# Clinical Phenotyping in the Prediction of Pediatric Acute Kidney Injury

Michael G. Semanik

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2016

Committee:

Meliha Yetisgen

David Crosslin

Sangeeta Hingorani

Program Authorized to Offer Degree:

Biomedical and Health Informatics

University of Washington

**Abstract**

Clinical Phenotyping in the Prediction of Pediatric Acute Kidney Injury

Michael G. Semanik

Chair of the Supervisory Committee:
Meliha Yetisgen, PhD, Associate Professor
Biomedical and Health Informatics

Predicting pediatric acute kidney injury is a difficult but important task. Accurate prediction would allow preventative measures to be taken before kidney injury occurs, decreasing the morbidity and mortality associated with this disease. This work describes the process of creating an "at risk for AKI" clinical phenotype from electronic health record data, which is then used to predict AKI in a retrospective data set. This predictive model has reasonable performance, with an F1 score of 0.67 and AUC of 0.75. In a subset of intensive care unit patients, the addition of unstructured data from clinician notes improves the model's F1 score to 0.72 and AUC to 0.77, suggesting a possible role for natural language processing in refining clinical phenotypes. Interpreting these models requires careful consideration of the information contained within each variable – specifically, the extent to which that information describes biologic processes within a patient or systemic processes within a hospital. Further evaluation of the use of clinical phenotyping in predicting pediatric AKI is necessary to confirm the utility of these models.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGEMENTS

The author would like express his heartfelt gratitude to the member of his committee, Drs. Meliha Yetisgen, David Crosslin, and Sangeeta Hingorani, for their invaluable contributions to this work. He would also like to thank Dr. Ari Pollack, Assaf Oron, Daksha Ranade, and each of the members of the UW-BioNLP group for their patience, insights, and general willingness to listen to his thoughts on this project at various and sundry times.

# Chapter 1. INTRODUCTION

Hospital acquired acute kidney injury (AKI) is an increasingly common and costly problem amongst pediatric inpatients. Although the hospital wide incidence of pediatric AKI has been difficult to determine, a review of the Kids' Inpatient Database (KID) found that an AKI discharge diagnosis occurred at a rate of 3.9 cases per 1000 admissions, meaning AKI affects more than 10,000 children annually in the United States [1]. Incident rates are even higher within various pediatric inpatient subpopulations, ranging from 10% of pediatric ICU patients [2] to 56% of cardiac surgery patients [3]. Furthermore, children diagnosed with AKI have worse outcomes, with studies suggesting that median length of stay increases by a week, and that mortality risk increases 5 to 25-fold [1,4,5]. Thus, early identification and prevention of hospital-acquired AKI is extremely desirable.

Since 2004, pediatric AKI has been diagnosed using criteria from one of four guidelines: RIFLE [6], pRIFLE [7], AKIN [8], or KDIGO [9]. Although the guidelines do not agree upon an exact definition of AKI, in general each characterizes AKI in terms of either a rising serum creatinine or a decrease in urine output. This characterization has the benefit of simplicity and practicality, as both serum creatinine and urine output measurements are routinely ordered and easily interpreted by clinicians. However, serum creatinine and urine output are not truly markers of injury, but rather indicate decreased renal function. Indeed, by the time creatinine has risen or urine output fallen, the kidney damage has already been done. This "late marker" problem has largely frustrated efforts to prevent AKI, and has spurred research and development into earlier, more injury-focused AKI biomarkers, such as NGAL [10], IL-18 [1], KIM-1 [12], and L-FABP [13]. Early results for these novel biomarkers are promising, but until they are

validated and accepted into clinical practice, providers remain frustratingly behind the curve when it comes to diagnosing and treating acute kidney injury.

The work described here attempts to address this deficiency. It is based on the hypothesis that the wealth of data routinely collected in electronic health records (EHRs) – specifically demographic information, procedural information, medications, laboratory values, vital signs, and the unstructured data found in clinician notes – can be combined to create an "at risk for AKI" clinical phenotype, which can then be used to predict which patients will develop AKI before functional markers change. Earlier identification will hopefully improve prevention efforts and ultimately decrease the morbidity and mortality associated with AKI.

# Chapter 2. RELATED WORK

The use of modeling to predict clinical outcomes is not new in pediatric medicine. Much work has been done in the pediatric intensive care population, where scores such as the PRISM-III [14] and PELOD-2 [15] combine clinical data to predict a patient's mortality risk. These scores generally perform quite well, with area-under-the-curve scores (AUCs) of 0.94 for the PRISM-III and 0.93 for the PELOD-2. Other models, broadly termed Pediatric Early Warning Scores (PEWS), have been used in emergency departments to predict which patients need to be admitted to the hospital generally and/or to the ICU specifically. A review of ten of these PEWS systems found that their sensitivities for pediatric ICU admission ranged from 0.61 to 0.94, with AUCs of 0.60 to 0.82 [16]. The sensitivities and AUCs were lower for hospital admission (0.28 to 0.86 and 0.56 – 0.68, respectively).

There has also been a recent trend in using EHR data to characterize specific clinical outcomes, known as clinical phenotyping. Chen et al provide an excellent review of the recent

advances in both supervised and unsupervised phenotype identification before presenting their own results, which use a latent Dirichlet allocation model to infer phenotypic types at two separate hospital systems [17]. The identified phenotypes were consistent within hospital systems, and showed less variance between hospital systems than ICD-9 coding, suggesting that clinical phenotyping may be a better way to categorize patients than reliance upon billing codes.

Clinical phenotyping has utility beyond providing better discharge diagnoses, however. Researchers have begun to leverage phenotypes in predictive modeling – that is, they seek to create an "at risk for a disease" phenotype as opposed to a "had a disease" phenotype. In pediatrics, much of this work has focused on sepsis, which like AKI is associated with significant morbidity and mortality. One such "at risk for sepsis" phenotype for use in neonates had an AUC of 0.80 [18], whereas a similar model for use in the broader intensive care population had an AUC of 0.87 [19]. These successes in phenotyping sepsis suggest that this approach can be applied to other pediatric disease states as well.

One such disease state is pediatric acute kidney injury. However, clinical phenotyping of AKI is in its infancy, and the few studies that exist have generally been confined to the cardiac surgery subpopulation [20,21,22], which is unique in that the exact timing of AKI is known (it is presumed to be the time of cardiac bypass). Thus, results from these studies may not be applicable the broader pediatric population. However, a few studies have attempted to characterize AKI in the overall ICU population. One such study develops the Renal Angina Index, which attempts to predict whether or not an ICU patient will develop severe AKI on day 3 of hospitalization based on three features from ICU admission – the change in estimated creatinine clearance from baseline, the amount of fluid overload, and the use of vasopressors [23]. This sketches a rough outline of a potential "at risk for AKI" phenotype, but is still

difficult to apply to the pediatric inpatient population at large (in which determination of a baseline creatinine clearance and fluid overload may be difficult). In the ICU population, however, the RAI performed well, with AUCs ranging from 0.74 – 0.81 in validation cohorts (although the positive predictive values of the score were lower, ranging from 0.18 – 0.39).

More recently, a Pediatric Early AKI Risk Score was developed, also for use in the pediatric ICU population [24]. This phenotype involved seven variables, and demonstrated a slight improvement in AUC compared to the Renal Angina Index (AUCs in the validation cohort were 0.81 – 0.86). However, positive predictive values were even lower (0.07). Use of either tool in practice, then, would generate a large number of false positives, potentially limiting their clinical utility.

An additional shortcoming of the above phenotyping methods is that they are time-limited; both are designed to predict the likelihood of AKI on Day 3 of hospitalization using data from ICU admission, but neither has been validated at other time points. Phenotypes that are temporally related to the event of interest rather than tied to a single point during a hospitalization would likely prove more robust in clinical practice. Given the above studies' temporal restrictions – as well as their population restrictions and difficulties with positive predictive values – better phenotypes of pediatric AKI risk are still required.

In a general sense, creating a better AKI risk estimate requires identifying better predictors. These predictors may exist as either *structured data* (i.e., data entered into discrete fields within an EHR, such as vital signs, lab values, demographic information, procedural information, and medication doses) or *unstructured data* (i.e., free-text data, such as the assessment and plan of a clinician's note). Each of the pediatric AKI risk scores described above use some form of structured data, but contain only a few predictors (three for the RAI and seven

for the Pediatric Early AKI Risk Score). Therefore, additional structured predictors, the incorporation of unstructured predictors, or some combination thereof could create a more granular "at risk for AKI" phenotype with better predictive performance.

However, utilizing unstructured data is not straightforward, and requires at least some basic natural language processing (NLP). For AKI this is somewhat uncharted territory, as there have been no studies evaluating the use of NLP in predicting AKI in either adults or children. Fortunately, NLP has shown promise in other medical conditions: one study utilizing both structured and unstructured EHR data successfully identified multiple sclerosis patients with a sensitivity of 82.7% and positive predictive value of 92.1% (though determination of disease severity proved more difficult, in that the model's predicted severity scores had only moderate correlation – $R^2 = 0.38$ – with the true severity scores) [25]. Another study was able to show excellent prediction of ventilator associated pneumonia among adult ICU patients using NLP of chest x-ray reports obtained prior to the event of interest, with a sensitivity of 91.4% and positive predictive value of 77.2% [26]. Therefore, there is reason to believe that the use of unstructured data may augment an AKI prediction system as well.

The aims of this work, then, are three-fold: firstly, to use structured clinical data to identify a robust "at risk for AKI" phenotype for all pediatric inpatients; secondly, to evaluate the performance of that phenotype in predicting acute kidney injury; and finally, to add unstructured data from clinician notes to improve the phenotype's predictive performance. The outcome will hopefully be a well-defined clinical phenotype with enhanced clinical utility (i.e., high sensitivity and positive predictive value) that can be used at nearly any point during a patient's hospital stay.

# Chapter 3. METHODS

The following provides details regarding the dataset used in this work, as well as the development of the models used to test the predictive performance of an "at risk for AKI" phenotype.

## 3.1 STUDY POPULATION

Patients admitted to Seattle Children's Hospital from 3/1/2012 – 2/28/2015 who met the following criteria were included in the dataset:

1) Patients had at least 2 creatinine values recorded in the Seattle Children's Hospital electronic health record between 9/1/2011 and 2/28/2015;
2) Patients were at least 12 months of age;
3) Patients were admitted to the hospital for at least 72 hours;
4) Patients were not admitted to the Inpatient Psychiatric Unit or Rehabilitation Unit;
5) Patients did not have a diagnosis of end-stage renal disease (ICD-9-CM codes 585.6, V45.11, V56.31, V56.32, V56.1, V56.2) or renal transplant (ICD-9-CM codes V42.0, 996.81).

If a patient was admitted to the hospital multiple times during the study period, only data from the first hospitalization was used.

Electronic clinician notes were only available for ICU patients; therefore, admission notes and progress notes were obtained for a randomly selected subset of ICU patients. Because these patients were part of the original dataset, they also met the above inclusion criteria.

## 3.2 DEFINING AKI

Baseline creatinine was defined as the lowest creatinine within the period ranging from 6 months prior to hospital admission to 14 days after hospital admission.

Several definitions of AKI were used, all based upon KDIGO guidelines [8]. The first, termed "KDIGO Stage 1", defines AKI as either a creatinine greater than or equal to 1.5 times

the baseline creatinine, or an absolute increase in creatinine of 0.3 mg/dL, whichever is lower. The second, "Modified KDIGO Stage 1", uses only the absolute increase in creatinine of 0.3 mg/dL as the AKI cutoff. The third, "KDIGO Stage 2", defines AKI as a creatinine greater than or equal to twice the baseline. Finally, "KDIGO Stage 3" defines AKI as greater than or equal to thrice the baseline. Urine output-based definitions of AKI were not used due to incomplete urine output records.

## 3.3 MODEL DESIGN

Because AKI is defined as an increase in serum creatinine, the ultimate question this work attempts to answer is "which variables are associated with an increased serum creatinine?" To be useful, these variables have to be obtained in a discrete timeframe – a Data Collection Window – that must occur some length of time – the Prediction Window – before the creatinine in question is measured. This concept is demonstrated in Figure 3.1.

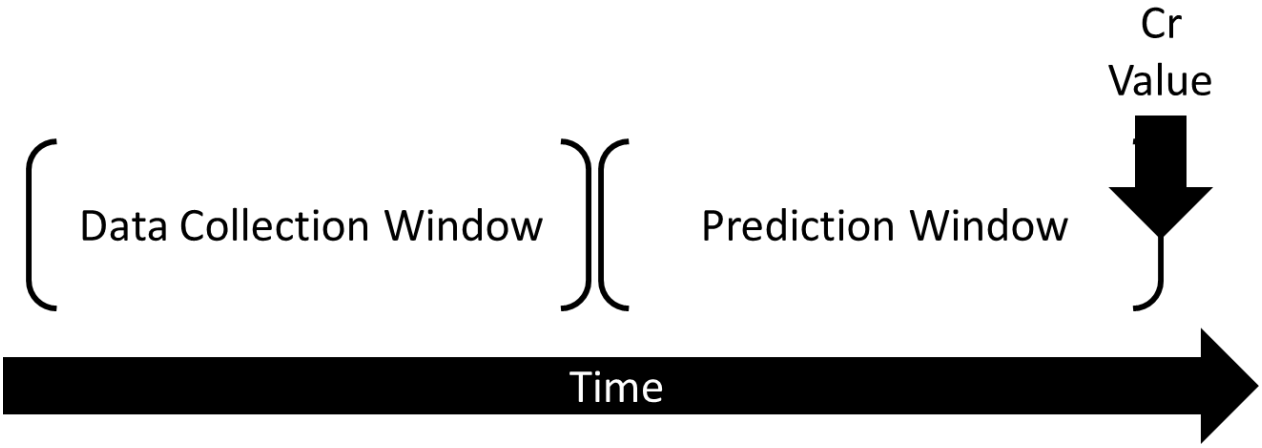


Figure 3.1. The data associated with each creatinine value is collected during a 24 – 48 hour Data Collection Window, which occurs at least 24 – 48 hours before the creatinine value's timestamp (the Prediction Window). The length of the Data Collection Window and the Prediction Length always sum to 72 hours.

Although it is generally agreed that serum creatinine is a "late marker" of injury, *how* late remains unclear, so three models were created in which the Prediction Length was variably set to

24, 36, or 48 hours. The length of the Data Collection Window also varied between 24, 36, and 48 hours, such that the total amount of time considered always summed to 72 hours. This ensured that the same creatinine values were considered in each model. Finally, to be included in the dataset, a creatinine value's Data Collection Window could extend to no more than 12 hours before hospital admission (which allowed data from emergency department visits prior to hospital admission to be captured). This restriction effectively excluded any creatinine values obtained prior to 60 hours of hospitalization.

Each creatinine value in the dataset was treated as an individual case. However, because multiple creatinine values are often obtained for an individual patient, a distinction must be made between patient-level variables (those variables which have the same value for every creatinine associated with that patient), and creatinine-level variables (those variables which differ in value for each creatinine associated with a patient). Of the 147 structured variables considered, only four were patient-level variables: age (at hospital admission), gender, admitting service, and whether or not the patient had experienced an AKI episode in the six months prior to hospital admission (termed "Prior AKI"). The remaining 143 variables were specific to an individual creatinine value, and can be subdivided into five categories – medications, procedures, labs, vital signs, and other – that are further delineated in Table 4.2 and Appendix A. As mentioned, unstructured variables were obtained only for a subset of ICU patients, and are considered in the "Unstructured Data Models" section below.

## 3.4 STATISTICS

Because every creatinine value was considered as a separate case, it is possible for a single patient to have multiple creatinine values consistent with AKI. For example, consider the following patient:

Patient AB
Date of Admission: 2/1/2001 Date of Discharge: 2/8/2001
Baseline creatinine (from 1/1/2001): 0.3

Creatinine Values:

| | 2/1/2001 | 2/2/2001 | 2/3/2001 | 2/4/2001 | 2/5/2001 | 2/6/2001 | 2/7/2001 |
|---|---|---|---|---|---|---|---|
| **Serum Cr** | 0.4 | 0.4 | 0.6 | 0.7 | 0.4 | 0.4 | 0.7 |
| **AKI?*** | n/a** | n/a** | Yes | Yes | No | No | Yes |

* Stage 1 KDIGO Criteria are used to define AKI for the purposes of this example.
** These values are not considered because their Data Collection Windows do not fall within 12 hours of hospital admission.

As can be seen, multiple creatinine values meet AKI criteria and should be classified as such. However, there is almost certainly a clinical distinction between *new-onset AKI* (that is, the first creatinine values crossing the AKI threshold, such as the values from 2/3/2001 and 2/7/2001 in this example) and *ongoing AKI* (subsequent creatinine values that meet criteria, such as the value from 2/4/2001). To account for this, a binary interaction term was introduced. This term was positive if the creatinine value in question had no other elevated creatinine value obtained during the Data Collection Window (i.e., the creatinine value represented new-onset AKI). The aforementioned 147 variables were each multiplied by this interaction term, producing a total of 294 variables that were then used as predictors in the final structured data models. Because the interaction term was binary, 147 variables applied to all creatinine values and 147 applied only to new-onset AKI. Any missing values in the dataset were imputed using medians.

Patients were randomly stratified into a training set (75% of the data) and a test set (the remaining 25%) using a patient-level identifier. All statistical analyses were performed in R, using the glmnet package. Specifically, an elastic-net logistic regression (alpha = 0.5) was performed, with AKI as the response variable and the aforementioned 294 variables as

predictors. 10-fold cross-validation was used on the training set. Models were tuned to maximize the F1 score (the harmonic mean of the sensitivity and positive predictive value). The maximum F1 score was targeted to ensure that the model valued positive prediction value and sensitivity over specificity. This was felt to be important because AKI is a relatively rare event; with rare events, it is relatively easy to generate a high specificity but more difficult to produce a high positive predictive value. Targeting high sensitivities and positive predictive values also ensures that the number of false negatives and false positives are minimized.

Because there is no standard way of calculating 95% confidence intervals for odds ratios produced by elastic-net logistic regression, a standard logistic regression was run with the features selected by the elastic-net, and these odds ratios and 95% confidence intervals are reported. Values are considered significant at $p < 0.05$.

Unstructured models were constructed and analyzed in the same way, and are described in more detail below.

## 3.5 UNSTRUCTURED DATA MODELS

As mentioned, electronic clinician notes were obtained for a random subset of ICU patients (the only patient population for which clinician notes was available during the timespan of the study). Two types of electronic notes were considered: ICU admission notes and ICU progress notes. For the purposes of generating features, however, both admission notes and progress notes were treated as single corpus. This corpus was processed as follows:

1) Deidentification was performed using the National Library of Medicine's NLM-Scrubber [27];
2) The billing information contained in each note was removed;
3) Section headers were removed (for a full list, please see Appendix B);
4) Documents were chunked into sentences using the Punkt tokenizer from the Natural Language Toolkit (NLTK) for Python;
5) Stop words were removed (for a full list, please see Appendix B);

6) Punctuation was removed using regular expressions, and all words were converted to lower case;
7) Unigrams, bigrams, and trigrams were generated, and their frequency tabulated;
8) Any unigrams, bigrams, or trigrams occurring in less than 2.5% of patients were excluded (the 2.5% threshold gave the best performance, results for other thresholds are reported in Appendix C);
9) Feature selection was performed and n-grams ranked using chi-squared. Ultimately the top 225 n-grams were included in the final model, as this threshold provided the best performance. The performances of other thresholds are reported in Appendix C.

The above steps, with the exception of deidentification, were performed in Python 2.7.11. The frequencies associated with each n-gram were then converted to a binary variable that indicated whether or not that n-gram occurred in notes associated with each individual creatinine value. N-gram data was then imported into R, where it was combined with the structured variables described above to create a "combination model" consisting of both structured and unstructured data. Once again, an interaction variable was associated with each of the selected n-grams to differentiate between new-onset and ongoing AKI. Cases remained divided into the same training (75%) and test (25%) sets used for generation of the structured models. The same elastic-net logistic regression model (alpha = 0.5) was run, with AKI as the outcome and the structured variables and n-grams as predictors. 10-fold cross-validation was once again performed on the training set, and models were once again tuned to maximize F1 scores. Odds ratios and 95% confidence intervals are reported from a standard logistic regression model using the features selected from the elastic-net, and are considered significant at $p < 0.05$.

# Chapter 4. RESULTS

Results of both purely structured data models and models consisting of structured and unstructured data are reported in this section.

## 4.1 STRUCTURED DATA MODEL RESULTS

The best performing structured model was that which had a Data Collection Window length of 48 hours, a Prediction Window length of 24 hours, and used KDIGO Stage 1 criteria to define AKI. The results of this model will be discussed here, and details regarding other models can be found in Appendix D. A total of 2428 patients with 10636 creatinine values were included in this model, of which 3565 (33.5%) had AKI using KDIGO Stage 1 criteria. Table 4.1 compares several characteristics of the training and test sets; in general, the two sets were similar with regards to the proportion of overall creatinine values meeting AKI criteria, the proportion of new-onset AKI values, and the proportion of ongoing AKI values. The two sets also had similar mean ages (10.1 years for the training set, 9.6 years for the test set) and male to female ratios (57.0% male for both). The mean age of the overall cohort was 10.0 years, and 57.0% of the overall cohort was male.

Table 4.1. Summary of training and test sets.

| | **Mean Age** | **% Male** | **% AKI** | **% AKI New-onset** | **% AKI Ongoing** |
|---|---|---|---|---|---|
| **Training (n = 8121)** | 10.1 years | 57.0% | 33.4% | 13.1% | 20.3% |
| **Test (n = 2515)** | 9.6 years | 57.0% | 33.8% | 12.2% | 21.6% |

The structured variables used as predictors in each model and the percent of missing data for each variable are presented in Table 4.2. The only variable with greater than 50% missing values was glucose standard deviation, at 55.5%. On average, only 6.3% of the dataset was missing (and therefore imputed as median values). For labs and vital signs, if the mean values were missing, the values for "% High" and "% Low" were classified as missing as well.

Table 4.2. Summary of variables in the AKI prediction model. The number and percent of missing values are reported for the dataset with a 48 hour Data Collection Window and 24 hour Prediction Window; there are minor differences in the other models.

| Variable | # Missing (%) | Variable | # Missing (%) |
|---|---|---|---|
| **Medications (36)** | | *Hematocrit* | |
| Acyclovir | 0 (0.0%) | Mean | 2095 (19.7%) |
| Aspirin | 0 (0.0%) | Std. Dev. | 4451 (41.8%) |
| Bleomycin | 0 (0.0%) | % High | 2095 (19.7%) |
| Captopril | 0 (0.0%) | % Low | 2095 (19.7%) |
| Carboplatin | 0 (0.0%) | *Platelet Count* | |
| Ceftazidime | 0 (0.0%) | Mean | 2329 (21.9%) |
| Cisplatin | 0 (0.0%) | Std. Dev. | 4674 (43.9%) |
| Cyclosporine | 0 (0.0%) | % High | 2329 (21.9%) |
| Cytarabine | 0 (0.0%) | % Low | 2329 (21.9%) |
| Enalapril | 0 (0.0%) | *Other* | |
| Epinephrine | 0 (0.0%) | Vanc level mean | 531 (0.05%) |
| Furosemide | 0 (0.0%) | Vanc level % High | 531 (0.05%) |
| Ganciclovir | 0 (0.0%) | Gent level mean | 0 (0%) |
| Gentamicin | 0 (0.0%) | Gent level % High | 0 (0%) |
| Ibuprofen | 0 (0.0%) | Tacro level mean | 72 (0.7%) |
| Indomethacin | 0 (0.0%) | Cyclosporine level mean | 111 (1.0%) |
| Ioversol | 0 (0.0%) | Electrolytes frequency | 1576 (14.8%) |
| Ketorolac | 0 (0.0%) | CBC frequency | 2351 (22.1%) |
| Lisinopril | 0 (0.0%) | **Vital Signs (47)** | |
| Losartan | 0 (0.0%) | *Heart Rate* | |
| Meloxicam | 0 (0.0%) | Mean | 142 (1.3%) |
| Mesalamine | 0 (0.0%) | Std. Dev. | 145 (1.4%) |
| Methotrexate | 0 (0.0%) | % High | 142 (1.3%) |
| Naproxen | 0 (0.0%) | % Low | 142 (1.3%) |
| Neomycin | 0 (0.0%) | Mean change | 72 (0.7%) |
| Pamidronate | 0 (0.0%) | Std. Dev. Change | 76 (0.7%) |
| piperacillin-tazobactam | 0 (0.0%) | % High change | 72 (0.7%) |
| Sirolimus | 0 (0.0%) | % Low change | 72 (0.7%) |
| Tacrolimus | 0 (0.0%) | *Respiratory Rate* | |
| Tobramycin | 0 (0.0%) | Mean | 142 (1.3%) |
| Torsemide | 0 (0.0%) | Std. Dev. | 144 (1.4%) |
| Valacyclovir | 0 (0.0%) | % High | 142 (1.3%) |
| Valganciclovir | 0 (0.0%) | % Low | 142 (1.3%) |
| Valsartan | 0 (0.0%) | Mean change | 72 (0.7%) |
| Vancomycin | 0 (0.0%) | Std. Dev. Change | 74 (0.7%) |

| | | | |
|---|---|---|---|
| Number of meds | 0 (0.0%) | % High change | 72 (0.7%) |
| **Procedures (6)** | | % Low change | 72 (0.7%) |
| Central Line | 0 (0.0%) | *Systolic Blood Pressure* | |
| Cardiac Catheterization | 0 (0.0%) | Mean | 143 (1.3%) |
| Heart Surgery | 0 (0.0%) | Std. Dev. | 155 (1.4%) |
| ECMO | 0 (0.0%) | % High | 143 (1.3%) |
| CT/MRI | 0 (0.0%) | % Low | 143 (1.3%) |
| Procedure time | 0 (0.0%) | Mean change | 73 (0.7%) |
| **Labs (48)** | | Std. Dev. Change | 85 (0.8%) |
| *Sodium* | | % High change | 73 (0.7%) |
| Mean | 1498 (14.1%) | % Low change | 73 (0.7%) |
| Std. Dev. | 3302 (31.0%) | *Diastolic Blood Pressure* | |
| % High | 1498 (14.1%) | Mean | 143 (1.3%) |
| % Low | 1498 (14.1%) | Std. Dev. | 155 (1.4%) |
| *Potassium* | | % High | 143 (1.3%) |
| Mean | 1511 (14.2%) | % Low | 143 (1.3%) |
| Std. Dev. | 3306 (31.1%) | Mean change | 73 (0.7%) |
| % High | 1511 (14.2%) | Std. Dev. Change | 85 (0.8%) |
| % Low | 1511 (14.2%) | % High change | 73 (0.7%) |
| *Chloride* | | % Low change | 73 (0.7%) |
| Mean | 1535 (14.4%) | *Temperature* | |
| Std. Dev. | 3422 (32.2%) | Mean | 159 (1.5%) |
| % High | 1535 (14.4%) | Std. Dev. | 188 (1.8%) |
| % Low | 1535 (14.4%) | % High | 159 (1.5%) |
| *Bicarbonate* | | % Low | 159 (1.5%) |
| Mean | 1552 (14.6%) | Mean change | 94 (0.9%) |
| Std. Dev. | 3371 (31.7%) | Std. Dev. Change | 110 (1.0%) |
| % High | 1552 (14.6%) | % High change | 94 (0.9%) |
| % Low | 1552 (14.6%) | % Low change | 94 (0.9%) |
| *Blood Urea Nitrogen* | | *Oxygen Saturation* | |
| Mean | 1877 (17.6%) | Mean | 1799 (16.9%) |
| Std. Dev. | 4057 (38.1%) | Std. Dev. | 2168 (20.3%) |
| % High | 1877 (17.6%) | % Low | 1799 (16.9%) |
| % Low | 1877 (17.6%) | Mean change | 1184 (11.1%) |
| *Glucose* | | Std. Dev. Change | 1408 (13.2%) |
| Mean | 4489 (42.2%) | % Low change | 1184 (11.1%) |
| Std. Dev. | 5896 (55.5%) | *Other* | |
| % High | 4489 (42.2%) | Vitals Frequency | 87 (0.1%) |
| % Low | 4489 (42.2%) | **Other (6)** | |
| *White Blood Cell Count* | | FiO2 Mean | 0 (0.0%) |
| Mean | 2322 (21.8%) | FiO2 % High | 0 (0.0%) |
| Std. Dev. | 4687 (44.1%) | ICU Admission | 0 (0.0%) |

| | | | |
|---|---|---|---|
| % High | 2322 (21.8%) | Mean GCS | 0 (0.0%) |
| % Low | 2322 (21.8%) | % Low GCS | 0 (0.0%) |
| *Hemoglobin* | | Mean Weight | 5 (0.0%) |
| Mean | 2266 (21.3%) | **Patient Level (5)** | |
| Std. Dev. | 4631 (43.5%) | Age (at admission) | 0 (0.0%) |
| % High | 2266 (21.3%) | Gender | 0 (0.0%) |
| % Low | 2266 (21.3%) | Admitting Service | 0 (0.0%) |
| | | Prior AKI | 0 (0.0%) |

In the best performing structured model, the elastic-net regression eliminated 183 of the 294 predicator variables. Of the remaining 111 variables, 36 had odds ratios with statistically significant 95% confidence intervals, and these are shown in Table 4.3. The model performed well overall, with an F1 score of 0.67 and AUC of 0.75, and correctly identified over 70% of the creatinine values meeting AKI criteria. Of the creatinine values predicted to have AKI, almost two-thirds were true positives. Specific performance metrics for this model are shown in Table 4.4.

Table 4.3. Variables included in the best performing model with significant odds ratios and 95% confidence intervals. Variables are divided into those associated with new-onset AKI, and those associated with all creatinine values (new-onset AKI and ongoing AKI).

| **Variables Associated with all Creatinine Values** | **OR (95% CI)** | **Interaction Variables Associated with New-onset AKI** | **OR (95% CI)** |
|---|---|---|---|
| **Increasing Risk** | | | |
| Prior AKI | 2.497 (2.114 - 2.949) | CICU Admit | 8.707 (3.843 - 19.731) |
| Central Line | 2.247 (1.535 - 3.29) | HemOnc Admit | 1.363 (1.020 - 1.821) |
| Gen Surg Admit | 1.358 (1.076 - 1.713) | High SBP (Cut 2) (per 10%) | 1.281 (1.092 - 1.504) |
| BMT Admit | 1.319 (1.022 - 1.701) | Mean Platelets (per 100,000) | 1.146 (1.089 - 1.206) |
| lisinopril (per 5 mg) | 1.261 (1.044 - 1.523) | Mean Tacro Level | 1.127 (1.057 - 1.201) |
| ICU Transfer | 1.243 (1.051 - 1.471) | Chloride Std. Dev. | 1.123 (1.057 - 1.193) |
| carboplatin (per 100 mg) | 1.225 (1.015 - 1.478) | ioversol (per 10 mL) | 1.039 (1.010 - 1.068) |
| Low DBP (Cut 1) (per | 1.184 (1.014 - 1.381) | Mean Vanc Level | 1.036 (1.011 - 1.062) |

| | | | |
|---|---|---|---|
| 10%) | | | |
| Hemoglobin Std. Dev. | 1.159 (1.038 - 1.295) | | |
| neomycin (per 100 mg) | 1.094 (1.011 - 1.184) | | |
| High Sodium (per 10%) | 1.065 (1.020 - 1.113) | | |
| furosemide (per 10 mg) | 1.035 (1.011 - 1.060) | | |
| methotrexate (per 100 mg) | 1.029 (1.006 - 1.052) | | |
| Mean HR (Cut 1) (per 5 bpm) | 1.029 (1.008 - 1.05) | | |
| Mean RR (Cut 1) | 1.022 (1.000 - 1.044) | | |
| Mean BUN | 1.021 (1.012 - 1.031) | | |
| Mean WBC (per 1,000) | 1.003 (1.000 - 1.005) | | |
| **Decreasing Risk** | | | |
| Bicarb Std. Dev. | 0.942 (0.890 - 0.998) | Mean BUN | 0.975 (0.962 - 0.988) |
| Sodium Std. Dev. | 0.937 (0.888 - 0.990) | SBP Std. Dev. (Cut 2) | 0.963 (0.942 - 0.983) |
| Mean SBP (Cut 1) (per 5 mmHg) | 0.933 (0.905 - 0.962) | High Chloride (per 10%) | 0.944 (0.911 - 0.979) |
| Rheumatology Admit | 0.440 (0.211 - 0.918) | Mean GCS | 0.919 (0.892 - 0.948) |
| CICU Admit | 0.257 (0.133 - 0.496) | lisinopril (per 5 mg) | 0.735 (0.573 - 0.942) |
| | | Central Line | 0.333 (0.199 - 0.558) |

Table 4.4. Performance measures of the best-performing structured data elastic-net regression model with a Data Collection Window of 48 hours, a Prediction Window of 24 hours, and AKI defined using KDIGO Stage 1 criteria.

| | **No AKI** | **AKI** | |
|---|---|---|---|
| **Predicts No AKI** | 1314 | 239 | NPV: 0.85 |
| **Predicts AKI** | 351 | 611 | PPV: 0.64 |
| | Specificity: 0.79 | Sensitivity: 0.72 | |

F1 Score: 0.67 AUC: 0.75

## 4.2 RESULTS FOR MODELS CONTAINING UNSTRUCTURED DATA

The subset used to generate the models combining structured and unstructured data consisted of 195 patients with 353 creatinine values. A summary of this subset can be found in Table 4.5. In

the subset, the differences in between the training set and test set are slightly more pronounced, likely because of the smaller overall size and the random selection process. There are also more males (59.2% versus 57.0%) in the subset than in the overall dataset, and the patients are older (10.8 years compared to 10.0 years).

Table 4.5. Summary of ICU subset.

| | Mean Age | % Male | % AKI | % AKI New-onset | % AKI Ongoing |
|---|---|---|---|---|---|
| **Training (n = 283)** | 10.7 years | 58.0% | 45.2% | 19.4% | 25.8% |
| **Test (n = 70)** | 11.1 years | 64.3% | 50.0% | 18.6% | 31.4% |

The structured model described above was applied to the 70 creatinine values in the n-gram test set, providing a baseline performance against which the n-gram model could be measured (shown in Table 4.6). The baseline performance was similar to the performance in the overall dataset, with a slightly improved F1 score of 0.70 (compared to 0.67 for the overall cohort) and slightly worse AUC of 0.70 (compared to 0.75). The lower AUC was due to a lower specificity in the ICU subset, which likely resulted from the fact that the proportion of AKI cases was higher in the subset than in the overall dataset.

Table 4.6. Performance measures of the elastic-net regression model in the ICU Subset, without n-grams.

| | No AKI | AKI | |
|---|---|---|---|
| **Predicts No AKI** | 25 | 11 | NPV: 0.69 |
| **Predicts AKI** | 10 | 24 | PPV: 0.71 |
| | Specificity: 0.71 | Sensitivity: 0.69 | |

F1 Score: 0.70 AUC: 0.70

The 225 highest ranking (by chi-squared criteria) n-gram binomial variables were then combined with the structured data, and the model run again. This combination model improved

upon the baseline model's performance in every respect, with summary measures of an F1 score of 0.76 and AUC of 0.77. Detailed performance measures for the combination model are presented in Table 4.7. The statistically significant n-grams utilized by the model and their associated odds ratios and 95% confidence intervals are shown in Table 4.8.

Table 4.7. Performance measures of the elastic-net regression model in the ICU Subset, with n-grams.

| | **No AKI** | **AKI** | |
|---|---|---|---|
| **Predicts No AKI** | 29 | 10 | NPV: 0.74 |
| **Predicts AKI** | 6 | 25 | PPV: 0.81 |
| | Specificity: 0.83 | Sensitivity: 0.71 | |

F1 Score: 0.76 AUC: 0.77

Table 4.8. N-grams included in the model, with corresponding odds ratios. N-grams are divided into those associated with the new-onset AKI and those associated with all creatinine values (new-onset and ongoing AKI).

| **Variables Associated with all Creatinine Values** | **OR (95% CI)** | **Interaction Variables Associated with New-onset AKI** | **OR (95% CI)** |
|---|---|---|---|
| "IV fluids maintenance" | 0.069 (0.013 – 0.356) | "morphine" | 0.168 (0.042 – 0.672) |
| "2L nasal" | 18.954 (2.276 – 157.819) | | |
| "white blood cell" | 5.219 (1.643 – 16.579) | | |
| "dose" | 2.819 (1.093 – 7.272) | | |

# Chapter 5. ERROR ANALYSIS

The models described above demonstrated reasonable performance, with F1 scores ranging from 0.67 to 0.76 and AUCs ranging from 0.70 to 0.77. However, they are not perfect, and consideration of where they went wrong may provide suggestions for improvement. This

analysis will focus on the results of the structured model because it had more errors available for review, but its findings are relevant to the n-gram model as well.

The structured model produced a total of 351 false negatives and 239 false positives. These errors had two things in common: firstly, most errors occurred near the AKI threshold (84.3% of false negatives and 73.6% of false positives involved creatinine values that were less than 0.2 mg/dL away from the threshold), suggesting that the model rarely misclassified obvious cases. Secondly, most errors – 72.4% of false negatives and 75.8% of false positives - involved creatinine values for which the baseline creatinine was 0.3 mg/dL or less. This is meaningful because this model used the KDIGO Stage 1 definition for AKI: either an increase in 0.3 mg/dL or 1.5 times the baseline creatinine, whichever is lower. Thus for a baseline creatinine of 0.1 mg/dL the AKI threshold is 0.15 mg/dL; for a baseline of 0.2 mg/dL the AKI threshold is 0.3 mg/dL; and for a baseline creatinine of 0.3 mg/dL the AKI threshold is 0.45 mg/dL. It may be that these small increases in serum creatinine - only 0.1 or 0.2 mg/dL - represent routine variation in creatinine measurement rather than true AKI. (Laboratory measurements of creatinine are not perfect, and a small amount of variation would not be surprising.) It is certainly plausible, then, that the "errors" in these cases are not representative of the clinical phenotype the model is designed to identify.

If this were true, then removal of cases with lower baseline creatinine values should increase the performance of the model. Table 5.1 shows the model's performance if creatinine values with a baseline creatinine of 0.3 mg/dL or less are excluded. The F1 score for this model increased from 0.67 to 0.72, and the AUC from 0.75 to 0.81. Indeed, every metric (sensitivity, specificity, positive predictive value, and negative predictive value) improved, suggesting that not every minor fluctuation in creatinine is meaningful.

Table 5.1. Performance Measures of the Elastic-net Regression Model with Data Collection Window 48 hours, Prediction Window 24 hours, KDIGO Stage 1 AKI definition, and the exclusion of any creatinine value with a baseline creatinine of 0.3 mg/dL or less.

| | **No AKI** | **AKI** | |
|---|---|---|---|
| **Predicts No AKI** | 718 | 75 | NPV: 0.91 |
| **Predicts AKI** | 189 | 345 | PPV: 0.65 |
| | Specificity: 0.79 | Sensitivity: 0.82 | |

F1 Score: 0.72 AUC: 0.81

Error analysis was also performed at a patient level, which demonstrated that misclassifications generally fell into one of several patterns. There were three primary patterns for false negatives and three for false positives, which are shown in Figure 5.1.

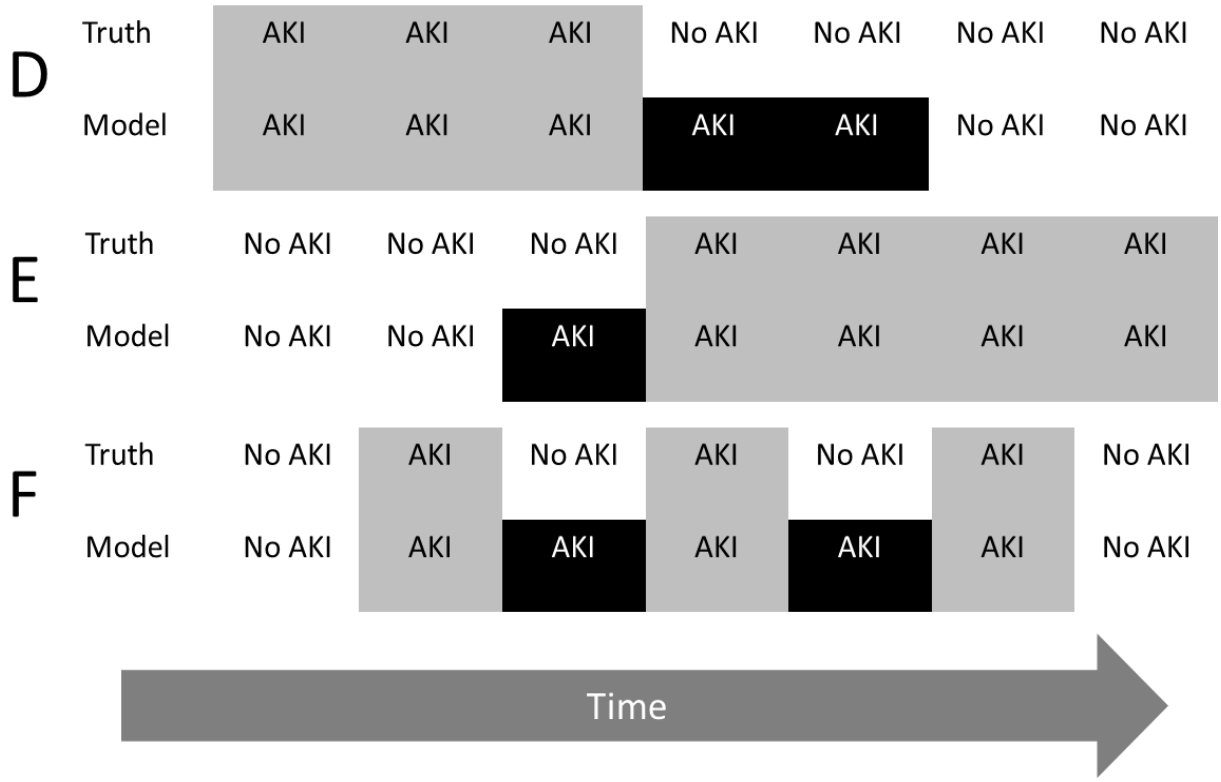


Figure 5.1. Different patterns of patient-level errors produced by the structured data model. A-C represent false negative patterns, whereas D-F represent false positive patterns. In A, the model is delayed in predicting AKI onset by 24 – 48 hours. In B, the model predicts resolution of AKI 24 – 48 hours before it actually happens. C demonstrates the scenario in which the model does not recognize that a single creatinine has met AKI criteria. In D, the model predicts that AKI will last 24 – 48 hours longer than it does. In E, the model predicts that AKI begins 24 – 48 hours before it actually does. F demonstrates the scenario in which creatinine values are intermittently crossing the AKI threshold, but the model predicts the entire series of creatinine values represents AKI. Error patterns A-B and D-E are likely more significant clinically, as patterns C and F do not occur in patients who develop prolonged substantial AKI.

The most common false negative error pattern involved delayed onset of AKI prediction (pattern A in Figure 2). This occurred in 116 cases, or 33.0% of all false negatives. Pattern B, in which the model predicted early AKI resolution, occurred in 50 cases (14.2% of false negatives). Pattern C occurred almost exclusively in patients with a baseline creatinine of 0.3 mg/dL or less, and was seen in 99 cases (28.2% of false negatives). The remaining 86 false negatives (24.5%) did not fit a particular pattern.

The most common false positive error pattern is demonstrated by Pattern D, and involved the model predicting prolonged AKI episodes. This occurred in 117 cases, or 49.0% of all false positives. Pattern E, in which the model predicted AKI onset early, occurred in 26 cases (10.9% of false positives). Pattern F also occurred most frequently in patients with baseline creatinine values of 0.3 mg/dL or less, and was seen in 49 cases (20.5% of the total false positives). The remaining 47 false positives (19.7%) did not fit an obvious pattern.

Analysis of these patterns indicates that the model has the most difficulty with determination of AKI onset and cessation. There are several possible explanations for this. Firstly, AKI has a wide range of potential causes, including poor renal perfusion, the direct nephrotoxic effect of medications, hyperfiltration of nephrotoxic substances (such as uric acid), and autoimmune disease. It is quite possible that different insults cause damage at varying rates, leading to the errors in predicting AKI onset and cessation seen in the model. Secondly, it is also possible that each cause has its own risk phenotype, such that the phenotype produced by the model is actually an amalgam of features from several sub-phenotypes. This would explain why the model was able to identify most but not all cases of kidney injury, and suggests that further study into quantifying AKI risk in specific disease states is required.

# Chapter 6. DISCUSSION

The models described above were able to predict a rise in creatinine 24 hours before it happened with reasonable performance. Especially encouraging was that the positive predictive values in all models were well above 50%, ensuring that the number of generated false positives was less than the number of correctly identified elevated creatinine values. This represents a significant improvement over previous methods of predicting acute kidney injury, which produced high

sensitivities but low positive predictive values. Nor was specificity sacrificed to improve positive predictive value, as the AUCs for the above models (0.75 for the overall model and 0.77 for the n-gram subset) are comparable to the AUCs obtained by previous methods.

This improvement in performance was likely due to two primary factors: the use of F1 scores as a preferred summary measure and overall model design. As mentioned, F1 scores are the harmonic mean of sensitivity and positive predictive value, and thus emphasize these components equally. This is in contrast to AUC scores, which emphasize sensitivity and specificity. When considering lower percentage outcomes such as AKI, it is easier to generate a high specificity (for instance, by simply assuming every case to be negative) than it is to produce a high positive predictive value. Tuning the model to select for higher F1 scores, then, proved essential to improving performance.

Model design also likely contributed to the performance improvement. Rather than relying on data collected at the time of admission, the use of a Data Collection Window ensured that the information associated with each creatinine value was both timely and relevant. The models described here also contained a wide variety of structured data, and specifically incorporated a large amount of vital sign data, which has not been used in previous studies. Overall, there were a greater number of variables associated with each creatinine value in this model compared to previous models, which also may have impacted performance.

## 6.1 BIOLOGIC VERSUS SYSTEMIC INFORMATION

However, the use of more variables is also potentially problematic, in that as the number of variables increases, so does the probability that some of those variables are associated with AKI purely by chance. Thus, it is essential that predictors included in the model make sense clinically. Determining the sensibility of each predictor is a difficult task at baseline, and is in

some cases a matter of expert opinion. This determination is further complicated by the fact that the meaning of an individual predictor is not always straightforward, because the information contained within that variable exists on a continuum between *systemic information* and *biologic information*. That is, each predictor begs the question, "To what extent is information captured regarding biologic processes happening within a patient, and to what extent is information captured regarding the hospital system in which patients find themselves?" The answer to this question depends on the predictor in question, but in general, each variable fits into one of three classes:

1) Mostly systemic – variables that primarily capture information about the hospital system, but contain some information about biologic processes (e.g., admitting service or medications)
2) Partly systemic – variables that carry a roughly equal amount information about the hospital system and biologic processes (e.g., laboratory values)
3) Mostly biologic – variables that primarily capture information about biologic processes, but contain some information about systemic processes (e.g., vital signs and demographics)

No variable is purely biologic, in that everything provides some information about the hospital system. Even demographics, which certainly provide biologic information about a patient, can also provide systemic information (e.g., the demographics of a children's hospital are significantly different from those of a Veteran's Administration hospital). Vital signs also provide a good deal of insight into biologic processes, but which vital signs are monitored and how frequently that monitoring occurs is systemic. Alternatively, although medications undoubtedly have a biological effect on patients, the majority of the information they convey may be systemic to the extent that medication use serves as a proxy for a specific disease process (e.g., chemotherapeutic agents are used to treat patients with cancer). N-gram data could span all three categories depending on the n-gram in question, but would also never be purely biologic.

With that in mind, Table 6.1 reorganizes the significant predictors from the structured data model into these variable types.

Table 6.1. Variables included in the best performing structured model with significant odds ratios and 95% confidence intervals. Variables are divided into those associated with new-onset AKI, and those associated with all creatinine values (new-onset and ongoing AKI), and are organized according to how much systemic information they provide.

| Variables Associated with all Creatinine Values | OR (95% CI) | Interaction Variables Associated with New-onset AKI | OR (95% CI) |
|---|---|---|---|
| **Mostly Systemic** | | | |
| Central Line | 2.247 (1.535 - 3.29) | CICU Admit | 8.707 (3.843 - 19.731) |
| Gen Surg Admit | 1.358 (1.076 - 1.713) | HemOnc Admit | 1.363 (1.020 - 1.821) |
| BMT Admit | 1.319 (1.022 - 1.701) | ioversol (per 10 mL) | 1.039 (1.010 - 1.068) |
| lisinopril (per 5 mg) | 1.261 (1.044 - 1.523) | lisinopril (per 5 mg) | 0.735 (0.573 - 0.942) |
| ICU Transfer | 1.243 (1.051 - 1.471) | Central Line | 0.333 (0.199 - 0.558) |
| carboplatin (per 100 mg) | 1.225 (1.015 - 1.478) | | |
| neomycin (per 100 mg) | 1.094 (1.011 - 1.184) | | |
| furosemide (per 10 mg) | 1.035 (1.011 - 1.060) | | |
| methotrexate (per 100 mg) | 1.029 (1.006 - 1.052) | | |
| Rheumatology Admit | 0.440 (0.211 - 0.918) | | |
| CICU Admit | 0.257 (0.133 - 0.496) | | |
| **Partly Systemic** | | | |
| Hemoglobin Std. Dev. | 1.159 (1.038 - 1.295) | Mean Platelets (per 100,000) | 1.146 (1.089 - 1.206) |
| High Sodium (per 10%) | 1.065 (1.020 - 1.113) | Mean Tacro Level | 1.127 (1.057 - 1.201) |
| Mean BUN | 1.021 (1.012 - 1.031) | Chloride Std. Dev. | 1.123 (1.057 - 1.193) |
| Mean WBC (per 1,000) | 1.003 (1.000 - 1.005) | Mean Vanc Level | 1.036 (1.011 - 1.062) |

| | | | |
|---|---|---|---|
| Bicarb Std. Dev. | 0.942 (0.890 - 0.998) | Mean BUN | 0.975 (0.962 - 0.988) |
| Sodium Std. Dev. | 0.937 (0.888 - 0.990) | High Chloride (per 10%) | 0.944 (0.911 - 0.979) |
| **Mostly Biologic** | | | |
| Low DBP (Cut 1) (per 10%) | 1.184 (1.014 - 1.381) | High SBP (Cut 2) (per 10%) | 1.281 (1.092 - 1.504) |
| Mean HR (Cut 1) (per 5 bpm) | 1.029 (1.008 - 1.05) | SBP Std. Dev. (Cut 2) | 0.963 (0.942 - 0.983) |
| Mean RR (Cut 1) | 1.022 (1.000 - 1.044) | | |
| Mean SBP (Cut 1) (per 5 mmHg) | 0.933 (0.905 - 0.962) | | |

The mostly systemic variables generally make sense clinically. Patients admitted to the Bone Marrow Transplant service, General Surgery service, or transferred to the ICU are probably systemically ill, and it is not surprising they would be at increased risk for AKI. There are two service-related variables that are more perplexing. The first is being admitted to the Rheumatology service, which is associated with a decreased risk of AKI. This may be because the majority of patients with rheumatologic diseases affecting the kidney are admitted to the Nephrology service instead, such that most patients admitted to Rheumatology have pristine kidney function. However, it is harder to explain how those admitted directly to the Cardiac ICU have a decreased risk of developing AKI; one would expect the opposite. Interestingly, the interaction term for Cardiac ICU admission shows a significantly increased risk, suggesting that those in the Cardiac ICU are at elevated risk for new-onset AKI, but that episodes of AKI there are not ongoing. This signal may simply be the result of a small sample size; only 1.7% of the dataset, or 183 creatinine values, were directly admitted to the Cardiac ICU. It may also be that AKI in the Cardiac ICU is generally due to the acute tubular necrosis secondary to cardiac bypass and/or poor cardiac output, and that patients recover more quickly from this type of AKI.

Turning to the medication variables, the known nephrotoxins lisinopril, carboplatin, neomycin, furosemide, and methrotrexate are associated with an increased risk of AKI. Ioversol, also believed to be nephrotoxic, is associated with new-onset AKI, though not with AKI overall. It is interesting that lisinopril is associated with a decreased risk of new-onset AKI, suggesting that while the drug may not contribute to AKI onset, patients on lisinopril have longer episodes of AKI. This is certainly reasonable clinically.

The partly systematic variables consist entirely of labs, which are biologic in that they reflect what is happening within a patient, but systemic in that not every lab is checked on every patient every day. Patients in the ICU, for instance, undergo more lab draws than patients outside of it. Thus, while it is satisfying that the lab values generally correspond with what one would expect clinically – higher white blood cell counts, indicative of systemic illness, and high sodium levels, indicative of volume depletion, should be associated with AKI – there is still the possibility these variables are conveying more systemic than biologic information. For instance, does having a variable chloride level truly put one at risk for kidney injury, or do patients who develop AKI have their chloride levels checked more frequently? Similarly, dose having variable hemoglobin values somehow affect kidney function, or is it representative of receiving a blood transfusion?

Like lisinopril, several lab values are associated with AKI overall but show decreased risk of new-onset AKI. This makes sense for BUN and chloride, which can certainly show derangements with AKI, but may be normal prior to AKI onset. The opposite may be true for vancomycin and tacrolimus levels, which would be high at AKI onset but may normalize later as clinicians respond to higher levels. Overall, then, the partly systemic data provided by lab values is clinically sound.

Finally, the mostly biologic variables consist of vital signs and demographic information. These also make sense clinically – one would expect, for example, that episodes of AKI during prior hospital admissions would put one at risk for future AKI. It also makes sense that higher heart rates and respiratory rates, signs of a sicker patient, are associated with AKI. The systolic blood pressure variables are more difficult to interpret. It would appear that the higher one's systolic blood pressure is during Cut 1 – the later 24 hours of the Data Collection Window – the less likely one is to either develop new-onset AKI or have ongoing AKI. This makes sense in that a higher blood pressures suggests better renal perfusion, which makes injury less likely. However, there are also associations between a high systolic blood pressure and a more variable systolic blood pressure in Cut 2 (the earlier 24 hours of the Data Collection Window) and an increased risk of new-onset AKI. Perhaps these represent direct sequelae of whatever injury causes AKI, during which a distressed kidney releases renin and causes a brief spike in blood pressure. However, this is purely speculation, and requires confirmation in other studies before being accepted.

Overall, the associations between the majority of the variables in the structured model and AKI can be reasonably explained clinically. The few variables that are more difficult to explain may represent novel AKI associations, or may simply be due to chance. Of course, all predictors need to be validated in future studies.

## 6.2 UNSTRUCTURED DATA

Analysis of the models containing unstructured data requires additional considerations. Firstly, the use of natural language processing of clinician notes to predict the onset of acute kidney injury has not been previously reported in the literature, so there is no available baseline against which the n-gram results can be compared. Therefore, the value of adding n-grams to the

structured data model was determined by comparing the performance of a baseline (structured data only) model to that of a combination (structured data plus n-grams) model. This comparison favored the inclusion of n-grams, suggesting that combining structured and unstructured data holds potential for more specific clinical phenotyping.

However, several limitations apply specifically to the n-gram portion of this study. Firstly, only ICU notes were included, meaning that any results involving n-grams can only be applied to an ICU population. Secondly, only 195 patients (with 353 creatinine values) were included in the subset analysis, which is a relatively small sample size. A smaller number of patients may increase the possibility that certain n-grams were associated with AKI purely by chance, although it is equally plausible that the study was not powered well enough to detect less potent associations.

The n-gram model was able to detect five potentially meaningful signals: "IV fluids maintenance", "morphine", "2L nasal", "white blood cell", and "dose". The first two were associated with a decreased risk of AKI, which certainly makes sense for "IV fluids maintenance": because maintenance IV fluids are a "default" setting, their use may mean that a clinician is not concerned about a patient getting too much (or too little) fluid. The "morphine" signal is harder to interpret, but many patients are admitted to the ICU for post-operative monitoring, and these patients may require adequate pain control (with morphine) but have a lower chance of developing AKI.

The three n-grams associated with an increased risk for AKI may reflect increasing clinical severity; for instance, patients requiring oxygen via nasal cannula (which is likely what "2L nasal" refers to) are sicker than those breathing room air. Patients for whom a "white blood cell" count is mentioned are also likely to be critically ill; the structured model corroborates this

in that an increasing white blood cell count is associated with an increased risk for AKI. Finally, a provider mentioning "dose" in a note may be referring to clinically significant medications that patient has received. Overall, then, it is reasonable to assume that the methods used to construct this n-gram model have detected clinically relevant signals, and may detect further signals in a larger dataset.

## 6.3 REPRODUCIBILITY

These models were produced with data from a single institution, which raises the question: how would they perform elsewhere? To suggest an answer requires expanding upon the concept of biologic versus systemic information to create *biologic* and *systemic phenotypes*.

Biologic phenotypes are potentially superior to primarily systemic phenotypes for two reasons. Firstly, they are more likely to be reproducible at other institutions, since they capture a clinical phenotype that is inherent to patients, not hospital systems. Secondly, they require very little upkeep, since a biologic phenotype will presumably not vary over time. Systemic phenotypes, however, will naturally evolve alongside hospital systems. Indeed, any implementation of a systemic model predicting AKI would need to be routinely monitored to ensure that the changes it introduces to the hospital system do not invalidate its results.

Because the models discussed in this work contain both biologic and systemic variables, they capture a mixed biologic-systemic phenotype. Therefore, it is unlikely that use of these exact variables would produce equivalent results at other institutions. However, the extent to which this is a problem is debatable – if a system works well at single institution, does reduced performance at another institution make it less valid? Furthermore, the process by which these models were created – the construction of Data Collection Windows and Prediction Windows relative to a clinical event of interest – could easily be applied to any hospital system. Using this

process to predict AKI at other institutions would likely result in models that perform similarly and contain similar biologic variables, but different systemic variables.

Given this, the best way to test a particular would be prospectively – that is, incorporate the model into an EHR and let it predict the onset of AKI in real time. Reasonable performance in a prospective test would suggest the model lends itself well to clinical use. Poor performance would suggest that variables in that model should be reconsidered.

# Chapter 7. FUTURE WORK

The results of this work suggest that it is possible to develop a clinical phenotype describing pediatric patients at risk for the development of AKI. However, validation of both model design and model performance is necessary before the clinical phenotype described herein is incorporated into clinical practice. This validation could occur using retrospective data at other institutions, prospective data at Seattle Children's Hospital and/or other institutions, or, ideally, a combination of the two.

Replicating the model design at other institutions has the added benefit of suggesting answers to one of questions raised by this work: namely, to what extent are individual variables representing biologic versus systemic information? As discussed, models developed using data from other institutions are likely to identify different systemic variables but similar biologic variables. Therefore, comparing multiple models should provide insight into which variables are primarily biologic (i.e., which are predictive at multiple hospitals) and which are primarily systemic (i.e., which are predictive only at a single institution). Attempts to clarify whether systemic differences are due to proscribed hospital policies or variations in provider training and preferences may also prove useful, and should be investigated.

The role of natural language processing in clinical phenotyping also needs to be further explored. The n-grams identified in this work suggest that structured data regarding a patient's intravenous fluids and respiratory support should be included in future models. The addition of medications that are not known to be nephrotoxic – such as morphine – should be considered as well. Ideally, more such signals would be identified with the use of a larger dataset (and therefore a larger corpus). Furthermore, there may be a role for NLP in extracting structured information that may not otherwise be available. For instance, if a patient is transferred to a hospital from an outside facility, routine structured data such as vital signs, medications, and labs would not be available in the receiving hospital's EHR, but may be transcribed in a clinician's note. If this information could be reliably extracted and added to the predictive model, it could significantly improve performance.

Finally, the concept of a Data Collection Window and Prediction Window is not unique to a rising serum creatinine. Indeed, the methods described in this work could – and hopefully will – be applied to any clinical event with a clearly defined onset. It may be that the future of medicine involves routine use of "at risk for a disease" clinical phenotypes, greatly improving clinicians' ability to identify and prevent problems before they happen.

# Chapter 8. CONCLUSION

Predicting pediatric acute kidney injury remains difficult task, but this work represents a step forward in the creation of a robust clinical phenotype describing patients at risk for developing AKI. This phenotype is temporally related to AKI onset, and when used in a predictive model it demonstrates reasonable performance, with an F1 score of 0.67 and AUC of 0.75. This performance improved in an ICU subset following the addition of unstructured data from

clinician notes, with the model's F1 score increasing to 0.72 and AUC increasing to 0.77. This suggests that natural language processing may have a role in refining clinical phenotypes. However, interpreting the results of these modes requires careful consideration of the information contained within each variable – specifically, the extent to which that information describes biologic processes within a patient or systemic processes within a hospital. Further evaluation of the use of clinical phenotyping in predicting pediatric AKI is necessary to confirm the utility of these models and to help differentiate which signals and primarily biologic and which are primarily systemic. Ultimately, a thoroughly vetted and comprehensive "at risk for AKI" phenotype will allow clinicians to identify and potentially prevent the onset of kidney injury, substantially reducing the morbidity and mortality associated with this disease.

# BIBLIOGRAPHY

[1] Sutherland SM, Ji J, Sheikhi FH, Widen E, Tian L, Alexander SR, Ling XB. AKI in hospitalized children: epidemiology and clinical associations in a national cohort. Clin J Am Soc Nephrol. 2013 Oct;8(10):1661-9. doi: 10.2215/CJN.00270113. Epub 2013 Jul 5. PubMed PMID: 23833312; PubMed Central PMCID: PMC3789331.

[2] Schneider J, Khemani R, Grushkin C, Bart R. Serum creatinine as stratified in the RIFLE score for acute kidney injury is associated with mortality and length of stay for children in the pediatric intensive care unit. Crit Care Med. 2010 Mar;38(3):933-9. doi: 10.1097/CCM.0b013e3181cd12e1. PubMed PMID: 20124891.

[3] Ricci Z, Di Nardo M, Iacoella C, Netto R, Picca S, Cogo P. Pediatric RIFLE for acute kidney injury diagnosis and prognosis for children undergoing cardiac surgery: a single-center prospective observational study. Pediatr Cardiol. 2013 Aug;34(6):1404-8. doi: 10.1007/s00246-013-0662-z. Epub 2013 Feb 22. PubMed PMID: 23430323.

[4] Plötz FB, Bouma AB, van Wijk JA, Kneyber MC, Bökenkamp A. Pediatric acute kidney injury in the ICU: an independent evaluation of pRIFLE criteria. Intensive Care Med. 2008 Sep;34(9):1713-7. doi: 10.1007/s00134-008-1176-7. Epub 2008 Jun 3. PubMed PMID: 18521567.

[5] Alkandari O, Eddington KA, Hyder A, Gauvin F, Ducruet T, Gottesman R, Phan V, Zappitelli M. Acute kidney injury is an independent risk factor for pediatric intensive care unit mortality, longer length of stay and prolonged mechanical ventilation in critically ill children: a two-center retrospective cohort study. Crit Care. 2011 Jun 10;15(3):R146. doi: 10.1186/cc10269. PubMed PMID: 21663616; PubMed Central PMCID: PMC3219018.

[6] Bellomo R, Ronco C, Kellum JA, Mehta RL, Palevsky P; Acute Dialysis Quality Initiative workgroup. Acute renal failure - definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. Crit Care. 2004 Aug;8(4):R204-12. Epub 2004 May 24. Review. PubMed PMID: 15312219; PubMed Central PMCID: PMC522841.

[7] Akcan-Arikan A, Zappitelli M, Loftis LL, Washburn KK, Jefferson LS, Goldstein SL. Modified RIFLE criteria in critically ill children with acute kidney injury. Kidney Int. 2007 May;71(10):1028-35. Epub 2007 Mar 28. PubMed PMID: 17396113.

[8] Mehta RL, Kellum JA, Shah SV, Molitoris BA, Ronco C, Warnock DG, Levin A; Acute Kidney Injury Network. Acute Kidney Injury Network: report of an initiative to improve

outcomes in acute kidney injury. Crit Care. 2007;11(2):R31. PubMed PMID: 17331245; PubMed Central PMCID: PMC2206446.

[9] Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group. KDIGO Clinical Practice Guideline for Acute Kidney Injury. Kidney inter., Suppl. 2012; 2: 1–138

[10] Devarajan P. Neutrophil gelatinase-associated lipocalin (NGAL): a new marker of kidney disease. Scand J Clin Lab Invest Suppl. 2008;241:89-94. doi: 10.1080/00365510802150158. Review. PubMed PMID: 18569973; PubMed Central PMCID: PMC2528839.

[11] Lin X, Yuan J, Zhao Y, Zha Y. Urine interleukin-18 in prediction of acute kidney injury: a systemic review and meta-analysis. J Nephrol. 2015 Feb;28(1):7-16. doi: 10.1007/s40620-014-0113-9. Epub 2014 Jun 5. Review. PubMed PMID: 24899123; PubMed Central PMCID: PMC4322238.

[12] Shao X, Tian L, Xu W, Zhang Z, Wang C, Qi C, Ni Z, Mou S. Diagnostic value of urinary kidney injury molecule 1 for acute kidney injury: a meta-analysis. PLoS One. 2014 Jan 3;9(1):e84131. doi: 10.1371/journal.pone.0084131. eCollection 2014. PubMed PMID: 24404151; PubMed Central PMCID: PMC3880280.

[13] Xu Y, Xie Y, Shao X, Ni Z, Mou S. L-FABP: A novel biomarker of kidney disease. Clin Chim Acta. 2015 May 20;445:85-90. doi: 10.1016/j.cca.2015.03.017. Epub 2015 Mar 20. Review. PubMed PMID: 25797895.

[14] Pollack MM, Patel KM, Ruttimann UE. PRISM III: an updated Pediatric Risk of Mortality score. Crit Care Med. 1996 May;24(5):743-52. PubMed PMID: 8706448.

[15] Leteurtre S, Duhamel A, Salleron J, Grandbastien B, Lacroix J, Leclerc F; Groupe Francophone de Réanimation et d'Urgences Pédiatriques (GFRUP). PELOD-2: an update of the PEdiatric logistic organ dysfunction score. Crit Care Med. 2013 Jul;41(7):1761-73. doi: 10.1097/CCM.0b013e31828a2bbd. PubMed PMID: 23685639.

[16] Seiger N, Maconochie I, Oostenbrink R, Moll HA. Validity of different pediatric early warning scores in the emergency department. Pediatrics. 2013 Oct;132(4):e841-50. doi: 10.1542/peds.2012-3594. Epub 2013 Sep 9. PubMed PMID: 24019413.

[17] Chen Y, Ghosh J, Bejan CA, Gunter CA, Gupta S, Kho A, Liebovitz D, Sun J, Denny J, Malin B. Building bridges across electronic health record systems through inferred phenotypic topics. J Biomed Inform. 2015 Jun;55:82-93. doi: 10.1016/j.jbi.2015.03.011. Epub 2015 Apr 1. PubMed PMID: 25841328; PubMed CentralPMCID: PMC4464930.

[18] Bekhof J, Reitsma JB, Kok JH, Van Straaten IH. Clinical signs to identify late-onset sepsis in preterm infants. Eur J Pediatr. 2013 Apr;172(4):501-8. doi: 10.1007/s00431-012-1910-6. Epub 2012 Dec 28. PubMed PMID: 23271492.

[19] Saptharishi LG, Jayashree M, Singhi S. Development and validation of the "Pediatric Risk of Nosocomial Sepsis (PRiNS)" score for health care-associated infections in a medical pediatric intensive care unit of a developing economy-a prospective observational cohort study. J Crit Care. 2016 Apr;32:152-8. doi: 10.1016/j.jcrc.2015.11.016. Epub 2015 Nov 24. PubMed PMID: 26785993.

[20] Jorge-Monjas P, Bustamante-Munguira J, Lorenzo M, Heredia-Rodríguez M, Fierro I, Gómez-Sánchez E, Hernandez A, Álvarez FJ, Bermejo-Martin JF, Gómez-Pesquera E, Gómez-Herreras JI, Tamayo E. Predicting cardiac surgery-associated acute kidney injury: The CRATE score. J Crit Care. 2016 Feb;31(1):130-8. doi: 10.1016/j.jcrc.2015.11.004. Epub 2015 Nov 6. PubMed PMID: 26700607

[21] Basu RK, Wong HR, Krawczeski CD, Wheeler DS, Manning PB, Chawla LS, Devarajan P, Goldstein SL. Combining functional and tubular damage biomarkers improves diagnostic precision for acute kidney injury after cardiac surgery. J Am Coll Cardiol. 2014 Dec 30;64(25):2753-62. doi: 10.1016/j.jacc.2014.09.066. Erratum in: J Am Coll Cardiol. 2015 Mar 24;65(11):1158-9. PubMed PMID: 25541128; PubMed Central PMCID: PMC4310455.

[22] Duthie FA, McGeehan P, Hill S, Phelps R, Kluth DC, Zamvar V, Hughes J, Ferenbach DA. The utility of the additive EuroSCORE, RIFLE and AKIN staging scores in the prediction and diagnosis of acute kidney injury after cardiac surgery. Nephron Clin Pract. 2014;128(1-2):29-38. doi: 10.1159/000357675. Epub 2014 Oct 24. PubMed PMID: 25358798.

[23] Basu RK, Zappitelli M, Brunner L, Wang Y, Wong HR, Chawla LS, Wheeler DS, Goldstein SL. Derivation and validation of the renal angina index to improve the prediction of acute kidney injury in critically ill children. Kidney Int. 2014 Mar;85(3):659-67. doi: 10.1038/ki.2013.349. Epub 2013 Sep 18. PubMed PMID: 24048379; PubMed Central PMCID: PMC4659420.

[24] Sanchez-Pinto LN, Khemani RG. Development of a Prediction Model of Early Acute Kidney Injury in Critically Ill Children Using Electronic Health Record Data. Pediatr Crit Care Med. 2016 Apr 27. [Epub ahead of print] PubMed PMID: 27124567.

[25] Xia Z, Secor E, Chibnik LB, Bove RM, Cheng S, Chitnis T, Cagan A, Gainer VS, Chen PJ, Liao KP, Shaw SY, Ananthakrishnan AN, Szolovits P, Weiner HL, Karlson EW, Murphy SN, Savova GK, Cai T, Churchill SE, Plenge RM, Kohane IS, De Jager PL. Modeling disease severity in multiple sclerosis using electronic health records. PLoS

One. 2013 Nov 11;8(11):e78927. doi: 10.1371/journal.pone.0078927. eCollection 2013. PubMed PMID: 24244385; PubMed Central PMCID: PMC3823928.

[26] C.A. Bejan, L. Vanderwende, H.L. Evans, M.M. Wurfel, M. Yetisgen-Yildiz. On-time clinical phenotype prediction based on narrative reports. Proceedings of the American Medical Informatics Association Fall Symposium (AMIA'13), Washington DC. November, 2013

[27] Kayaalp, M., Browne, A.C., Dodd, Z.A, Sagan, P., McGee, T., McDonald, C.J. (2015). Challenges and Insights in Using HIPAA Privacy Rule for Clinical Text Annotation. *Proceedings of the Annual American Medical Informatics Association Fall Symposium*.

# Appendix A: Details about Variables Included in the Model

*Medications*

The summative doses (in mg) of the 35 medications administered during the Data Collection Window were included in the model. The total number of medications received in Data Collection Window was also recorded as a separate variable. For example, if Patient A received 800 mg of acyclovir, 2000 mg of vancomycin, and 20 mg of furosemide during the Data Collection Window, the value of acyclovir would be 800, the value of vancomycin would be 2000, the value of furosemide would be 20, the value of the remaining 32 medications would be 0, and the value of the number of medications would be 3.

*Procedures*

For the purposes of model construction, "procedures" indicate those procedures that require general anesthesia (with the exception of emergent ECMO cannulation). Thus, if patients received CT scans or MRIs without anesthesia, these were not included in the dataset. Each procedure was tagged with a start time and end time, so procedures were only associated with a specific creatinine if the entirety of the procedure fell within that creatinine's Data Collection Window. The length of the procedure (in minutes) was also recorded, and entered into the model as a continuous variable.

Procedure names were associated with procedure variables as follows:

Central Line: Insertion Cath Dialysis, Insertion Cath Intravenous, Insertion Cath Other, Insertion Cath Other US Guided, Insertion Catheter Apheresis, Insertion Central Venous Line, Insertion PICC Line, zzCath Placement Central Line GEN

Cardiac Catheterization: Card Cath Heart Right, Card Cath w/ Bal Aortic Angioplasty, Card Cath w/ Stent Placement, Cardiac Catheterization

Heart Surgery: Aortopexy, Arterial Switch Procedure, Fontan, Konno Procedure, Repair Aortic Arch, Repair Aortic Coarctation Pump Standby, Repair Aortic Valve, Repair Atrial Septal Defect, Repair Atrioventricular Valve, Repair AV Canal, Repair Coarctation, Repair Cor Triatrium, Repair Interrupted Aortic Arch, Repair Mitral Valve, Repair Partial Anom Pulm Venous, Repair Pulmonary Artery, Repair Pulmonary Valve, Repair RV Outflow Tract Obstruction, Repair Sub Aortic Stenosis, Repair Supravalvular Aortic Stenosis, Repair Tricuspid Valve, Repair Ventricular Septal Defect, Replacement Aortic Valve, Replacement Mitral Valve, Replacement Pulmonary Valve, Ross Konno Procedure, RV to PA Conduit, zzOH ASD Secundum Repair LT 8 yrs, zzOH Fontan LT 8 yrs, zzOH Anom Pulm Venous Partial Rep LT 8yr, zzOH Left Ventric Asst Dev Plcmt GT 8yrs, zzOH Damus Kaye Stansel Procedure LT 8 yrs, zzOH Ventricular Septal Defect Repair LT 8 yrs, zzOH Aortic Valve Repair LT 8 yrs, zzOH ASD Sinus Venosus Repair LT 8 yrs, zzOH Conduit Replace RV to PA LT 8 yrs, zzOH

Pulmonary Valve Replacement GT 8 yrs, zzOH AV Septal Defect Repair LT 8 yrs, zzOH Coronary Osteoplasty GT 8 yrs, zzOH Sub Aortic Stenosis Repair LT 8 yrs, zzOH Conduit Replace RV to PA LT 8 yrs, zzOH Left Ventric Asst Dev Plcmt GT 8 yrs, zzOH Konno Procedure GT 8 yrs, zzOH Left Ventric Asst Dev Plcmt LT 8 yrs, zzOH Tetralogy of Fallot Repair LT 8 yrs

CT/MRI scans: CT scan, MRI

ECMO: ECMO Cannulation, zzECMO Cannulation GEN

*Labs*

Because not every patient has every lab drawn every day, lab values represented the largest source of missing data in the model. As such, if greater than 50% of cases were missing a specific lab test, that lab test was excluded. This resulted in the inclusion of only 10 general laboratory values (sodium, potassium, chloride, bicarbonate, blood urea nitrogen, glucose, white blood cell count, hemoglobin, hematocrit, and platelet count), and 4 drug levels (vancomycin, gentamicin, tacrolimus, and cyclosporine).

Four measures were obtained for each general laboratory value: the mean of the measurements obtained during the Data Collection Window, the standard deviation of those measurements, the proportion of values above the upper limit of normal, and the proportion of values below the upper limit of normal. Missing data was imputed using median values.

The upper and lower limits of normal for each general laboratory value were obtained from the Seattle Children's Hospital EHR, and are recorded in the table below.

Table A.1. Normal ranges for labs included in the model, extracted from the SCH EHR.

| Lab Value | Normal Lower Bound | Normal Upper Bound |
|---|---|---|
| **Sodium** | 135 | 145 |
| **Potassium** | 3.5 | 5.5 |
| **Chloride** | 96 | 109 |
| **Bicarbonate** | 18 | 27 |
| **Blood Urea Nitrogen** | 6 | 20 |
| **WBC Count** | 4.5 – 6 (varies with age) | 11 – 15.5 (varies with age) |
| **Hemoglobin** | 10.5 – 13.5 (varies with age and gender) | 13.5 – 17.5 (varies with age and gender) |
| **Hematocrit** | 33 – 41 (varies with age and gender) | 39 – 53 (varies with age and gender) |
| **Platelet Count** | 150 – 250 (varies with age) | 450 – 600 (varies with age) |

The mean value of each drug level was also calculated. Patients who did not receive vancomycin, gentamicin, tacrolimus, or cyclosporine during the Data Collection Window were

assumed to have a level of zero. Because drug levels are often checked infrequently, there were very few cases in which more than one value occurred during the Data Collection Window, so standard deviations were not performed. The proportion of values above a given threshold were also obtained for vancomycin and gentamicin, but not for tacrolimus and cyclosporine, since the target level for these drugs varies widely depending on indication. The threshold for vancomycin was set at 15, and at 2 for gentamicin.

Finally, the frequency with which electrolytes and a complete blood count were obtained during the Data Collection Window was recorded. The electrolyte frequency was determined by averaging the frequencies of sodium, potassium, chloride, and bicarbonate checks, and the complete blood count frequency was determined by averaging the frequencies of the white blood cell count, hemoglobin, hematocrit, and platelet count checks.

*Vital Signs*

Six vital signs were included in the model: heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, temperature, and oxygen saturation. Because vital signs are checked relatively frequently on hospitalized patients, there was enough data to divide the Data Collection Window into two halves, Cut 1 and Cut 2 (with Cut 1 being the earlier half of the window).

Means, standard deviations, and the proportion of vitals above and below the normal range were calculated for each vital in each cut. The difference between Cut 1 and Cut 2 was then calculated for the means, standard deviations, and proportions. Finally, the frequency of each vital was calculated, and an overall Vitals Frequency determined by averaging the frequency of heart rate, respiratory rate, blood pressure, temperature, and oxygen saturation checks.

The normal range for each vital sign was determined in two ways. Firstly, aged-based normal bounds were obtained from the Seattle Children's Hospital EHR. Secondly, the mean and standard deviation for each vital sign during the entirety of the Data Collection Window was calculated, and the upper threshold of normal was set at two standard deviations above the mean, and the lower threshold at two standard deviations below the mean. The two methods of calculating a normal range were then compared, with the lower of the two upper bounds and the higher of the two lower bounds used as the final cutoff points. (Upper bounds were ignored for oxygen saturation, since 100% saturation would still be considered normal in most instances.)

The normal bounds from the Seattle Children's Hospital EHR are in the table below.

Table A.2. Normal ranges for vital signs included in the model, extracted from the SCH EHR.

| Vital Sign | Normal Lower Bound | Normal Upper Bound |
|---|---|---|
| **Heart Rate** | 60 – 90 (varies with age) | 130 – 150 (varies with age) |
| **Respiratory Rate** | 14 – 20 (varies with age) | 20 – 40 (varies with age) |
| **Systolic Blood Pressure** | 72 – 90 (varies with age) | 106 – 140 (varies with age) |
| **Diastolic Blood Pressure** | 55 – 65 (varies with age) | 65 – 90 (varies with age) |
| **Temperature** | 35.5 | 38.0 |
| **Oxygen Saturation** | 85% | n/a |

*Other*

FiO2: The mean FiO2 values and proportion of FiO2 values above 21% were calculated for each case. If FiO2 was not recorded, it was assumed to be 21%.

GCS: The mean Glasgow Coma Scores and proportion of Glasgow Coma Scores below 13 were calculated for each case. If GCS was not recorded, it was assumed to be 15.

ICU Admission: A binary variable; positive if a patient was admitted to the ICU during the Data Collection Window.

Mean weight: The mean weight during the Data Collection Window was calculated for each case.

*Patient Level Variables*

Age: Recorded in days, and determined at the time of hospital admission.

Gender: The patient's gender, determined at the time of hospital admission.

Admitting Service: This is a factorial variable with one of 13 values:

0 – General Medicine
1 – Pediatric ICU
2 – Cardiac ICU
3 – Bone Marrow Transplant
4 – Hematology/Oncology
5 – Cardiac Surgery
6 – General Surgery
7 – Neurosurgery
8 – Oral Surgery
9 – Orthopedics
10 – Otolaryngology
11 – Plastic Surgery
12 – Urology

13 – Cardiology
14 – Nephrology
15 – Pulmonary
16 – Gastroenterology
17 – Rheumatology

Prior AKI: As described in the manuscript, this is a binary variable that is set to true if a patient has experienced an episode of AKI in the six months prior to hospital admission.

# Appendix B: Information Regarding Unstructured Models

*Section Headers*

CHIEF COMPLAINT
HISTORY OF PRESENT ILLNESS
PROBLEM LIST
ALLERGIES
PAST MEDICAL HISTORY
BIRTH HISTORY
MEDICATIONS
SOCIAL HISTORY
FAMILY HISTORY
REVIEW OF SYSTEMS
IMMUNIZATIONS
PHYSICAL EXAMINATION
LABORATORIES
GLOBAL ASSESSMENT
IMPRESSION AND PLAN
ANALYSIS
CONSULTATIONS
PROCEDURES
RESPIRATORY
FLUIDS/ELECTROLYTES/NUTRITION
CARDIOVASCULAR
RENAL
HEMATOLOGY
HEMATOLOGIC
HEMATOLOGY-ONCOLOGY
ONCOLOGY
ONCOLOGIC
ENDROCRINE
INFECTIOUS DISEASE
GASTROINTESTINAL
NEUROLOGIC
NEUROLOGY
SOCIAL
PAIN MANAGEMENT
PSYCHIATRIC
PERSONALNAME (tag from deindentification)
DATE (tag from deindentification)
ALPHANUMERICID (tag from deindentification)

*Stop Word List*

Seattle
Children's
Hospital
She
He
Her
His
Him
A
About
Again
All
Almost
Also
Although
Always
Among
An
And
Another
Any
Are
As
At
Be
Because
Been
Before
Being
Between
Both
But
By
Can
Could
Did
Do
Does
Done
Due
During
Each
Either
Enough
Especially
Etc
For
Found
From
Further
Had
Has
Have
Having
Here
How
However
I
If
In
Into
Is
It
Its
Itself
Just
Made
Mainly
Make
May
Might
Most
Mostly
Must
Nearly
Of
Often
On
Our
Overall
Perhaps
Quite
Rather
Really
Regarding
Seem
Seen
Several
Should
Since
So
Some
Such
Than
That
The
Their
Theirs
Them
Then
There
Therefore
These
They
This
Those
Thus
To
Upon
Various
Was
We
Were
What
When
Which
While
Who
With
Would

# Appendix C: Performance of the Structured Data and N-gram Model at Various Thresholds

Table C.1. Performance of the model at different N-gram inclusion thresholds.

| Threshold* | Best F1 Score | Best AUC |
|---|---|---|
| 1% | 0.74 | 0.76 |
| 2.5% | 0.76 | 0.77 |
| 5% | 0.73 | 0.74 |
| 7.5% | 0.72 | 0.73 |
| 10% | 0.72 | 0.73 |

*Threshold refers to the percentage of patients an n-gram had to appear in to be included in the final feature set. For example, 1% indicates that the n-gram must be present in at least 1% of patients.

Table C.2. Performance of the 2.5% threshold model at various feature counts.

| **Feature Count** | **F1 Score** | **AUC** |
|---|---|---|
| 25 | 0.71 | 0.71 |
| 50 | 0.70 | 0.70 |
| 75 | 0.72 | 0.73 |
| 100 | 0.73 | 0.74 |
| 125 | 0.74 | 0.74 |
| 150 | 0.74 | 0.74 |
| 175 | 0.74 | 0.74 |
| 200 | 0.75 | 0.76 |
| 225 | 0.76 | 0.77 |
| 250 | 0.76 | 0.77 |
| 275 | 0.76 | 0.77 |
| 300 | 0.76 | 0.77 |
| 325 | 0.76 | 0.77 |
| 350 | 0.76 | 0.77 |
| 375 | 0.76 | 0.77 |
| 400 | 0.76 | 0.77 |
| 425 | 0.76 | 0.77 |
| 450 | 0.76 | 0.77 |
| 475 | 0.76 | 0.77 |
| 500 | 0.76 | 0.77 |
| 525 | 0.76 | 0.77 |
| 550 | 0.76 | 0.77 |
| 575 | 0.71 | 0.73 |
| 600 | 0.71 | 0.73 |
| 625 | 0.71 | 0.73 |
| 650 | 0.71 | 0.73 |
| 675 | 0.71 | 0.73 |
| 700 | 0.71 | 0.73 |

# Appendix D: Performance of Other Structured Models

Table D.1. Results using a 36 hour Data Collection Window, 36 hour Prediction Length, and AKI defined using KDIGO Stage 1 criteria.

| | **No AKI** | **AKI** | |
|---|---|---|---|
| **Predicts No AKI** | 1340 | 273 | NPV: 0.83 |
| **Predicts AKI** | 325 | 577 | PPV: 0.64 |
| | Specificity: 0.80 | Sensitivity: 0.68 | |

F1 Score: 0.66 AUC: 0.74

Table D.2. Results using a 24 hour Data Collection Window, 48 hour Prediction Length, and AKI defined using KDIGO Stage 1 criteria.

| | **No AKI** | **AKI** | |
|---|---|---|---|
| **Predicts No AKI** | 1249 | 285 | NPV: 0.81 |
| **Predicts AKI** | 416 | 565 | PPV: 0.58 |
| | Specificity: 0.75 | Sensitivity: 0.66 | |

F1 Score: 0.62 AUC: 0.71

Table D.3. Results using a 48 hour Data Collection Window, 24 hour Prediction Length, and AKI defined using Modified KDIGO Stage 1 criteria.

| | **No AKI** | **AKI** | |
|---|---|---|---|
| **Predicts No AKI** | 2084 | 118 | NPV: 0.95 |
| **Predicts AKI** | 114 | 199 | PPV: 0.64 |
| | Specificity: 0.95 | Sensitivity: 0.63 | |

F1 Score: 0.63 AUC: 0.79

Table D.4. Results using a 48 hour Data Collection Window, 24 hour Prediction Length, and AKI defined using Modified KDIGO Stage 1 criteria, excluding creatinine values with a baseline of 0.3 mg/dL or less.

| | **No AKI** | **AKI** | |
|---|---|---|---|
| **Predicts No AKI** | 839 | 108 | NPV: 0.89 |
| **Predicts AKI** | 116 | 264 | PPV: 0.69 |
| | Specificity: 0.88 | Sensitivity: 0.71 | |

F1 Score: 0.70 AUC: 0.79

Table D.5. Results using a 48 hour Data Collection Window, 24 hour Prediction Length, and AKI defined using KDIGO Stage 2 criteria.

| | **No AKI** | **AKI** | |
|---|---|---|---|
| **Predicts No AKI** | 1880 | 147 | NPV: 0.93 |
| **Predicts AKI** | 200 | 288 | PPV: 0.59 |
| | Specificity: 0.90 | Sensitivity: 0.66 | |

F1 Score: 0.62 AUC: 0.78

Table D.6. Results using a 48 hour Data Collection Window, 24 hour Prediction Length, and AKI defined using KDIGO Stage 2 criteria, excluding creatinine values with a baseline of 0.3 mg/dL or less.

| | **No AKI** | **AKI** | |
|---|---|---|---|
| **Predicts No AKI** | 1033 | 133 | NPV: 0.89 |
| **Predicts AKI** | 81 | 80 | PPV: 0.50 |
| | Specificity: 0.93 | Sensitivity: 0.38 | |

F1 Score: 0.43 AUC: 0.65

Table D.7. Results using a 48 hour Data Collection Window, 24 hour Prediction Length, and AKI defined using KDIGO Stage 3 criteria.

| | **No AKI** | **AKI** | |
|---|---|---|---|
| **Predicts No AKI** | 2049 | 393 | NPV: 0.84 |
| **Predicts AKI** | 31 | 42 | PPV: 0.58 |
| | Specificity: 0.99 | Sensitivity: 0.10 | |

F1 Score: 0.17 AUC: 0.54

Table D.8. Results using a 48 hour Data Collection Window, 24 hour Prediction Length, and AKI defined using KDIGO Stage 3 criteria, excluding creatinine values with a baseline of 0.3 mg/dL or less.

| | **No AKI** | **AKI** | |
|---|---|---|---|
| **Predicts No AKI** | 1092 | 183 | NPV: 0.86 |
| **Predicts AKI** | 22 | 30 | PPV: 0.58 |
| | Specificity: 0.98 | Sensitivity: 0.14 | |

F1 Score: 0.23 AUC: 0.56

Table D.9. Results using a 48 hour Data Collection Window, 24 hour Prediction Length, and AKI defined using KDIGO Stage 1 criteria, new-onset AKI only.

| | **No AKI** | **AKI** | |
|---|---|---|---|
| **Predicts No AKI** | 1128 | 188 | NPV: 0.86 |
| **Predicts AKI** | 405 | 132 | PPV: 0.25 |
| | Specificity: 0.74 | Sensitivity: 0.41 | |

F1 Score: 0.31 AUC: 0.57

Table D.10. Results using a 48 hour Data Collection Window, 24 hour Prediction Length, and AKI defined using KDIGO Stage 1 criteria, ongoing AKI only.

| | **No AKI** | **AKI** | |
|---|---|---|---|
| **Predicts No AKI** | 75 | 94 | NPV: 0.44 |
| **Predicts AKI** | 142 | 527 | PPV: 0.79 |
| | Specificity: 0.35 | Sensitivity: 0.85 | |

F1 Score: 0.82 AUC: 0.60