Opportunities to Improve Feasibility, Effectiveness and Costs Associated with a Total Joint Replacements High-Volume Hospital Registry

Michele Ulivi[a], Valentina Meroni[a], Luca Orlandini[a], Lorenzo Prandoni[b], Nicolò Rossi[b], Giuseppe M. Peretti[a,c], Linda Greta Dui[d], Laura Mangiavini[a,c*], Simona Ferrante[d*]

Affiliations

[a] IRCCS Istituto Ortopedico Galeazzi, Via Riccardo Galeazzi, 4, 20161, Milan, Italy

[b] Residency Programme in Orthopedics and Traumatology, University of Milan, Via Festa del Perdono 7, 20122, Milan, Italy

[c] Department of Biomedical Sciences for Health, University of Milan, Via Festa del Perdono 7, 20122, Milan, Italy

[d] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Via Giuseppe Ponzio, 34, 20133, Milan Italy

** Laura Mangiavini and Simona Ferrante contributed equally to this paper as senior authors*

Corresponding author

Linda Greta Dui, MS

NearLab, Dept. Electronics, Information and Bioengineering, Politecnico di Milano

Via Giuseppe Colombo, 40, 20133 – Milano - IT

E-mail: lindagreta.dui@polimi.it

Fax: +39-0223999003

Phone: +39-0223999690

Abstract

Background. Clinical registries are powerful tools for collecting uniform data longitudinally, thus making it possible to evaluate the outcome of patients affected by a specific pathology. In the context of total joint arthroplasty, registries serve also as post-market surveillance. Adoption of registries is a heavy burden for clinical settings in terms of resources and infrastructures. Excessive workload leads to incomplete data collection which undermines the effectiveness of a registry and consequently the workload needs to be optimised.

Methods. Starting from the use case of the Istituto Ortopedico Galeazzi, the time and personnel dedicated to the registry was estimated. Analysis of the data collected in the first years enabled us to propose a methodology for workload reduction. Different Machine Learning models were leveraged to predict patients with excellent satisfaction to reduce the number of assessments in their clinical post-operative follow-up. Moreover, feature selection was used to identify any unnecessary clinical scale to collect.

Results. Given an acceptance rate of 3,500 patients per year, 22 doctors and 6 non-medical employees were required to adopt a registry properly. Among the tested models, the Naïve Bayes gave the best performance (AUPRC=0.81) in predicting patient satisfaction at six months. Moreover, we found that the 12-item Short Form was poorly informative in predicting satisfaction at six-months.

Conclusions. In this study machine learning was leveraged to provide a methodology to reduce workload in the use of pathology registries. Such workload reduction can have a considerable impact at a larger scale, and improve registry feasibility in high-volume hospitals.

Keywords

Total Joint Replacements; Registry; Workload; Optimisation; Machine Learning; High Volume Hospital; Patient Reported Outcome Measures; Clinical-Based Outcome Measures.

Introduction

Progressive ageing of the population in developed countries has led to an increase in the incidence of chronic diseases [1]. Arthritis represents one of the most common degenerative pathologies [2] and it is one of the major causes of pain and functional disability, thus constituting a heavy burden for Health Systems [3]. Total joint arthroplasty is a safe and efficient procedure to ease the pain and improve articular function in patients with a severe grade of arthritis [4]. Quality of life expectations after total joint replacement are very high [5]. Nonetheless, this clinical procedure is linked with the occurrence of both short- and long-term adverse events, such as, thromboembolism, infections, articular stiffness, instability, implant mobilisation and failure [6]. Therefore, a careful short- and long-term monitoring of patients undergoing total joint arthroplasty becomes a necessity [6,7]. Indeed, stakeholders such as insurance companies, medical staff, and manufacturers have great interest in collecting clinical data on total joint arthroplasty outcome and follow up [8]. Moreover, the newly released EU Medical Device Regulation (article 108) [9] strongly encourages the adoption of registries for implantable devices, to collect comparable information on long-term safety of implantable devices. Registries are pathology-specific databases, where a large number of patients' clinical outcomes are collected, also in a long-term follow-up perspective. Thus they are a keystone in clinical practice as they provide the so-called "real world"

therapeutic situation [9,10]. Indeed, registries represent a valid tool to address the need for proactive post-market surveillance, not only in the orthopaedic field but in many other specialty areas [11].

Registry adoption presents some challenges. It requires an increase in manpower allocation to comply with operational aspects mostly related to patient contact and data collection. This resource requirement is not always compatible and feasible for hospitals or health systems. When insufficient personnel are dedicated, the large-scale registries are incomplete and totally lose their effectiveness and reliability in the longitudinal monitoring of patients.

To overcome this difficulty and comply with the new regulation, there is a need for workload quantification and optimisation. Clinical workload optimisation can be seen as a classical customer segmentation optimisation problem, using patient outcomes as an evaluation metric. These optimisation difficulties are typically overcome using Machine Learning (ML) methodologies. ML applied to registry data is gaining interest in the medical field and it has been applied in different situations such as the prediction of patient outcomes [12,13], mortality, and complications [14]. Huber and colleagues [12] predicted patient-reported improvement (based on minimal important difference) after hip or knee replacement from two years of data collection. They achieved their best Area Under the Receiving Characteristic curve (AUROC) predicting the Visual Analogue Scale of pain (VAS) and the Oxford Hip and Knee Score (Q score) with Extreme Gradient Boosting (XGBoost) [15], with results between 0.70 and 0.87. Zupan and colleagues [13] predicted long-term outcome after hip prosthesis implantation. They proposed a Naïve Bayes classifier [16] paired with a hierarchical decision model, where expert knowledge injection was possible. They evaluated the model using the accuracy metric (56.3%).

Harris and colleagues [14] leveraged large scale registries to build risk-prediction models for 30-days mortality and complications after total joint arthroplasty. They used Lasso Regression [17] on demographic and clinical variables, and evaluated the performance using the C-statistic. They achieved their best results in predicting renal complication (0.78), cardiac complication (0.73), and death (0.73). To this extent, ML is very promising, as it provides performant computational methods as a guide towards a better standard of care.

In this study, we focus on the experience of the Istituto Ortopedico Galeazzi (IOG). At IOG, a registry has been adopted by the Ortopedia Ricostruttiva Articolare della Clinica Ortopedica (Joint Reconstruction Orthopaedics, ORACO) division since 2013. This electronic registry monitors patients undergoing joint arthroplasty. In fact, it is a single collector of information related to surgical operations, such as type of implanted device, collected pre- and post-operative diagnostic imaging (e.g., X-rays, CT and RMN, as well as any other imaging technique). Moreover, it stores Patient Reported Outcomes Measures (PROMs) and clinical scales, and helps in their collection.

This study has two main objectives: first, to assess overall time needed for collection of data with the registry, with Istituto Ortopedico Galeazzi as a reference; second, to provide methods to reduce the workload, by means of Machine Learning techniques.

Methods

In this section, we address workload estimation and optimisation. The term “workload” refers to the amount of time that people of the ORACO division devoted to the use of the registry. In our specific context, we refer to the number of visits performed by medical staff, which includes Clinical-Based Outcome Measures (CBOMs) administration, and to

the amount of time dedicated to Patient-Reported Outcome Measures (PROMs) completion by non-medical staff.

The described analyses were performed on data collected after both Ethical Committee approval and the collection of written informed consent by all the participants involved.

To be included into the registry, patients must be 18 years old or more and in need of a hip or knee arthroplasty. The only exclusion criterion was to present severe comorbidities which might prevent patients from returning to the hospital for follow-up visits.

Workload estimation

Table 1 summarises the monitoring procedure adopted at IOG for patients undergoing hip and knee surgery. The procedure included seven visits, covering ten-year time span. Each visit was scheduled at the given time, indicating also a tolerance interval acceptable for each of them.

FOLLOW-UP STEP	TOLERANCE
Pre-op phase	-
Three months	15 days
Six months	15 days
One year	30 days
Two years	30 days
Five years	One year
Ten years	One year

Table 1: Protocol structure

Regarding the visits, each one was conducted by two residents. One of them actually visited the patient while the other completed the CBOMs. We estimated that it is possible

to visit four patients per hour, but residents can spend only half of their working day on this activity. Thus, we considered four hours per day, five days per week, forty-six weeks per year.

The registry was personalised to collect data from the IOG hospital. Two CBOMs were chosen: the Harris Hip Score (HHS) [18], and the Knee Society Score (KSS) [19]. Such CBOMs were intended to assess the dysfunction of patients who underwent hip or knee surgery respectively. They were administered by a qualified healthcare professional, i.e., the resident. Concerning the PROMs, all patients completed the Visual Analogue Scale for Pain (VAS) [20], and the 12-item Short Form (SF-12) [21]. In addition, hip patients completed the Hip disability and Osteoarthritis Outcome Score (HOOS-PS) [22], and knee patients completed the Knee injury and Osteoarthritis Outcome Score (KOOS-PS) [23]. From the first follow-up on, the Satisfaction rate for the patient was added. As these questionnaires are PROMs, intended to be representative of the patients' voices without doctor mediation [24], they were usually collected before the visit. An automatic alert was sent to those patients who provided an e-mail address when they entered the registry but when this method was ineffective (e.g., the elderlies) completion of the PROMs was aided by IOG operators. They helped patients *in loco*, providing tablets with digital forms and supporting the completion of the tests, or they interviewed patients through a phone call. Collecting time for all the PROMs, considering technical needs and time for calls and recalls, was fifteen minutes (four patients per hour). The operators could devote all their working time to this task (eight hours), five days per week, forty-six weeks per year.

For a very practical estimation of the hospital-level effort needed to manage all the follow-up visits, a projection was made based on the described protocol and patient volume. The projection started from the assumption that 3,500 new patients are admitted

each year on the basis of historical data from the IOG. Due to the old age of the population that undergoes total joint replacement, we also adjusted the projection by a *censoring* factor of 2% at the five-year follow-up, and 5% at ten years [25]. To further approximate a realistic effort, we considered the real-life dropout experienced by the ORACO division of the IOG, as reference sample. Their data were stored in the registry, and they included longitudinal assessment of patients undergoing hip and knee surgery since 2013. We fitted their form completion rate to the IOG workload, to build a data-driven description of a realistic number of visits.

On these bases, we computed the number of required visits and interviews necessary to fulfil the protocol, considering both the *per protocol* effort and the real-life dropout-adjusted effort.

Workload optimisation

To achieve workload reduction, two approaches can be adopted: delete a whole follow-up visit or delete just some assessments in each visit. In the first approach, prediction models were used to forecast each patient's need for further examination, with the aim of decreasing the number of required follow-ups. In the second approach, feature selection was performed to identify whether some questionnaires can be removed from follow-up visits, without any impact on the general assessment.

Regardless of the approach, the first step was to select the subgroup of patients that potentially can avoid some assessments, from among the largest homogeneous groups in the database. The Satisfaction questionnaire was used to stratify the population. Specifically, the question "How do you describe the result of your surgical operation?" was used to dichotomise the population into the "excellent group" (answers *Excellent*)

and “non excellent” group (*Very good, Good, Quite good*, and *Bad* answers). The excellent group was considered as the sample that does not need to be visited in the follow up.

Descriptive statistics was used to select the candidate time point to be deleted from the “excellent” patients’ follow-up programme. The time point that significantly overcame the threshold for an excellent result (90 points [26], for the Harris Hip score) in terms of clinical-based outcome measure was selected. Rather than PROMs, we considered the CBOMs as they closely reflect the real clinical situation, rather than patient perceptions.

To test the potentiality of Machine Learning in predicting patient satisfaction at the selected time point given the previous assessments we investigated the following ML techniques: Logistic classifier, Support Vector Machines (SVM), Random Forest (RF), and Naïve Bayes (NB) [27]. Model predictors were: demographic variables, such as age at the operation, gender, BMI (pre-operatory, at three months, and its change); variables related to the operation, such as the season of the year (a categorical variable with four levels), and the length in minutes of the operation; the PROMs scores (i.e., the VAS, the SF-12, Mental and Physical components, the Hoos-ps), in the pre-operatory phase, their corresponding scores at three months, and their differences; the Satisfaction rate at three months; and the clinical questionnaire, again in the pre-operatory phase, at three months and its difference. The total number of available features was 38. Categorical variables with more than two levels (i.e., Satisfaction rate at three months) were replaced by dummy variables, with the One-Hot Encoding technique [27]. All the other variables were treated as continuous. We checked for collinearity between variables through a Spearman correlation, considering 0.95 as deletion threshold, which ensures a Variance Inflation Factor less than 10 [28]. To avoid data imputation, which may lead to spurious

information, we excluded patients without a complete predictors set. We stratified the dataset into training and test set with a proportion of 80:20, and we kept the proportions of the two classes as in the original dataset both in the training and test sets. The continuous variables were standardised considering training set statistics, to avoid data leakage [29]. A ten-fold cross-validation was used to train the models, as internal verification. Then, to achieve external verification, we tested the trained model on the previously held out test set. This procedure was repeated five times, with random training/test splits, to better approximate the real performance of the models. As the class of interest was under-represented, the evaluation metric for our predictions was the Area Under the Precision-Recall Curve (AUPRC) [30].

Concerning feature selection, we chose the Correlation-based Feature Subset selection [31], with a Best First search method and ten-fold cross-validation. The attributes regarded as important by the algorithm at least in one of the ten folds were selected and we repeated the training on the new set of features on the most performing model. To determine whether the predictive power of the reduced set of features could be compared to that of the complete set of features, we performed a Wilcoxon matched paired test on the model AUPRC.

Feature engineering was performed in R 3.3.3, feature selection and the predictions were made in Weka 3.8.

Results

Workload estimation

Figure 1 reports the number of hip and knee patients enrolled in the registry by the ORACO division from January 2013 to July 2019. The number of patients is decreasing

as some of them have not reached each time point yet. In absolute terms (Figure 1 – Panel A), 1,386 ORACO patients started the follow-up programme (the pre-operatory column in the graph). At the operation, ORACO hip and knee patients have a mean age of 71 ± 10 years old. 1,239 reached the three-month follow-up step, 1,177 the six-month step, 1,076 the twelve-month step, 893 the two-year step, and 450 the five-year step. Given these figures, Figure 1 – Panel B reports the dropout percentages, based on the number of forms collected by the ORACO division. The completion rate follows an exponential decay, with an average of 71% the first year (mediated on 3, 6, and 12 months), 43% at 24 months, and 34% at 5 years. Given this trend, we considered a completion rate of 30% at 10 years, for the projection.

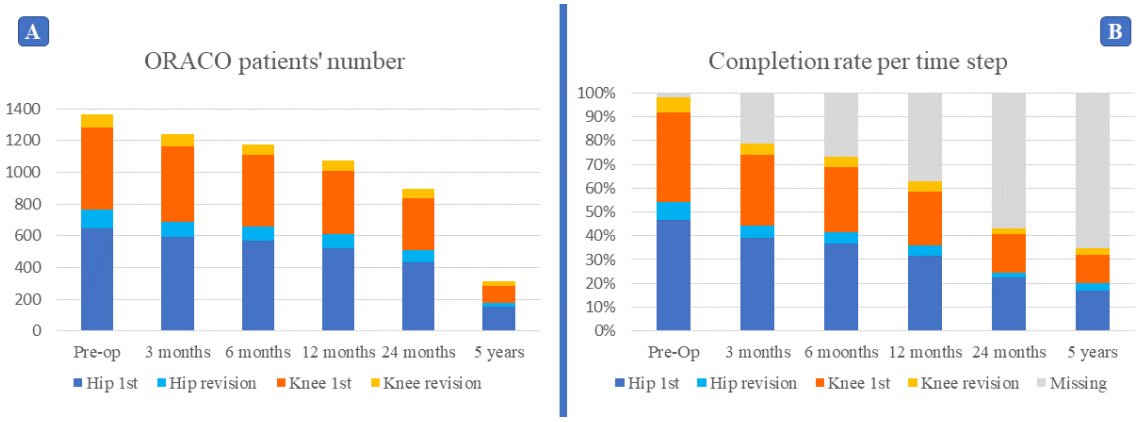

Figure 1: ORACO patients follow-up situation. Panel A: number of patients per follow-up step for the ORACO division from 2013 to 2019, divided by specialty (hip and knee, first operation and revision). Panel B: forms completion rate for the ORACO division, normalised by the number of patients in each time point.

Considering the whole hospital effort of 3,500 patients, i.e., the entire hip and knee IOG workload, Figure 2, Panel A, shows the *per protocol* projection of the number of visits per year. It starts from the very beginning of a registry (year one) and ends when patients who entered the follow-up at year one successfully complete all the steps (year eleven).

Solid bars represent the actual visits required. According to the protocol structure reported in Table 1, in year one the expected visits are: pre-operatory, 3 months, and 6 months, which brings the total to 10,500 visits. One year after the beginning of the registry (year two), in addition to the new patients 10,500 visits, the previous year's patients reach the 12-month follow-up, so that the total number of required visits becomes 14,000. They reach a plateau after ten years of registry activity (year eleven). Even adjusting for the *censoring*, we reach the impressive number of 24,255 visits per year. Given that the proposed protocol offers the opportunity to anticipate or delay a visit within a specified tolerance time interval (see Table 1), we also estimated the maximum workload generated by the worst case tolerance effect. Indeed, if all the visits of the following year are anticipated and all the visits of the previous year are delayed, we reach the excessive maximum of 39,000 follow-up visits in one year (dashed bars in Figure 1). Given this estimation of the number of visits, Figure 2, Panels B and C, represents the number of workers who must be devoted to this apparatus, suggesting the great amount of resources needed, both in terms of people and infrastructures (dedicated rooms, phones, etc.). When the registry is at regime, from 7 to 11 pairs of residents (i.e., from 14 to 21 students) must spend half of their working day administering questionnaires and follow-up visits. On the other hand, from 4 to 6 full-time employees must be dedicated to *in loco* completion and calls and recalls. Nonetheless, the ORACO division's experience confirmed that there is a substantial real-life dropout (see Figure 1, Panel B). Figure 2, Panel D-F, reports the same projection corrected by the dropout observed in the ORACO division. Given this fact, the maximum number of patients is 23,523, which corresponds to 5 to 7 pairs of doctors and 3 to 4 operators.

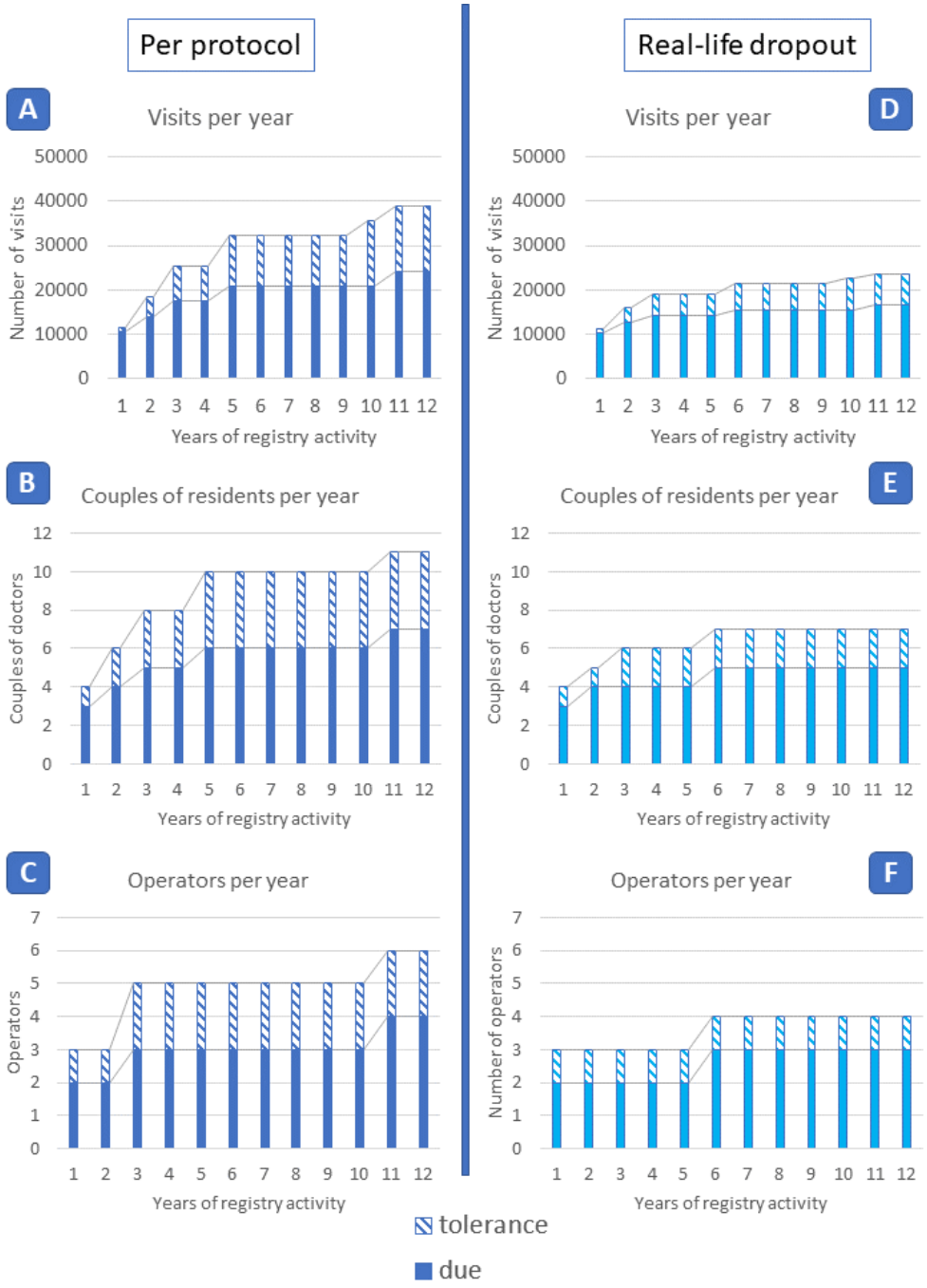

Figure 2: Workload and personnel involvement estimation. Panel A: Estimation of the required visits from the start up to the regime phase. Panel B: Estimation of the required pairs of doctors to carry out the expected visits and CBOMs. Panel C: Estimation of the operators required to complete the PROMs. Panel D, E and F: Estimation of the required visits, pairs of doctors, and operators, projecting ORACO's real-life dropout on the whole hospital workload. In all panels

the solid bars represent the follow-up executed at the due time, the dashed bars introduce the tolerance effect.

Workload optimisation

To predict the outcomes and optimise the workload, we considered the largest homogeneous group in our database: hip replacements, first operation. Figure 3, Panel A, shows a trend with improvement on the HHS scale for all hip patients. From six months on a plateau is reached and the score is significantly above the 90-point threshold, as the 95% confidence interval of the median (the notch in the box-and-whiskers plot) is completely above this threshold. Thus, we selected six months as the candidate time point to be discarded from the assessment. To further confirm this choice, we dichotomized the HHS score using the six months Satisfaction score. The number of hip patients who reached the six-month time point and satisfied the condition of a complete predictors set was 213. More specifically, the dataset comprised 129 non-excellent patients (84 females, 45 males, 69.5 ± 11.8 years old) and 84 excellent patients (45 females, 39 males, 71.2 ± 11.0 years old). As shown in Figure 3, Panel B, at three months only the "non-excellent" group did not significantly reach the HHS-based excellent result (the 95% confidence interval of the median is below the 90 points threshold).

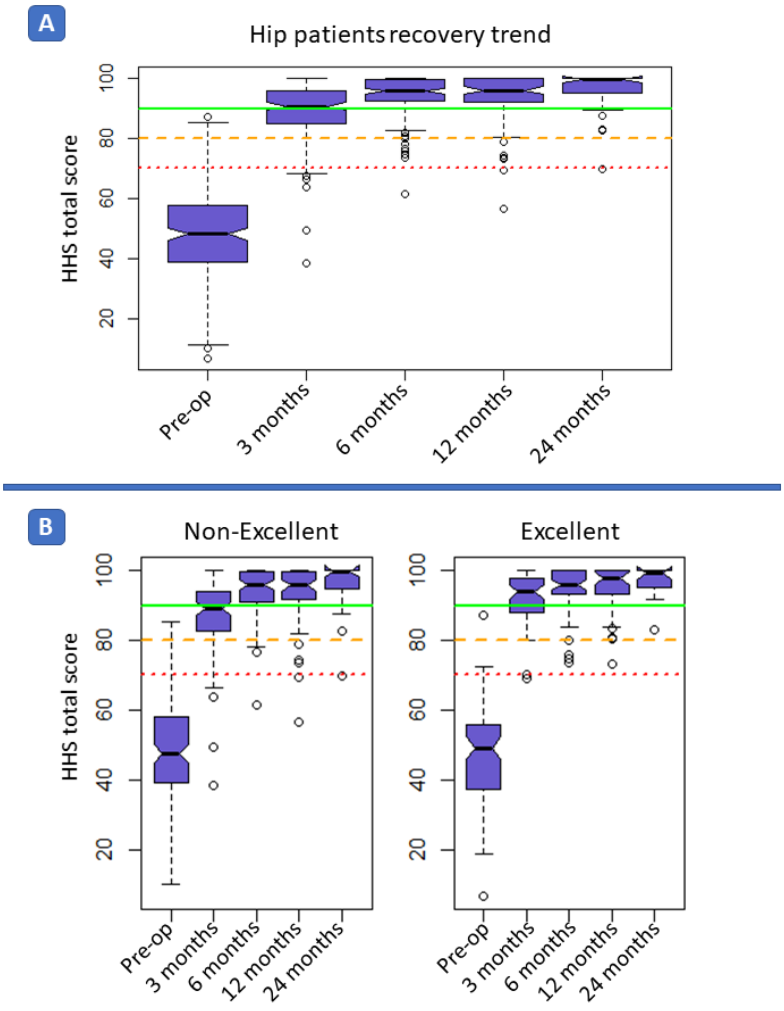

Figure 3: Harris Hip Score trend. Panel A: all ORACO patients considered for the prediction. Panel B: stratification based on the six-month Satisfaction. Y-axis: Harris Hip Score. X-axis: follow-up step. The score is reported with box-and-whiskers plot, where the box is the inter-quartile range, the horizontal line is the median and the notch is its 95% confidence interval. The horizontal lines are the poor (red, dotted line), good (orange, dashed line) and excellent (green, solid line) score thresholds, as in Nilsdotter et al [25].

We now report the results of the predictions on the selected patients group, to suggest those who do not need to be visited at six months. In the feature engineering phase, none of the predictors was deleted for collinearity. The first step of model selection was performed on a Logistic classifier with a Ridge correction of 0.1; on a Support Vector

Machine with a one-degree polynomial kernel and a complexity parameter equal to 2; on a Random Forest with trees of three attributes and depth equal to five; and on a Naïve Bayes without kernel estimation. The weighted AUPRC for the training and test set are reported in Table 2, rows 1 to 4, with their median and inter quartile range on five repetitions. The results are presented as class performance, and the weighted average. Given these results, we proceeded with the Naïve Bayes classifier.

	Internal verification (cross-validation training)			**External verification (hold out testing)**		
	Non-excellent	**Excellent**	**Weighted average**	**Non-excellent**	**Excellent**	**Weighted average**
Logistic	0.77 (0.05)	0.60 (0.06)	0.70 (0.04)	0.82 (0.07)	0.74 (0.23)	0.78 (0.07)
SVM	0.70 (0.02)	0.51 (0.04)	0.62 (0.01)	0.62 (0.18)	0.52 (0.32)	0.63 (0.14)
RF	0.80 (0.03)	0.61 (0.06)	0.72 (0.03)	0.81 (0.09)	0.69 (0.25)	0.75 (0.15)
NB	0.84 (0.02)	0.68 (0.10)	0.76 (0.04)	0.86 (0.11)	0.74 (0.31)	0.81 (0.11)
NB (feature selection)	0.84 (0.02)	0.69 (0.06)	0.77 (0.02)	0.83 (0.12)	0.79 (0.21)	0.81 (0.12)

Table 2: Training and test set performance, in terms of median Area Under the Precision-Recall Curve and its inter quartile range. Columns: single class and weighted average performance. Rows: models. The last line reports the performance of the Naïve Bayes with a subset of features.

Figure 4 reports the results of feature selection. The most important feature, selected in each of the ten repetitions of the algorithm, was the three months "excellent" answer to the Satisfaction questionnaire. Other important features were: gender, Hoos-ps at three months and its increment, VAS at three months and its increment, surgery length, HHS

total and sub-scores increment, and the “quite good” Satisfaction answer at three months. The SF-12 score never appears in the feature selection, nor does the BMI. The other scores appeared both as three-month scores and as delta scores, thus it is not possible to eliminate their collection in one of the two steps. The Wilcoxon matched pairs test on the test set AUPRC (five repetitions) revealed that there was not any difference ($p = 1$) between the model trained on the complete set of attributes or on the selected attributes only. This result is reported in the last row of Table 2.

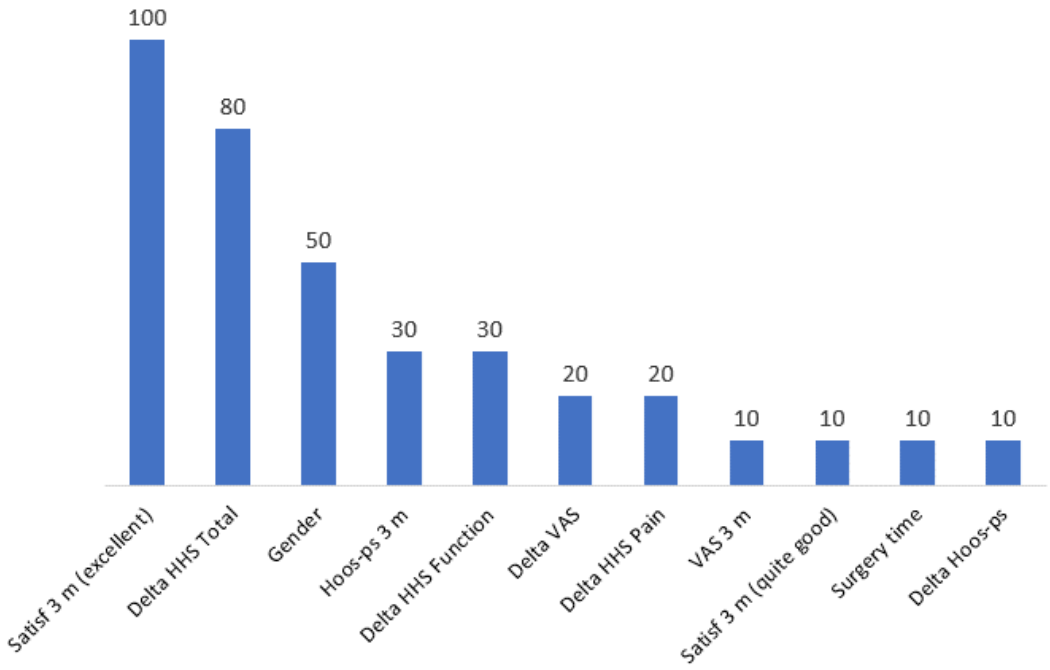

Figure 4: Feature importance. The x-axis reports the feature selected by the algorithm, the y-axis the percentage of folds each feature was selected.

Considering that the “excellent patients” represented 39% of our dataset, the workload was recomputed to estimate the achieved reduction. Each year we can exclude 1,365 patients from the six-month visit. This action would be translated into a maximum number of visits of 32,000, instead of 39,000. Concerning the two doctors needed to perform each of such visits, they decrease from a maximum of 11 to a maximum of 9 pairs. As for non-

medical operators, we additionally consider that it would be possible to interview one patient more per hour (from 4 to 5) because of the SF-12 removal from the protocol, thus resulting in a maximum number of 4 operators needed at registry regime, instead of 6.

Discussion

In this work, we analysed the feasibility of maintaining a total joint arthroplasty high-volume registry. The importance of registries to collect information about the operation and the follow-up is ubiquitously recognized [9]. Nonetheless, the excessive medical and non-medical workload might reduce their application, and undermine their effectiveness.

To address this problem, we presented a projection of the effort needed to maintain a total joint arthroplasty registry, starting from the real case of the ORACO division of the Istituto Ortopedico Galeazzi experience. Moreover, we provided methods to reduce such workload, with the final aim of improving feasibility, effectiveness, and associated costs.

Our projection of a full-operative registry shows that the great number of new incoming patients and previously operated follow-ups prevent protocol adherence. Such a workload does not consist only of medical examinations. In fact, collecting Patient Reported Outcome Measures also requires dedicated non-clinical personnel and infrastructures, such as rooms, tablets, and phones. The most relevant consequence of the excessive workload is the high number of dropouts. Electronic registries may aid in PROMs collection by sending automatic email alerts in the due follow-up time. However, in total joint arthroplasty, this feature is still ineffective, considering the old age of the patients (71 ± 10 years old). Nonetheless, the help of the electronic platform alone might smooth

the data collection process and enable analytics, but cannot avoid the need for visits, PROMs and clinical scales collection.

To improve feasibility operatively we proposed a workload optimisation. We focused both on reducing the need for one of the follow-up steps for non-critical patients and on reducing the number of PROMs collected for each step. For this purpose we built a machine learning model to predict which patients can avoid the six-month follow-up as they would report excellent satisfaction. The performance of our model can be compared to the State of the Art of hip surgery outcome prediction. For instance, Zupan and colleagues [13] reached their best performance in predicting long-term outcome with a Naïve Bayes classifier at 56.3% of accuracy; Huber et al. reached an area under the ROC curve of 0.87, when classifying patients who surpassed the minimal important difference for the VAS scale [12]. Given the precision and recall in estimating patient satisfaction at six months, we suggest reducing the number of control visits at this time point, and devote them only to those patients who are predicted to achieve "non-excellent" satisfaction. Additionally, we found that the SF-12 questionnaire was poorly informative in predicting six-month satisfaction. Therefore, we suggest reducing the number of PROMs collected, eliminating the SF-12. This methodology has the potential of creating a big impact at scale, as they reduce the number of visits and shorten the evaluation. Indeed, applying the proposed workload reduction to the IOG use case would produce an 18% reduction of control visits and pairs of doctors, and a 33% reduction of operators needed to comply with the protocol.

Besides the reported use case, the main value of our work is to provide a general method to achieve workload optimisation in the context of pathology registries. In fact, in other clinical settings we may wish to collect different outcome measures, or perform periodic

evaluation at different time points. Notwithstanding this, preliminary data collected in the registry setting up phase should be used to perform a custom optimisation before the workload becomes excessive.

A limit of these analyses is that we cannot include in the model those patients who did not complete all the follow-up steps. In particular, we could not access the health status of those who did not answer the six-month follow-up. Thus, these results are affected by Non-Response bias by design [32]. Nonetheless, we can assume that the workload reduction proposed will make it possible to visit and collect scales from more patients, thus progressively reducing Non-Response influence.

Conclusions

In conclusion, in this work we quantified the difficulties of a total joint replacement registry maintenance, and we propose two actions to optimise the workload: first, to reduce the number of control visits, and devote them only to patients who will not achieve an "excellent" satisfaction predicted by a machine learning model; second, to reduce the number of collected PROMs, eliminating the less informative in such prediction. We believe that these actions have the potential to improve the feasibility, effectiveness, and costs associated with a total joint replacements high-volume hospital registry.

Conflict of interest statement

Linda Greta Dui was previously employed in the company who developed the electronic registry. None of other authors have conflicts of interest to disclose.

Authors' contribution

MU: Patients enrolment, patient treatment, study conceptualization, supervision of the methodology, paper drafting, revision.

VM: Patients enrolment, patient treatment, data collection, paper drafting.

LO: Patients enrolment, patient treatment, paper drafting, revision.

LP: Data collection, paper drafting.

NR: Data collection, paper drafting.

GMP: Study conceptualisation, supervision of the methodology, paper drafting, revision.

LGD: Data analysis and statistics, machine learning models, paper drafting.

SF: Study conceptualisation, supervision of the methodology used, paper drafting and revision.

LM: Study conceptualisation, supervision of the methodology, paper drafting, revision.

All authors have approved the final article.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] World Health Organization, Facing the facts: The impact of chronic disease in Canada., Prev. Chronic Dis. A Vital Investment. (2005). https://doi.org/10.1017/CBO9781107415324.004.

[2] M. Cross, E. Smith, D. Hoy, S. Nolte, I. Ackerman, M. Fransen, L. Bridgett, S. Williams, F. Guillemin, C.L. Hill, L.L. Laslett, G. Jones, F. Cicuttini, R. Osborne,

T. Vos, R. Buchbinder, A. Woolf, L. March, The global burden of hip and knee osteoarthritis: Estimates from the Global Burden of Disease 2010 study, Ann. Rheum. Dis. (2014). https://doi.org/10.1136/annrheumdis-2013-204763.

[3] C.J.L. Murray, T. Vos, R. Lozano, M. Naghavi, A.D. Flaxman, C. Michaud, M. Ezzati, K. Shibuya, J.A. Salomon, S. Abdalla, V. Aboyans, J. Abraham, I. Ackerman, R. Aggarwal, S.Y. Ahn, M.K. Ali, M.A. AlMazroa, M. Alvarado, H.R. Anderson, L.M. Anderson, K.G. Andrews, C. Atkinson, L.M. Baddour, A.N. Bahalim, S. Barker-Collo, L.H. Barrero, D.H. Bartels, M.G. Basáñez, A. Baxter, M.L. Bell, E.J. Benjamin, D. Bennett, E. Bernabé, K. Bhalla, B. Bhandari, B. Bikbov, A. Bin Abdulhak, G. Birbeck, J.A. Black, H. Blencowe, J.D. Blore, F. Blyth, I. Bolliger, A. Bonaventure, S. Boufous, R. Bourne, M. Boussinesq, T. Braithwaite, C. Brayne, L. Bridgett, S. Brooker, P. Brooks, T.S. Brugha, C. Bryan-Hancock, C. Bucello, R. Buchbinder, G. Buckle, C.M. Budke, M. Burch, P. Burney, R. Burstein, B. Calabria, B. Campbell, C.E. Canter, H. Carabin, J. Carapetis, L. Carmona, C. Cella, F. Charlson, H. Chen, A.T.A. Cheng, D. Chou, S.S. Chugh, L.E. Coffeng, S.D. Colan, S. Colquhoun, K.E. Colson, J. Condon, M.D. Connor, L.T. Cooper, M. Corriere, M. Cortinovis, K. Courville De Vaccaro, W. Couser, B.C. Cowie, M.H. Criqui, M. Cross, K.C. Dabhadkar, M. Dahiya, N. Dahodwala, J. Damsere-Derry, G. Danaei, A. Davis, D. De Leo, L. Degenhardt, R. Dellavalle, A. Delossantos, J. Denenberg, S. Derrett, D.C. Des Jarlais, S.D. Dharmaratne, M. Dherani, C. Diaz-Torne, H. Dolk, E.R. Dorsey, T. Driscoll, H. Duber, B. Ebel, K. Edmond, A. Elbaz, S. Eltahir Ali, H. Erskine, P.J. Erwin, P. Espindola, S.E. Ewoigbokhan, F. Farzadfar, V. Feigin, D.T. Felson, A. Ferrari, C.P. Ferri, E.M. Fèvre, M.M. Finucane, S. Flaxman, L. Flood, K. Foreman, M.H.

Forouzanfar, F.G.R. Fowkes, M. Fransen, M.K. Freeman, B.J. Gabbe, S.E. Gabriel, E. Gakidou, H.A. Ganatra, B. Garcia, F. Gaspari, R.F. Gillum, G. Gmel, D. Gonzalez-Medina, R. Gosselin, R. Grainger, B. Grant, J. Groeger, F. Guillemin, D. Gunnell, R. Gupta, J. Haagsma, H. Hagan, Y.A. Halasa, W. Hall, D. Haring, J.M. Haro, J.E. Harrison, R. Havmoeller, R.J. Hay, H. Higashi, C. Hill, B. Hoen, H. Hoffman, P.J. Hotez, D. Hoy, J.J. Huang, S.E. Ibeanusi, K.H. Jacobsen, S.L. James, D. Jarvis, R. Jasrasaria, S. Jayaraman, N. Johns, J.B. Jonas, G. Karthikeyan, N. Kassebaum, N. Kawakami, A. Keren, J.P. Khoo, C.H. King, L.M. Knowlton, O. Kobusingye, A. Koranteng, R. Krishnamurthi, F. Laden, R. Lalloo, L.L. Laslett, T. Lathlean, J.L. Leasher, Y.Y. Lee, J. Leigh, D. Levinson, S.S. Lim, E. Limb, J.K. Lin, M. Lipnick, S.E. Lipshultz, W. Liu, M. Loane, S. Lockett Ohno, R. Lyons, J. Mabweijano, M.F. MacIntyre, R. Malekzadeh, L. Mallinger, S. Manivannan, W. Marcenes, L. March, D.J. Margolis, G.B. Marks, R. Marks, A. Matsumori, R. Matzopoulos, B.M. Mayosi, J.H. McAnulty, M.M. McDermott, N. McGill, J. McGrath, M.E. Medina-Mora, M. Meltzer, Z.A. Memish, G.A. Mensah, T.R. Merriman, A.C. Meyer, V. Miglioli, M. Miller, T.R. Miller, P.B. Mitchell, C. Mock, A.O. Mocumbi, T.E. Moffitt, A.A. Mokdad, L. Monasta, M. Montico, M. Moradi-Lakeh, A. Moran, L. Morawska, R. Mori, M.E. Murdoch, M.K. Mwaniki, K. Naidoo, M.N. Nair, L. Naldi, K.M.V. Narayan, P.K. Nelson, R.G. Nelson, M.C. Nevitt, C.R. Newton, S. Nolte, P. Norman, R. Norman, M. O'Donnell, S. O'Hanlon, C. Olives, S.B. Omer, K. Ortblad, R. Osborne, D. Ozgediz, A. Page, B. Pahari, J.D. Pandian, A. Panozo Rivero, S.B. Patten, N. Pearce, R. Perez Padilla, F. Perez-Ruiz, N. Perico, K. Pesudovs, D. Phillips, M.R. Phillips, K. Pierce, S. Pion, G. V. Polanczyk, S. Polinder, C.A. Pope, S. Popova, E. Porrini, F.

Pourmalek, M. Prince, R.L. Pullan, K.D. Ramaiah, D. Ranganathan, H. Razavi, M. Regan, J.T. Rehm, D.B. Rein, G. Remuzzi, K. Richardson, F.P. Rivara, T. Roberts, C. Robinson, F. Rodriguez De Leòn, L. Ronfani, R. Room, L.C. Rosenfeld, L. Rushton, R.L. Sacco, S. Saha, U. Sampson, L. Sanchez-Riera, E. Sanman, D.C. Schwebel, J.G. Scott, M. Segui-Gomez, S. Shahraz, D.S. Shepard, H. Shin, R. Shivakoti, D. Silberberg, D. Singh, G.M. Singh, J.A. Singh, J. Singleton, D.A. Sleet, K. Sliwa, E. Smith, J.L. Smith, N.J.C. Stapelberg, A. Steer, T. Steiner, W.A. Stolk, L.J. Stovner, C. Sudfeld, S. Syed, G. Tamburlini, M. Tavakkoli, H.R. Taylor, J.A. Taylor, W.J. Taylor, B. Thomas, W.M. Thomson, G.D. Thurston, I.M. Tleyjeh, M. Tonelli, J.A. Towbin, T. Truelsen, M.K. Tsilimbaris, C. Ubeda, E.A. Undurraga, M.J. Van Der Werf, J. Van Os, M.S. Vavilala, N. Venketasubramanian, M. Wang, W. Wang, K. Watt, D.J. Weatherall, M.A. Weinstock, R. Weintraub, M.G. Weisskopf, M.M. Weissman, R.A. White, H. Whiteford, N. Wiebe, S.T. Wiersma, J.D. Wilkinson, H.C. Williams, S.R.M. Williams, E. Witt, F. Wolfe, A.D. Woolf, S. Wulf, P.H. Yeh, A.K.M. Zaidi, Z.J. Zheng, D. Zonies, A.D. Lopez, Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010, Lancet. (2012). https://doi.org/10.1016/S0140-6736(12)61689-4.

[4] L. Busija, L. Bridgett, S.R.M. Williams, R.H. Osborne, R. Buchbinder, L. March, M. Fransen, Osteoarthritis, Best Pract. Res. Clin. Rheumatol. (2010). https://doi.org/10.1016/j.berh.2010.11.001.

[5] A. Garratt, L. Schmidt, A. Mackintosh, R. Fitzpatrick, Quality of life measurement: Bibliographic study of patient assessed health outcome measures, Br. Med. J. (2002). https://doi.org/10.1136/bmj.324.7351.1417.

[6] R. Sharma, C. Vannabouathong, S. Bains, A. Marshall, S.J. MacDonald, J. Parvizi, M. Bhandari, Meta-analyses in joint arthroplasty: A review of quantity, quality, and impact, J. Bone Jt. Surg. - Ser. A. (2011). https://doi.org/10.2106/JBJS.J.01289.

[7] P. Deshpande, Bl. Sudeepthi, S. Rajan, C. Abdul Nazir, Patient-reported outcomes: A new era in clinical research, Perspect. Clin. Res. (2011). https://doi.org/10.4103/2229-3485.86879.

[8] M. Ulivi, L.C. Orlandini, V. Meroni, M.D.M. Lombardo, G.M. Peretti, Clinical Performance, Patient Reported Outcome, and Radiological Results of a Short, Tapered, Porous, Proximally Coated Cementless Femoral Stem: Results up to Seven Years of Follow-Up, J. Arthroplasty. (2018). https://doi.org/10.1016/j.arth.2017.11.046.

[9] E. Parliament, Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, (2017).

[10] D.J. Berry, M. Kessler, B.F. Morrey, Maintaining a hip registry for 25 years: Mayo clinic experience, in: Clin. Orthop. Relat. Res., 1997. https://doi.org/10.1097/00003086-199711000-00007.

[11] H. Malchau, G. Garellick, D. Berry, W.H. Harris, O. Robertson, J. Kärrlholm, D. Lewallen, C.R. Bragdon, L. Lidgren, P. Herberts, Arthroplasty implant registries over the past five decades: Development, current, and future impact, J. Orthop. Res. (2018). https://doi.org/10.1002/jor.24014.

[12] M. Huber, C. Kurz, R. Leidl, Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning, BMC Med.

Inform. Decis. Mak. (2019). https://doi.org/10.1186/s12911-018-0731-6.

[13] B. Zupan, J. Demšar, D. Smrke, K. Božikov, V. Stankovski, I. Bratko, J.R. Beck, Predicting patient's long-term clinical status after hip arthroplasty using hierarchical decision modelling and data mining, Methods Inf. Med. (2001). https://doi.org/10.1055/s-0038-1634460.

[14] A.H.S. Harris, A.C. Kuo, Y. Weng, A.W. Trickey, T. Bowe, N.J. Giori, 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). https://doi.org/10.1097/CORR.0000000000000601.

[15] T. Chen, T. He, xgboost : eXtreme Gradient Boosting, R Packag. Version 0.4-2. (2015).

[16] H. Zhang, The optimality of Naive Bayes, in: Proc. Seventeenth Int. Florida Artif. Intell. Res. Soc. Conf. FLAIRS 2004, 2004.

[17] R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, J. R. Stat. Soc. Ser. B. (1996). https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[18] W.H. HARRIS, Traumatic Arthritis of the Hip after Dislocation and Acetabular Fractures, J. Bone Jt. Surg. (1969). https://doi.org/10.2106/00004623-196951040-00012.

[19] N. Caplan, D.F. Kader, Rationale of the knee society clinical rating system, in: Class. Pap. Orthop., 2014. https://doi.org/10.1007/978-1-4471-5451-8_48.

[20] D.D. Price, P.A. McGrath, A. Rafii, B. Buckingham, The validation of visual analogue scales as ratio scale measures for chronic and experimental pain, Pain.

(1983). https://doi.org/10.1016/0304-3959(83)90126-4.

[21] B. Gandek, J.E. Ware, N.K. Aaronson, G. Apolone, J.B. Bjorner, J.E. Brazier, M. Bullinger, S. Kaasa, A. Leplege, L. Prieto, M. Sullivan, Cross-validation of item selection and scoring for the SF-12 Health Survey in nine countries: Results from the IQOLA Project, J. Clin. Epidemiol. (1998). https://doi.org/10.1016/S0895-4356(98)00109-7.

[22] M. Klässbo, E. Larsson, E. Mannevik, Hip disability and osteoarthritis outcome score: An extension of the Western Ontario and McMaster Universities Osteoarthritis Index, Scand. J. Rheumatol. (2003). https://doi.org/10.1080/03009740310000409.

[23] E.M. Roos, H.P. Roos, L.S. Lohmander, C. Ekdahl, B.D. Beynnon, Knee Injury and Osteoarthritis Outcome Score (KOOS) - Development of a self-administered outcome measure, J. Orthop. Sports Phys. Ther. (1998). https://doi.org/10.2519/jospt.1998.28.2.88.

[24] A.D.L. Patrick, G.H. Guyatt, C. Acquadro, Chapter 17 : Patient-reported outcomes, Heal. (San Fr. (2008).

[25] N. Balakrishnan, Progressive censoring methodology: an appraisal, Test. (2007). https://doi.org/10.1007/s11749-007-0061-y.

[26] A. Nilsdotter, A. Bremander, Measures of hip function and symptoms: Harris Hip Score (HHS), Hip Disability and Osteoarthritis Outcome Score (HOOS), Oxford Hip Score (OHS), Lequesne Index of Severity for Osteoarthritis of the Hip (LISOH), and American Academy of Orthopedic Surgeons (AAOS) Hip and Knee Questionnaire, Arthritis Care Res. (2011). https://doi.org/10.1002/acr.20549.

[27] A. Ghatak, Machine Learning with R, 2017. https://doi.org/10.1007/978-981-10-6808-9.

[28] J.F. Hair, R.E. Anderson, R.L. Tatham, W.C. Black, Multivariate data analysis, 5th ed., 1998. https://doi.org/10.14267/CJSSP.2016.02.04.

[29] S. Kaufman, S. Rosset, C. Perlich, Leakage in data mining: Formulation, detection, and avoidance, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2011. https://doi.org/10.1145/2020408.2020496.

[30] J. Davis, M. Goadrich, The relationship between precision-recall and ROC curves, in: ACM Int. Conf. Proceeding Ser., 2006. https://doi.org/10.1145/1143844.1143874.

[31] M. a. Hall, L. a. Smith, Practical feature subset selection for machine learning, Comput. Sci. (1998).

[32] F. Cabitza, L.G. Dui, G. Banfi, PROs in the wild: Assessing the validity of patient reported outcomes in an electronic registry, Comput. Methods Programs Biomed. (2019). https://doi.org/10.1016/j.cmpb.2019.01.009.