

*Article*

# Predicting Acute Kidney Injury: A Machine Learning Approach Using Electronic Health Records

Sheikh S. Abdullah <sup>1</sup>, Neda Rostamzadeh <sup>1</sup>, Kamran Sedig <sup>1,\*</sup>, Amit X. Garg <sup>2</sup> and Eric McArthur <sup>3</sup>

<sup>1</sup> Insight Lab, Western University, London, ON N6A 3K7, Canada; sabdul9@uwo.ca (S.S.A.); nrostamz@uwo.ca (N.R.)

<sup>2</sup> Department of Medicine, Epidemiology and Biostatistics, Western University, London, ON N6A 3K7, Canada; amit.garg@lhsc.on.ca

<sup>3</sup> ICES, London, ON N6A 3K7, Canada; eric.mcarthur@ices.on.ca

\* Correspondence: sedig@uwo.ca; Tel.: +1-519-661-2111 (ext. 86612)

Received: 13 June 2020; Accepted: 3 August 2020; Published: 5 August 2020



**Abstract:** Acute kidney injury (AKI) is a common complication in hospitalized patients and can result in increased hospital stay, health-related costs, mortality and morbidity. A number of recent studies have shown that AKI is predictable and avoidable if early risk factors can be identified by analyzing Electronic Health Records (EHRs). In this study, we employ machine learning techniques to identify older patients who have a risk of readmission with AKI to the hospital or emergency department within 90 days after discharge. One million patients' records are included in this study who visited the hospital or emergency department in Ontario between 2014 and 2016. The predictor variables include patient demographics, comorbid conditions, medications and diagnosis codes. We developed 31 prediction models based on different combinations of two sampling techniques, three ensemble methods, and eight classifiers. These models were evaluated through 10-fold cross-validation and compared based on the AUROC metric. The performances of these models were consistent, and the AUROC ranged between 0.61 and 0.88 for predicting AKI among 31 prediction models. In general, the performances of ensemble-based methods were higher than the cost-sensitive logistic regression. We also validated features that are most relevant in predicting AKI with a healthcare expert to improve the performance and reliability of the models. This study predicts the risk of AKI for a patient after being discharged, which provides healthcare providers enough time to intervene before the onset of AKI.

**Keywords:** acute kidney injury; electronic health records; data mining; automated analysis; imbalanced data; prediction models; risk stratification

## 1. Introduction

Acute kidney injury (AKI) is common among patients admitted to hospitals, affecting approximately 10% of hospitalized patients and more than 25% of patients in the intensive care unit [1,2]. AKI is defined as an abrupt loss of kidney function over a short period of time [2]. AKI may lead to prolonged hospital stays, lower chance of survival, and a higher risk of developing chronic kidney disease. Over the last 10–15 years, the incidence rate of AKI has increased in the United States [3,4], the United Kingdom [5] and Canada [6,7]. The growing incidence rate of AKI is associated with the changing spectrum of diseases. There is an increasing body of evidence proving that patients with extrarenal complications and multiple comorbidities are at a greater risk of developing AKI [8,9]. Aikar et al. [10] have shown that the high comorbidity rate, measured by the Deyo-Charlson comorbidity index, is associated with AKI. As a patient's number of comorbid conditions grows, there is a rise

in associated physician visits, healthcare utilization, medication intake and hospitalizations [11], ultimately leading to an increase in healthcare expenditure. Given the associated risk and expense, a promising strategy is required to improve the care for AKI patients. However, a UK-based report published in 2009 demonstrated significant under-recognition of AKI, leading to delayed recognition, inadequate treatment and ineffective monitoring [12,13].

Thus, there is a rising demand for techniques that can be used for the detection of AKI. However, the complex pathophysiology and etiology of AKI make the diagnosis and management of this disease challenging. There are different guidelines such as RIFLE [14], AKIN [15], WRF [16] and KDIGO [17] for AKI diagnosis. Most of these guidelines rely on a rise in serum creatinine (i.e., a laboratory test) alone as the gold standard. However, serum creatinine-based guidelines are often not ideal for the diagnosis of AKI among older patients because the age-related deteriorations in glomerular filtration rates affect the baseline measure [18]. Another limitation of this measurement is the fact that serum creatinine may vary with muscle mass since it is a product of muscle catabolism [19]. In addition, serum creatinine-based guidelines require a pre-morbid serum creatinine value to be used as a baseline creatinine, which may not be available for all patients [20]. Although some guidelines also rely on urine output to diagnose AKI, it is only monitored for patients with reduced kidney function [18]. Despite these challenges, even if AKI can be diagnosed properly, the clinicians often fail to intervene due to a lack of time and treatment options. The treatments of AKI are primarily focused on avoiding nephrotoxic medications and administering supportive care [17]. Although more advanced treatments have been identified in recent years, their effectiveness has not been proven in clinical trials yet [21]. Thus, interventions often have poor results if a patient has developed AKI already [22,23]. So, it is more effective to predict AKI prior to its diagnosis. A number of recent studies have shown that AKI is predictable and avoidable if early risk factors can be identified using Electronic Health Records (EHRs). For instance, Kate et al. (2016) have revealed that it is possible to predict up to 30% of AKI cases in the hospital setting using the patient data stored in EHRs [18].

EHRs contain patient medical records, such as comorbid conditions, medications, laboratory test results, diagnosis codes, demographics and discharge summaries, which can be used for the risk profiling of patients [20,24]. With the evolution of EHRs and the widespread use of information technology systems, these medical records are available nowadays for subsequent reuses [25–28]. EHRs offer an opportunity to employ machine learning techniques in order to recognize risk factors associated with AKI and identify patients at risk of developing AKI. Several clinical decision support systems have been developed in recent years for earlier detection of AKI using machine learning techniques [29–35]. However, many of these systems suffer from various performance and design related issues, such as lack of predictive power, substantial trade-offs between sensitivity and specificity, a limited number of machine learning techniques, small population size, lack of predictors and limited patient populations [20,34].

This study is designed to predict AKI among hospitalized and emergency department patients using machine learning techniques. We incorporate ICES' healthcare administrative datasets containing one million older patients' medical records who visited the hospital or emergency department between 2014 and 2016. We developed 31 prediction models based on different combinations of two sampling techniques, three ensemble methods and eight classifiers. Our study differs from other studies in several ways: (1) we developed prediction models for patients who are at risk of developing AKI within 90 days timeframe after being discharged from the hospital or emergency department; (2) we included a large number of predictors to train the models; and (3) we validated the important features of each model with healthcare experts through formative evaluations to improve the performance and reliability of the models. The rest of this paper is organized as follows. Section 2 describes the methodology employed for the design of the study. Section 3 presents the experimental results. Finally, Section 4 includes the discussion, and Section 5 describes the limitations of the study.

## 2. Materials and Methods

We discuss the data sources and methodology in this section, which includes the design settings, design flow, data integration, cohort entry criteria, input features, outcomes and proposed machine learning techniques.

### 2.1. Study Design and Setting

We conducted a population-based retrospective cohort study in older patients who visited a hospital or emergency department between 1 April 2014 and 31 March 2016, using health administrative databases stored at ICES. These datasets were connected using unique encoded identifiers and analyzed at ICES. The use of datasets in this study is authorized under section 45 of Ontario's Personal Health Information Protection Act, which does not need review by a Research Ethics Board.

Ontario has a population of about 13 million residents with universal access to physician services and hospital care, which includes 1.9 million people aged 65 years or older. We suppressed the results of this study in cells with five or fewer patients to comply with ICES privacy regulations and minimize the possibility of reidentification of patients.

### 2.2. Workflow

Figure 1 shows the basic workflow of the study described in this paper. In the first step, we created an integrated dataset from five different health administrative databases. The data sources are discussed in Section 2.3. Next, we describe the inclusion and exclusion criteria in Section 2.4. The features in the comorbidity, prescription, demographic and diagnosis codes data were encoded and transformed into suitable forms for analysis in the preprocessing stage, which is discussed in Section 2.7. The analysis techniques and results are presented in Sections 2.8 and 3, respectively.

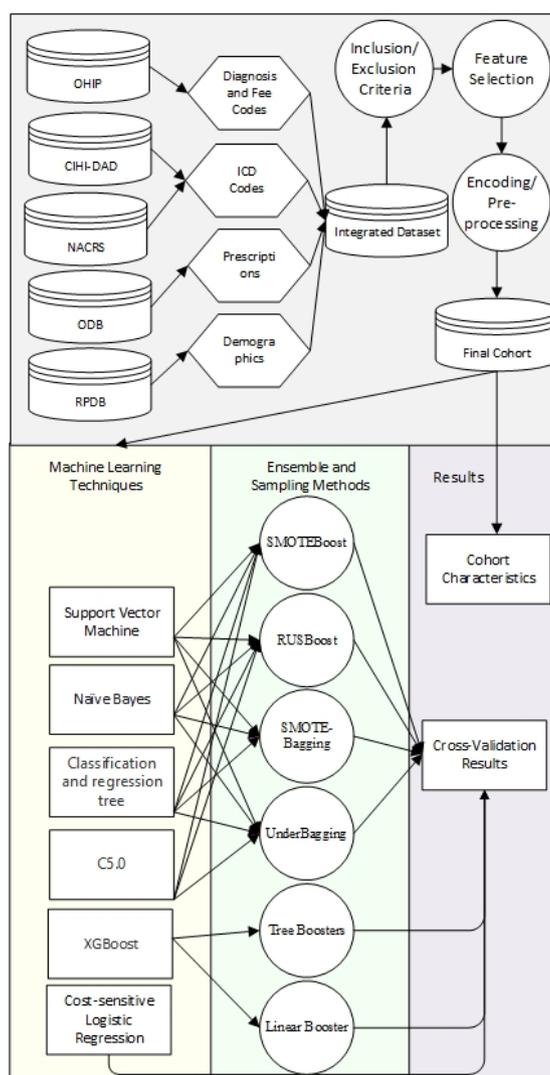
### 2.3. Data Sources

We ascertained patient characteristics, drug prescriptions, outcome and medical history data from 5 administrative databases (as shown in Table A1 Appendix A). These datasets are linked using a unique identifier, which is derived from health card numbers. We collected vital statistics from the Ontario Registered Persons Database (RPDB) [36], which includes demographic data of all residents in Ontario who have a valid health card. We utilized the Ontario Drug Benefit (ODB) Program database [37] to get prescription medication use data. The ODB database holds all the outpatient prescription records dispensed to older patients, which has an error rate of less than 1% [38]. We ascertained baseline comorbidity, emergency department visit and hospital admission data from the National Ambulatory Care Reporting System (NACRS) (i.e., for the emergency department) [39] and the Canadian Institute for Health Information Discharge Abstract Database (CIHI-DAD) (i.e., for hospital admissions) [40]. We applied the ICD-10 (i.e., International Classification of Diseases, post-2002) [41] codes to identify baseline comorbidities within the look-back window. In addition, baseline comorbidity data were acquired from the Ontario Health Insurance Plan (OHIP) database [42], which holds claim records for physician services. All the coding definitions for the comorbidity databases are provided in Table A2.

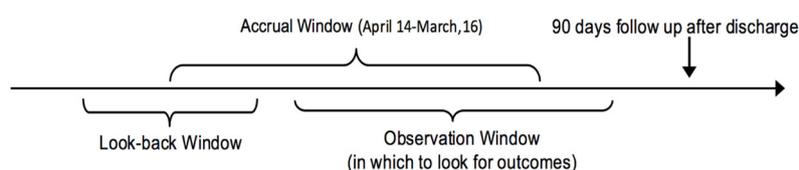
### 2.4. Cohort Entry Criteria

We identified a cohort of individuals aged 65 years or older who visited the emergency department or were admitted to hospital between 2014 and 2016 (Figure 2). The hospital admission or emergency department discharge dates were taken as the cohort entry or index date. If a patient had multiple hospital admissions and emergency department visits, we chose the first incident. We excluded patients with invalid or missing age, sex and/or health card number. In addition, we excluded patients who (1) previously underwent a kidney transplant or dialysis treatment, as AKI is usually no longer relevant once patients develop end-stage kidney disease; (2) left the emergency department or hospital without being seen by a physician or against medical advice; and (3) developed AKI during emergency

department visit or hospital admission, as they are already under observation. The diagnosis codes for the exclusion criteria are presented in Table A3.



**Figure 1.** Workflow diagram of the presented study where different colors are used to represent three main parts (data integration and preprocessing, analysis and validation). The figure shows how different combinations are formed using two sampling techniques (i.e., under sampling and synthetic minority over-sampling technique), three ensemble methods (i.e., boosting, bagging and XGBoost), and eight machine learning classifiers.



**Figure 2.** Provides an overview of data creation plan and how we prepared the final cohort.

We identified 2,305,783 hospitalization and 12,347,256 emergency department visit records in CIHI-DAD and NACRS, respectively. Next, a total of 5,635,909 unique individuals were identified using RPDB. There were 1,007,993 individuals included in the cohort after excluding patients with

invalid age, sex, and/or health card number and selecting patients aged 65 years or older. Finally, a total of 905,442 individuals were included in the final cohort after applying the other exclusion criteria.

### 2.5. Input Features

All these features from different data sources were integrated using the encoded identifiers derived by ICES using patient health card numbers. For each patient, we generated new features and aggregated multiple values (rows) of a single feature into one by considering the latest values of that feature. There were totals of 307,624, 768,293, 898,538 and 891,176 unique observations in the aggregated CIHI-DAD, NACRS, ODB and OHIP databases, respectively. We identified patients transferred from the emergency department to the hospital (appeared in both CIHI-DAD and NACRS) and removed duplicates by considering the first incident. We identified a total number of 1878 unique diagnosis codes (using CIHI-DAD, NACRS and OHIP) and 595 distinct medications (using ODB) for 905,442 individuals who were included in the final cohort. We used the Chi-Square test for feature selection and then filtered the selected features with a healthcare expert. The final combined dataset included a total of 86 unique features. The cohort contained 11 comorbidity features—namely, chronic kidney disease, diabetes mellitus, cerebrovascular disease, coronary artery disease, hypertension, chronic liver disease, major cancers, peripheral vascular disease, heart failure and kidney stones. We applied a 5-year look-back window to detect these baseline comorbidities. There were four demographics features—namely, sex, age, region and income quintile. We included 55 medications that were prescribed to the patients within 120 days before the first hospital admission or emergency department visit. These medications belonged to 13 distinct drug classes—namely, ACE-inhibitors (blood pressure and heart failure), beta-blockers (blood pressure), alpha-adrenergic blocking agents (blood pressure), angiotensin-receptor blockers (blood pressure), calcium blockers (blood pressure), macrolides (antibiotics), fluoroquinolones (antibiotics), potassium-sparing diuretics (weak diuretic), other diuretics, nonsteroidal anti-inflammatory agents (pain relievers), oral hypoglycemic (diabetes mellitus) and immunosuppressive agents (immune system activity).

The cohort also included 16 ICD-10 diagnosis codes that were identified during the index hospitalization or emergency department visit. The codes were related to delirium, mycoplasma pneumoniae, disorders of fluid, electrolyte and acid-base balance (e.g., hyperosmolality and hypernatraemia, hypo-osmolality and hyponatraemia, acidosis, alkalosis, mixed disorder of acid-base balance, hyperkalaemia, hypokalaemia, fluid overload, and other disorders of electrolyte and fluid balance), atrial fibrillation, anemia, femur fracture, valve disorders, atherosclerotic cardiovascular disease, diseases of the digestive system (e.g., paralytic ileus, intussusception, volvulus, gallstone ileus, other impaction of intestine, intestinal adhesions with obstruction, and other and unspecified intestinal obstruction ileus), Certain infectious and parasitic diseases (e.g., sepsis due to *Staphylococcus aureus*, other specified *Staphylococcus*, *Haemophilus influenzae*, *Escherichia coli*, *Pseudomonas*, *Serratia marcescens*, other Gram-negative organisms, Gram-negative Septicaemia and *Enterococcus*), dehydration and other volume depletion, abnormal function (e.g., abnormal results of function tests of central nervous system, peripheral nervous system and special senses, pulmonary function tests, cardiovascular function tests, kidney function tests, liver function tests, thyroid function tests, other endocrine function tests and electrocardiogram suggestive of ST-segment elevation myocardial infarction, abnormal cardiovascular function tests, and other abnormal results of cardiovascular function tests), chronic pulmonary (e.g., chronic obstructive pulmonary disease with acute lower respiratory infection and acute exacerbation and other specified chronic obstructive pulmonary disease), dementia, glomerular disorders (e.g., glomerular disorders in infectious and parasitic diseases, neoplastic diseases, blood diseases and disorders involving the immune mechanism, diabetes mellitus, other endocrine, nutritional and metabolic diseases, and systemic connective tissue disorders) and hyperplasia of prostate.

## 2.6. Outcome: Identification of AKI

Machine learning models were built to predict AKI within 90 days after being discharged from the hospital or emergency department. Positive cases were those in which patients revisited hospital or emergency department with AKI within 90 days after being discharged, and negative cases were the ones wherein hospitalizations or emergency department visits with AKI never took place. There were totals of 899,449 negative and 5993 positive cases in the dataset. There were no recurrent AKI examples (i.e., excluded 25,084 patients) in the data because we excluded the cases wherein AKI or dialysis was acquired during the index hospital stay or emergency department visit.

The incidence of AKI was detected using the Canadian Institute for Health Information Discharge Abstract Database and National Ambulatory Care Reporting System based on the ICD-10 (International Classification of Diseases—Tenth Revision) diagnostic codes (i.e., ICD-10 code of AKI is “N17”).

## 2.7. Data Preprocessing

The features in the cohort were transformed into a format and scale that was suitable for the machine learning techniques. For each feature described in Section 2.5, the last recorded value before the first hospital admission or emergency department visit was captured. Medication, diagnosis code and comorbidity features were set to either “Y” or “N.” If a patient had a certain comorbid condition or was prescribed a medication, then its corresponding value was taken as “Y.” Instead of reporting individual ages, we calculated age group features for the patients. If a patient’s age was within the specified range of an age group, we set the value to “1” for that corresponding feature. The sex feature took either “M” or “F” if the information was available in the dataset. Patients with invalid age or sex were removed from the cohort. The region feature took either “R” or “U” to represent rural or urban, respectively. The income feature took an integer value ranging between 1 and 5 to represent the income quintile of a particular patient.

## 2.8. Analysis Using Machine Learning Techniques

We employed both traditional and state-of-art analysis techniques to build trust with end-users and, at the same time, allow them to explore complex relationships in the dataset. We developed 31 AKI prediction models based on combinations of eight classifiers—namely, classification and regression tree (CART) [43], C5.0 [44], naïve Bayes (NB) [45], logistic regression [46], and support vector machine (SVM) with four different kernels (linear, polynomial, sigmoid and radial) [47], two sampling techniques (namely, under sampling and SMOTE) and three ensemble methods—namely, Boosting, Bagging and XGBoost. These techniques were chosen for several reasons, as follows: (1) They each represent different types of machine learning methods. For example, the decision tree is a rule-based, regression is a statistical, and naïve Bayes is a probability-based method. (2) Each of these methods has its own set of advantages and limitations. For instance, decision tree models are more human-interpretable but often fail to represent complex relationships among data elements. On the contrary, SVM is equipped to model complex non-linear relationships using different kernels, but is difficult to interpret. (3) Medical experts are more familiar with regression than other machine learning algorithms, which convinced us to include regression in this analysis.

### 2.8.1. Ensemble Methods

Since the number of negative cases was significantly higher than the number of positive cases, we considered the dataset as highly imbalanced. Traditional machine learning techniques, such as decision tree, support vector machine and so on, which are designed to optimize the overall accuracy, tend to achieve poor performance in this class imbalanced learning scenario because they try to minimize the overall error to which the minority class barely contributes. These techniques have shown high precision (i.e., a small number of false positives), reduced sensitivity (i.e., a higher number of false negatives) and low AUROC scores for our dataset, because they get biased toward the majority

class and fail to map minority class. An ensemble method offers a solution to this problem by combining several classification models to obtain better performance than the base classifiers [48]. To deal with the class imbalance issue in this study, we incorporated four different combinations of ensemble and sampling methods—namely, SMOTEBoost, SMOTE-Bagging, UnderBagging and RUSBoost, which are available in the “embc” package of R [49–51]. The RUSBoost was implemented using the “rus” function in the “ebmc” package. The weak learners in RUSBoost are trained on random under-sampled datasets [52]. Those learners are then combined to generate the final ensemble model. We used the “sbo” function to implement SMOTEBoost. SMOTE (Synthetic Minority Oversampling Technique) is a sampling technique that synthesizes new instances for the minority class using the *k*-nearest-neighbors algorithm [53]. SMOTEBoost returns several weak learners that are trained on SMOTE-generated datasets along with their error estimations [54]. The “sbag” function was used to implement SMOTEBagging, which combines SMOTE and random over-sampling to rebalance the dataset [44]. We used the “ub” function to implement the UnderBagging method. Unlike other ensemble methods discussed above, UnderBagging only incorporates random under-sampling to reduce the instances of the majority class in each bag to rebalance the class distribution. We configured this function in such a way that the amount of majority instances became equal compared to the minority instances (i.e., imbalance ratio = 1). We compared the models’ performance for different ensemble sizes (i.e., 10, 15, 20, 25 and 30) and used 20 weak learners for the algorithms. We used NB, SVM, CART and C50 as weak learners for the ensemble methods, which are discussed in the following subsections. Since ensemble methods are designed to combine several base models to obtain better performance than the weak learners, and these algorithms (i.e., NB, SVM, CART and C50) are used as weak learners in this study, we did not perform an explicit grid search to tune the hyperparameters.

#### Support Vector Machine

The objective of the SVM is to find an optimal separating hyperplane in a multi-dimensional space (i.e., depending on the number of features) that distinctly divides the instances of different classes. Although SVM models are often not human-interpretable, it has been proven to work well on prediction tasks involving a large number of features [18]. It has become popular in healthcare research recently because it is more effective in analyzing high dimensional EHRs. In addition, the regularization parameters of SVM kernels help users avoid over-fitting. Since the performance of the models widely varies depending on the selection of the kernel [55], and kernels are quite sensitive to over-fitting [56], one of the main challenges is to select an appropriate kernel. Thus, we tested the performance of four well-known kernel functions in this study—namely, linear, polynomial, sigmoid and radial.

#### Decision Tree

A decision tree is the representation of possible outcomes of a decision depending on certain conditions [44]. It is similar to a flowchart where every non-leaf node represents a test for a specific feature, and the leaf node represents a particular outcome. A decision tree reduces the ambiguity of complicated clinical decisions and requires reduced effort for data preparation compared to other techniques. It can be an effective technique for analyzing datasets with missing values because the tree-building process is not affected by the missing data [57]. We chose the decision tree mainly because it is easy to interpret and understand. Despite the advantages, decision tree models are often volatile, meaning that a minor alteration in the training data may cause a massive change in the structure of the tree. To overcome this issue, we included other types of base classifiers along with the decision tree and verified the structure of the generated tree with a healthcare expert. We incorporated two different algorithms to develop decision tree models in this study. The classification and regression tree (CART) were implemented using the “rpart” package [43], and the C5.0 classifier was implemented using the “C50” package in R [44].

## Naïve Bayes

NB is a simple probabilistic classifier founded on Bayes theorem [45], which is exceptionally fast to train compared to other complex techniques [55]. Classification of the new data using this technique only requires mathematical operations based on the feature probability. We chose NB mainly because it is less sensitive to missing data. However, since this technique is designed based on the assumption of feature independence, the performance may deteriorate when features in the training data are related. We used the “naive Bayes” package to implement the NB algorithm in this study [58].

### 2.8.2. Logistic Regression

Logistic regression draws a separating line among the classes using the training dataset, and then applies that line to classify the unknown data points. It is used to analyze the relationships between one dependent feature and one or more independent features. Logistic regression models are informative as they reveal the association among features in terms of odds ratios. Over the last few decades, logistic regression techniques have become very popular in healthcare studies [59]. Although logistic regression models are not designed to support imbalanced classification directly, they can be modified to work with skewed distributions. In order to adjust the regression coefficients while training with the imbalance data, we implemented a cost-sensitive regression model. We adjusted the weight of the minority class based on the cost of its misclassification compared to the cost of misclassifying the majority class. We used internal 10-fold cross-validation during training to determine the appropriate weight for the minority class.

### 2.8.3. XGBoost

XGBoost (i.e., eXtreme Gradient Boosting) is an advanced implementation of gradient boosted decision trees that can be used for ranking, regression and classification problems [60]. One of the main advantages of XGBoost is that it supports parallel computation, which makes it faster than other techniques of gradient boosting. Because of its time complexity and performance superiority, it has been widely used in healthcare research, such as analysis of EHRs [61] and cancer diagnosis [62]. We used the “xgboost” package to implement XGBoost in R. Since this implementation of XGBoost only works with numeric data, we converted the categorical features in our dataset into numerical vectors. The “xgboost” package includes both a tree learning algorithm and a linear model solver. We implemented both algorithms to compare their performances. This package also has a built-in mechanism to control the balance of positive and negative weights. To train the models with unbalanced data, we adjusted the “scale\_pos\_weight” parameter based on the ratio of the negative class to the positive class [63]. We performed a grid search on the parameters of XGBoost and tuned the regularization parameters using the best parameters from the grid search.

## 2.9. Tools and Technologies

We primarily used two different data analysis software: SAS and R. SAS was used to cut and process the cohort because ICES health administrative databases were stored in a SAS server [64]. We used SAS programming, SQL and predefined macros to prepare data for analysis. Then, we loaded the preprocessed dataset in R packages [65] for additional analysis using machine learning techniques. We chose R mainly because it (1) is installed on the ICES workstations already, (2) has a rich array of machine learning libraries, (3) is open-source and platform-independent, and (4) continuously provides updates with new libraries.

### 3. Results

This section presents the results of this study. We divided the results into two subsections. First, we provide an overview of the dataset in Section 3.1. The results of predictive models are presented in Section 3.2.

#### 3.1. Cohort Characteristics

A total of 905,442 participants were included in the derivation cohort, of which 5993 had AKI during their hospital admission or emergency department visit after being discharged from the index encounter. We excluded 25,084 patients who developed AKI during the index hospitalization or emergency department visit. Selected characteristics of the derivation cohort are presented in Table 1.

**Table 1.** All the patients in the cohort were aged 65 years or older, where the mean age was 70 years. Among the participants, about 56% were women. About 6% of patients were in long term care, and 16% were from rural areas. The pre-existing comorbidities were diabetes (38%), hypertension (88%), major cancer (16%), coronary artery disease (25%), cerebrovascular disease (3%), heart failure (14%), chronic kidney disease (9%), kidney stones (1%) and peripheral vascular disease (2%). Some of the commonly prescribed medications were rosuvastatin calcium (22%), atorvastatin calcium (24%), amlodipine besylate (19%), metformin hcl (16%) and hydrochlorothiazide (20%). Baseline characteristics of patients in the cohort who were admitted to the hospital or visited the emergency department between 2014 and 2016.

Characteristics	Patients Admitted to Hospital or Visited Emergency Department		
	Total Patients	AKI	No AKI
Cohort size	905,442	5993	899,449
<b>Age, yr, mean (SD)</b>			
65 to <70	181,088 (20%)	589	180,499
70 to <80	371,231 (41%)	1911	369,320
80 to <90	269,147 (30%)	2485	269,147
≥90	81,489 (9%)	1008	80,481
<b>Sex</b>			
Women	507,047 (56%)	2901	504,146
<b>Year of cohort entry (index date)</b>			
2014–2015	588,537 (65%)	3987	584,550
2015–2016	316,904 (34%)	2006	314,898
<b>Location</b>			
Rural residence	144,870 (16%)	501	144,369
<b>LTC</b>			
Long-term care	36,217 (4%)	745	35,472
<b>Income Quintile</b>			
1 (lowest)	172,035 (19%)	1306	170,729
2	189,143 (21%)	1318	187,825
3	182,588 (20%)	1173	181,415
4	181,086 (20%)	1154	179,932
5 (highest)	180,590 (20%)	1043	179,547

Table 1. Cont.

Characteristics	Patients Admitted to Hospital or Visited Emergency Department		
	Total Patients	AKI	No AKI
<b>Comorbid conditions (by codes)</b>			
Hypertension	814,604 (88%)	5784	808,820
Diabetes	358,472 (38%)	3306	355,166
Heart failure	125,136 (14%)	1821	123,315
Coronary artery disease	239,437 (26%)	2005	237,432
Chronic liver disease	33,359 (4%)	297	33,062
Cancer	145,286 (16%)	1016	144,270
Chronic kidney disease	86,442 (9%)	1854	84,588
Kidney stones	12,457 (1%)	93	12,364
Peripheral vascular disease	13,197 (2%)	158	13,039
Cerebrovascular disease	25,835 (3%)	282	25,553
<b>Hospital Diagnosis Codes</b>			
Disorders of fluid, electrolyte and acid-base balance (E87)	13,563 (1%)	962	12,601
Delirium (F05)	4996 (1%)	342	4654
Atrial fibrillation (I48.91)	34,120 (4%)	1978	32,142
Mycoplasma pneumoniae (B96)	6197 (1%)	434	5763
Anaemia (D64.9)	11,814 (1%)	791	11,023
Valve disorders (I35)	1261 (1%)	186	1075
Fracture of femur (S72)	7263 (1%)	231	7032
Atherosclerotic cardiovascular disease (I25.10)	21,472 (2%)	1256	20,216
Volume depletion (E86.9)	3739 (1%)	240	3499
Diseases of the digestive system (K00-K95)	4552 (1%)	264	4288
Abnormal functions of organs and systems (R94.8)	11,348 (2%)	725	10,623
Chronic pulmonary (J81.1)	24,217 (3%)	971	23,246
Hyperplasia of prostate (N40.1)	5047 (1%)	153	4894
Certain infectious and parasitic diseases (A00-B99)	1191 (1%)	105	1086
Dementia (F03. 90)	8714 (1%)	390	8324
Glomerular disorders (N08)	3988 (1%)	569	3419

### 3.2. Classification Results

We evaluated all of the machine learning models using 10-fold cross-validation [66]. The cohort was divided into 10 equal groups, wherein 9 groups were used for training, and the 10th group was used for testing. We repeated this process 10 times, using different parts for training and testing, and assessed the performance of the models for each fold. We then combined the results of these folds to calculate the evaluation scores. We measured the validity of the tests in terms of sensitivity and specificity. Sensitivity is the capacity of a test to classify an individual as “at-risk” correctly. It represents the probability of a test being positive when “AKI” is present. On the contrary, specificity refers to the ability to classify an individual as “risk-free” correctly. Since predicting AKI was a binary classification problem (i.e., AKI or Non-AKI), all of the machine learning techniques were capable of providing a confidence score along with the output. The trade-off between sensitivity and 1-specificity was achieved by altering the threshold on the confidence scores, generating the receiver operating characteristic (ROC) curve. We used the ROC space to compare the performances of alternative tests in terms of 1-specificity and sensitivity. Thus, we computed and reported sensitivity, specificity, and area under the receiver operating characteristic curve (AUROC). The AUROC ranged from 0.61 to 0.88 for predicting AKI among the 31 machine learning models. The average AUROC values of ensemble methods were higher than the cost-sensitive logistic regression model. Among the sampling-based

ensemble methods, the performances of the UnderBagging and RUSBoost methods were better than the SMOTE. We achieved the best result of AUROC 0.88 with (1) a combination of RUSBoost and SVM using a sigmoid kernel and (2) XGBoost using a tree learning algorithm. The AUROC of the linear boosting algorithm (XGBoost) was 0.84, which was higher than the cost-sensitive logistic regression but lower than the tree learning algorithm (XGBoost). Since it is a disease prediction problem, high sensitivity was more useful than specificity. The highest sensitivity was 0.90, which was achieved using SVM-sigmoid and SVM-radial kernels with RUSBoost and SMOTE-Bagging, respectively. The complete list of performance measures is presented in Table 2.

**Table 2.** Performances of the machine learning techniques grouped by four ensemble-based methods and results of cost-sensitive regression analysis. The table contains sensitivity, specificity and AUROC of the cross-validation AUC of all the prediction models.

Ensemble-Based Methods	Machine Learning Techniques	Sensitivity	Specificity	AUROC
NA	Logistic Regression	0.79	0.72	0.77 ± 0.038
SMOTEBoost	Classification and Regression Trees (CART)	0.77	0.69	0.74 ± 0.039
	C5.0	0.84	0.78	0.83 ± 0.036
	NB (Naïve Bayes)	0.61	0.89	0.75 ± 0.038
	Support Vector Machine (SVM) (linear)	0.84	0.74	0.79 ± 0.035
	SVM (polynomial)	0.78	0.82	0.81 ± 0.033
	SVM (sigmoid)	0.76	0.85	0.84 ± 0.035
SMOTE-Bagging	SVM (radial)	0.70	0.83	0.82 ± 0.034
	CART	0.60	0.71	0.68 ± 0.041
	C5.0	0.62	0.84	0.79 ± 0.036
	NB	0.69	0.73	0.72 ± 0.039
	SVM (linear)	0.76	0.84	0.81 ± 0.031
	SVM (polynomial)	0.82	0.73	0.80 ± 0.033
UnderBagging	SVM (sigmoid)	0.84	0.71	0.81 ± 0.030
	SVM (radial)	0.90	0.74	0.86 ± 0.029
	CART	0.71	0.83	0.79 ± 0.035
	C5.0	0.88	0.76	0.85 ± 0.032
	NB	0.58	0.72	0.61 ± 0.041
	SVM (linear)	0.77	0.84	0.83 ± 0.035
RUSBoost	SVM (polynomial)	0.85	0.71	0.84 ± 0.037
	SVM (sigmoid)	0.89	0.71	0.85 ± 0.034
	SVM (radial)	0.79	0.90	0.86 ± 0.033
	CART	0.78	0.74	0.76 ± 0.039
	C5.0	0.84	0.77	0.82 ± 0.028
	NB	0.68	0.72	0.71 ± 0.039
XGBoost	SVM (linear)	0.84	0.78	0.83 ± 0.035
	SVM (polynomial)	0.74	0.85	0.82 ± 0.037
	SVM (sigmoid)	0.90	0.79	0.88 ± 0.029
	SVM (radial)	0.71	0.87	0.85 ± 0.034
XGBoost	Tree boosting	0.89	0.81	0.88 ± 0.031
	Linear boosting	0.86	0.77	0.84 ± 0.033

#### 4. Discussion

In this study, we demonstrated how machine learning techniques could help with the prediction of AKI using administrative health databases stored at ICES. Several machine learning-based models have been developed in recent studies to predict AKI among ICU and post-operative patients [29–35]. However, most of these models only focus on a specific medical condition, and consider the risk factors associated with that condition. For instance, Go et al. (2010) examined how AKI affects the risk of chronic kidney disease, cardiovascular events and other patient-related outcomes in hospital settings [67]. The earlier AKI can be predicted, the better the chances are of preventing AKI and its associated cost. The features that have been used in most of the existing studies work better in predicting AKI if their values are recorded closer to the timing of AKI onset. However, it may not be

beneficial to detect AKI close to its onset because clinicians will not have enough time to intervene. Thus, there is a trade-off between accuracy and usefulness, which can be optimized by using the information available in EHRs. Although some studies have developed risk stratification models for AKI using EHRs [68,69], they can only predict hospital-acquired AKI and do not consider patients who are at risk of developing AKI after being discharged. To our best knowledge, there are no previous studies in the literature that predict the risk of AKI after being discharged from the hospital using both the historical and healthcare utilization data. Thus, this study is not only novel, but also clinically relevant, because it provides clinicians with the ability to intervene and treat patients before AKI causes irreversible damage.

We analyzed all AKI events that took place within 90 days after being discharged from the hospital or emergency department, and developed prediction models to identify high-risk patients. We decided to choose a 90-day timeframe for following up because (1) out of all AKI cases within six months after discharge, about 85% were acquired within this timeframe, and (2) it was a reasonable timeframe considering the trade-off between the models' usefulness (from a clinical point of view) and predictive power (from a machine learning point of view). Table 3 shows how many acquired AKI cases were identified within different time intervals. The machine learning models presented in this study can be adapted to make predictions at any other timeframes if needed.

**Table 3.** The number of AKI cases are grouped into six time periods.

Intervals	Readmission with AKI
1–3 days	415
4–7 days	534
8–14 days	888
15–30 days	1517
31–60 days	3579
61–90 days	1499

We incorporated eight different machine learning classifiers, three ensemble methods and two sampling techniques to develop 31 prediction models. Although each combination of machine learning techniques and ensemble-based methods performed reasonably well, the performance of SVM with sigmoid kernel and tree-based XGBoost produced better results than other techniques in general. The performances of all of the ensemble-based methods were consistent, and produced similar results for different base classifiers. The results shown in Table 2 indicate that the models agreed with each other.

To understand the models better, we explored the features that are important in each prediction model. We analyzed this information with a nephrologist to confirm the correctness of the models. We observed the odds ratio and p-value of the features in the regression model, the feature importance in the decision tree and XGBoost models, and the coefficients in the SVM-linear models in order to understand the associations between different features and AKI. The features included in this study can be divided into four categories—namely, demographics, comorbidities, medications and diagnosis codes.

In general, features from comorbidities and hospital diagnosis codes were more associated with AKI. Although the importance of the features varied based on the machine learning techniques, most of the features that stood out were common among these models. For instance, diabetes mellitus, hypertension, coronary artery disease, heart failure, major cancers, chronic liver disease, peripheral vascular disease and chronic kidney disease were the comorbidity features that were important in most of the prediction models. These comorbid conditions are already known to be associated with AKI in the literature [70–74]. The medication features that contributed to the higher risk of AKI include furosemide, allopurinol, hydrochlorothiazide, atorvastatin, metolazone, sunitinib malate, spironolactone, dexamethasone, chlorthalidone, atenolol, dexamethasone and oseltamivir

phosphate. These medications are known to be nephrotoxic [3,75–79]. Delirium, anaemia, mycoplasma, fluid disorders, atrial fibrillation, atherosclerotic cardiovascular disease, mycoplasma pneumoniae, hyperplasia of prostate, glomerular disorders and valve disorders were the features belonging to the diagnosis codes that were associated with increasing the risk of AKI in the prediction models. Several studies in the literature associate these medical conditions with AKI [80–84]. Among the demographic features, age, sex, location (i.e., urban or rural residence) and long-term care were found to be associated with AKI in most of the prediction models. Similar to comorbidity, medication and diagnosis code, these demographic features are already known to be associated with AKI [85–87] in the literature, which more conclusively proves the correctness of the prediction models. Through a comprehensive analysis of ICES's healthcare administrative datasets, this study shows that AKI is predictable using EHRs. Successful implementation of these prediction models in a healthcare setting can potentially reduce the risk of AKI among older patients.

## 5. Limitations and Future Work

The paper should be evaluated with respect to several limitations. First, our models were trained and tested on a cohort of older patients (65 years or older), which limits the generalizability of the models. Second, we excluded patients with missing or invalid demographics information. This may affect the performance of the models if the excluded data includes any interesting or rare cases. Third, the models are based on a cohort containing Ontario patients only, which limits this study to a specific geographic location. Fourth, the proposed prediction models are trained and tested on a specific patient cohort. It is essential to test the models' performances with real-time medical data before applying them in a clinical setting. Fifth, since we developed 31 prediction models, and many of them have different mechanisms of identifying feature importance, the interrelationships produced by these models are very complex. This paper only identifies the most significant predictors, but does not incorporate any ranking system for predictors. Finally, we identified episodes of AKI using ICD-10 codes, which may not include undetected cases in hospital settings. Moreover, since AKI was identified using the diagnosis code, this study does not consider the severity of AKI. Our future work concerns a deeper analysis of severe AKI that requires dialysis.

## 6. Conclusions

AKI is characterized by a sharp decline in renal function, and is associated with increased health-related costs and mortality. AKI is avoidable and may be preventable through an earlier prediction using risk factors available in EHRs. This study is designed to identify older patients who are discharged from the hospital or emergency department, and are at risk of developing AKI within 90 days after discharge. We employed eight traditional and state-of-art machine learning algorithms, along with two sampling techniques and three ensemble methods, to build AKI prediction models. The performances of these models were consistent, and a maximum AUROC of 0.88 was achieved through 10-fold cross-validation. We analyzed the models with a healthcare expert and identified features that are most relevant in predicting AKI. Most of these features are already known to be AKI-associated, which proves the correctness and feasibility of the prediction models. This study predicts the risk of AKI for a patient after being discharged from the hospital or emergency department, which provides healthcare providers enough time to intervene, monitor them more carefully, and avoid prescribing nephrotoxic medications for such patients.

**Author Contributions:** Conceptualization, S.S.A., N.R., K.S., A.X.G. and E.M.; methodology, S.S.A. and N.R.; software, S.S.A. and N.R.; validation, S.S.A., N.R., A.X.G. and E.M.; formal analysis, S.S.A. and N.R.; data curation, S.S.A., N.R. and E.M.; writing—original draft preparation, S.S.A. and N.R.; writing—review and editing, S.S.A., N.R., K.S., A.X.G. and E.M.; supervision, K.S. and A.X.G. All authors have read and agree to the published version of the manuscript.

**Funding:** This research received no external funding.

**Acknowledgments:** We would like to thank all authors and publishers who shared images of their tools with us.

**Conflicts of Interest:** The authors declare that there is no conflict of interest. Amit Garg is supported by Adam Linton, Chair in Kidney Health Analytics and a Clinician Investigator Award from the Canadian Institutes of Health Research (CIHR).

**Appendix A**

**Table A1.** List of databases held at ICES.

Data Source	Description	Study Purpose
Canadian Institute for Health Information Discharge Abstract Database and National Ambulatory Care Reporting System	The Canadian Institute for Health Information Discharge Abstract Database and National Ambulatory Care Reporting System collect diagnostic and procedural variables for inpatient stays and ED visits, respectively. Diagnostic and inpatient procedural coding use the 10th version of the Canadian Modified International Classification of Disease system 10th Revision (after 2002).	Cohort creation, description, exposure and outcome estimation
Ontario Drug Benefits	The Ontario Drug Benefits database includes a wide range of outpatient prescription medications available to all Ontario citizens over the age of 65. The error rate in the Ontario Drug Benefits database is less than 1%.	Medication prescriptions, description and exposure
Registered Persons Database	The Registered Persons Database captures demographic (sex, date of birth, postal code) and vital status information on all Ontario residents. Relative to the Canadian Institute for Health Information Discharge Abstract Database in-hospital death flag, the Registered Persons Database has a sensitivity of 94% and a positive predictive value of 100%.	Cohort creation, description and exposure
Ontario Health Insurance Plan	The Ontario Health Insurance Plan database contains information on Ontario physician billing claims for medical services using fee and diagnosis codes outlined in the Ontario Health Insurance Plan Schedule of Benefits. These codes capture information on outpatient, inpatient and laboratory services rendered to a patient.	Cohort creation, stratification, description, exposure and outcome

**Table A2.** Coding definitions for co-morbid conditions.

Variable	Database	Code	Set Code
Major cancer	Canadian Institute for Health Information Discharge Abstract Database	International Classification of Diseases 9th Revision	150, 154, 155, 157, 162, 174, 175, 185, 203, 204, 205, 206, 207, 208, 2303, 2304, 2307, 2330, 2312, 2334
		International Classification of Diseases 10th Revision	971, 980, 982, 984, 985, 986, 987, 988, 989, 990, 991, 993, C15, C18, C19, C20, C22, C25, C34, C50, C56, C61, C82, C83, C85, C91, C92, C93, C94, C95, D00, D010, D011, D012, D022, D075, D05
	Ontario Health Insurance Plan	Diagnosis	203, 204, 205, 206, 207, 208, 150, 154, 155, 157, 162, 174, 175, 183, 185
Chronic liver disease	Canadian Institute for Health Information Discharge Abstract Database	International Classification of Diseases 9th Revision	4561, 4562, 070, 5722, 5723, 5724, 5728, 573, 7824, V026, 571, 2750, 2751, 7891, 7895
	Ontario Health Insurance Plan	International Classification of Diseases 10th Revision	B16, B17, B18, B19, I85, R17, R18, R160, R162, B942, Z225, E831, E830, K70, K713, K714, K715, K717, K721, K729, K73, K74, K753, K754, K758, K759, K76, K77
		Diagnosis	571, 573, 070
Coronary artery disease (excluding angina)	Canadian Institute for Health Information Discharge Abstract Database	Fee code	Z551, Z554
		Canadian Classification of Diagnostic, Therapeutic and Surgical Procedures	4801, 4802, 4803, 4804, 4805, 481, 482, 483
		Canadian Classification of Health Interventions	1IJ50, 1IJ76
		International Classification of Diseases 9th Revision	412, 410, 411
		International Classification of Diseases 10th Revision	I21, I22, Z955, T822

Table A2. Cont.

Variable	Database	Code	Set Code
	Ontario Health Insurance Plan	Diagnosis	410, 412
		Fee code	R741, R742, R743, G298, E646, E651, E652, E654, E655, Z434, Z448
Diabetes	Canadian Institute for Health Information Discharge Abstract Database	International Classification of Diseases 9th Revision	250
		International Classification of Diseases 10th Revision	E10, E11, E13, E14
	Ontario Health Insurance Plan	Diagnosis	250
	Ontario Health Insurance Plan	Fee code	Q040, K029, K030, K045, K046
Heart failure	Canadian Institute for Health Information Discharge Abstract Database	Canadian Classification of Diagnostic, Therapeutic and Surgical Procedures	4961, 4962, 4963, 4964
		Canadian Classification of Health Interventions	1HP53, 1HP55, 1HZ53GRFR, 1HZ53LAFR, 1HZ53YFR
		International Classification of Diseases 9th Revision	I500, I501, I509, I255, J81
		International Classification of Diseases 10th Revision	I21, I22, Z955, T822
	Ontario Health Insurance Plan	Diagnosis	428
	Ontario Health Insurance Plan	Fee code	R701, R702, Z429
Hypertension	Canadian Institute for Health Information Discharge Abstract Database	International Classification of Diseases 9th Revision	401, 402, 403, 404, 405
		International Classification of Diseases 10th Revision	I10, I11, I12, I13, I15
	Ontario Health Insurance Plan	Diagnosis	401, 402, 403
Kidney stones	Canadian Institute for Health Information Discharge Abstract Database	International Classification of Diseases 9th Revision	5920, 5921, 5929, 5940, 5941, 5942, 5948, 5949, 27411
		International Classification of Diseases 10th Revision	N200, N201, N202, N209, N210, N211, N218, N219, N220, N228
Peripheral vascular disease	Canadian Institute for Health Information Discharge Abstract Database	Canadian Classification of Diagnostic, Therapeutic and Surgical Procedures	5125, 5129, 5014, 5016, 5018, 5028, 5038, 5126, 5159
		Canadian Classification of Health Interventions	1KA76, 1KA50, 1KE76, 1KG50, 1KG57, 1KG76MI, 1KG87, 1IA87LA, 1IB87LA, 1IC87LA, 1ID87LA, 1KA87LA, 1KE57
		International Classification of Diseases 9th Revision	4402, 4408, 4409, 5571, 4439, 444
		International Classification of Diseases 10th Revision	I700, I702, I708, I709, I731, I738, I739, K551
	Ontario Health Insurance Plan	Fee code	R787, R780, R797, R804, R809, R875, R815, R936, R783, R784, R785, E626, R814, R786, R937, R860, R861, R855, R856, R933, R934, R791, E672, R794, R813, R867, E649
	Ontario Health Insurance Plan	Diagnosis	430, 431, 432, 4340, 4341, 4349, 435, 436, 3623
Cerebrovascular disease (stroke or transient ischemic attack)	Canadian Institute for Health Information Discharge Abstract Database	International Classification of Diseases 9th Revision	430, 431, 432, 4340, 4341, 4349, 435, 436, 3623
		International Classification of Diseases 10th Revision	I62, I630, I631, I632, I633, I634, I635, I638, I639, I64, H341, I600, I601, I602, I603, I604, I605, I606, I607, I609, I61, G450, G451, G452, G453, G458, G459, H340
Chronic kidney disease	Canadian Institute for Health Information Discharge Abstract Database	International Classification of Diseases 9th Revision	4030, 4031, 4039, 4040, 4041, 4049, 585, 586, 5888, 5889, 2504
		International Classification of Diseases 10th Revision	E102, E112, E132, E142, I12, I13, N08, N18, N19
	Ontario Health Insurance Plan	Diagnosis	403, 585

**Table A3.** Diagnostic codes for exclusion criteria.

Variable	Database	Code Set	Code
Dialysis	Canadian Institute for Health Information Discharge Abstract Database	Canadian Classification of Diagnostic, Therapeutic and Surgical Procedures	5127, 5142, 5143, 5195, 6698
		Canadian Classification of Health Interventions	1PZ21, 1OT53DATS, 1OT53HATS, 1OT53LATS, 1SY55LAFT, 7SC59QD, 1KY76, 1KG76MZXXA, 1KG76MZXXN, 1JM76NC, 1JM76NCXXN
		International Classification of Diseases 9th Revision	V451, V560, V568, 99673
		International Classification of Diseases 10th Revision	T824, Y602, Y612, Y622, Y841, Z49, Z992
	Ontario Health Insurance Plan	Fee code	R850, G324, G336, G327, G862, G865, G099, R825, R826, R827, R833, R840, R841, R843, R848, R851, R946, R943, R944, R945, R941, R942, Z450, Z451, Z452, G864, R852, R853, R854, R885, G333, H540, H740, R849, G323, G325, G326, G860, G863, G866, G330, G331, G332, G861, G082, G083, G085, G090, G091, G092, G093, G094, G095, G096, G294, G295
Kidney transplant	Canadian Institute for Health Information Discharge Abstract Database	Canadian Classification of Health Interventions	1PC85
	Ontario Health Insurance Plan	Fee code	S435, S434

## References

- Selby, N.M.; Crowley, L.; Fluck, R.J.; McIntyre, C.W.; Monaghan, J.; Lawson, N.; Kolhe, N.V. Use of Electronic Results Reporting to Diagnose and Monitor AKI in Hospitalized Patients. *Clin. J. Am. Soc. Nephrol.* **2012**, *7*, 533–540. [[CrossRef](#)] [[PubMed](#)]
- Porter, C.J.; Juurlink, I.; Bisset, L.H.; Bavakunji, R.; Mehta, R.L.; Devonald, M.A.J. A real-time electronic alert to improve detection of acute kidney injury in a large teaching hospital. *Nephrol. Dial. Transplant.* **2014**, *29*, 1888–1893. [[CrossRef](#)] [[PubMed](#)]
- Wu, X.; Zhang, W.; Ren, H.; Chen, X.; Xie, J.; Chen, N. Diuretics associated acute kidney injury: Clinical and pathological analysis. *Ren. Fail.* **2014**, *36*, 1051–1055. [[CrossRef](#)]
- Nadkarni, G.N.; A Patel, A.; Ahuja, Y.; Annapureddy, N.; Agarwal, S.K.; Simoes, P.K.; Konstantinidis, I.; Kamat, S.; Archdeacon, M.; Thakar, C.V. Incidence, Risk Factors, and Outcome Trends of Acute Kidney Injury in Elective Total Hip and Knee Arthroplasty. *Am. J. Orthop.* **2016**, *45*, E12–E19. [[PubMed](#)]
- Kolhe, N.V.; Muirhead, A.W.; Wilkes, S.R.; Fluck, R.J.; Taal, M.W. The epidemiology of hospitalised acute kidney injury not requiring dialysis in England from 1998 to 2013: Retrospective analysis of hospital episode statistics. *Int. J. Clin. Pr.* **2016**, *70*, 330–339. [[CrossRef](#)]
- Liu, S.; Joseph, K.; Bartholomew, S.; Fahey, J.; Lee, L.; Allen, A.C.; Kramer, M.S.; Sauve, R.; Young, D.C.; Liston, R.M. Temporal trends and regional variations in severe maternal morbidity in Canada, 2003 to 2007. *J. Obstet. Gynaecol. Can.* **2010**, *32*, 847–855. [[CrossRef](#)]
- Mehrabadi, A.; Liu, S.; Bartholomew, S.; A Hutcheon, J.; A Magee, L.; Kramer, M.S.; Liston, R.M.; Joseph, K. Hypertensive disorders of pregnancy and the recent increase in obstetric acute renal failure in Canada: Population based retrospective cohort study. *BMJ* **2014**, *349*, g4731. [[CrossRef](#)]
- Mehta, R.L.; Pascual, M.T.; Soroko, S.; Savage, B.R.; Himmelfarb, J.; Ikizler, T.A.; Paganini, E.P.; Chertow, G.M. Spectrum of acute renal failure in the intensive care unit: The PICARD experience. *Kidney Int.* **2004**, *66*, 1613–1621. [[CrossRef](#)]
- Siddiqui, N.F.; Coca, S.G.; Devereaux, P.; Jain, A.K.; Li, L.; Luo, J.; Parikh, C.R.; Paterson, M.; Philbrook, H.T.; Wald, R.; et al. Secular trends in acute dialysis after elective major surgery—1995 to 2009. *Can. Med. Assoc. J.* **2012**, *184*, 1237–1245. [[CrossRef](#)]
- Waikar, S.S.; Curhan, G.C.; Wald, R.; McCarthy, E.P.; Chertow, G.M. Declining Mortality in Patients with Acute Renal Failure, 1988 to 2002. *J. Am. Soc. Nephrol.* **2006**, *17*, 1143–1150. [[CrossRef](#)]

11. Zulman, D.M.; Asch, S.M.; Martins, S.B.; Kerr, E.A.; Hoffman, B.B.; Goldstein, M.K. Quality of Care for Patients with Multiple Chronic Conditions: The Role of Comorbidity Interrelatedness. *J. Gen. Intern. Med.* **2013**, *29*, 529–537. [[CrossRef](#)] [[PubMed](#)]
12. Ali, T.; Khan, I.; Simpson, W.; Prescott, G.J.; Townend, J.; Smith, W.; MacLeod, A. Incidence and Outcomes in Acute Kidney Injury: A Comprehensive Population-Based Study. *J. Am. Soc. Nephrol.* **2007**, *18*, 1292–1298. [[CrossRef](#)] [[PubMed](#)]
13. Bagshaw, S.M.; George, C.; Bellomo, R. Changes in the incidence and outcome for early acute kidney injury in a cohort of Australian intensive care units. *Crit. Care* **2007**, *11*, R68. [[CrossRef](#)] [[PubMed](#)]
14. Eriksen, B.O.; Hoff, K.R.S.; Solberg, S. Prediction of acute renal failure after cardiac surgery: Retrospective cross-validation of a clinical algorithm. *Nephrol. Dial. Transplant.* **2003**, *18*, 77–81. [[CrossRef](#)]
15. Palevsky, P.M.; Liu, K.D.; Brophy, P.D.; Chawla, L.S.; Parikh, C.R.; Thakar, C.V.; Tolwani, A.J.; Waikar, S.S.; Weisbord, S.D. KDOQI US Commentary on the 2012 KDIGO Clinical Practice Guideline for Acute Kidney Injury. *Am. J. Kidney Dis.* **2013**, *61*, 649–672. [[CrossRef](#)] [[PubMed](#)]
16. Gottlieb, S.S.; Abraham, W.; Butler, J.; Forman, D.E.; Loh, E.; Massie, B.M.; O'Connor, C.M.; Rich, M.W.; Stevenson, L.W.; Young, J.; et al. The prognostic importance of different definitions of worsening renal function in congestive heart failure. *J. Card. Fail.* **2002**, *8*, 136–141. [[CrossRef](#)]
17. Clinical Practice Guideline KDIGO Clinical Practice Guideline for Acute Kidney Injury. *Kidney Int. Suppl.* **2012**, *2*, 1–138.
18. Kate, R.J.; Perez, R.M.; Mazumdar, D.; Pasupathy, K.S.; Nilakantan, V. Prediction and detection models for acute kidney injury in hospitalized older adults. *BMC Med. Inform. Decis. Mak.* **2016**, *16*, 39. [[CrossRef](#)]
19. Delanaye, P.; Pottel, H.; Cavalier, E. Serum Creatinine: Not So Simple! *Nephron* **2017**, *136*, 302–308. [[CrossRef](#)] [[PubMed](#)]
20. Mohamadlou, H.; Lynn-Palevsky, A.; Barton, C.; Chettipally, U.; Shieh, L.; Calvert, J.; Saber, N.R.; Das, R. Prediction of Acute Kidney Injury with a Machine Learning Algorithm Using Electronic Health Record Data. *Can. J. Kidney Health Dis.* **2018**, *5*, 5. [[CrossRef](#)]
21. Pozzoli, S.; Simonini, M.; Manunta, P. Predicting acute kidney injury: Current status and future challenges. *J. Nephrol.* **2017**, *31*, 209–223. [[CrossRef](#)] [[PubMed](#)]
22. Mehta, R.L. Management of acute kidney injury: It's the squeaky wheel that gets the oil! *Clin. J. Am. Soc. Nephrol.* **2011**, *6*, 2102–2104. [[CrossRef](#)] [[PubMed](#)]
23. Lieske, J.C.; Chawla, L.; Kashani, K.; Kellum, J.A.; Koyner, J.L.; Mehta, R.L. Biomarkers for Acute Kidney Injury: Where Are We Today? Where Should We Go? *Clin. Chem.* **2014**, *60*, 294–300. [[CrossRef](#)] [[PubMed](#)]
24. Rostamzadeh, N.; Abdullah, S.S.; Sedig, K. Data-Driven Activities Involving Electronic Health Records: An Activity and Task Analysis Framework for Interactive Visualization Tools. *Multimodal Technol. Interact.* **2020**, *4*, 7. [[CrossRef](#)]
25. Delamarre, D.; Bouzillé, G.; Dalleau, K.; Courtel, D.; Cuggia, M. Semantic integration of medication data into the EHOP Clinical Data Warehouse. *Stud. Health Technol. Inform.* **2015**, *210*, 702–706.
26. Abramson, E.L.; Barrón, Y.; Quaresimo, J.; Kaushal, R. Electronic Prescribing Within an Electronic Health Record Reduces Ambulatory Prescribing Errors. *Jtr. Comm. J. Qual. Patient Saf.* **2011**, *37*, 470–478. [[CrossRef](#)]
27. Abdullah, S.S.; Rostamzadeh, N.; Sedig, K.; Garg, A.X.; McArthur, E. Visual Analytics for Dimension Reduction and Cluster Analysis of High Dimensional Electronic Health Records. *Informatics* **2020**, *7*, 17. [[CrossRef](#)]
28. Abdullah, S.S.; Rostamzadeh, N.; Sedig, K.; Garg, A.X.; McArthur, E. Multiple Regression Analysis and Frequent Itemset Mining of Electronic Medical Records: A Visual Analytics Approach Using VISA\_M3R3. *Data* **2020**, *5*, 33. [[CrossRef](#)]
29. Rashidi, H.H.; Sen, S.; Palmieri, T.L.; Blackmon, T.; Wajda, J.; Tran, N.K. Early Recognition of Burn- and Trauma-Related Acute Kidney Injury: A Pilot Comparison of Machine Learning Techniques. *Sci. Rep.* **2020**, *10*, 1–9. [[CrossRef](#)]
30. Tran, N.K.; Sen, S.; Palmieri, T.L.; Lima, K.; Falwell, S.; Wajda, J.; Rashidi, H.H. Artificial intelligence and machine learning for predicting acute kidney injury in severely burned patients: A proof of concept. *Burn* **2019**, *45*, 1350–1358. [[CrossRef](#)]

31. E Davis, S.; A Lasko, T.; Chen, G.; Siew, E.D.; Matheny, M.E. Calibration drift in regression and machine learning models for acute kidney injury. *J. Am. Med. Inform. Assoc.* **2017**, *24*, 1052–1061. [[CrossRef](#)] [[PubMed](#)]
32. Cheng, P.; Waitman, L.R.; Hu, Y.; Liu, M. Predicting Inpatient Acute Kidney Injury over Different Time Horizons: How Early and Accurate? *AMIA Annu. Symp. Proced.* **2017**, *2017*, 565–574.
33. Ibrahim, N.E.; McCarthy, C.P.; Shrestha, S.; Gaggin, H.K.; Mukai, R.; Magaret, C.A.; Rhyne, R.F.; Januzzi, J.L. A clinical, proteomics, and artificial intelligence-driven model to predict acute kidney injury in patients undergoing coronary angiography. *Clin. Cardiol.* **2019**, *42*, 292–298. [[CrossRef](#)] [[PubMed](#)]
34. Gameiro, J.; Branco, T.; Lopes, J.A. Artificial Intelligence in Acute Kidney Injury Risk Prediction. *J. Clin. Med.* **2020**, *9*, 678. [[CrossRef](#)] [[PubMed](#)]
35. Abdullah, S.S.; Rostamzadeh, N.; Sedig, K.; Lizotte, D.J.; Garg, A.X.; McArthur, E. Machine Learning for Identifying Medication-Associated Acute Kidney Injury. *Informatics* **2020**, *7*, 18. [[CrossRef](#)]
36. Registered Persons Database (RPDB)—Ontario Data Catalogue. Available online: <https://data.ontario.ca/dataset/registered-persons-database-rpdb> (accessed on 25 July 2020).
37. Ontario Drug Benefit (ODB) Database—Ontario Data Catalogue. Available online: <https://data.ontario.ca/dataset/ontario-drug-benefit-odb-database> (accessed on 25 July 2020).
38. Levy, A.R.; O'Brien, B.J.; Sellors, C.; Grootendorst, P.; Willison, N. Coding accuracy of administrative drug claims in the Ontario Drug Benefit database. *Can. J. Clin. Pharmacol.* **2003**, *10*, 67–71.
39. National Ambulatory Care Reporting System Metadata (NACRS) CIHI. Available online: <https://www.cihi.ca/en/national-ambulatory-care-reporting-system-metadata-nacrs> (accessed on 25 July 2020).
40. Discharge Abstract Database Metadata (DAD) CIHI. Available online: <https://www.cihi.ca/en/discharge-abstract-database-metadata-dad> (accessed on 25 July 2020).
41. ICD-10 Version: 2019. Available online: <https://icd.who.int/browse10/2019/en> (accessed on 25 July 2020).
42. Data Available through DASm. Available online: <https://www.ices.on.ca/DAS/Data> (accessed on 25 July 2020).
43. Wilkinson, L. Classification and Regression Trees. Available online: [http://cda.psych.uiuc.edu/multivariate\\_fall\\_2013/systat\\_cart\\_manual.pdf](http://cda.psych.uiuc.edu/multivariate_fall_2013/systat_cart_manual.pdf) (accessed on 5 August 2020).
44. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Elsevier: Amsterdam, The Netherlands, 2014.
45. Lewis, D.D. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Machine Learning: ECML-98*; Nédellec, C., Rouveirol, C., Eds.; Springer: Berlin/Heidelberg, Germany, 1998; pp. 4–15.
46. Bahnsen, A.C.; Aouada, D.; Ottersten, B. Example-Dependent Cost-Sensitive Logistic Regression for Credit Scoring. In Proceedings of the 2014 13th International Conference on Machine Learning and Applications, Detroit, MI, USA, 3–5 December 2014; pp. 263–269.
47. Cristianini, N.; Shawe-Taylor, J. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Available online: </core/books/an-introduction-to-support-vector-machines-and-other-kernelbased-learning-methods/A6A6F4084056A4B23F88648DDBFDD6FC> (accessed on 23 April 2020).
48. Dietterich, T.G. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*; Springer: Berlin/Heidelberg, Germany, 2000; pp. 1–15.
49. Freund, Y.; E Schapire, R. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [[CrossRef](#)]
50. Wang, S.; Yao, X. *Diversity Analysis on Imbalanced Data Sets by Using Ensemble Models*; IEEE: Piscataway, NJ, USA, 2009.
51. Barandela, R.; Valdovinos, R.; Rosas, R.M.V.; Sánchez, J.S. New Applications of Ensembles of Classifiers. *Pattern Anal. Appl.* **2003**, *6*, 245–256. [[CrossRef](#)]
52. Seiffert, C.; Khoshgoftaar, T.M.; Van Hulse, J.; Napolitano, A. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *IEEE* **2009**, *40*, 185–197. [[CrossRef](#)]
53. Chawla, N.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
54. Galar, M.; Fernandez, A.; Barrenechea, E.; Bustince, H.; Herrera, F. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Trans. Syst. Man Cybern. Part C Rev.* **2011**, *42*, 463–484. [[CrossRef](#)]
55. Tomar, D.; Agarwal, S. A survey on Data Mining approaches for Healthcare. *Int. J. BioSci. BioTechnol.* **2013**, *5*, 241–266. [[CrossRef](#)]
56. Cawley, G.C.; Talbot, N.L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107.

57. Xie, N.; Liu, Y. Notice of Retraction: Review of decision trees. In Proceedings of the 2010 3rd International Conference on Computer Science and Information Technology, Chengdu, China, 9–11 July 2010; pp. 105–109.
58. McCallum, A.; Nigam, K. A Comparison of Event Models for Naive Bayes Text Classification. In *AAAI-98 Workshop on Learning for Text Categorization*; AAAI Workshop: Madison, WI, USA, 1998; pp. 41–48.
59. Ismail, B.; Anil, M. Regression methods for analyzing the risk factors for a life style disease among the young population of India. *Indian Heart J.* **2014**, *66*, 587–592. [[CrossRef](#)] [[PubMed](#)]
60. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: San Francisco, CA, USA, 2016; pp. 785–794.
61. Wang, C.; Wang, S.; Shi, F.; Wang, Z. Robust Propensity Score Computation Method based on Machine Learning with Label-corrupted Data. *arXiv*, 2018; arXiv:1801.03132.
62. Wang, C.-W.; Lee, Y.-C.; Calista, E.; Zhou, F.; Zhu, H.; Suzuki, R.; Komura, D.; Ishikawa, S.; Cheng, S.-P. A benchmark for comparing precision medicine methods in thyroid cancer diagnosis using tissue microarrays. *Bioinformatics* **2018**, *34*, 1767–1773. [[CrossRef](#)]
63. Wang, C.; Deng, C.; Wang, S. Imbalance-XGBoost: Leveraging Weighted and Focal Losses for Binary Label-Imbalanced Classification with XGBoost. Available online: <https://arxiv.org/abs/1908.01672> (accessed on 5 August 2020).
64. SAS Enterprise BI Server. Available online: [https://www.sas.com/en\\_ca/software/enterprise-bi-server.html](https://www.sas.com/en_ca/software/enterprise-bi-server.html) (accessed on 19 February 2020).
65. RStudio Open Source & Professional Software for Data Science Teams. Available online: <https://rstudio.com/> (accessed on 19 February 2020).
66. Japkowicz, N.; Shah, M. *Evaluating Learning Algorithms: A Classification Perspective*; Cambridge University Press: Cambridge, UK, 2011.
67. Go, A.S.; Parikh, C.R.; Ikizler, T.A.; Coca, S.G.; Siew, E.D.; Chinchilli, V.M.; Hsu, C.-Y.; Garg, A.X.; Zappitelli, M.; Liu, K.D.; et al. The assessment, serial evaluation, and subsequent sequelae of acute kidney injury (ASSESS-AKI) study: Design and methods. *BMC Nephrol.* **2010**, *11*, 22. [[CrossRef](#)]
68. Matheny, M.E.; Miller, R.A.; Ikizler, T.A.; Waitman, L.R.; Denny, J.C.; Schildcrout, J.S.; Dittus, R.S.; Peterson, J.F. Development of Inpatient Risk Stratification Models of Acute Kidney Injury for Use in Electronic Health Records. *Med. Decis. Mak.* **2010**, *30*, 639–650. [[CrossRef](#)]
69. Kane-Gill, S.L.; Sileanu, F.E.; Murugan, R.; Trietley, G.S.; Handler, S.; Kellum, J.A. Risk factors for acute kidney injury in older adults with critical illness: A retrospective cohort study. *Am. J. Kidney Dis.* **2014**, *65*, 860–869. [[CrossRef](#)]
70. Dylewska, M.; Chomicka, I.; Małyszko, J. Hypertension in Patients with Acute Kidney Injury. *Wiad. Lek.* **2019**, *72*, 2199–2201. [[CrossRef](#)]
71. Hsu, R.K.; Hsu, C.-Y. The Role of Acute Kidney Injury in Chronic Kidney Disease. *Semin. Nephrol.* **2016**, *36*, 283–292. [[CrossRef](#)] [[PubMed](#)]
72. Girman, C.J.; Kou, T.D.; Brodovicz, K.; Alexander, C.M.; O’Neill, E.A.; Engel, S.; Williams-Herman, D.E.; Katz, L. Risk of acute renal failure in patients with Type 2 diabetes mellitus. *Diabet. Med.* **2012**, *29*, 614–621. [[CrossRef](#)] [[PubMed](#)]
73. Olsson, D.; Sartipy, U.; Braunschweig, F.; Holzmann, M.J.; Hertzberg, D. Acute Kidney Injury Following Coronary Artery Bypass Surgery and Long-term Risk of Heart Failure. *Circ. Hear. Fail.* **2013**, *6*, 83–90. [[CrossRef](#)]
74. Rydén, L.; Sartipy, U.; Evans, M.; Holzmann, M.J. Acute Kidney Injury After Coronary Artery Bypass Grafting and Long-Term Risk of End-Stage Renal Disease. *Circulation* **2014**, *130*, 2005–2011. [[CrossRef](#)] [[PubMed](#)]
75. Chao, C.-T.; Tsai, H.-B.; Wu, C.-Y.; Lin, Y.-F.; Hsu, N.-C.; Chen, J.-S.; Hung, K.-Y. Cumulative Cardiovascular Polypharmacy Is Associated with the Risk of Acute Kidney Injury in Elderly Patients. *Medicine* **2015**, *94*, e1251. [[CrossRef](#)]
76. Ho, K.M.; Power, B.M. Benefits and risks of furosemide in acute kidney injury. *Anaesthesia* **2010**, *65*, 283–293. [[CrossRef](#)] [[PubMed](#)]
77. Verdoodt, A.; Honoré, P.P.M.; Jacobs, R.; De Waele, E.; Van Gorp, V.; De Regt, J.; Spapen, H.D. Do statins induce or protect from acute kidney injury and chronic kidney disease: An update review in 2018. *J. Transl. Intern. Med.* **2018**, *6*, 21–25. [[CrossRef](#)] [[PubMed](#)]

78. Pierson-Marchandise, M.; Gras, V.; Moragny, J.; Micallef, J.; Gaboriau, L.; Picard, S.; Choukroun, G.; Masmoudi, K.; Liabeuf, S. The drugs that mostly frequently induce acute kidney injury: A case—Noncase study of a pharmacovigilance database. *Br. J. Clin. Pharmacol.* **2017**, *83*, 1341–1349. [[CrossRef](#)]
79. Perez-Ruiz, F. Treatment with Allopurinol is Associated with Lower Risk of Acute Kidney Injury in Patients with Gout: A Retrospective Analysis of a Nested Cohort. *Rheumatol. Ther.* **2017**, *4*, 419–425. [[CrossRef](#)] [[PubMed](#)]
80. Kocięcka, M.Z.; Dabrowski, M.; Stepinska, J. Acute kidney injury after transcatheter aortic valve replacement in the elderly: Outcomes and risk management. *Clin. Interv. Aging* **2019**, *14*, 195–201. [[CrossRef](#)] [[PubMed](#)]
81. Ng, R.R.G.; Tan, G.H.J.; Liu, W.; Ti, L.K.; Chew, S.T.H. The Association of Acute Kidney Injury and Atrial Fibrillation after Cardiac Surgery in an Asian Prospective Cohort Study. *Medicine* **2016**, *95*, e3005. [[CrossRef](#)] [[PubMed](#)]
82. Godin, M.; Bouchard, J.; Mehta, R.L. Fluid Balance in Patients with Acute Kidney Injury: Emerging Concepts. *Nephron Clin. Pr.* **2013**, *123*, 238–245. [[CrossRef](#)] [[PubMed](#)]
83. Carrara, C.; Abbate, M.; Sabadini, E.; Remuzzi, G. Acute Kidney Injury and Hemolytic Anemia Secondary to *Mycoplasma pneumoniae* Infection. *Nephron* **2017**, *137*, 148–154. [[CrossRef](#)] [[PubMed](#)]
84. Siew, E.D.; Fissell, W.H.; Tripp, C.M.; Blume, J.D.; Wilson, M.D.; Clark, A.J.; Vincz, A.J.; Ely, E.W.; Pandharipande, P.P.; Girard, T.D. Acute Kidney Injury as a Risk Factor for Delirium and Coma during Critical Illness. *Am. J. Respir. Crit. Care Med.* **2017**, *195*, 1597–1607. [[CrossRef](#)]
85. Evans, R.D.R.; Hemmilä, U.; Craik, A.; Mtekatoka, M.; Hamilton, F.; Kawale, Z.; Kirwan, C.J.; Dobbie, H.; Dreyer, G. Incidence, aetiology and outcome of community-acquired acute kidney injury in medical admissions in Malawi. *BMC Nephrol.* **2017**, *18*, 21. [[CrossRef](#)]
86. Neugarten, J.; Golestaneh, L. Female sex reduces the risk of hospital-associated acute kidney injury: A meta-analysis. *BMC Nephrol.* **2018**, *19*, 314. [[CrossRef](#)]
87. Yokota, L.G.; Sampaio, B.M.; Rocha, E.P.; Balbi, A.; Prado, I.R.S.; Ponce, D. Acute kidney injury in elderly patients: Narrative review on incidence, risk factors, and mortality. *Int. J. Nephrol. Renov. Dis.* **2018**, *11*, 217–224. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).