https://doi.org/10.1055/s-0041-1740232.
ISSN 0716-4548.

de Estadísticas e Información de Salud, DEIS, in Spanish), the data of 8,970 hospital discharges of patients who had undergone THA for osteoarthritis between 2016 and 2018 were obtained. A total of 15 variables available in the DEIS registry, in addition to the percentage of poverty in the patient's borough of origin were included to predict the probability that a patient would have a shortened (< 3 days) or prolonged (> 3 days) stay after surgery. By using machine learning techniques, 8 prediction algorithms were trained with 80% of the sample. The remaining 20% was used to validate the predictive capabilities of the models created from the algorithms. The optimization metric was evaluated and ranked using the area under the receiver operating characteristic curve (AUC-ROC), which corresponds to how well a model can distinguish between two groups.

Results The XGBoost algorithm had the best performance, with an average AUC-ROC of 0.86 (standard deviation [SD]: 0.0087). Secondly, we observed that the linear support vector machine (SVM) algorithm obtained an AUC-ROC of 0.85 (SD: 0.0086). The relative importance of the explanatory variables showed that the region of residence, the administrative health service, the hospital where the patient was operated on, and the care modality are the variables that most determine the length of stay.

Discussion The present study developed machine learning algorithms based on free-access Chilean big data, which helped create and validate a tool that demonstrates an adequate discriminatory capacity to predict shortened versus prolonged hospital stay in elderly patients undergoing elective THA.

Conclusion The algorithms created through the use of machine learning allow to predict the hospital stay in Chilean patients undergoing elective total hip arthroplasty.

Resumen

Introducción La predicción de la estadía hospitalaria luego de una artroplastia total de cadera (ATC) electiva es crucial en la evaluación perioperatoria de los pacientes, con un rol determinante desde el punto de vista operacional y económico. Internacionalmente, se han empleado macrodatos (*big data*, en inglés) e inteligencia artificial para llevar a cabo evaluaciones pronósticas de este tipo. El objetivo del presente estudio es desarrollar y validar, con el empleo del aprendizaje de máquinas (*machine learning*, en inglés), una herramienta capaz de predecir la estadía hospitalaria de pacientes chilenos mayores de 65 años sometidos a ATC por artrosis.

Material y Métodos Empleando los registros electrónicos de egresos hospitalarios anonimizados del Departamento de Estadísticas e Información de Salud (DEIS), se obtuvieron los datos de 8.970 egresos hospitalarios de pacientes sometidos a ATC por artrosis entre los años 2016 y 2018. En total, 15 variables disponibles en el DEIS, además del porcentaje de pobreza de la comuna de origen del paciente, fueron incluidos para predecir la probabilidad de que un paciente presentara una estadía acortada (< 3 días) o prolongada (> 3 días) luego de la cirugía. Utilizando técnicas de aprendizaje de máquinas, 8 algoritmos de predicción fueron entrenados con el 80% de la muestra. El 20% restante se empleó para validar las capacidades predictivas de los modelos creados a partir de los algoritmos. La métrica de optimización se evaluó y ordenó en un *ranking* utilizando el área bajo la curva de característica operativa del receptor (*area under the receiver operating characteristic curve*, AUC-ROC, en inglés), que corresponde a cuan bien un modelo puede distinguir entre dos grupos.

Resultados El algoritmo XGBoost obtuvo el mejor desempeño, con una AUC-ROC promedio de 0,86 (desviación estándar [DE]: 0,0087). En segundo lugar, observamos que el algoritmo lineal de máquina de vector de soporte (*support vector machine*, SVM, en inglés) obtuvo una AUC-ROC de 0,85 (DE: 0,0086). La importancia relativa de las

Palabras clave
- estadía hospitalaria
- aprendizaje de máquinas
- artroplastia total de cadera

variables explicativas demostró que la región de residencia, el servicio de salud, el establecimiento de salud donde se operó el paciente, y la modalidad de atención son las variables que más determinan el tiempo de estadía de un paciente.

Discusión El presente estudio desarrolló algoritmos de aprendizaje de máquinas basados en macrodatos chilenos de libre acceso, y logró desarrollar y validar una herramienta que demuestra una adecuada capacidad discriminatoria para predecir la probabilidad de estadía hospitalaria acortada *versus* prolongada en adultos mayores sometidos a ATC por artrosis.

Conclusión los algoritmos creados a traves del empleo del aprendizaje de máquinas permiten predecir la estadía hospitalaria en pacientes chilenos operado de artroplastia total de cadera electiva.

Introduction

In Chile, total hip arthroplasty (THA) for the treatment of severe osteoarthritis is guaranteed by law for patients over 65 years of age.[1] However, little is known about the results of THA in this particular group of patients, and there is no national scientific publication (to our knowledge) that addresses the issue of hospital stay, which has a leading role in the era of value-based arthroplasty.

In the world and particularly in the United States, a sustained decrease in the length of hospital stay of patients after THA has been observed, without increased risks.[2] Recently it has even been proven that the outpatient modality can be successful in a select group of patients.[3,4] The length of hospital stay for patients over 65 years of age in the United States (2015-2016) averaged 1.8 days.[5] In Chile, these data have not been published.

Several tactics can be used to reduce the hospital stay after THA, including standardized management protocols,[6,7] and other tactics that go hand in hand with the prediction of potential perioperative complications.[8,9] Among the challenges of THA in our country, we have described the relevance of keeping our perioperative approach updated and with the same standards as those of the leading countries on the subject.[10]

As we advance in the global crisis caused by the COVID-19 pandemic, elective surgery is performed with a reduced hospital stay, without compromising patients' safety.[11,12] Surgeons should be able to predict the occurrence of possible complications, as well as to determine the possible length of hospital stay in their patients.

Machine learning is one of the branches of artificial intelligence[13], and it is understood as the manner in which computer algorithms (that is, machines) can "learn" relationships or complex patterns based on empirical data and, therefore, produce mathematical models that link a large number of covariates to a target variable of interest.[14]

In the medical field, among other applications, this means being able to predict, based on data extracted from specialized electronic records, risk scores (in the form of regression and prognosis) to help clinicians make more efficient and accurate decisions; therefore, machine learning can be a support tool in clinical decision making. Specifically in arthroplasty, studies[15–17] involving this technology have gained momentum, providing assistance to solve complex problems that we face in our practice.[18]

Our hypothesis is that the machine learning process can predict the length of hospital stay in patients undergoing THA, which has a dual purpose in clinical practice: 1) to help improve the group with a high probability of a short stay, further reducing their stay; and 2) to identify the group with a low probability of a short stay, to improve their perioperative care and eventually bring them safely to the short stay group.

The objective of the present study is to develop and validate, using machine learning, a tool capable of predicting the length of hospital stay of patients over 65 years of age undergoing THA for osteoarthritis.

Materials and Methods

Funding

The present research project and manuscript were funded by the 2020 Research Grant of the Chilean Society of Orthopaedics and Traumatology.

Data Source and Study Population

The present is a registry study. Databases of hospital discharges for the years 2016, 2017 and 2018 were collected from the website of the Department of Health Statistics and Information (Departamento de Estadísticas e Información en Salud, DEIS, in Spanish) of the Chilean Ministry of Health.[19] Each of these databases contains de-identified records of all hospital discharges from both public and private centers in our country, including 39 columns with data related to each of the individualized hospital discharges. Each of these records contains characterisctics pertaining to demographics, the hospital center, discharges, diagnosis, etc. In the studied period, the data of 4,944,017 hospital discharges were collected. Considering the 39 aforementioned columns, the total volume of individual variables to be discriminated and evaluated was of 192,816,663.

Considering that the data of each particular case is deidentified and comes from a public database (the identification is an alphanumeric code with no personal data, not linkable to an individual patient), the present study did not require authorization from the ethics committee.

From the primary data source, a derived database was created, including only patients aged $\geq$ 65 years who underwent THA (or total hip endoprosthesis) for osteoarthritis. These cases are covered under the Explicit Guarantees in Health.[1] These cases were selected through codes 2104129 (*Total hip endoprosthesis, does not include prosthesis*) and 2104229 (*Total hip endoprosthesis, includes prosthesis*) of the Chilean National Health Fund (Fondo Nacional de Salud, FONASA, in Spanish), which correspond to the M16 diagnosis (coxarthrosis) on the International Classification of Diseases, 10th revision (ICD-10), with all its secondary classifications. Patients with any kind of health insurance and from all parts of Chile operated between 2016 and 2018 were included. Procedures coded as 2104129 and 2104229 performed for a diagnosis of proximal femur fracture (S72 diagnosis on the ICD-10) and cases that were discharged from the hospital categorized as "deceased" were excluded. The sample included all the cases registered in our country for the indicated period.

Clinically-Relevant Outcome (Variable to Predict)

According to literature,[20] hospital stays longer than three days can be considered prolonged in the context of elective THA. In the present study, *short stay* will be defined as shorter than or equal to three days, and *prolonged stay*, as those longer than three days, it must be considered that, for the studied period, the experience in outpatient THA was limited to certain groups in our country.[4]

A prediction of the length of hospital stay was made as a binary variable, described as a function of two classes based on the days of hospitalization. Thus, the variable to be modeled takes two possible values: "short stay" or "prolonged stay".

Predictive variables

From the group of 39 individual variables for each of the DEIS hospital discharges corresponding to the study population, 21 were chosen (►**Table 1**) because they were considered relevant by the group of authors at the time of data processing. The data records were complete for each of the variables. Of these, 16 variables were used when performing a predictive process for hospital discharge. In addition, the variable "percentage of poverty in the borough" extracted from the database of the Chilean Ministry of Social Development was included.[21] There were no missing data in the registry used, so it was not necessary to perform imputation techniques.[22] It is important to note that the DEIS database contains variables collected for epidemiological purposes, and does not capture enough data at the level of individual patients. Consequentely, this model excludes variables such as comorbidities, functionality, and surgical details that could certainly influence the length of hospital stay.

Data Preparation (Sample Balance)

For the correct processing of the nominal variables, they were transformed using one-hot encoding, that is, multiple dichotomous columns that represented the existence or not of a particular characteristic for each specific hospital discharge. In terms of the processing of continuous variables, their scale was standardized in the range between 0 and 1, with 0 corresponding to the minimum value in the original data, and 1, to the maximum for each of them. Furthermore, given that there is a higher proportion of cases with 3 or more days, it was necessary to balance the training sample[23] following an oversampling procedure of the underrepresented class.[24]

Training and Testing of the Classification Algorithms

For the present study, different algorithms and hyperparameter configurations available in computer code libraries for the Python programming language were tested. In particular, seven algorithms available in the sklearn package were tested (logistic regression, decision tree classifier, linear support vector machine, naive bayes, random forest classifier, adaboost, and multilayer perceptron). Although a detailed description of the operation of each algorithm is outside the scope of the objectives of this article, the intuition behind this selection refers to the trade-off between predictive power and the possible interpretation and transparency capacity of the models created (so that the evaluation of the predictors of the model are not under the influence of the authors once they have been integrated into the project). In the literature on machine learning, it is common to group algorithms depending on whether they use systems of mathematical equations as a fundamental modeling strategy, or whether they generate computational decision rules, the latter tending to be easier to interpret. The most advanced models, for example, random forest or multilayer perceptron (a type of artificial neural network) can contain thousands of decision rules or mathematical equations, potentially having millions of parameters to estimate and interpret. Thus, the algorithms of logistic regression, support vector machines, naive bayes, and multilayer perceptron are based on systems of mathematical equations. On the other hand, the decision trees, random forest, and adaboost algorithms generate a set of computational decision rules.

As aforementioned, as the number of equations or decision rules generated by the algorithms increases, it is typically expected that the predictive performance of the algorithm improves. However, increasing the complexity of the model by adding equations or rules also increases the difficulty of human interpretation of the models created.

Table 1 Variables available in the database of the Department of Health Statistics and Information (Departamento de Estadísticas e Información de Salud, DEIS, in Spanish)

Item from the DEIS hospital discharge database				
No	**Variable name**	**Description**	**Datatype**	**Used in the model**
1	ID_PACIENTE	Unique and anonymous identifier of the patient	Text	Just to discard duplicates
2	ESTABLECIMIENTO_SALUD	Hospital code	Number	Included as a possible predictor
3	GLOSA_ESTABLECIMIENTO_SALUD	Hospital name	Text	Not included in the model
4	PERTENENCIA_ESTABLECIMIENTO_SALUD	Hospital classification (part of the National Health Services System or not)	Text	Included as a possible predictor
5	SEREMI	SEREMI (Regional Ministerial Health Department) code	Number	Included as a possible predictor
6	SERVICIO_DE_SALUD	Health Service code	Number	Included as a possible predictor
7	SEXO	Code of the biological sex of the patient	Number	Included as a possible predictor
8	FECHA_NACIMIENTO	Patient's birthdate	Date	Not included in the model
9	EDAD_CANT	Numerical record of the patient's age at admission	Number	Included as a possible predictor
10	TIPO_EDAD	Unit of measurement of age, according to the modality described in values	Number	Not included in the model
11	EDAD_AÑOS	Age in years of the patient at the time of admission	Number	Not included in the model
12	PUEBLO_ORIGINARIO	Code of the town of origin code	Number	Not included in the model
13	PAIS_ORIGEN	Code of the country of origin	Number	Not included in the model
14	GLOSA_PAIS_ORIGEN	Classification of the country of origin	Text	Used to exclude foreign patients
15	COMUNA_RESIDENCIA	Code of the borough of residence	Text	Included as a possible predictor
16	GLOSA_COMUNA_RESIDENCIA	Name of the borough of residence	Text	Not included in the model
17	REGION_RESIDENCIA	Code of the region of residence	Text	Included as a possible predictor
18	GLOSA_REGION_RESIDENCIA	Name of the region of residence	Text	Not included in the model
19	PREVISION	Patient's health insurance code at the time of admission	Number	Included as a possible predictor
20	BENEFICIARIO	FONASA beneficiary code	Text	Included as a possible predictor
21	MODALIDAD	FONASA modality Code	Number	Included as a possible predictor
22	PROCEDENCIA	Code of origin of the patient at the time of admission	Number	Not included in the model
25	ANO_EGR	Year of discharge	Number	Not included in the model

Table 1 (*Continued*)

Item from the DEIS hospital discharge database				
No	**Variable name**	**Description**	**Datatype**	**Used in the model**
26	FECHA_EGR	Date of discharge	Date	Not included in the model
27	AREA_FUNCIONAL_EGRESO	Code of the level of care or functional area from which the patient was discharged	Number	Included as a possible predictor
28	DIAS_ESTAD	Days of total stay	Number	Variable that was the objective
29	CONDICION_EGRESO	Code of the condition at patient discharge	Number	Used to exclude discharges resulted from decease
30	DIAG1	International Classification of Diseases, 10th revision (ICD-10) code of the main diagnosis	Text	Included as a possible predictor
31	GLOSA_DIAG1	Classification of the main diagnosis	Text	Included as a possible predictor
32	DIAG2	Code of the external cause	Text	Not included in the model
33	GLOSA_DIAG2	Classification of the external cause	Text	Not included in the model
34	INTERV_Q	Surgical intervention code	Number	Used to exclude discharges without associated surgery
35	CODIGO_INTERV_Q_PPAL	FONASA main surgical intervention code	Text	Used to identify cases
36	GLOSA INTERV_Q_PPAL	Classification of the main surgical intervention	Text	Included as a possible predictor
37	PROCED	Procedure code	Number	Not included in the model
38	CODIGO_PROCED_PPAL	FONASA main procedure code	Text	Not included in the model
39	GLOSA_PROCED_PPAL	Classification of the main procedure	Text	Not included in the model
*40	% POBREZA COMUNA	Poverty rate in the borough	Number	Included as a possible predictor

Note: *Data obtained from the website of the Chilean Ministry of Social Development.

Therefore, it is also possible to group the algorithms into "open boxes" or "closed boxes". According to this classification, the algorithms of logistic regression, decision trees, support vector machines, and naive bayes are considered more of the "open-box" type, generating fewer or more equations according to the order in which they were listed, and the algorithms random forest, adaboost, and multilayer perceptron, as "closed boxes", generating fewer or more decision rules according to the order in which they were listed.

In addition, due to their good level of performance in other similar binary classification tasks, an additional family of algorithms, called gradient boosting trees, was included, which would also belong to the group of "closed boxes" that generates a large number of computational rules, which was implemented through the XGBoost package (an open-source software library).

The model was tested using 80% of the available data, and the remaining 20% was reserved to confirm the predictive capabilities of the model. This part of the data is traditionally called a test sample. Additionally, a resampling process, or boostrapping of 100 iterations, was carried out in order to obtain confidence intervals of the adjustment and performance figures of the selected models.

Evaluation and Adjustment of Models

To evaluate the performance of the algorithms and predictive models, we used their discrimination power (quantified as the area under the receiver operating characteristic curve, AUC-ROC[25]) in the data.

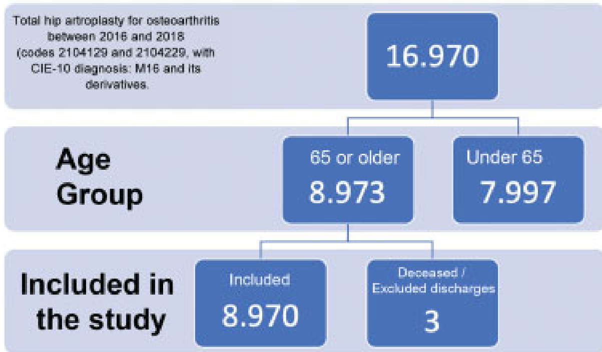

Fig. 1 Total hip arthroplasty due to arthrosis between 2016 and 2018 (codes 2104129 and 2104229, with ICD-10 diagnosis: M16 and its derivatives).

The optimization metrics were evaluated and ranked using AUC-ROC, which corresponds to how well a model can distinguish between two groups. The level of discrimination was classified as excellent (0.9–1), good (0.8–0.89), fair (0.7–0.79), poor (0.6–0.69), and failed (0.5–0.59).[26]

Other traditional metrics for classification problems are also reported, which include: "accuracy": the ratio of the correct number of predictions over the total samples; "average precision": average accuracy of the predictions based on the percentage of positive predictions that are correct; "precision": a measurement of the accuracy of a prediction based on the percentage of positive predictions that are correct; "recall": measurements of the percentage of positive scientific predictions against possible positives in the dataset; and "F1": harmonic average of precision and recall, with the best value being 1 (perfect precision), and the worst, 0. For each of the above, the estimated confidence intervals are also reported based on the resampling or bootstrapping procedure.

Model Report

The report of the model in the present manuscript uses international recommendations for this type of study,[27,28] with the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) checklist.[28]

Results

In total, 8,970 cases were included (►**Figure 1**): 5,662 women (63.12%) and 3,308 (36.88%) men. Their median age was of 72 years, with an interquartile range of 9 years, and a range between 65 and 97 years (►**Figure 2**).

The final sample included 6,746 (75.21%) FONASA patients, 1,599 (17.82%) patients from private healthcare insurers (instituciones de salud previsional, ISAPRES, in Spanish), and 625 (6.97%) patients from other health insurers. Of the FONASA patients, 286 (4.2%) were type-A beneficiaries, 4,801 (71.2%), type-B beneficiaries, 469 (6.9%) , type-C beneficiaries, and 1,191 (13.3%), type-D beneficiaries. In this same group of FONASA patients, 5,321 (78.9%) were operated on under the institutional-care modality, and 1,425 (21.1%), under the free-choice modality.

The 4 most frequent diagnoses were M169 (6,124 cases; 68.27%), M161 (1,623 cases; 18.09%), M160 (862 cases; 9.61%), and M167 (176 cases; 1.96%).

The 5 most frequent boroughs of origin of the patient were Las Condes (426 cases; 4.75%), Viña del Mar (365 cases; 4.07%), La Florida (253 cases; 2.82%), Puente Alto (239 cases; 2.66%), and Santiago (235 cases; 2.62%), which corresponds to 16.92% of the total number of cases in Chile.

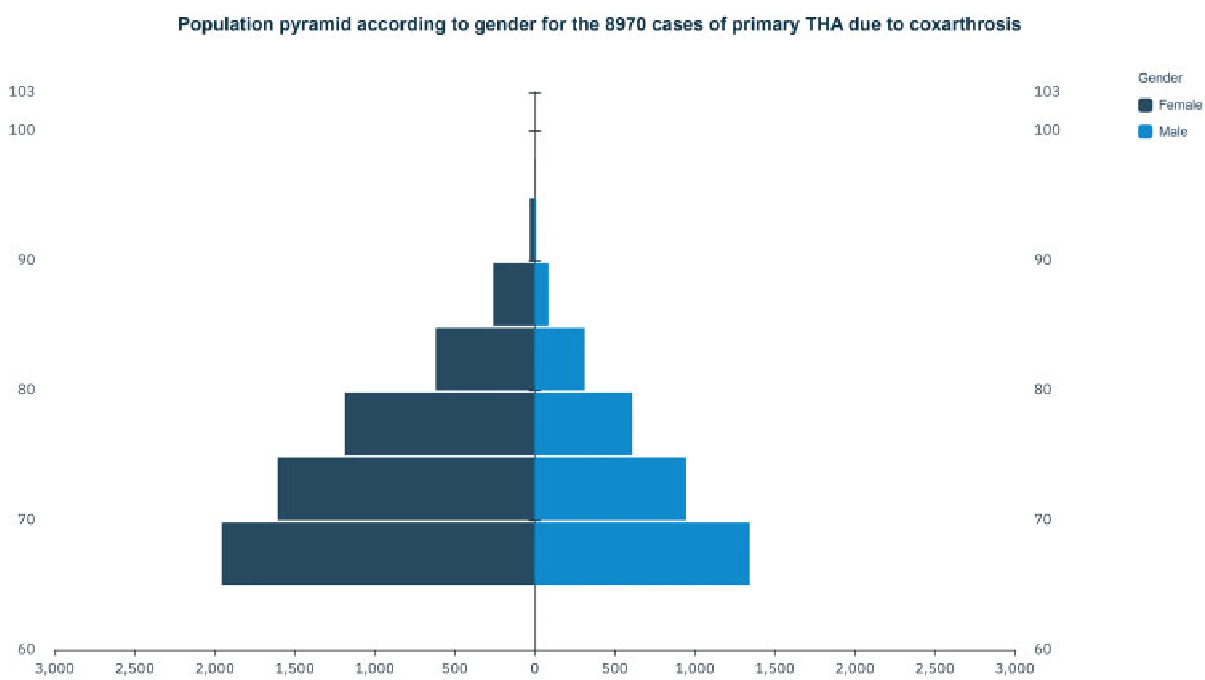

Fig. 2 Population pyramid according to gender for the 8970 cases of primary THA due to coxarthrosis.

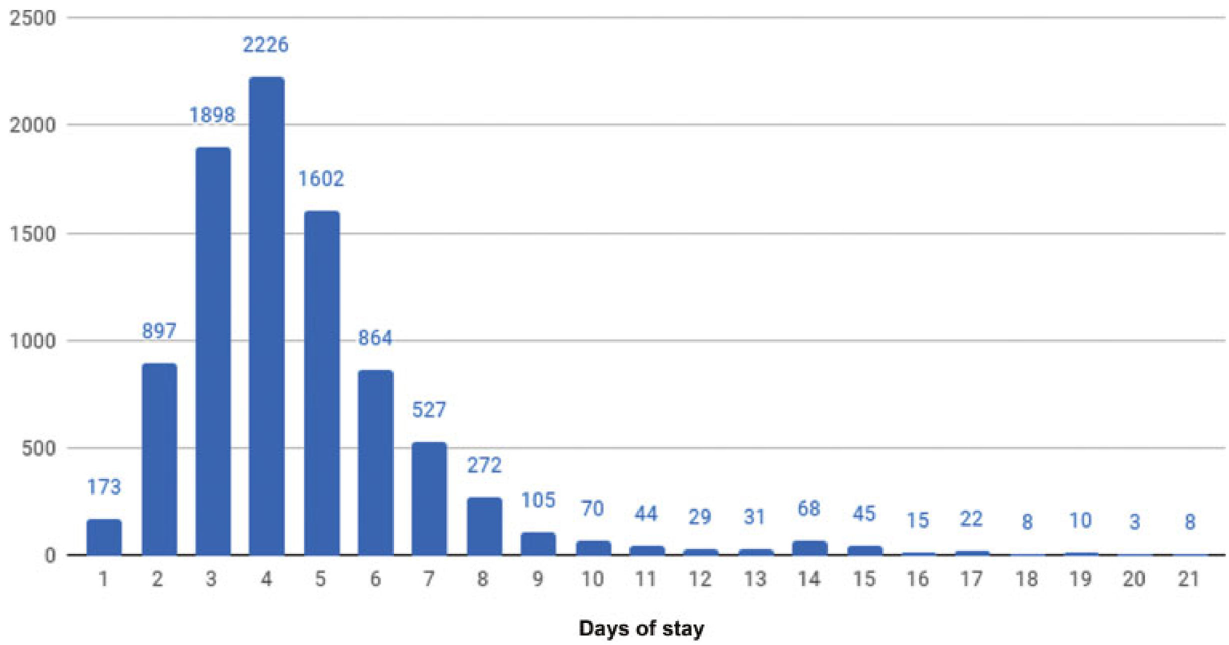

Fig. 3 Days of stay.

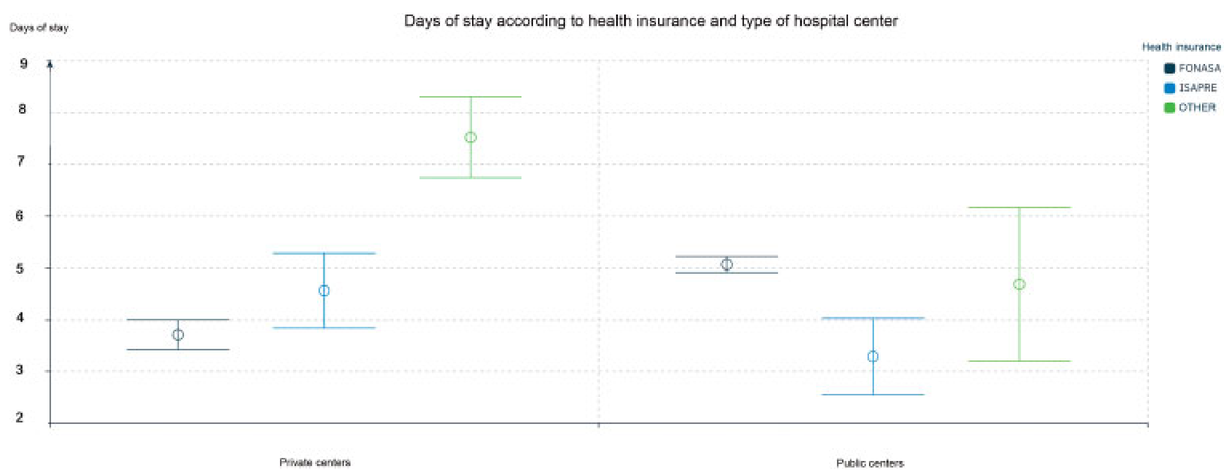

Fig. 4 Days of stay according to health insurance and type of hospital center.

One hundred hospital centers performed THAs in patients with osteoarthritis in the period studied. A total of 5,133 (81.88%) cases were operated on in centers that are part to the National Health Services System, and 1,136 cases (18.12%) were operated on in private centers.

The median number of days of stay was 4, with an interquartile range of 2 days and a range between 1 and 143 days. The histogram of days of stay is shown in ►**Figure 3**.

The days of stay categorized by type of hospital and health insurance are shown in ►**Figure 4**.

In total, 2,968 patients had a short stay (33.09%), and 6,002 had a prolonged stay (66.91%).

Performance of the Decision Algorithms

Eight algorithms were evaluated both in the training and the testing samples; however, these were ordered in a ranking according to their performance in the test sample. The latter is considered a better measurement of the performance of the model when applied in real scenarios. Among them, the XGBoost algorithm had the best performance, with an average AUC-ROC of 0.86 (SD: 0.0087). This means that the XGBoost algorithm had the best performance when discriminating between short and long hospital stays (longer or shorter than three days). Secondly, we observe that the Linear-SVM algorithm showed a very close AUC-ROC, of 0.8568 (SD: 0.0086), but with a lower SD.

Table 2 Results of the adjustment of the models in the training and test sample

Results of the training sample								
Bootstrap of 100 samples. Standard deviation is reported in parentheses								
	Overall accuracy	Class recall 0	Class recall 1	Class precision 0	Class precision 1	f1 score 0	f1 score 1	Area under the curve
XGBoost – Gradient-Boosted Trees	81.56%	77.44%	86.05%	84.76%	79.24%	80.92%	82.50%	90.46%
	(0.86%)	(1.40%)	(1.34%)	(1.20%)	(1.00%)	(0.94%)	(0.85%)	(0.77%)
Support Vector Machines	81.19%	78.76%	83.94%	83.07%	79.81%	80.86%	81.82%	89.55%
	(0.38%)	(0.62%)	(0.68%)	(0.57%)	(0.44%)	(0.39%)	(0.39%)	(0.27%)
AdaBoost	79.65%	76.79%	83.11%	81.98%	78.17%	79.30%	80.56%	88.16%
	(0.43%)	(0.75%)	(0.93%)	(0.74%)	(0.47%)	(0.41%)	(0.45%)	(0.27%)
Logistic Regression	81.13%	78.32%	84.37%	83.37%	79.56%	80.76%	81.89%	89.62%
	(0.42%)	(0.61%)	(0.79%)	(0.68%)	(0.44%)	(0.42%)	(0.45%)	(0.27%)
Random Forest	79.40%	74.91%	83.68%	82.15%	76.96%	78.34%	80.16%	86.99%
	(1.15%)	(2.07%)	(1.88%)	(1.62%)	(1.44%)	(1.37%)	(1.20%)	(0.91%)
Neural Net – Multilayer Perceptron	89.99%	91.03%	88.79%	89.04%	90.84%	90.02%	89.80%	97.19%
	(0.50%)	(1.21%)	(0.69%)	(0.57%)	(1.09%)	(0.62%)	(0.54%)	(0.31%)
Decision Tree	66.04%	63.32%	68.33%	74.35%	70.46%	61.45%	64.69%	74.05%
	(2.33%)	(27.95%)	(25.14%)	(14.06%)	(10.91%)	(13.47%)	(8.31%)	(2.03%)
Naive Bayes	65.07%	38.05%	94.97%	88.33%	60.56%	53.07%	73.94%	67.51%
	(1.60%)	(3.89%)	(0.68%)	(0.89%)	(1.38%)	(3.81%)	(0.89%)	(1.73%)
Test Sample Results								
Bootstrap of 100 samples. Standard deviation is reported in parentheses								
	Overall accuracy	Class recall 0	Class recall 1	Class precision 0	Class precision 1	f1 score 0	f1 score 1	Area under the curve
XGBoost – Gradient-Boosted Trees	81.74%	75.62%	80.23%	88.56%	61.97%	81.56%	69.90%	86.01%
	(0.87%)	(1.60%)	(2.24%)	(1.19%)	(1.73%)	(0.92%)	(1.31%)	(0.87%)
Support Vector Machines	81.35%	77.21%	78.81%	88.05%	63.12%	82.26%	70.07%	85.68%
	(0.37%)	(1.40%)	(1.98%)	(1.08%)	(1.86%)	(0.90%)	(1.48%)	(0.86%)
AdaBoost	79.95%	75.81%	79.98%	88.45%	62.06%	81.63%	69.87%	85.55%
	(0.40%)	(1.33%)	(1.81%)	(0.99%)	(1.61%)	(0.83%)	(1.26%)	(0.90%)
Logistic Regression	81.34%	76.60%	78.49%	87.81%	62.40%	81.81%	69.51%	85.16%
	(0.43%)	(1.33%)	(1.88%)	(1.03%)	(1.73%)	(0.87%)	(1.39%)	(0.90%)
Random Forest	79.30%	72.70%	77.43%	86.70%	58.43%	79.06%	66.56%	82.32%
	(1.23%)	(2.32%)	(2.88%)	(1.54%)	(2.33%)	(1.56%)	(2.04%)	(1.36%)
Neural Net – Multilayer Perceptron	89.91%	82.12%	64.44%	82.37%	64.07%	82.24%	64.23%	82.07%
	(0.58%)	(1.16%)	(2.43%)	(1.13%)	(1.77%)	(0.81%)	(1.70%)	(0.95%)
Decision Tree	65.82%	62.70%	66.65%	83.63%	53.84%	66.05%	54.06%	72.58%
	(2.47%)	(28.09%)	(25.86%)	(8.78%)	(12.33%)	(17.75%)	(4.52%)	(2.15%)
Naive Bayes	66.51%	36.80%	90.04%	88.14%	41.39%	51.81%	56.69%	64.35%
	(1.70%)	(4.05%)	(1.36%)	(1.63%)	(1.73%)	(4.14%)	(1.59%)	(1.94%)

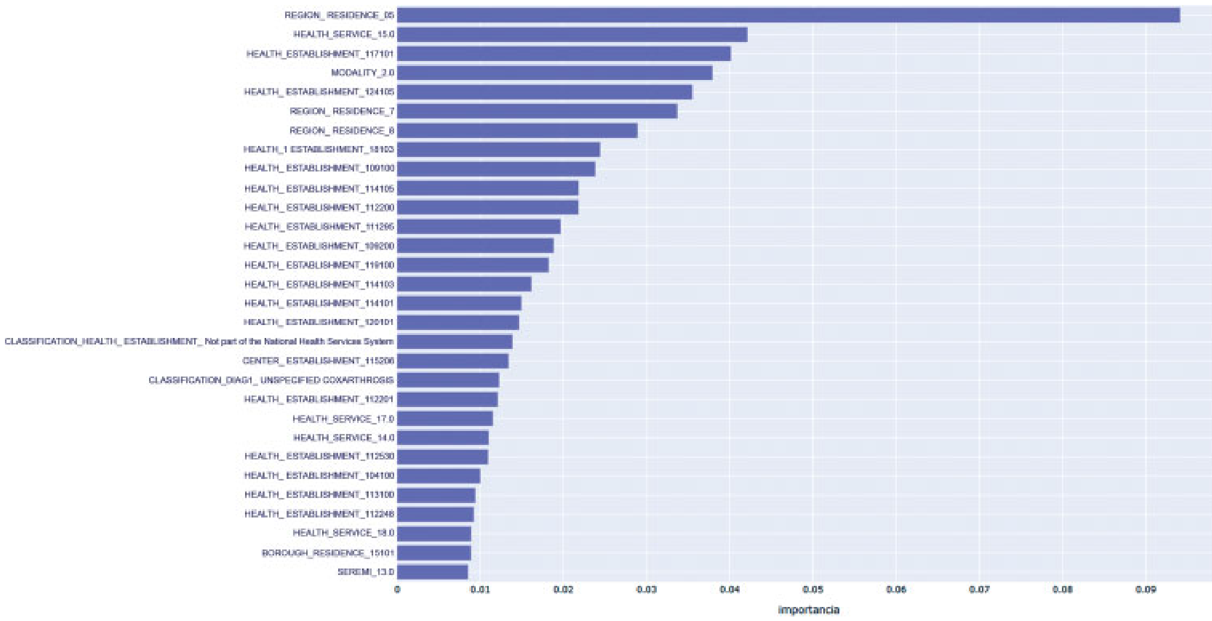

Fig. 5 Relative importance of the 30 most important variables of the model for length of stay.

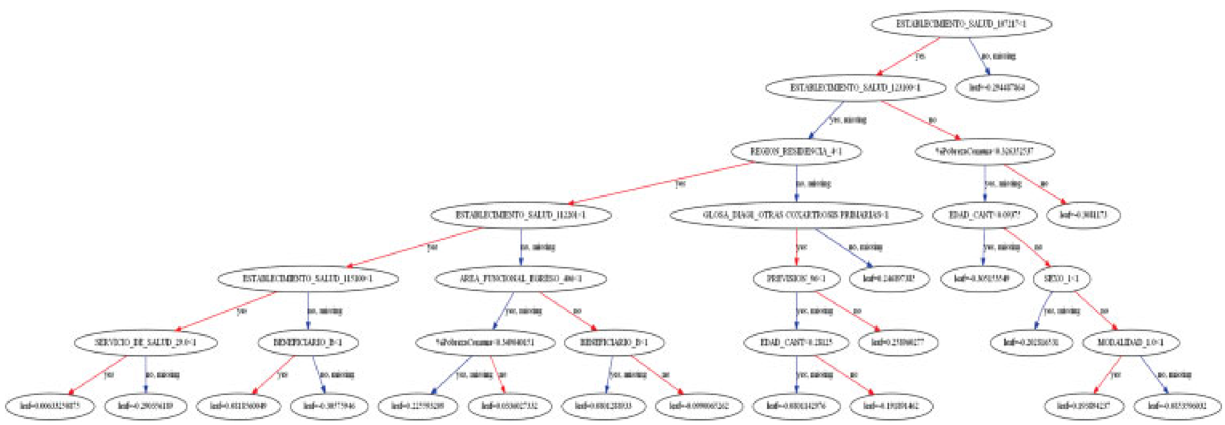

Fig. 6 A representative classification tree of the XGBoost algorithm.

►**Table 2** shows the different classification metrics for each of the evaluated algorithms. Following the concept of accuracy (ratio of the correct number of predictions over the total of samples), the XGBoost algorithm was able to correctly predict 81.74% of the time when a case corresponded to a short or long stay.

To also inquire about the relative importance of the explanatory variables, the importance score assigned by the algorithm to the thirty most important variables is reported in ►**Figure 5.** In this sense, the fact that the region of residence, the health service, the health center where the patient was operated on, and the care modality are the variables that most determine the length of stay of patients.

►**Figure 6** shows a representative classification tree of the XGBoost algorithm.

Discussion

Our research project successfully developed and validated a model to predict the length of hospital stay in Chilean patients over 65 years of age undergoing THA using artificial intelligence in its machine learning modality and big data of Chilean origin. The XGBoost algorithm had the best predictive performance by discriminating when the hospital stay is classified as shortened or prolonged (longer or shorter than three days). We also found that the five most important factors in this prediction, all freely accessible in the ministerial database, are the region of residence, the health service, hospital, and the FONASA modality of care. The accuracy of the algorithm in terms of classification is good.

According to Ramkumar et al.,[29] machine learning could be described as a software that perform tasks automatically based on a data source without an explicit programming. This technology has rapidly been incorporated into medicine, and it represents the natural extension of traditional statistical methods. Specifically in the arthroplasty literature, there are several recent publications that use machine learning to create prediction models of length of hospital stay and payments related to surgeries,[29] probability of complications,[26] satisfaction[30] etc. All of these publications, like the present one, use extensive databases that can be considered big data.[31]

Our study has several limitations and some notable aspects. The first limitation is that it is a registry study; therefore, there is a possible collection and coding bias that could finally alter the results, especially considering that the ICD-10 and FONASA codes are used to identify the studied cases. Despite this observation, we believe that since it is a ministerial database, with all the rigor that this implies, it is solid enough to overcome this limitation. Secondly, none of the database studies contains enough information at the patient level.[32] This is especially important in our work, considering that most of the studies carried out in the Northern Hemisphere using this methodology use variables at the patient level, including comorbidities and, in some cases, functionality.[16,26,30] We consider this to be the main flaw in our work; however, the database used is the only one that allowed us to freely access big data at the national level. Despite this observation, it is necessary to emphasize that the role of the individual characteristics of the patient may not be the most relevant one in explaining the length of hospital stay in elective arthroplasty. Kang et al.[33] demonstrated, in a series of two thousand patients, that the main determinants of prolonged stay in arthroplasty are social factors: admission to hospital the day before surgery, and late start with postoperative rehabilitation. In paralel, Burn et al.[34] showed that, although the individual factors of the patients are relevant to explain the length of hospital stay in arthroplasty, between 1997 and 2014 in the United Kingdom, a reduction in the length of stay was achieved due to the improvement in the efficiency of the practices, given that the profile of the patients remained stable. Further reinforcing the fact that the individual characteristics of the patients are secondary when explaining variability at the time of hospital discharge, the Cleveland Clinic OME Arthroplasty Group demonstrated (using American big data) that, in elective THA patients, "while the factors related to patients explain some variation in the hospital stay, the main culprits are the factors related to the procedure, specifically the hospital"[35] where the patient was operated on, with the surgical approach used also having a determinant role. This mentioned evidence helps to understand the results of our study and to weigh the lack of individual variables as a non-critical limitation of our model. Thirdly, considering that the COVID-19 pandemic could have influenced the practice of THA[11] in Chile in terms of its postoperative period and earlier discharge from hospital,[12,36] we believe that the data corresponding to the years 2016-2018 may not be completely representative of the scenario that we are going to experience in 2021. However, the fundamentals of our algorithm can be used to evaluate the results of after THA hospital discharges registered for the year 2020 and beyond.

The question that arises is: is this calculator useful in our scenario? The evaluation of the possibility of early or late discharge from a highly-frequent surgery guaranteed by law is of total relevance in public policies. Calculating the different possibilities of early discharge for a FONASA patient who undergoes surgery in hospital A versus hospital B, or clinic X, is useful to visualize the variability that exists in practices. When generating bundled-payment models, it is important to predict whether the patient operated on in Hospital A will have a longer hospital stay than in Hospital B. The usefulness of the "bedside" calculator may be limited by the absence of free-access clinical big data in Chile, but, on the other hand, the usefulness from the perspective of evaluating the performance of institutions is very high. As we stated in the objectives of the study, the identification of groups with a high probability of a shortened stay (certain patients in some hospitals) can help institutions to further improve their practices. On the other hand, the identification of hospitals that are not efficient in the management of their hospital stays may help the authorities to allocate resources in order to improve their practices.

Regarding the strengths of our study, we believe that the first and most important is the achievement of a multidisciplinary effort involving four experts, two of them surgeons and two engineers with formal education in artificial intelligence, who performed the first study involving big data and artificial intelligence in our specialty in Chile.

Conclusion

In the present study, we developed machine-learning algorithms based on free-access Chilean big data, and we were able to validate a tool that demonstrates an adequate discriminatory capacity to predict the probability of a shortened versus prolonged hospital stay in elderly patients undergoing THA for osteoarthritis.

References

1 Problema de salud GES N° 12: Endoprótesis total de cadera en personas de 65 años y más con artrosis de cadera con limitación funcional severa. Orientación en Salud Superintendencia de Salud, Gobierno de Chile n.d. http://www.supersalud.gob.cl/difusion/665/w3-article-586.html (accessed December 18, 2020).

2 Grosso MJ, Neuwirth AL, Boddapati V, Shah RP, Cooper HJ, Geller JA. Decreasing Length of Hospital Stay and Postoperative Complications After Primary Total Hip Arthroplasty: A Decade Analysis From 2006 to 2016. J Arthroplasty 2019;34(03):422–425. Doi: 10.1016/j.arth.2018.11.005

3 Goyal N, Chen AF, Padgett SE, et al. Otto Aufranc Award: A Multicenter, Randomized Study of Outpatient versus Inpatient Total Hip Arthroplasty. Clin Orthop Relat Res 2017;475(02): 364–372. Doi: 10.1007/s11999-016-4915-z

4 Paredes O, Ñuñez R, Klaber I. Successful initial experience with a novel outpatient total hip arthroplasty program in a public health system in Chile. Int Orthop 2018;42(08):1783–1787. Doi: 10.1007/s00264-018-3870-6

5 Greenky MR, Wang W, Ponzio DY, Courtney PM. Total Hip Arthroplasty and the Medicare Inpatient-Only List: An Analysis of Complications in Medicare-Aged Patients Undergoing Outpatient Surgery. J Arthroplasty 2019;34(06):1250–1254. Doi: 10.1016/j.arth.2019.02.031

6 Featherall J, Brigati DP, Faour M, Messner W, Higuera CA. Implementation of a Total Hip Arthroplasty Care Pathway at a High-Volume Health System: Effect on Length of Stay, Discharge Disposition, and 90-Day Complications. J Arthroplasty 2018;33(06):1675–1680. Doi: 10.1016/j.arth.2018.01.038

7 Ripollés-Melchor J, Abad-Motos A, Díez-Remesal Y, et al; Postoperative Outcomes Within Enhanced Recovery After Surgery Protocol in Elective Total Hip and Knee Arthroplasty (POWER2) Study Investigators Group for the Spanish Perioperative Audit and Research Network (REDGERM) Association Between Use of Enhanced Recovery After Surgery Protocol and Postoperative Complications in Total Hip and Knee Arthroplasty in the Postoperative Outcomes Within Enhanced Recovery After Surgery Protocol in Elective Total Hip and Knee Arthroplasty Study (POWER2). JAMA Surg 2020;155(04):e196024–e196024. Doi: 10.1001/jamasurg.2019.6024

8 Manning DW, Edelstein AI, Alvi HM. Risk Prediction Tools for Hip and Knee Arthroplasty. J Am Acad Orthop Surg 2016;24(01):19–27. Doi: 10.5435/JAAOS-D-15-00072

9 Sconza C, Respizzi S, Grappiolo G, Monticone M. The Risk Assessment and Prediction Tool (RAPT) after Hip and Knee Replacement: A Systematic Review. Joints 2019;7(02):41–45. Doi: 10.1055/s-0039-1693459

10 Diaz Ledezma C, Radovic I. What's new in hip arthroplasty? South American perspective. Recent Advances in Orthopedics-2. Jaypee Brothers Medical Publishers (P) Ltd; 2018

11 Parvizi J, Gehrke T, Krueger CA, et al; International Consensus Group (ICM) and Research Committee of the American Association of Hip and Knee Surgeons (AAHKS) Resuming Elective Orthopaedic Surgery During the COVID-19 Pandemic: Guidelines Developed by the International Consensus Group (ICM). J Bone Joint Surg Am 2020;102(14):1205–1212. Doi: 10.2106/JBJS.20.00844

12 Donell ST, Thaler M, Budhiparama NC, et al. Preparation for the next COVID-19 wave: The European Hip Society and European Knee Associates recommendations. Knee Surg Sports Traumatol Arthrosc 2020;28(09):2747–2755. Doi: 10.1007/s00167-020-06213-z

13 Myers TG, Ramkumar PN, Ricciardi BF, Urish KL, Kipper J, Ketonis C. Artificial Intelligence and Orthopaedics: An Introduction for Clinicians. J Bone Joint Surg Am 2020;102(09):830–840. Doi: 10.2106/JBJS.19.01128

14 Cabitza F, Locoro A, Banfi G. Machine Learning in Orthopedics: A Literature Review. Front Bioeng Biotechnol 2018;6:75. Doi: 10.3389/fbioe.2018.00075

15 Bini SA. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? J Arthroplasty 2018;33(08):2358–2361. Doi: 10.1016/j.arth.2018.02.067

16 Haeberle HS, Helm JM, Navarro SM, et al. Artificial Intelligence and Machine Learning in Lower Extremity Arthroplasty: A Review. J Arthroplasty 2019;34(10):2201–2203. Doi: 10.1016/j.arth.2019.05.055

17 Ramkumar PN, Haeberle HS, Bloomfield MR, et al. Artificial Intelligence and Arthroplasty at a Single Institution: Real-World Applications of Machine Learning to Big Data, Value-Based Care, Mobile Health, and Remote Patient Monitoring. J Arthroplasty 2019;34(10):2204–2209. Doi: 10.1016/j.arth.2019.06.018

18 Anderson AB, Grazal CF, Balazs GC, Potter BK, Dickens JF, Forsberg JA. Can Predictive Modeling Tools Identify Patients at High Risk of Prolonged Opioid Use After ACL Reconstruction? Clin Orthop Relat Res 2020;478(07):0–1618. Doi: 10.1097/CORR.0000000000001251

19 Departamento de Estadisticas e Información de Salud. n.d. https://deis.minsal.cl/#publicaciones (accessed December 18, 2020).

20 Farley KX, Anastasio AT, Premkumar A, Boden SD, Gottschalk MB, Bradbury TL. The Influence of Modifiable, Postoperative Patient Variables on the Length of Stay After Total Hip Arthroplasty. J Arthroplasty 2019;34(05):901–906. Doi: 10.1016/j.arth.2018.12.041

21 Observatorio Social - Ministerio de Desarrollo Social y Familia. n.d. http://observatorio.ministeriodesarrollosocial.gob.cl/pobreza-comunal-2017#basedatos (accessed February 4, 2021).

22 Mackinnon A. The use and reporting of multiple imputation in medical research - a review. J Intern Med 2010;268(06):586–593. Doi: 10.1111/j.1365-2796.2010.02274.x

23 López V, Fernández A, García S, Palade V, Herrera F. An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using data Intrinsic Characteristics. Inf Sci 2013;250:113–141. Doi: 10.1016/j.ins.2013.07.007

24 del Rio S, Benitez JM, Herrera F. Analysis of Data Preprocessing Increasing the Oversampling Ratio for Extremely Imbalanced Big Data Classification. 14th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom, vol. 2, 2015, p. 180–5. https://doi.org/10.1109/Trustcom.2015.579

25 Cerda J, Cifuentes L. Uso de curvas ROC en investigación clínica: Aspectos teórico-prácticos. Rev Chilena Infectol 2012;29(02):138–141. Doi: 10.4067/S0716-10182012000200003

26 Harris AHS, Kuo AC, Weng Y, Trickey AW, Bowe T, Giori NJ. Can Machine Learning Methods Produce Accurate and Easy-to-use Prediction Models of 30-day Complications and Mortality After Knee or Hip Arthroplasty? Clin Orthop Relat Res 2019;477(02):452–460. Doi: 10.1097/CORR.0000000000000601

27 Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet 2019;393(10181):1577–1579. Doi: 10.1016/S0140-6736(19)30037-6

28 Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015;350:g7594. Doi: 10.1136/bmj.g7594

29 Ramkumar PN, Navarro SM, Haeberle HS, et al. Development and Validation of a Machine Learning Algorithm After Primary Total Hip Arthroplasty: Applications to Length of Stay and Payment Models. J Arthroplasty 2019;34(04):632–637. Doi: 10.1016/j.arth.2018.12.030

30 Kunze KN, Polce EM, Sadauskas AJ, Levine BR. Development of Machine Learning Algorithms to Predict Patient Dissatisfaction After Primary Total Knee Arthroplasty. J Arthroplasty 2020;35(11):3117–3122. Doi: 10.1016/j.arth.2020.05.061

31 MeSH Browser. n.d. https://meshb.nlm.nih.gov/record/ui?ui=D000077558 (accessed December 19, 2020).

32 Grauer JN, Leopold SS. Editorial: large database studies–what they can do, what they cannot do, and which ones we will publish. Clin Orthop Relat Res 2015;473(05):1537–1539. Doi: 10.1007/s11999-015-4223-z

33 Kang HW, Bryce L, Cassidy R, Hill JC, Diamond O, Beverland D. Prolonged length of stay (PLOS) in a high-volume arthroplasty unit. Bone Jt Open 2020;1(08):488–493. Doi: 10.1302/2633-1462.18.BJO-2020-0047.R1

34 Burn E, Edwards CJ, Murray DW, et al. Trends and determinants of length of stay and hospital reimbursement following knee and hip

replacement: evidence from linked primary care and NHS hospital records from 1997 to 2014. BMJ Open 2018;8(01):e019146. Doi: 10.1136/bmjopen-2017-019146
35 Girbino KL, Klika AK, Barsoum WK, et al; Cleveland Clinic OME Arthroplasty Group. Understanding the Main Predictors of Length of Stay After Total Hip Arthroplasty: Patient-Related or Procedure-Related Risk Factors? J Arthroplasty 2021;36(05):1663–1670.e4. Doi: 10.1016/j.arth.2020.11.029
36 Athey AG, Cao L, Okazaki K, et al. Survey of AAHKS International Members on the Impact of COVID-19 on Hip and Knee Arthroplasty Practices. J Arthroplasty 2020;35(7S):S89–S94. Doi: 10.1016/j.arth.2020.04.053

