Development and evaluation of ensemble-based classification models for predicting unplanned hospital readmissions after hysterectomy

Isabella Eigner
Institute of Information Systems
Friedrich-Alexander-University Erlangen-Nuremberg
Nuremberg, Germany
Email: isabella.eigner@fau.de

Daniel Reischl
Institute of Information Systems
Friedrich-Alexander-University Erlangen-Nuremberg
Nuremberg, Germany
Email: daniel.reischl@fau.de

Freimut Bodendorf
Institute of Information Systems
Friedrich-Alexander-University Erlangen-Nuremberg
Nuremberg, Germany
Email: freimut.bodendorf@fau.de

Abstract

Unplanned hospital readmissions are a key indicator of quality in healthcare and can lead to high unnecessary costs for the hospital due to additional required resources or reduced payments by insurers or governments. Predictive analytics can support the identification of patients at high-risk for readmission early on to enable timely interventions. In Australia, hysterectomies present the 2nd highest observed readmission rates of all surgical procedures in public hospitals. Prior research so far only focuses on developing explanatory models to identify associated risk factors for past patients. In this study, we develop and compare 24 prediction models using state-of-the-art sampling and ensemble methods to counter common problems in readmission prediction, such as imbalanced data and poor performance of individual classifiers. The application and evaluation of these models are presented, resulting in an excellent predictive power with under- and oversampling and an additional slight increase in performance when combined with ensemble methods.

Keywords Predictive analytics, readmissions, hysterectomy, ensemble learning, sampling

1 Introduction

The Australian Institute for Health and Welfare (AIHW) tracks 28-day unplanned readmission rates for seven surgical procedure groups, i.e. hip replacements, knee replacements, tonsillectomy and adenoidectomy, cataract surgery, appendectomy, prostatectomy, and hysterectomy (AIHW, 2017b). The service rate in Australia (i.e., number of separations per 1,000 population) for hysterectomies (3.3) is only surpassed by cataract surgeries (9.3). These procedures, however, show very low readmission rates overall (0.3%). Besides tonsillectomy and adenoidectomy that show readmission rates of 3.4% on average, hysterectomy procedures have the 2nd highest rate of unplanned readmissions in Australia (3.3%) (AIHW, 2017a). Research has shown that hysterectomies are association with a high complication risk, however, the influencing risk factors are not fully known (Daugbjerg *et al.*, 2014). In addition, Australia has one of the highest frequencies of hysterectomy procedures as compared to other OECD countries (262.2 procedures per 100,000 females) (OECD, 2018). Thus, analyzing hysterectomies as one of the most frequent and risk-prone procedures for unplanned hospital readmissions using Australian healthcare data offers great potential for generating useful insights and furthermore reducing unnecessary costs.

According to a systematic review by Artetxe *et al.* (2018) on predictive models for hospital readmission risk, machine learning methods can improve the prediction ability over traditional statistical approaches. Such contributions to this academic field are aimed at first aligning complex and sensitive information across multiple sources, using, among others, administrative, insurance, clinical, and government registry data. This information is thereafter used to identify patients in need of additional healthcare resources by means of various intervention methods (Billings *et al.*, 2013). To identify patients at risk of readmission, predictive analytics has developed into a popular research area in medicine and healthcare management (Zhou *et al.*, 2016; Kansagara *et al.*, 2011). The task of readmission prediction presents multiple challenges that have to be dealt with during the data pre-processing and analysis. Since the population of readmitted cases is usually low with respect to non-readmissions (3.3 % for hysterectomy in Australia on average), the analyst has to deal with an imbalanced class distribution. Furthermore, planned and unplanned readmissions need to be clearly separated as to avoid noise in the training set. While no universal definition for unplanned readmissions is available, the AIHW characterises them as “readmissions where the principal diagnosis indicates an adverse event.” (AIHW, 2017b). For this study, a readmission is defined as a revisit to the hospital that is directly related to the index admission, takes place in acute care, and where the time span between the discharge date of the index admission and the admission date of the revisit does not exceed 28 days.

Although the importance of predictive analyses in Information Systems (IS) is apparent (Gregor, 2006; Shmueli and Koppius, 2011) and contributions of IS in healthcare have been numerously demonstrated in the past (Haried *et al.*, 2017), research on predictive analytics in healthcare is still scarce in IS literature (Bardhan *et al.*, 2015). Furthermore, while studies in the past rather focus on explanatory modeling and hypothesis testing, the importance and major differences of building powerful prediction models have recently become apparent (Shmueli and Koppius, 2011). In readmission prediction research in general, especially the initial conditions targeted in the Hospital Readmission Reduction Program (HRRP) (acute myocardial infarction, heart failure, and pneumonia) are addressed in a variety of studies that are, however, primarily conducted in the US (Weinreich *et al.*, 2016; Baechle *et al.*, 2017; Castillo *et al.*, 2017; Amarasingham *et al.*, 2010; Frizzell *et al.*, 2017; Au *et al.*, 2012). Readmissions in Australian hospitals, especially focusing on the AIHW procedure groups are a novel and promising research area that will increasingly affect the Australian healthcare system as unplanned readmissions are more and more focused by insurers and the government (Health Innovation & Reform Council, 2013). This study presents a novel approach to identify patients at risk for 28-day readmission after a hysterectomy utilizing different sampling and ensemble methods. A dataset of 3,466 hysterectomy episodes at a private not-for-profit Australian hospital group is used to build and evaluate 24 different prediction models. The results of a literature review on common risk factors from previous research on predictive models for hospital readmissions in general as well as classification methods and their predictive power that are typically used in readmission prediction are used as a base for this study. Furthermore, the review indicates that diagnosis-specific prediction models perform better than general risk prediction models ([Anonymous], 2018). Thus, decision trees (DT), support vector machines (SVM), and artificial neural networks (ANN) are combined with under- and oversampling as well as bagging and boosting techniques to evaluate a potential increase in predictive performance through the use of sampling and ensemble methods. In addition to the general risk factors identified in previous research on predicting readmissions, disease-specific risk factors are identified in studies presenting explanatory models on hospital readmissions after hysterectomy. The process to build empirical models presented by Shmueli and Koppius (2011) is used to guide the subsequent analysis. The modeling section of this paper is structured according to the process depicted in Table 1. Lastly, the implications and limitations of this study are presented.

Goal	*Build and compare prediction models to identify patients at risk for 28-day readmission after hysterectomy utilizing sampling and ensemble methods*
Data collection & study design	*Observational data / retrospective study*
Data preparation & EDA	*3,466 hysterectomy episodes; readmission rate 4.8 %*
Variables	*25 attributes*
Methods	*Decision tree / Artificial neural net / Support vector machine*
Evaluation	*Recall / AUC / F2-measure*

Table 1: Analysis process

2 Theoretical Background

2.1 Hospital Readmissions

The Australian government defines readmissions as "unplanned and unexpected hospital readmissions to the same public hospitals within 28-days for selected surgical procedures" (AIHW, 2017b). Although readmissions are a central theme in the Australian healthcare sector, criteria to specify whether an admission counts as a readmission vary among the different states or insurers. Rates are measured within a 28-day or 30-day time frame from the index admission. In Western Australia, an admission is labelled an unplanned readmission if the previous admission occurred within a time frame of 28 days and the patient is admitted for the same or a related condition or a complication following the index admission (Government of Western Australia, Department of Health, 2017). Since 2006, the Australian Institute of Health and Welfare (AIHW) has been tracking 28-day readmission rates (AIHW, 2017b). Monitoring of unplanned readmission rates across Australia is executed through the instalment of the National Healthcare Agreement (NHA) which contains unplanned readmission rates as a quality of care indicator. The calculation for the report, however, is limited to public hospitals. Here, readmissions are defined by the following criteria that have to be fulfilled to qualify for the inclusion in the statistic (AIHW, 2017b):

- The admission has to follow a separation from the same hospital where the patient was either treated with a knee replacement (TKA), hip replacement (THA), tonsillectomy and adenoidectomy (T&A), hysterectomy (HRT), prostatectomy (PRO), cataract surgery (CAT) or appendectomy (APP).
- The second admission has to occur within 28 days of the previous separation.
- A principal diagnosis has to have one of the following codes: T80/88, T98.3, E89, G97, H59, H95, I97, J95, K91, M96 or N99. These diagnosis codes include complications, sequelae of complications, and post-procedural disorders.

For these procedures, unplanned readmission rates of 2.09% (APP), 0.32% (CAT), 1.92% (THA), 2.31% (TKA), 3.34% (hysterectomy), 2.65% (PRO), and 3.47% (T&A) can be observed in Australian hospitals on average (AIHW 2017a, p. 225). Excluded from penalisation are planned readmissions as well as obstetrical delivery, transplant surgery, maintenance, chemotherapy, rehabilitation, and non-acute readmissions for a scheduled procedure. If a hospital exceeds the readmission rate benchmark, which is calculated on the risk-adjusted national average for the conditions above, funding rates are reduced. The corresponding expected risk adjustment, to account for region specific populations, is calculated by taking several factors, such as the hospital specific distribution of patient's age, gender and previous conditions as well as clinical risk factors using data from the preceding three financial years, into account. In summary, the AIHW definition of readmissions explicitly excludes planned admissions, considers only a specific list of surgeries, and focuses on a 28-day period.

2.2 Imbalanced Data

Imbalanced data, also known as skewed data, has a strong unequal distribution of the minority and majority classes (SUN *et al.*, 2009). In the case of hospital readmissions, the minority class is represented by unplanned readmissions. The main issue with handling imbalanced data is that traditional classifiers tend to perform best with an equal class distribution while the relevant information from the minority class might be overlooked with regards to the majority class (SUN *et al.*, 2009). There are a number of different approaches to handle imbalanced data (Galar *et al.*, 2012; SUN *et al.*, 2009; Kotsiantis, 2007; Longadge and Dongre, 2013; Chawla, 2005; He and Garcia, 2009) that can either be grouped in **algorithm approaches**, **data level approaches**, or a **combination of both**. Algorithm approaches manipulate the classifier to give a higher attention to the minority class. Since adaptations of the classifier need expert knowledge of the algorithm and the data domain (SUN *et al.*, 2009), this approach is not suitable for this study. Data level approaches, which are also known as external approaches (Haixiang *et al.*, 2017), change the data dimensions and can be further distinguished into **feature selection** and **resampling** (Haixiang *et al.*, 2017; Kotsiantis, 2007). The feature selection performed in this study is mainly based on risk factors derived from the literature. Resampling methods manipulate the number of entities to reduce the skew of the data. Resampling can be divided into *undersampling* and *oversampling*, where undersampling reduces the entities from the majority class, while oversampling creates additional entities of the minority class (Kotsiantis, 2007; Galar *et al.*, 2012). From the variety of over- and undersampling methods presented in literature (Galar *et al.*, 2012; Haixiang *et al.*, 2017) this study focuses on the most prominent techniques, namely *random undersampling* and the *synthetic minority oversample technique (SMOTE).* Random undersampling, which is one of the most commonly applied undersampling techniques (Haixiang *et al.*, 2017), is the process of randomly removing entities of the majority class to reduce the data imbalance (Galar *et al.*, 2012). The most commonly used oversampling technique is SMOTE (Chawla *et al.*, 2011) and its derivations (Haixiang *et al.*, 2017).

2.3 Ensemble Learning

Hybrid methods of predicting imbalanced data include **cost-sensitive learning** and **ensemble learning**. Cost-sensitive learning follows the approach of manipulating the algorithm to weight the minority class higher and improve the classifier performance. Cost-sensitive approaches have the downside that the actual costs of misclassification must be known (SUN *et al.*, 2009). In this study, the costs of misclassification—namely, the costs of unplanned readmissions—are not known. Ensemble learning is performed when the results of several classifiers are combined to predict a future observation (Galar *et al.*, 2012). Ensemble learning can either be performed by combining different classifiers or by applying variations of the same classifier (Haixiang *et al.*, 2017). This study aims to benchmark the performance of traditional classifiers to single classifier ensembles. Single classifier ensembles are grouped into parallel ensembles ("*bagging*") and iterative ensembles ("*boosting*"). Parallel ensembles train different base classifiers simultaneously, while iterative approaches train one base classifier after another (Haixia ng *et al.*, 2017). **Bagging**, which is short for "bootstrapped aggregating," is introduced by Breiman (1996) and combines several base classifiers into one classifier by bootstrapping the data into several different bags. Then, for each of the bags, the base classifier is trained and applied to the application set. Subsequently, the differently trained classifiers vote as to which class a new entity belongs, and a majority vote of the classifiers determines in which class the observation fits best. The most prominent **boosting** method, **AdaBoost** ("adaptive boosting") (Freund and Schapire, 1997) is based on the principle of boosting introduced by Schapire (1990) and uses the base principle of improving the algorithm in every iteration to achieve a higher performance. Single classifier ensembles for imbalanced data combine either resampling methods or cost-sensitive approaches with the traditional ensemble methods, bagging or boosting (Galar *et al.*, 2012). Combining resampling with traditional ensembles resamples the data in bagging approaches after bootstrapping, while AdaBoost-based ensembles resample the data at each iteration before training the base classifier. This study focuses on random undersampling and SMOTE sampling. If undersampling is combined with bagging, the literature uses the term "underbagging" (Galar *et al.*, 2012), while the combination of SMOTE sampling and bagging is referred to as "BaggingSMOTE" (Błaszczyński and Stefanowski, 2015). The combination of AdaBoost with random undersampling is called "RUSBoost" in the literature (Seiffert *et al.*, 2008), and the combination of AdaBoost with SMOTE as SMOTEBoost (Lavračetal.,ÈChawlaetal.,È(Chawla *et al.*, 2003).

3 Data Analysis

3.1 Goal Definition

The goal of this study is to develop a prediction model that identifies patients at discharge with substantial risks of unplanned readmission to the hospital after hysterectomy. Since the data is imbalanced and this study aims to predict as many potential risk patients as possible, the receiver operating curve (ROC) in combination with the area under the curve (AUC) score is used. To identify whether a model proves sufficient, a target AUC score is defined. The AUC score of 0.50 can be described as random guessing, while an AUC score of 1.0 means that the prediction is 100% correct (Fawcett, 2006). Models with good discrimination powers are models with AUC scores of at least 0.80 (Kleinbaum and Klein, 2010, p. 357). Thus, this study aims to achieve an AUC score above 0.80. This is also aligned with current prediction models, which have AUC scores between 0.53 and 0.83 (Zhou *et al.*, 2016).

3.2 Data Collection and Study Design

This study uses observational data from a large not-for-profit private health-care group in Victoria containing clinical, demographic, and financial information from anonymized patient episodes between the years 2011 to 2015. In total, the dataset comprises 642,407 patient episodes, where each entity describes a single episode at the hospital. For each episode, 483 different attributes are provided, which can be categorized into social demographic (e.g., age and nationality), financial (e.g., hospital charges), medical (e.g., diagnoses and procedures), and hospital-related (e.g., admission ward, length of stay) factors. The diagnosis data is coded in an Australian modification of the 10th version of the International Classification for Diseases (ICD-10-AM). The World Health Organization introduced ICD to name global health trends and statistics and is the international standard for reporting diseases and health conditions (World Health Organization, 2018). The procedure data is coded in the Australian Classification of Health Interventions (ACHI). To extract all relevant episodes for the procedure groups under study, the respective ACHI codes provided by the AIHW are used. Patients who died before or after discharge (n = 10) from the hospital as well as patients who were discharged at their own risk (n = 4) and patients who were admitted after December 3, 2015 (n = 75) are excluded from this dataset. Since unplanned readmissions are not directly flagged in the data, episodes that led to a readmission are marked as such, if the difference between the admission date of the revisit and discharge date of the episode is within the range of 0 to 28 days. Planned readmissions are excluded in this step. The final dataset counts a total of 3,466 hysterectomies. 166 episodes (4.8 %) lead to a 28-day readmission to the same hospital group.

In addition to general risk factors derived from studies on predicting all-cause patient readmissions, studies providing explanatory models for procedure-specific readmissions are analysed to identify specific risk factors for hysterectomies. A summary of all potential risk factors according to the respective sources is illustrated in Table 2.

Study	Data / Patients	Time-frame	Country	Risk factors	Proc-edures	Readmission rate
Kreuninger *et al.*, 2018	3,981 (benign)	60-day	USA	Type of procedure, perioperative complications	L, A, V, R	1.9 – 3.5 %
Lonky *et al.*, 2017	3,106 (benign)	90-day	USA	Race, type of procedure, blood loss, operative complications, length of stay	L, A, V	3.5 %
Philp *et al.*, 2017	119 (cervix cancer)	30-day	Canada	Age, operation time, blood loss, intraoperative complications	L	5%
Penn *et al.*, 2016	40,580 (benign)	30-day	USA (NSQIP)	Medical and surgical complication	L, A, V	2.8%
Lee *et al.*, 2016	1.649 (benign + malignant)	30-day	USA	Complications, prior abdominal surgeries, malignancy, length of stay, blood loss	L, A, V	6%
Fitch *et al.*, 2016	21,926	30-day	CCED	Type of procedure (inpatient vs. outpatient)	L, A, V	2.1 – 3.05%

Jennings *et al.*, 2015	8,890 (benign)	30-day	USA (NSQIP)	Comorbidities, substance abuse, operation time, complications	L	3.1%
Dessources *et al.*, 2015	41,196 (benign + malignant)	30-day	USA (NSQIP)	Complications, comorbidities, age	L, A, V	3.4 – 6.1%
Catanzarite *et al.*, 2015	21,228 (benign)	30-day	USA (NSQIP)	Complications, return to theatre, age, comorbidities, smoking, prior surgeries, operation time, status (inpatient vs. outpatient), length of stay	L, A, V	2.7 – 3.8%
Daugbjerg *et al.*, 2014	22,150 (benign)	30-day	Denmark	Socioeconomic status	L, A, V	6%
Liang *et al.*, 2013	395 (endometrial cancer)	90-day	USA	Length of stay, postoperative complications	R	7.6%
Summitt *et al.*, 1994	113	2 weeks	USA	Medication, complication, blood loss	L, V	3.8 %.

Table 2: Identified studies on readmission risk factors after a hysterectomy

3.3 Data Preparation

The next step of developing a prediction model is to prepare the data, which mainly includes data cleaning, handling missing values, and splitting the data (Shmueli and Koppius, 2011). The term "data cleaning" describes the process of detecting and removing data errors and inconsistencies. A way to identify unclean data is to perform data profiling, which uses metadata to discover errors in the data. Errors discovered can be illegal values, misspellings, missing values, varying value representations, and duplicates (Rahm and Do, 2000). Next, the data is split into training and validation sets. The training set is a part of the data that is used to train the prediction model, and the validation set is used to determine the performance of the model. For this study, a cross-validation approach with 10-fold validation is used.

3.4 Exploratory Data Analysis

The goal of the exploratory data analysis is to analyze the dataset visually and numerically to ensure that the data is suitable for the prediction model as well as reducing the dimensions (Shmueli and Koppius, 2011). Because too many predictors can decrease the performance of a prediction model, the dimensions are systematically reduced in this step. Especially procedure and diagnosis data are reduced to relevant factors as well as redundancies within lengths of stay and admission dates are resolved. The numerical distribution gives an insight into how the two cohorts differ. While patients who were not readmitted visited the hospital 0.22 times on average, readmitted patients visited the hospital an average of 1.94 times within the last three months. Additionally, the procedure "radical abdominal hysterectomy with radical excision of pelvic lymph nodes" (10% /6%) and "total abdominal hysterectomy with removal of adnexa" (26%/ 23%) show a higher readmission rate within the study group. Additionally, patients being readmitted stay longer in the hospital (6.5/6.07) and longer in the operating room (160/136) and have a higher number of diagnoses (2.43/2.13), complications (1.49/0.95) and procedures (5/4.32). Furthermore, readmitted patients more often have low haemoglobin levels (25 %/14%), have less histology performed (34%/40%) and are released from the hospital later in the day (57%/48%).

3.5 Choice of Variables

After reducing dimensions, the next step is to select which variables to use for the prediction model. Two factors are especially relevant for the variable selection. At first, the variables must have a measurement quality, which means variables that do not assist in predicting unplanned readmissions are not relevant for the model. The second factor is the ex-ante availability. This signifies that the predictor must be accessible at point of prediction. In this study, the parameter must be available at time of discharge (Shmueli and Koppius, 2011). A feature is seen as beneficial if it is correlated with the prediction flag but is not redundant to any other relevant feature (Yu and Liu, 2003). This means that the variables must have the ability to predict readmissions while they not being highly correlated with each other. Since correlations above 0.70 are seen as highly correlated (Asuero *et al.*, 2007), features with a correlation above 0.70 are removed. Additionally, in this step, variables that only include low information or no information are also excluded for predicting unplanned readmissions. Thus, columns only containing one constant value are removed as well as columns with a variance below 0.05. In regard to ex-ante availability, three additional features containing post-discharge information are removed since these factors are not accessible ex-ante. The resulting risk factors can be grouped into hospital-related and hospital-visit-related factors, socio-demographics, the patient history, as well as laboratory data. In summary, this leads to 25 different factors accessible at discharge.

Attribute	Type	Description (dc = distinct count)
Patient id	Categorical	Patient unique identifier (dc = 3,462)
Episode id	Categorical	Unique episode identifier (dc = 3,446)
Led_to_readmission	Boolean	Label attribute to be predicted
Admission year	Categorical	dc = 5 (2011 - 2015)
Admission ward	Categorical	dc = 31
Admission patient classification	Categorical	dc = 20; type of patient (e.g., surgical, medical)
Age	Numeric	mean = 55.35; range = 21 - 98
Blood usage	Boolean	Yes (13.3%) No (86.7%)
Campus	Categorical	dc = 6
Cancer	Boolean	Yes (27.44%) No (72.56%)
Complications	Numeric	mean = 0.98; range = 0 - 9
Diagnosis count	Numeric	mean = 3.50; range = 0 - 10
Discharge ward	Categorical	dc = 34
Discharge patient classification	Categorical	dc = 20; type of patient (e.g., surgical, medical)
Haemoglobin low	Boolean	Yes (14.83%) No (85.17%)
Histology	Boolean	Yes (39.47%) No (60.53%)
Hysterectomy procedure	Categorical	dc = 14;
Length of stay	Numeric	mean = 5.10; range = 0 - 69
Metastatic cancer	Boolean	Yes (7.59%) No (92.41%)
Procedure count	Numeric	mean = 4.35; range = 1 – 10
Separation after 10am	Boolean	Yes (48.21%) No (51.79%)
Total number of beds	Numeric	mean = 2.32; range = 0 – 11
Total number of wards	Numeric	mean = 1.03; range = 0 – 5
Total time in theatre	Numeric	mean = 137.48; range = 0 – 771
Visits past 6 months	Numeric	mean = 0.49; range = 0 – 27

Table 3: Final feature set

3.6 Choice of Potential Methods

Next, it must be determined which algorithms are used. According to Shmueli and Koppius (2011) either data-driven, shrinkage, or ensemble methods can be used. Based on the literature review, combinations of resampling with bagging and boosting show different strengths, yet there is no way to determine which of the algorithms are most suitable for predicting unplanned readmissions. Underbagging is used in the only study incorporating the problem of imbalanced data to predict unplanned hospital readmissions, while RUSBoost is similar to SMOTEBoost, showing performance improvement in many cases. Yet the most comprehensive study of Galar *et al.* (2012) highlights that a combination of SMOTE and bagging is the best bagging approach. This is why this study applies SMOTEBoost, BaggingSMOTE, underbagging, and RUSBoost to the data. Since all of these ensembles are combinations of either bagging or boosting with random undersampling or SMOTE-sampling, these components are also tested in combination with the base classifier. The base classifiers applied are the most frequently used classifiers in combination with ensemble learning— namely, DT, SVM, and ANN.
This leads to 24 different combinations being benchmarked within this study. First, the base classifiers DT, SVM, and ANN are applied; next, the traditional ensemble methods AdaBoost and bagging are applied in combination with them. Next, the base classifiers are benchmarked in combination with random undersampling, underbagging, and RUSBoost. Finally, models incorporating oversampling are applied (namely, SMOTE-sampling, BaggingSMOTE, and SMOTEBoost). RUS, SMOTE-sampling, and underbagging are based on the library ImbalancedLearn, which is an extension of Scikit-Learn

(Lemaitre *et al.*, 2016). For RUSBoost and SMOTEBoost, the algorithms developed by Johnson (2017) are utilized, which are based on the Scikit-Learn implementation of AdaBoost. Because the ANN implementation of Scikit-Learn does not support class weighting (which is necessary to apply boosting approaches to the data) ANN cannot be applied in combination with AdaBoost, RUSBoost, and SMOTEBoost. This leads to a benchmark of 24 different algorithms to predict unplanned hospital readmissions.

3.7 Evaluation, Validation, and Model Selection

The next step is to validate and select the best model. To evaluate the models, accuracy is usually used as a performance measure. For validation, Shmueli and Koppius (2011) state that the performance can be measured by applying the model to a holdout set or by using a cross-validation approach. This study uses a tenfold cross validation. Finally, as part of the model selection, different predictors should be assed to improve the model performance. To compare the different results, the data is finally prepared to fit the need of the base classifiers. Thus, the performances of the 24 models are assessed and evaluated. For all models, a random seed is used to make the results reproducible. The seed is set to 11 and ensures that the retraining of a model with the same parameters shows identical results. For each classifier, a grid search is performed to attain the best parameter settings. For the DT, a maximum depth of 8, the gini index as the split criterion and the default setting for selecting the best split is chosen. The SVM is implemented using the support vector classification (SVC) algorithm with a radial basis function (RBF) kernel, setting the penalty parameter C to 32768 and the gamma to 0.000488. The ANN implementation utilizes the Multi-layer Perceptron (MLP) classifier with 8 neurons in the hidden layer, the lbfgs solver for weight optimization, the rectified linear unit (relu) function as the activation function, and an alpha of 0.01. For evaluation, we include the area under the curve (AUC) score, which presents a popular measure in healthcare analyses. In addition, we investigate the recall, which represents the ratio of all correctly predicted readmissions (predicted positives) to the true readmissions (true positives). To be able to easily track the cost of our prediction models with regards to the false positive rate, the F-score is included in the evaluation that considers both precision and recall. For this measure, the β is set to 2, to allow a higher weight for the recall (Sattar *et al.*, 2006).

Overall, underbagged decision trees show the most promising results for identifying unplanned readmissions with an AUC of 0.94, detecting almost all positive cases of the data set. On the other hand, bagging without prior resampling leads to an unsatisfactory predictive power across all three classifiers. Here, the issue of imbalanced data becomes apparent, where the prediction models tend to sort all examples to the majority class. Surprisingly, a standard SVM with no prior sampling, bagging, or boosting approach leads to a high AUC score of 0.87, whereas sampling and ensemble methods rather worsen the results. Table 4 summarizes the results for each prediction model, where models with an AUC > 0.9, AUC > 0.75 and AUC < 0.6 are highlighted in green, orange and red respectively. The best and worst models are highlighted in bold in addition.

	Decision Tree			**SVM**			**Neural Net**		
	AUC	**Recall**	**F2**	**AUC**	**Recall**	**F2**	**AUC**	**Recall**	**F2**
Standard	0.73	0.49	0.48	**0.87**	**0.83**	**0.62**	0.55	0.10	0.12
RUS	0.93	0.97	0.70	0.78	0.77	0.44	0.70	0.63	0.34
SMOTE	0.91	0.92	0.69	0.84	0.78	0.59	0.71	0.63	0.36
Bagging	**0.68**	**0.38**	**0.40**	0.51	0.02	0.02	**0.5**	**0**	**0**
Underbagged	**0.94**	**0.99**	**0.71**	0.81	0.72	0.53	0.73	0.67	0.39
Overbagged	0.93	0.95	0.71	0.62	0.28	0.27	**0.79**	**0.65**	**0.54**
AdaBoost	0.69	0.40	0.41	**0.5**	**0**	**0**	**0.5**	**0**	**0**
RUS Boost	0.93	0.96	0.70	**0.5**	**0**	**0**	**0.5**	**0**	**0**
Smote Boost	0.68	0.40	0.40	**0.5**	**0**	**0**	**0.5**	**0**	**0**

Table 4: Model evaluation

4 Conclusion

We conclude that the task of identifying patients at risk of readmission is highly complex and risk factors depend heavily on the presented context. Furthermore, the issue of imbalanced data and the poor performance of individual classifiers need to be considered in readmission prediction. Due to these restrictions, we present and compare prediction models to determine readmissions after a hysterectomy procedure utilizing both sampling and ensemble methods. To this end, individual classifiers with no prior sampling, individual classifiers using under- and oversampled data, as well as bagged and boosted classifiers with and without prior sampling are built and evaluated. This way, the suitability of sampling and ensemble methods for the task at hand is analyzed by investigating a potential increase in predictive performance. Another advantage of our approach is the inclusion of both empirical evidence from past studies to construct relevant attributes as well as the investigation of all variables already collected onsite. This way, the resulting feature set is developed rigorously while keeping an open mind about further relevant risk factors not yet considered in the past. Overall, 24 prediction models are evaluated performing with AUC scores ranging from 0.5 to 0.94. Decision trees show the overall best performance considering all evaluation metrics. Support vector machines still yield satisfactory AUC rates of maximum 0.87 while neural nets perform considerably poorly with a maximum AUC score of 0.79. By including the F2-measure into the evaluation metrics, the cost of correct predictions can also be implicitly observed. Overall, combining RUS and bagging with decision trees are recommended for this context. The results indicate that the influence of sampling is higher than performing bagging or boosting, as the ensemble methods only slightly increase the predictive performance after sampling.

The implications of this study are evident for research and practice. With unplanned hospital readmissions as a key indicator of healthcare quality and associated unnecessary costs, the early identification of high-risk patients can support timely interventions to reduce avoidable readmission. This offers the opportunity for cost-reduction and an increased quality of healthcare services for hospitals and practitioners. On the other hand, this study gives an indication of the suitability of implementing ensemble methods in research and practice. As the number of studies investigating Australian healthcare data in general and readmission prediction after hysterectomies specifically is still limited, this study addresses an important research gap that can motivate further research in this area. Nevertheless, identifying patients at risk of readmission is a continuously challenging task due to the individual factors that influence patient care outcome in different healthcare settings.

Lastly, some limitations of this study also need to be mentioned. The data used to develop the predictive models in this paper are extracted retrospectively from a single private hospital group comprising multiple campuses in Australia. Furthermore, due to a lot of missing values, many features were excluded and thus didn't provide additional substantial findings on relevant risk factors. As a result, potentially important features (e.g., BMI) also could not be harnessed. Finally, domain experts (e.g., gynaecologists) should be consulted for a qualitative evaluation and interpretation of the results in future studies. This way, the suitability of the final feature set can be further strengthened or expanded.

5 References

AIHW (2017a), "Admitted patient care 2015–16: Australian hospital statistics", available at: http://www.aihw.gov.au/WorkArea/DownloadAsset.aspx?id=60129559534.

AIHW (2017b), "National Healthcare Agreement: PI 23–Unplanned hospital readmission rates, 2017", available at: http://meteor.aihw.gov.au/content/index.phtml/itemId/630049.

Amarasingham, R., Moore, B.J., Tabak, Y.P., Drazner, M.H., Clark, C.A., Zhang, S., Reed, W.G., Swanson, T.S., Ma, Y. and Halm, E.A. (2010), "An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data", *Medical care*, Vol. 48 No. 11, pp. 981–988.

[Anonymous, 2018] Details omitted for double-blind reviewing.

Artetxe, A., Beristain, A. and Graña, M. (2018), "Predictive models for hospital readmission risk. A systematic review of methods", *Computer Methods and Programs in Biomedicine*, Vol. 164, pp. 49–64.

Asuero, A.G., Sayago, A. and González, A.G. (2007), "The Correlation Coefficient. An Overview", *Critical Reviews in Analytical Chemistry*, Vol. 36 No. 1, pp. 41–59.

Au, A.G., McAlister, F.A., Bakal, J.A., Ezekowitz, J., Kaul, P. and van Walraven, C. (2012), "Predicting the risk of unplanned readmission or death within 30 days of discharge after a heart failure hospitalization", *American heart journal*, Vol. 164 No. 3, pp. 365–372.

Baechle, C., Agarwal, A., Behara, R. and Zhu, X. (2017), "A cost sensitive approach to predicting 30-day hospital readmission in COPD patients", in Fotiadis, D.I. (Ed.), *4th IEEE EMBS International Conference on Biomedical and Health Informatics, Orland, FL, USA*, IEEE, Piscataway, NJ, pp. 317–320.

Bardhan, I., Oh, J.-h., Zheng, Z. and Kirksey, K. (2015), "Predictive Analytics for Readmission of Patients with Congestive Heart Failure", *Information Systems Research*, Vol. 26 No. 1, pp. 19–39.

Billings, J., Georghiou, T., Blunt, I. and Bardsley, M. (2013), “Choosing a model to predict hospital admission: an observational study of new variants of predictive models for case finding”, *BMJ Open*, Vol. 3 No. 8, e003352.
Błaszczyński, J. and Stefanowski, J. (2015), “Neighbourhood sampling in bagging for imbalanced data”, *Neurocomputing*, Vol. 150, pp. 529–542.
Breiman, L. (1996), *Machine Learning*, Vol. 24 No. 2, pp. 123–140.
Castillo, A., Edriss, H., Selvan, K. and Nugent, K. (2017), “Characteristics of Patients With Congestive Heart Failure or Chronic Obstructive Pulmonary Disease Readmissions Within 30 Days Following an Acute Exacerbation”, *Quality management in health care*, Vol. 26 No. 3, pp. 152–159.
Catanzarite, T., Vieira, B., Qin, C. and Milad, M.P. (2015), “Risk Factors for Unscheduled 30-day Readmission after Benign Hysterectomy”, *Southern medical journal*, Vol. 108 No. 9, pp. 524–530.
Chawla, N.V. (2005), “Data Mining for Imbalanced Datasets. An Overview”, in Maimon, O.Z. and Rokach, L. (Eds.), *Data mining and knowledge discovery handbook*, Springer, New York, NY, pp. 853–867.
Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2011), “SMOTE: Synthetic Minority Over-sampling Technique”, *Journal Of Artificial Intelligence Research*.
Chawla, N.V., Lazarevic, A., Hall, L.O. and Bowyer, K.W. (2003), “SMOTEBoost. Improving Prediction of the Minority Class in Boosting”, in Lavrač, N., Gamberger, D., Todorovski, L. and Blockeel, H. (Eds.), *Knowledge Discovery in Databases: PKDD 2003: 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings*, *Lecture Notes in Computer Science*, Vol. 2838, Springer, Berlin, Heidelberg, pp. 107–119.
Daugbjerg, S.B., Cesaroni, G., Ottesen, B., Diderichsen, F. and Osler, M. (2014), “Effect of socioeconomic position on patient outcome after hysterectomy”, *Acta obstetricia et gynecologica Scandinavica*, Vol. 93 No. 9, pp. 926–934.
Dessources, K., Hou, J.Y., Tergas, A.I., Burke, W.M., Ananth, C.V., Prendergast, E., Chen, L., Neugut, A.I., Hershman, D.L. and Wright, J.D. (2015), “Factors associated with 30-day hospital readmission after hysterectomy”, *Obstetrics and gynecology*, Vol. 125 No. 2, pp. 461–470.
Fawcett, T. (2006), “An introduction to ROC analysis”, *Pattern Recognition Letters*, Vol. 27 No. 8, pp. 861–874.
Fitch, K., Huh, W. and Bochner, A. (2016), “Open vs. Minimally Invasive Hysterectomy. Commercially Insured Costs and Readmissions”, *Managed care (Langhorne, Pa.)*, Vol. 25 No. 8, pp. 40–47.
Freund, Y. and Schapire, R.E. (1997), “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting”, *Journal of Computer and System Sciences*, Vol. 55 No. 1, pp. 119–139.
Frizzell, J.D., Liang, L., Schulte, P.J., Yancy, C.W., Heidenreich, P.A., Hernandez, A.F., Bhatt, D.L., Fonarow, G.C. and Laskey, W.K. (2017), “Prediction of 30-Day All-Cause Readmissions in Patients Hospitalized for Heart Failure: Comparison of Machine Learning and Other Statistical Approaches”, *JAMA Cardiology*, Vol. 2 No. 2, pp. 204–209.
Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. and Herrera, F. (2012), “A Review on Ensembles for the Class Imbalance Problem. Bagging-, Boosting-, and Hybrid-Based Approaches”, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 42 No. 4, pp. 463–484.
Government of Western Australia, Department of Health (2017), “Admission, Readmission, Admission, Readmission, Discharge and Transfer Reference Manual”, available at: http://www.health.wa.gov.au/circularsnew/attachments/1242.pdf (accessed 16 August 2017).
Gregor, S. (2006), “The Nature of Theory in Information Systems”, *MIS Q*, Vol. 30 No. 3, pp. 611–642.
Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H. and Bing, G. (2017), “Learning from class-imbalanced data. Review of methods and applications”, *Expert Systems with Applications*, Vol. 73, pp. 220–239.
Haried, P., Claybaugh, C. and Dai, H. (2017), “Evaluation of health information systems research in information systems research: A meta-analysis”, *Health informatics journal*, 1460458217704259.
He, H. and Garcia, E.A. (2009), “Learning from Imbalanced Data”, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21 No. 9, pp. 1263–1284.
Health Innovation & Reform Council (2013), *Hospital Readmission Findings*.
Jennings, A.J., Spencer, R.J., Medlin, E., Rice, L.W. and Uppal, S. (2015), “Predictors of 30-day readmission and impact of same-day discharge in laparoscopic hysterectomy”, *American journal of obstetrics and gynecology*, Vol. 213 No. 3, 344.e1-7.
Johnson, R. (2017), *ND DIAL: Imbalanced Algorithms*.
Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M. and Kripalani, S. (2011), “Risk prediction models for hospital readmission: a systematic review”, *JAMA*, Vol. 306 No. 15, pp. 1688–1698.

Kleinbaum, D.G. and Klein, M. (2010), “Assessing Discriminatory Performance of a Binary Logistic Model: ROC Curves”, in Kleinbaum, D.G., Klein, M. and Pryor, E.R. (Eds.), *Logistic Regression: A Self-Learning Text*, *Statistics for Biology and Health*, 3. ed., Springer Science+Business Media LLC, New York, NY, pp. 345–387.

Kotsiantis, S.B. (2007), “Supervised Machine Learning: A Review of Classification Techniques”, *Informatica*, Vol. 31 No. 3.

Kreuninger, J.A., Cohen, S.L., Meurs, E.A.I.M., Cox, M., Vitonis, A., Jansen, F.W. and Einarsson, J.I. (2018), “Trends in readmission rate by route of hysterectomy - a single-center experience”, *Acta obstetricia et gynecologica Scandinavica*, Vol. 97 No. 3, pp. 285–293.

Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H., Chawla, N.V., Lazarevic, A., Hall, L.O. and Bowyer, K.W. (Eds.) (2003), *SMOTEBoost: Improving Prediction of the Minority Class in Boosting: Knowledge Discovery in Databases: PKDD 2003*, Springer Berlin Heidelberg.

Lee, M.S., Venkatesh, K.K., Growdon, W.B., Ecker, J.L. and York-Best, C.M. (2016), “Predictors of 30-day readmission following hysterectomy for benign and malignant indications at a tertiary care academic medical center”, *American journal of obstetrics and gynecology*, Vol. 214 No. 5, 607.e1-607.e12.

Lemaitre, G., Nogueira, F. and Aridas, C.K. (2016), *Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning*, available at: http://arxiv.org/pdf/1609.06570.

Liang, M.I., Rosen, M.A., Rath, K.S., Clements, A.E., Backes, F.J., Eisenhauer, E.L., Salani, R., O'Malley, D.M., Fowler, J.M. and Cohn, D.E. (2013), “Reducing readmissions after robotic surgical management of endometrial cancer. A potential for improved quality care”, *Gynecologic oncology*, Vol. 131 No. 3, pp. 508–511.

Longadge, R. and Dongre, S. (2013), *Class Imbalance Problem in Data Mining Review*, available at: http://arxiv.org/pdf/1305.1707.

Lonky, N.M., Mohan, Y., Chiu, V.Y., Park, J., Kivnick, S., Hong, C. and Hudson, S.M. (2017), “Hysterectomy for benign conditions. Complications relative to surgical approach and other variables that lead to post-operative readmission within 90 days of surgery”, *Women's health (London, England)*, Vol. 13 No. 2, pp. 17–26.

OECD (2018), *Health care utilisation*, OECD Publishing.

Penn, C.A., Morgan, D.M., Rice, L.W., Harris, J.A., Rauh-Hain, J.A. and Uppal, S. (2016), “Timing of and Reasons for Unplanned 30-Day Readmission After Hysterectomy for Benign Disease”, *Obstetrics and gynecology*, Vol. 128 No. 4, pp. 889–897.

Philp, L., Covens, A., Vicus, D., Kupets, R., Pulman, K. and Gien, L.T. (2017), “Feasibility and safety of same-day discharge after laparoscopic radical hysterectomy for cervix cancer”, *Gynecologic oncology*, Vol. 147 No. 3, pp. 572–576.

Rahm, E. and Do, H.H. (2000), “Data Cleaning: Problems and Current Approaches”, in *Zinc Industry*, Woodhead Pub, S.l., Introduction/page i-Introduction/page ii.

Sattar, A., Kang, B.-h., Sokolova, M., Japkowicz, N. and Szpakowicz, S. (Eds.) (2006), *Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation: AI 2006: Advances in Artificial Intelligence*, Springer Berlin Heidelberg.

Schapire, R.E. (1990), “The strength of weak learnability”, *Machine Learning*, Vol. 5 No. 2, pp. 197–227.

Seiffert, C., Khoshgoftaar, T.M., van Hulse, J. and Napolitano, A. (2008), “RUSBoost. Improving classification performance when training data is skewed”, in *19th International Conference on Pattern Recognition, 2008: ICPR 2008 ; 8 - 11 Dec. 2008, Tampa, Florida, USA, Tampa, FL, USA, 8/12/2008 - 11/12/2008*, IEEE, Piscataway, NJ, pp. 1–4.

Shmueli, G. and Koppius, O. (2011), “Predictive Analytics in Information Systems Research”, *MIS Q*, Vol. 35 No. 3, pp. 553–572.

Summitt, R.L., Stovall, T.G., Lipscomb, G.H., Washburn, S.A. and Ling, F.W. (1994), “Outpatient hysterectomy. Determinants of discharge and rehospitalization in 133 patients”, *American journal of obstetrics and gynecology*, Vol. 171 No. 6, 1480-4; discussion 1484-7.

SUN, Y., Wong, A.K. and Kamel, M. (2009), “Classification of imbalanced data. A review”, *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 23 No. 04, pp. 687–719.

Weinreich, M., Nguyen, O.K., Wang, D., Mayo, H., Mortensen, E.M., Halm, E.A. and Makam, A.N. (2016), “Predicting the Risk of Readmission in Pneumonia. A Systematic Review of Model Performance”, *Annals of the American Thoracic Society*, Vol. 13 No. 9, pp. 1607–1614.

Yu, L. and Liu, H. (2003), “Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution”, paper presented at Proceedings, Twentieth International Conference on Machine Learning.

Zhou, H., Della, P.R., Roberts, P., Goh, L. and Dhaliwal, S.S. (2016), "Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review", *BMJ Open*, Vol. 6 No. 6, e011060.

Copyright