Predicting Renal Function Post Nephrectomy for Renal Masses

Anna Dominic amd9200@nyu.edu
Chen Chen cc4865@nyu.edu
Julia Manasson manasj02@nyu.edu
Siri Desiraju scd4156@nyu.edu

Abstract

Renal masses often require excision by partial or radical nephrectomy, which results in the removal of healthy tissue and has the potential for renal function impairment. In this study we introduce a predictive model that integrates pre-operative patient characteristics with the type of surgical intervention to predict short-term postoperative renal function. Features were extracted from structured and unstructured electronic health record data at NYU Langone Health. Patients were predominantly white males between the ages of $\sim$50 and 70 years, and a substantial proportion were overweight with comorbidities of hypertension and diabetes. After extensive cleaning and preprocessing of the data, postoperative renal function was modeled using Multiple Linear Regression and XGBoost with and without recursive feature elimination, yielding a mean absolute percentage error of $\sim$18-23$\%$. Upon additional improvements, our model could be employed to predict how well a patient does several weeks after surgery, aiding both clinician and patient in the selection of the most appropriate treatment strategy.

1 Introduction

Kidneys are vital organs that enable a multitude of functions, spanning the gamut of waste product and toxin filtration, drug metabolism, blood pressure control, electrolyte and immune regulation, and erythropoiesis.[1] Many factors contribute to renal function impairment, including the development of a renal mass. A substantial proportion of renal masses are benign,[2−3] but many are malignant, and it is estimated that in 2023 there were 81,800 new cases and 14,890 deaths from kidney cancer in the United States.[4] The management of renal masses differs depending on age, comorbid conditions, tumor pathology, presence of a solitary kidney, as well as patient and physician preference. Two available surgical options are partial nephrectomy (PN), where the mass and surrounding tissue are excised, and radical nephrectomy (RN), where the entire kidney is excised. [2−3] Yet, surgical interventions inevitably remove healthy tissue, which can result in renal function impairment. Although PN is designed to preserve renal tissue, it can be a significantly more complex surgery than RN, translating into longer operating times, higher intraoperative blood loss, and higher rates of postsurgical complications.[5] Given the essential role of kidneys in the maintenance of homeostasis, the ability to predict renal function post nephrectomy is key to choosing the optimal treatment strategy.

2 Related Work

There have been several prior attempts to model postoperative renal function and survival following PN and RN, utilizing a range of approaches such as logistic regression, hierarchical linear mixed models, and hazard ratios.[6−10] Models were built from large patient cohorts and some were externally validated, ensuring robustness. However, most of these approaches are limited by retrospective single-center designs, lack data on renal tumor complexity and intraoperative factors, exclude patients with severe pre-existing renal failure and benign tumors, and have poorly defined postoperative timeframes.

Preprint. Under review.

3 Problem definition and algorithm

3.1 Task

Building on earlier studies, we aimed to leverage machine learning models to predict short-term postoperative renal function in patients undergoing PN or RN. We used structured and unstructured electronic health record (EHR) data to construct a model that predicts either estimated glomerular filtration rate (eGFR) or serum creatinine (Cr) two weeks post-surgery. The features incorporated into our model have known effects on kidney function, or were used in prior studies, and include preoperative kidney function, the type of nephrectomy (partial or radical) performed, and a range of patient characteristics: age, sex, race, Body Mass Index (BMI), tobacco use, and the presence of comorbid conditions hypertension (HTN) and diabetes (DM). To model these relationships, we started with a baseline Multiple Linear Regression algorithm, and ultimately selected XGBoost for its robust performance in handling diverse data types and its adeptness at dealing with missing data.

3.2 Algorithm

We initially built our model using Multiple Linear Regression, but due to its limitations gravitated to XGBoost (Extreme Gradient Boosting), a sophisticated ensemble learning technique known for its ability to enhance predictive accuracy. In this section, we provide an overview of our XGBoost algorithm.

3.2.1 Algorithm overview

XGBoost operates on the principle of boosting, wherein a multitude of weak learners, typically decision trees, are sequentially trained to correct the errors of their predecessors. The model gradually assembles a strong predictive ensemble, where each subsequent learner focuses on minimizing the residuals left by the preceding ones. This iterative process results in a highly accurate and robust predictive model.

3.2.2 Algorithm steps

1. Initialization: Begin with a base model, often a shallow decision tree.
2. Compute residuals: Calculate residuals between the predicted and actual values.
3. Train weak learner: Fit a new weak learner to the residuals, emphasizing areas where the previous model performed poorly.
4. Update predictions: Combine the predictions of all learners, giving higher weight to those that contributed more during training.
5. Iterate: Repeat the process for a predefined number of iterations or until a specified performance threshold is reached.

3.3 Assumptions

The algorithm operates under certain assumptions to ensure its effectiveness and reliability:

- Data preprocessing: The input dataset has been preprocessed to handle missing values effectively and categorical features are appropriately encoded for model compatibility.
- Standardization: Continuous features are standardized to ensure consistent scaling across the dataset, facilitating algorithm convergence.

XGBoost involves a complex interplay of mathematical formulations. Here, we outline the essential equations governing the XGBoost algorithm:

Mathematical Formulas: Objective Function (Loss Function):

$$Obj = \sum_{i=1}^{n} loss(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Regularization Term (Penalty):

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$$

Prediction Calculation:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta h_t(x_i)$$

- $\hat{y}_i^{(t)}$ is the predicted value at iteration t.
- $h_t(x_i)$ is the output of the t-th weak learner for input x_i.
- η is the learning rate, controlling the step size during optimization.

4 Experimental Evaluation

4.1 Data

The data were obtained from NYU Langone Health (NYULH) patient EHR records between the years 2011 and 2023. SQL was used to extract patient-related data from a large data lake. Python scripts employing pandas and numpy libraries were used to clean and preprocess the data. A flowchart of how the data were processed is shown in Figure 1.

We did not have direct access to the data. Instead, a data dictionary was used to glean the structure of EHR records within NYULH and SQL queries generated to extract feature-related information, including demographic (age, sex, race), BMI, medical history (HTN, DM), social history (tobacco use, drug use), laboratory (Cr, eGFR), procedure note, operative note, and pathology report attributes. eGFR and Cr laboratory values were used to represent renal function. We pulled data that were restricted to patients undergoing nephrectomy, and obtained it in batches corresponding to the above-listed attributes. After cleaning and preprocessing, data tables were merged into a single dataframe that was used to train and evaluate our models.

We excluded entries corresponding to pediatric patients (<18 years), individuals with preoperative renal function indicative of end-stage renal disease (e.g., eGFR <15), and patients who did not have both preoperative and postoperative renal function. Furthermore, we required patients to have pathology reports issued within 31 days of their nephrectomy procedure in order to verify the type of nephrectomy performed. We only accessed pathology reports from 2019 and beyond.

For many patients the type of nephrectomy procedure was not specified or incorrectly specified in the extracted procedure name. We therefore relied on free-form text within pathology reports to generate procedure type classifications whenever possible. First, we excluded pathology reports not associated with nephrectomies (e.g., biopsies, non-renal procedures). Next, we generated key words that were specifically associated with RN (e.g., 'ureter', 'radical', 'total') or PN ('inked', 'partial'), and used them to classify the procedures. For individuals without corresponding pathology reports, we used the procedure name when it differentiated between partial or radical nephrectomy, having found it to be accurate in the majority of cases that were cross-referenced with pathology reports (match rates of 86% for PN and 95% for RN). Patients for whom we could not ascertain the type of nephrectomy by pathology report or procedure name were excluded. Patients with transplant kidneys or multiple nephrectomies were also excluded. The type of nephrectomy was maintained as a boolean value.

Several features extracted from the data lake had multiple longitudinal entries associated with different values and timestamps. This was the case for BMI, tobacco use, HTN, and DM. To establish a one-to-one relationship with each patient, we extracted feature values representing a patient's status closest to the date of nephrectomy and prior to its occurrence. BMI was maintained as a continuous variable, while tobacco use, HTN, and DM were represented as boolean variables (either presence or absence). Notably, there were several hundred subtypes of HTN and DM diagnoses in the data lake and presence of any subtype prior to nephrectomy constituted having the diagnosis. For the remaining features, race was grouped into five main categories and maintained as boolean variables. Sex was likewise maintained as a boolean variable, while age was maintained a continuous variable.

Unlike serum Cr, due to the presence of multiple eGFR calculation methods, it was not always possible to match the pre- and postoperative eGFR values. Furthermore, some eGFR values were '>60' and could not be used. However, eGFR is an accepted measure for defining stages of kidney

disease and employed in guideline recommendations. As a result, we analyzed the data as two separate cohorts: the eGFR cohort (more restrictive) and the Cr cohort (less restrictive).

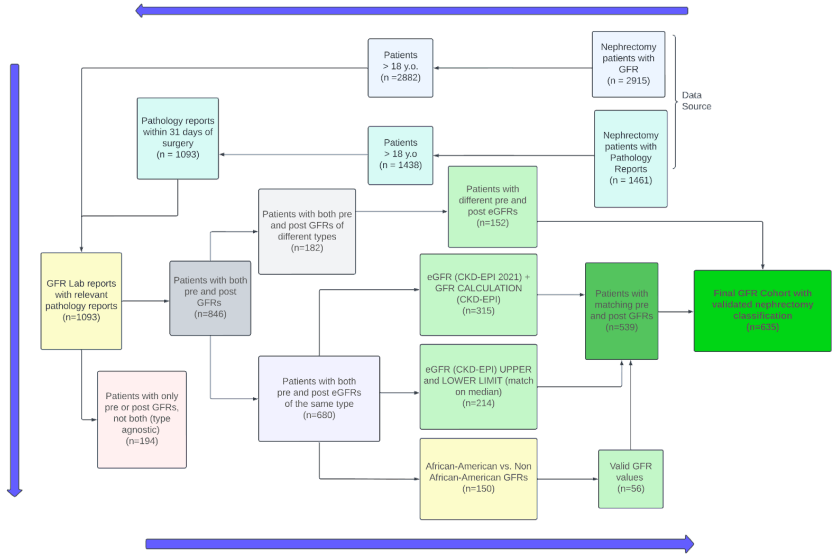

Figure 1: Flowchart of the data processing steps

4.2 Methodology

Multiple Linear Regression and XGBoost models were developed using Python's scikit-learn library and applied to the eGFR and Cr cohorts. The choice of XGBoost was motivated by its capacity to handle missing data, relative interpretability, efficiency with large datasets, and general applicability. Features considered in the modeling process were: age, sex, race, BMI, HTN, DM, tobacco use, nephrectomy type, and preoperative renal function represented by eGFR or Cr. The target variable was postoperative renal function represented by eGFR or Cr. The features (X) and target variable (y) were defined, with categorical features one-hot encoded to binary columns. To ensure uniform scaling, continuous features were standardized using StandardScaler. The dataset was split into training (80%), validation (10%) and testing (10%) sets.

A Multiple Linear Regression model was instantiated, trained on the training set, and subsequently used to predict the target variable on the test set. Model performance was assessed using Mean Squared Error (MSE), R-squared, and Mean Absolute Percentage Error (MAPE).

An XGBoost regression model was constructed with an exhaustive search for optimal hyperparameters using grid search. A grid of hyperparameters was specified, encompassing parameters such as colsample_bytree, learning_rate, max_depth, alpha, and n_estimators. Grid search was performed using cross-validation on the training set, optimizing for the negative MSE. The best model was selected based on the grid search results and evaluated on the validation set. The performance was assessed on the test set using MSE and MAPE.

Recursive Feature Elimination (RFE) was applied in tandem with an XGBoost regression to enhance feature selection. After performing grid search and predefining model hyperparameters, RFE was iteratively employed on the training set to identify the most informative features for the model. The optimal number of features was ascertained by assessing cross-validated performance (via MSE) across various feature subsets. The model was subsequently trained using the selected features on the training set and evaluated on the validation set. Both eGFR and Cr cohorts were further subdivided by race and the best performing model was executed on these groups, providing a nuanced understanding of model performance within distinct demographic categories.

4.3 Results

The baseline characteristics of patients in the the eGFR and Cr cohorts are shown in Table 1. In both cohorts, the patients were predominantly male, white, and overweight. Approximately 60% had HTN, 25-30% had DM, and 40-50% reported tobacco use.

	eGFR Cohort		Cr Cohort	
	PN (n=471)	RN (n=234)	PN (n=527)	RN (n=237)
Age (mean years ± SD)	61.2 ± 12.8	63.6 ± 13.2	61.4 ± 12.4	62.7 ± 13.7
Sex (%)				
Male	293 (62.2%)	141 (60.3%)	332 (63%)	154 (65%)
Female	178 (37.8%)	93 (39.7%)	195 (37%)	83 (35%)
Race/ethnicity (%)				
White	339 (72%)	146 (62.3%)	378 (71.8%)	140 (59.1%)
Black or African American	60 (12.8%)	42 (18%)	67 (12.7%)	43 (18.2%)
Hispanic or Latino or Spanish	19 (4%)	6 (2.6%)	17 (3.2%)	6 (2.5%)
Asian	10 (2.1%)	10 (4.3%)	9 (1.7%)	12 (5%)
American Indian or Native	1 (0.1%)	2 (0.8%)	1 (0.2%)	1 (0.4%)
Other	42 (9%)	28 (12%)	55 (10.4%)	35 (14.8%)
Diabetes (%)	128 (27.2%)	67 (28.6%)	143 (25.8%)	78 (27.3%)
Hypertension (%)	289 (61.3%)	143 (61.1%)	349 (63%)	180 (63%)
BMI (mean BMI ± SD)	29.6 ± 6.2	28.6 ± 5.7	29.8 ± 6.4	28.4 ± 5.5
Smoking (%)	197 (41.8%)	91 (38.9%)	235 (43.4%)	119 (47.4%)

Table 1: eGFR and Cr cohort baseline characteristics split by type of nephrectomy

Results for the Multiple Linear Regression and XGBoost models are shown in Tables 2-4. In general, the XGBoost models performed better for both the eGFR and Cr cohorts, underscoring the advantages derived from employing ensemble methods and capturing non-linear relationships within the data. The application of RFE further refined model performance. Interestingly, only preoperative serum creatinine, type of nephrectomy, and African-American race were selected as important features for the Cr cohort, while all features were found to be important for the eGFR cohort. The eGFR RFE-XGBoost model demonstrated a marginally lower MAPE compared to its non-RFE counterpart, and the Cr RFE-XGBoost model also achieved a notable reduction in MAPE, highlighting the efficacy of feature selection in the enhancement of predictive accuracy in the set of patients.

Metric	**eGFR**	**Cr**
MSE	233.526	1.277
R-squared	0.607	0.514
MAPE	23.70%	26.68%

Table 2: Performance metrics of the Multiple Linear Regression model

Metric	**eGFR**	**Cr**
MSE	166.819	0.805
MAPE	20.20%	22.98%
Best Hyperparameters	{ 'alpha': 1, 'colsample_bytree': 0.3, 'learning_rate': 0.2, 'max_depth': 3, 'n_estimators': 50 }	{ 'alpha': 1, 'colsample_bytree': 0.3, 'learning_rate': 0.2, 'max_depth': 3, 'n_estimators': 100 }

Table 3: Performance metrics of the XGBoost model using all features

Metric	**eGFR**	**Cr**
MSE	211.683	0.8557
MAPE	18.81%	20.55%
Features Selected	All except race indicators for Asian, Hispanic	BMI, pre-creatinine, Type of surgery (partial or radical), Race indicator for white

Table 4: Performance metrics of the XGBoost model using RFE

Results across specific patient demographics are shown in Table 5. The consistency in the robust performance of both eGFR and Cr models among white patients is evident and manifested by lower MSE and MAPE values. Conversely, the models' performance for patients of color (POC), particularly in the Cr cohort, reveals higher MSE and MAPE values. This observation implies potential complexities and challenges in accurately predicting renal function within this demographic subgroup, warranting further investigation and consideration of specific contextual factors.

Metric	eGFR (White)	eGFR (POC)	Cr (White)	Cr (POC)
MSE	204.7687	351.3271	0.2374	0.9837
MAPE	18.93%	20.94%	18.99%	20.06%
Features Selected	All	All	All except social history	All

Table 5: Performance metrics of the XGBoost model across patient demographics

Overall, the XGBoost models, especially when augmented with feature selection techniques, demonstrate superior predictive capabilities for both eGFR and Cr cohorts. The nuanced differences in performance between the two cohorts highlight the need for tailored modeling approaches based on the specific renal biomarker under consideration.

4.4 Discussion

To simulate a real-world situation, we applied our best model to two types of patients undergoing PN–one who had minimal change in renal function and one who had significant decline in renal function post surgery. Our model performed fairly well in the case where renal function remained stable (Cr actual 0.81, Cr predicted 0.65, eGFR actual 82.9, eGFR predicted 78.7). However, it performed quite poorly in the case where significant change in renal function was observed (Cr actual 1.76, Cr predicted 0.93, eGFR actual 49.5, eGFR predicted 90.1). This highlights one important limitation of the model, which is a training set that underrepresented cases of renal function instability post nephrectomy.

Unlike prior studies, our cohort was relatively more diverse, allowing the examination of model performance across different demographic categories. Furthermore, we had a clearly defined outcome timepoint, which is reflective of how well patients do immediately after surgery. Similar to prior studies, our approach was also based on a retrospective single-center design and did not include information about renal mass size, complexity, and location. Additionally, HTN and DM were treated as binary variables and there was no differentiation between a well-controlled vs poorly-controlled state, which can have significant impact on kidney function. Finally, our cohorts were restricted to patients where the type of nephrectomy could be confirmed from the procedure name or newer pathology report records, as well as those who had laboratories performed within NYULH, discounting a sizeable proportion of the available population.

Moving forward, our models could be improved by enlarging the cohort sizes through the incorporation of older pathology reports and laboratory values from notes, the addition of features related to tumor characteristics and intraoperative complications or lack thereof, the identification of subjects with solitary kidney prior to surgery, and the refining of HTN and DM features. Our models could further be adjusted to predict intermediate and long-term renal function in addition to short-term renal function.

5 Conclusion

EHR data is complex and messy. Despite these challenges, we were able to effectively extract data from a large data lake without having direct access, clean and preprocess the data in a biologically-meaningful manner, making critical decisions along the way, and generate an initial model with reasonable predictive power when presented with real-world cases. With modest improvements to the training data and features, this model has the potential to be deployed as a tool in selecting the most appropriate treatment strategy for patients with renal masses.

6 Lessons Learned

This project allowed us to experience, first-hand, the challenges and inconsistencies of EHR records, requiring extensive cleaning, preprocessing, meticulous scrutiny, and validation. Working with these data is an iterative process and requires significant domain knowledge to make medically-relevant assumptions and decisions. The data contain biases, which can limit its interpretability within the context of a larger population. Most of all, it requires patience and careful planning to achieve meaningful results.

References

[1] Eckardt KU, Coresh J, Devuyst O, Johnson RJ, Köttgen A, Levey AS, Levin A. Evolving importance of kidney disease: from subspecialty to global health burden. Lancet. 2013 Jul 13;382(9887):158-69. doi: 10.1016/S0140-6736(13)60439-0. Epub 2013 May 31. PMID: 23727165.

[2] Campbell SC, Clark PE, Chang SS et al: Renal Mass and Localized Renal Cancer: Evaluation, Management, and Follow-Up: AUA Guideline Part I. J Urol 2021; 206: 199.

[3] Campbell SC, Uzzo RG, Karam JA, et al: Renal Mass and Localized Renal Cancer: Evaluation, Management, and Follow-up: AUA Guideline: Part II. J Urol 2021; 206: 209.

[4] Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin. 2023 Jan;73(1):17-48. doi: 10.3322/caac.21763. PMID: 36633525.

[5] Huang R, Zhang C, Wang X, Hu H. Partial Nephrectomy Versus Radical Nephrectomy for Clinical T2 or Higher Stage Renal Tumors: A Systematic Review and Meta-Analysis. Front Oncol. 2021 Jun 10;11:680842. doi: 10.3389/fonc.2021.680842. PMID: 34178668; PMCID: PMC8222682.

[6] Roussel E, Laenen A, Bhindi B, De Dobbeleer A, Stichele AV, Verbeke L, Van Cleynenbreugel B, Sprangers B, Beuselinck B, Van Poppel H, Joniau S, Albersen M. Predicting short- and long-term renal function following partial and radical nephrectomy. Urol Oncol. 2023 Feb;41(2):110.e1-110.e6. doi: 10.1016/j.urolonc.2022.10.006. Epub 2022 Nov 10. PMID: 36372636.

[7] Bhindi B, Lohse CM, Schulte PJ, Mason RJ, Cheville JC, Boorjian SA, Leibovich BC, Thompson RH. Predicting Renal Function Outcomes After Partial and Radical Nephrectomy. Eur Urol. 2019 May;75(5):766-772. doi: 10.1016/j.eururo.2018.11.021. Epub 2018 Nov 23. PMID: 30477983.

[8] Aguilar Palacios D, Wilson B, Ascha M, Campbell RA, Song S, DeWitt-Foy ME, Campbell SC, Abouassaly R. New Baseline Renal Function after Radical or Partial Nephrectomy: A Simple and Accurate Predictive Model. J Urol. 2021 May;205(5):1310-1320. doi: 10.1097/JU.0000000000001549. Epub 2020 Dec 24. PMID: 33356481.

[9] Schmid M, Abd-El-Barr AE, Gandaglia G, Sood A, Olugbade K Jr, Ruhotina N, Sammon JD, Varda B, Chang SL, Kibel AS, Chun FK, Menon M, Fisch M, Trinh QD. Predictors of 30-day acute kidney injury following radical and partial nephrectomy for renal cell carcinoma. Urol Oncol. 2014 Nov;32(8):1259-66. doi: 10.1016/j.urolonc.2014.05.002. Epub 2014 Aug 14. PMID: 25129142.

[10] Chan VW, Abul A, Osman FH, Ng HH, Wang K, Yuan Y, Cartledge J, Wah TM. Ablative therapies versus partial nephrectomy for small renal masses - A systematic review and meta-analysis. Int J Surg. 2022 Jan;97:106194. doi: 10.1016/j.ijsu.2021.106194. Epub 2021 Dec 24. PMID: 34958968.

7 Student Contribution

Anna Dominic:

- Implemented and evaluated multiple linear regression and XGBoost models on eGFR and Creatinine cohorts, subgroup analysis and interpeted results.
- Employed techniques for predictive analysis, contributed to initial preprocessing.

Chen Chen:

- Subset GFR and creatinine cohorts according to exclusion criteria and validity of matching pathology report.
- Join GFR and creatinine cohorts with tobacco use, BMI and diagnosis of hypertension and diabetes to create final dataset for modeling

Julia Manasson:

- Wrote SQL queries to pull data from the data lake
- Classified nephrectomy procedures using pathology reports and procedure names
- Identified appropriate subcategories of HTN and DM diagnoses

Siri Desiraju :

- Wrote SQL queries to pull data from the data lake
- Classified nephrectomy procedures using pathology reports and procedure names
- Distilled the feature records to establish a one-to-one relationship, ensuring that each patient's feature values were derived from the closest date to the surgery.

All students have contributed equally in literature review, making the poster and the final report.

8 Acknowledgements

We are immensely grateful to Dr. Madhur Nayan MD PhD, whose mentorship has been invaluable throughout this capstone journey. His guidance and expertise have not only shaped the trajectory of our research but have also inspired a deeper understanding of the subject matter. Special thanks to our dedicated guides, Brian McFee, Assistant Professor of Music Technology and Data Science, whose insights propelled our project to new heights, and to Jacopo Cirrone, Elisha Cohen, and Saadia Gabriel, esteemed Data Science Faculty Fellows, whose unwavering support and encouragement have been instrumental in making this capstone a truly enriching and transformative learning experience. We would also like to thank NYU Langone for its resources.