KI REPORTS
KIReports.org
CLINICAL RESEARCH

Predicting Kidney Transplant Recipient Cohorts' 30-Day Rehospitalization Using Clinical Notes and Electronic Health Care Record Data

Michael Arenson[1,2], Julien Hogan[1], Liyan Xu[3], Raymond Lynch[1], Yi-Ting Hana Lee[1], Jinho D. Choi[3], Jimeng Sun[4], Andrew Adams[5] and Rachel E. Patzer[1,6]

[1]Department of Surgery, Division of Transplantation, Emory University School of Medicine, Atlanta, Georgia, USA; [2]Department of Pediatrics, Child Health Equity Center, UMass Chan Medical School, Worcester, Massachusetts, USA; [3]Department of Computer Science, Emory University, Atlanta, Georgia, USA; [4]Department of Computer Science, University of Illinois, Urbana-Champaign, Champaign, Illinois, USA; [5]Department of Surgery, Division of Transplantation, University of Minnesota, Minneapolis, Minnesota, USA; and [6]Department of Epidemiology, Rollins School of Public Health Emory University, Atlanta, Georgia, USA

Introduction: Rehospitalization after kidney transplant is costly to patients and health care systems and is associated with poor outcomes. Few prediction model studies have examined whether inclusion of clinical notes data from the electronic medical record (EMR) enhances prediction of rehospitalization.

Methods: In a retrospective, observational study of first-time, adult kidney transplant recipients at a large, urban hospital in southeastern United States (2005–2015), we examined 30-day rehospitalization (30DR) using structured EMR and unstructured (i.e., clinical notes) data. We used natural language processing (NLP) methods on 8 types of clinical notes and included terms in predictive models using unsupervised machine learning approaches. Both the area under the receiver operating curve and precision-recall curve (ROC and PRC, respectively) were used to determine and compare model accuracy, and 5-fold cross-validation tested model performance.

Results: Among 2060 kidney transplant recipients, 30.7% were readmitted within 30 days. Predictive models using clinical notes did not meaningfully improve performance over previous models using structured data alone (ROC 0.6821; 95% confidence interval [CI]: 0.6644, 0.6998). Predictive models built using solely clinical notes performed worse than models using both clinical notes and structured data. The data that contributed to the top performing models were not identical but both included structured data and progress notes (ROC 0.6902; 95% CI: 0.6699, 0.7105).

Conclusions: Including new features from clinical notes in risk prediction models did not substantially increase predictive accuracy for 30DR for kidney transplant recipients. Future research should consider pooling data from multiple institutions to increase sample size and avoid overfitting models.

Kidney Int Rep (2023) **8,** 489–498; https://doi.org/10.1016/j.ekir.2022.12.006
KEYWORDS: early readmission; kidney transplantation; machine learning; natural language processing; predicting readmission; risk prediction

Postdischarge rehospitalization after kidney transplant is a common and preventable problem that is both costly to patients and health care systems and is associated with poor outcomes. More than 50% of patients are hospitalized in the year following kidney transplantation, and posttransplant hospitalization is associated with higher rates of graft loss,[1,2] lower patient survival, and poor quality of life.[3] These poor outcomes are significantly more pronounced among historically marginalized groups. Meanwhile, the 30DR rate is widely utilized by policymakers and payers as an important surrogate metric of hospital quality and a strong correlate of mortality.[4] Reducing the frequency and costs associated with preventable 30DR is thought to be essential to improving the quality of the health care system.

Predictive analytics have previously been used to identify patients at risk of rehospitalization, identifying a number of demographic, socioeconomic, clinical, transplant surgery, health care utilization factors;

Correspondence: Rachel E. Patzer, Department of Surgery, Emory University School of Medicine, 101 Woodruff Circle, 5101 WMB, Atlanta, Georgia 30322, USA. E-mail: rpatzer@emory.edu
Received 20 May 2022; revised 4 December 2022; accepted 5 December 2022; published online 12 December 2022

and timing of readmission as important risk factors for early rehospitalization after kidney transplantation.[5-10] However, these data are often limited to surveillance data sets and are of variable quality and limited granularity. Generally, these structured data are static and do not reflect the dynamic nature of factors known to impact all levels of the end stage kidney disease care trajectory, including risk factors or protective factors identified during the pretransplant period, during the transplant surgery, and the perioperative period immediately following transplantation but before discharge from the hospital.[5,11-14]

In addition, studies in transplantation focused on hospitalization outcomes have targeted structured data, rather than utilizing potentially novel predictive variables such as those from free-text clinical notes within EMR. Clinical notes may represent an untapped data source that could provide novel predictor variables to enhance real-time prediction in medicine.[15] NLP is a tool that can be used to analyze EMR free-text documentation from patient's clinical notes and has been utilized in surgery.[16,17] For example, physician notes have been analyzed to predict mortality in patients admitted to the surgical intensive care unit[16] as well as identify postoperative complications.[17]

However, to our knowledge, no research studies have examined the potential accuracy of clinical notes in predicting rehospitalization among kidney transplant recipients. We hypothesized that incorporating unstructured data from clinical notes into predictive models would improve predictive accuracy of postdischarge hospitalization of kidney transplant recipients. Therefore, the objectives of our study were as follows: (i) to characterize kidney transplant-related clinical notes using NLP, and (ii) to determine if accuracy of a 30DR predictive model using traditional structured data elements from the EMR could be improved with the use of NLP-derived variables from patients' clinical notes. We examined predictors from the pretransplant time through the time of patient discharge, with the intent to eventually use this model in real time to inform clinical care prior to the time of patient discharge.

METHODS

Study Population and Data Sources

This was a retrospective observational analysis of adult kidney transplant recipients at a large institution in southeastern US who received their transplant between January 2005 and December 2015 and received follow-up care at the study institution. The data sources considered for inclusion in the study included pretransplant data available within the institution's EMR, as well as perioperative data available immediately after surgery but before discharge. Data was categorized according to whether it pertained to the donor, the recipient, or it related to the transplant process (e.g., transplant surgery). There were 2060 patients included in the primary analysis ($n = 2060$, year of first-time transplant 2005−2015) (Figure 1). Clinical notes (described below) were obtained from the pretransplant evaluation process as well as at the time of the index transplant hospitalization but before the date of discharge, with the intention that a predictive model for early hospital readmission would be most helpful to hospital staff before discharging the patient home from the hospital. Except for selection committee and progress notes, all other note types were included from 1 year before transplant and through the time of discharge.

Study Outcome

The outcome of interest was rehospitalization within 30 days of discharge from transplant. Rehospitalization was defined as the first unplanned hospital admission postdischarge from the patient's index hospitalization at the time of transplant. We assumed a hospitalization was unplanned if the patient was admitted through the institution's emergency department.

Data Variables and Definitions

A full list of structured variables and unstructured data sources can be found in Supplementary Table S1. The data were collected from the hospital's EMR. This was then linked to national surveillance data available on waitlisting and transplantation from the United Network for Organ Sharing data, to avoid excluding patients for missing structured data.

Unstructured Data Characteristics

Unstructured data comprised free-text clinical notes in the EMR. We obtained 8 note types available from the local institution's EMR database. We used a multidisciplinary team of clinicians to determine which clinical notes were both available within the EMR and clinically relevant. We considered all of the 8 types of clinical notes available within the EMR for this patient population, starting at the time of transplant evaluation (pretransplantation) and through the initial discharge from the hospital after the transplant surgery. These notes included the following: social work (n =1118), waitlist selection committee note ($n = 2033$), echo report ($n = 1110$), history and physical ($n = 1422$), consultation note ($n = 1354$), progress note ($n = 1415$), operative ($n = 1472$), and discharge summary ($n = 517$) (Supplementary Table S1). Only a small number of these patients did not have any unstructured data ($n =$ 18 had no clinical notes) but were still included in the

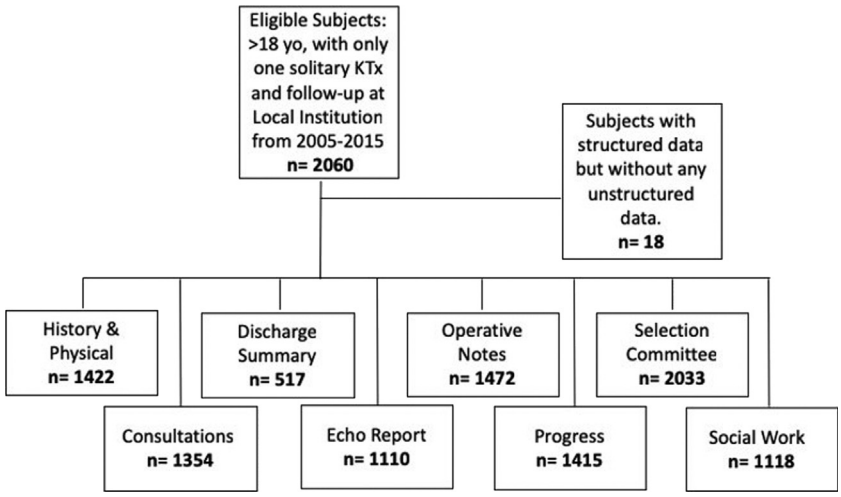

Figure 1. Inclusion and exclusion flow chart. Only patients with structured data were included in the analysis ($N = 2060$, years 2005–2015). Patients were not excluded if they were missing clinical notes ($n = 18$ had no clinical notes).

structured data analysis. Because of their file size and relevance, only those progress notes written between the time of admission for transplant surgery and subsequent discharge were analyzed.

Statistical Analysis

Patient demographic and clinical characteristics were conducted using SAS (Cary, NC). Baseline models were constructed using all available structured variables, including age, race, ethnicity, dialysis vintage, comorbidities, and other variables known to be associated with 30DR.[11,12,18] Variables were chosen based on statistical significance ($P \leq 0.05$) or previous identification in the literature of contributing to rehospitalization. In total, 80 variables were included in the structured model.

Characterization of transplant-related clinical notes was performed using R[19] and text mining methods.[20] Free text from individual patients' clinical notes of the same type were merged into one long free-text file for each patient. For example, all progress notes from one patient were combined, which allowed progress note text from separate patients to be compared. We used term frequency (TF), term frequency-inverse document frequency (TF-IDF).[21] We have tried more advanced methods in the past, including word embedding,[22] and neural network-based models such as long short-term memory and bidirectional encoder representations from transformers.[23] Based on these previous studies, we found that advanced methods give little meaningful improvement while largely increasing the complexity, especially if they are prone to overfitting, given the size of our data. Therefore, we identified the TF-IDF with linear model logistic regression, which has the advantage of being accurate, efficient, and easy to interpret.

Initially we analyzed all TF-IDF unigram, bigram, and trigrams in the model. Trigrams did not boost performance and easily overfit the model. Therefore, we used unigrams and bigrams in our final model. We discarded any unigrams or bigrams that appeared in fewer than 2 patients' notes.

We used an ensemble learning logistic regression method[24] to construct the models and combine modeling estimates. Ensemble learning is a general meta approach to machine learning that seeks better predictive performance by combining the predictions from multiple models. We created models for each note type using logistic regression, and then averaged the prediction of all the models. We have shown this method is typically more accurate than traditional concatenation approaches.[22]

We built several predictive models. The first baseline model consisted of structured data only and was built using boosting trees because of better performance over logistic regression. For each of the 8 note types, we used TF-IDF words as variables and created a model with logistic regression. Using these models, we examined the predictive accuracy of the notes by themselves and as an addition to the baseline structured data model. Finally, we combined variables from and analyzed separate combinations of notes (e.g., combining progress + consultation notes variables).

Internal validation and the area under the curve for both ROC and PRC were used to determine and compare model accuracy. Much like the ROC, the PRC is used for evaluating the performance of binary classification algorithms and reflects the average precision of the model. It is often used in situations where data sets are imbalanced or skewed, such as when the number of negative observations (e.g., patients not

Table 1. Baseline characteristics of kidney transplant recipients from Emory Transplant Center, stratified by readmission within 30 days posttransplant, 2005–2015

Characters	Study population $N = 2060$	Readmitted within 30 d $n = 633$ (30.7%)	Not readmitted within 30 d $n = 1427$ (69.3%)	*P*-value
Recipient factors				
Demographic				
Age, yr, median (IQR)	51 (40–60)	50 (40–60)	51 (40–60)	0.95
Race, n (%)				0.14
Caucasian or White	900 (43.7)	271 (42.8)	629 (44.1)	
Black or African American	970 (47.1)	315 (49.8)	655 (45.9)	
Others	71 (3.5)	15 (2.4)	56 (3.9)	
Unknown, unavailable or unreported	119 (5.8)	32 (5.1)	87 (6.1)	
Ethnicity, n (%)				0.26
Non-Hispanic or Latino	1587 (77.0)	474 (74.9)	1113 (78.0)	
Hispanic or Latino	68 (3.3)	21 (3.3)	47 (3.3)	
Unknown, unavailable, unreported	405 (19.7)	138 (21.8)	267 (18.7)	
Gender, n (%)				0.45
Male	1194 (58.0)	359 (56.7)	835 (58.5)	
Female	866 (42.0)	274 (43.3)	592 (41.5)	
Clinical				
Primary cause of ESKD, *n* (%)				<0.0001
Diabetes	544 (26.4)	223 (35.2)	321 (22.5)	
Primary GN	423 (20.5)	116 (18.3)	307 (21.5)	
Secondary	107 (5.2)	26 (4.1)	81 (5.7)	
Cystic/hereditary/congenital disease	211 (10.2)	45 (7.1)	166 (11.6)	
Hypertension	577 (28.0)	162 (25.6)	415 (29.1)	
Neoplasms/tumor	15 (0.7)	4 (0.6)	11 (0.8)	
Other	181 (8.8)	56 (8.9)	125 (8.8)	
Missing	2 (0.1)	1 (0.2)	1 (0.1)	
Prior transplants status, n (%)				<0.0001
Yes	163 (7.9)	63 (10.0)	100 (7.0)	
No	1,711 (83.1)	465 (73.5)	1,246 (87.3)	
Unknown	186 (9.0)	105 (16.6)	81 (5.7)	
Karnofsky status				0.0001
Required considerable assistance	740 (35.9)	269 (42.5)	471 (33.0)	
Normal activities with little effort	293 (14.2)	75 (11.9)	218 (15.3)	
Unknown, unavailable or unreported	1027 (49.9)	289 (45.7)	738 (51.7)	
Blood subtype, n (%)				<0.0001
O	852 (41.4)	234 (37.0)	618 (43.3)	
A	617 (30.0)	168 (26.5)	449 (31.5)	
B	306 (14.9)	93 (14.7)	213 (14.9)	
AB	99 (4.8)	33 (5.2)	66 (4.6)	
Unknown, unavailable or unreported	186 (9.0)	105 (16.6)	81 (5.7)	
Comorbidities (all diagnosed before transplant)				
Infectious and parasitic diseases	511 (24.8%)	177 (27.5%)	334 (23.6%)	0.05
Neoplasms	717 (34.8%)	218 (33.9%)	499 (35.2%)	0.56
Endocrine, nutritional and metabolic disease, and immunity disorders	1795 (87.1%)	583 (90.7%)	1,212 (85.5%)	0.001
Diseases of the blood and blood-forming organs	1412 (68.5%)	433 (67.3%)	979 (69.1%)	0.43
Mental disorders	609 (29.6%)	215 (33.4%)	394 (27.8%)	0.009
Diseases of the nervous system and sense organs	775 (37.6%)	285 (44.3%)	490 (34.6%)	<0.0001
Diseases of the circulatory system	2055 (99.8%)	642 (99.8%)	1,413 (99.7%)	0.59
Diseases of the respiratory system	850 (41.3%)	279 (43.4%)	571 (40.3%)	0.19
Diseases of the digestive system	1261 (61.2%)	414 (64.4%)	847 (59.8%)	0.05
Diseases of the genitourinary system	2059 (99.95%)	642 (99.8%)	1,417 (100.0%)	0.14
Complications of pregnancy, childbirth, and the puerperium	31 (1.5%)	6 (0.9%)	25 (1.8%)	0.15
Diseases of the skin and subcutaneous tissue	338 (16.4%)	108 (16.8%)	230 (16.2%)	0.75
Diseases of the musculoskeletal system and connective tissue	735 (35.7%)	244 (38.0%)	491 (34.7%)	0.15
Congenital anomalies	5001 (24.3%)	138 (21.5%)	363 (25.6%)	0.04
Certain conditions originating in the perinatal period	8 (0.4%)	4 (0.6%)	4 (0.3%)	0.25
Social (status up to transplant date)				
Alcohol use, n (%) ($n = 1624$)				0.0002

(Continued on following page)

Table 1. (Continued) Baseline characteristics of kidney transplant recipients from Emory Transplant Center, stratified by readmission within 30 days posttransplant, 2005–2015

Characters	Study population $N = 2060$	Readmitted within 30 d $n = 633$ (30.7%)	Not readmitted within 30 d $n = 1427$ (69.3%)	*P*-value
Deny	1103 (67.9)	360 (74.4)	743 (65.2)	
Past	93 (5.7)	30 (6.2)	63 (5.5)	
Current	428 (26.4)	94 (19.4)	334 (29.3)	
Smoking status, n (%) ($n = 1227$)				0.59
Never smoked	708 (57.7)	179 (54.9)	529 (58.7)	
Former smoker	421 (34.3)	121 (37.1)	300 (33.3)	
Light tobacco smoker	4 (0.3)	2 (0.6)	2 (0.2)	
Current someday	25 (2.0)	6 (1.8)	19 (2.1)	
Current everyday	69 (5.6)	18 (5.5)	51 (5.7)	
Donor factors				
Demographic				
Age, yr, median (IQR)	39 (25–49)	39 (23–49)	39 (27–49)	0.41
Type of transplant donor				0.002
Deceased	1386 (67.3)	458 (72.4)	928 (65.0)	
Living	643 (31.2)	164 (25.9)	479 (33.6)	
Pediatric	31 (1.5)	11 (1.7)	20 (1.4)	
Transplant Factors				
Length of hospital stay, d, median (IQR)	4 (4–6)	5 (4–7)	4 (4–6)	<0.0001
ABO compatible, n (%)				<0.0001
Yes	1930 (93.7)	570 (90.1)	1360 (95.3)	
No	24 (1.2)	4 (0.6)	20 (1.4)	
Unknown	106 (5.2)	59 (9.3)	47 (3.3)	
Labs (peritransplant)				
Creatinine at discharge posttransplant (g/dl), median (IQR), [mean of min–max]	1.9 (1.3–3.7), [1.1–11.0]	2.0 (1.3–4.2), [1.1–11.2]	1.9 (1.3–3.5), [1.0–10.9]	0.09
Missing, n (%)	17 (0.83)	5 (0.79)	12 (0.84)	
White blood cell at transplant (10E3MCL), median (IQR), [mean of min–max]	7.85 (5.8–10.4), [3.0–16.7]	7.5 (5.5–9.7), [2.6–17.8]	7.9 (5.9–10.6), [3.11–16.24]	0.001
Missing n (%)	54 (2.6)	18 (2.8)	36 (2.5)	
Hemoglobin A1C at transplant (percent), median (IQR), [min–max]	5.4 (4.9–6.3), [5.1–7.4]	5.6 (5.0–7.1), [5.1–7.8]	5.3 (4.9–6.0), [5.0–7.3]	<0.0001
Missing n (%)	1037 (50.3)	194 (46.5)	743 (52.1)	
Hemoglobin at discharge posttransplant (GMDL), median (IQR), [mean of min–max]	9.5 (8.6–10.6), [7.9–14.8]	9.4 (8.6–10.4), [7.5–14.8]	9.6 (8.6–10.7), [8.1–14.8]	0.05
Missing n (%)	14 (0.7)	4 (0.6)	10 (0.7)	
Transplant milestones				
No. of d from referral start to evaluation start, mean (SD) (n = number of subjects with nonmissing data)	113.5 (953.4)	89.6 (246.4) ($n = 449$)	123.1 (1117.3) ($n = 1130$)	0.34
No. of d from referral end to evaluation end, mean (SD) (n = number of subjects with nonmissing data)	268.7 (389.9)	314.4 (378.1) ($n = 449$)	250.8 (393.7) ($n = 1130$)	0.005
No. of d from evaluation start to evaluation end, mean (SD) (n = number of subjects with nonmissing data)	280.3 (320.6)	326.4 (353.0) ($n = 460$)	262.8 (305.8) ($n = 1186$)	0.001
No. of d from waitlist start to transplant, mean (SD) (n = number of subjects with nonmissing data)	799.1 (677.5)	819.8 (721.9) (n = 521)	787.7 (655.4) (n = 1300)	0.52

ESKD, end-stage kidney disease; GN, glomerlular nephritis; IQR, interquartile range.

rehospitalized) outweighs the number of positive observations (e.g., patients rehospitalized). A strong PRC has both high positive-predictive value (or, "precision") and high sensitivity ("recall").[25] We used 5-fold cross-validation to test the performance of each model, in which we created a training data subset with 80% of the patients (randomly chosen) and then use the remaining 20% of patients as a model testing set. The average ROC curve was calculated for all 5 cross-validation testing sets. The CI was calculated in a standard way such that a T-distribution with degree of freedom as 4 (because we split 5 fold) was adopted to estimate the mean and variance of the evaluation metrics, and 95% CI was calculated accordingly based on the estimated mean and variance. Analyses were performed using SAS, R, and Python.[26] The study was approved by the Institutional Review Board.

RESULTS

Of 2109 patients with locally performed kidney transplants whose data were accessible since January 2005, 2060 transplant patients met eligibility criteria for inclusion in the final cohort (Figure 1).

Table 2. ROC[a] and PRC[b] for each clinical notes using TF-IDF[c] followed by logistic regression C-statistics for each clinical notes using TF-IDF[a] followed by logistic regression

Note	n	ROC	95% CI	PRC	95% CI
Discharge summary	517	0.6262	0.5836, 0.6688	0.4777	0.4250, 0.5303
Progress	1415	0.6196	0.6002, 0.639	0.4370	0.4011, 0.4728
Selection conference	2033	0.594	0.5739, 0.6141	0.4198	0.3969, 0.4427
Consultations	1354	0.5934	0.559, 0.6278	0.4164	0.3710, 0.4619
History and physical	1422	0.5899	0.5566, 0.6232	0.4190	0.3848, 0.4532
Echo	1110	0.5274	0.4952, 0.5596	0.3277	0.2975, 0.3579
Operative	1472	0.527	0.4946, 0.5594	0.3725	0.3329, 0.4121
Social worker	1118	0.5263	0.4886, 0.564	0.3404	0.2795, 0.4012

CI, confidence interval; PRC, precision-recall curve; ROC, receiver operating curve; TF-IDF, term frequency-inverse document frequency.
[a]Area under the receiver operating curve.
[b]Area under the precision-recall curve. Higher numbers indicate better precision.
[c]Term frequency-inverse document frequency (TF-IDF) model on only unstructured notes using TF-IDF and logistic regression.

Recipient, Donor, and Transplant Characteristics

The study population was predominantly Black or African American (47%) and male (58%); the mean age was 51 years (Table 1). Of the final cohort, 633 (30.7%) were admitted within 30 days of hospital discharge from the initial transplant. Black or African American patients were slightly overrepresented among the population readmitted within 30 days (49.8% vs. 45.9%). The most common cause of end stage kidney disease among patients with 30DR in the study population was diabetes (35.2% vs. 22.5% with no 30DR), and hypertension (26% vs. 25% with no 30DR). Patients who were readmitted (vs. not readmitted) in 30 days were more likely to have a diagnosis of infectious or parasitic diseases (27.5% vs. 23.6%) before transplant; endocrine, nutritional, and metabolic disease, or immunity disorders (90.7% vs. 85.5%); and mental disorders (33.4% vs. 27.8%). In addition, patients with a Karnofsky score of "Required Considerable Assistance" were more likely to be readmitted (42.5% vs. 33%). Patients who had received a prior transplant had a higher likelihood of 30-day readmission (10% vs. 7%). In addition, patients who were readmitted within 30 days took longer to complete the transplant process (326.4 days [SD 353.0] vs. 262.8 days [SD 305.8] for those not readmitted).

With respect to donor factors, patients were more likely to be readmitted within 30 days if their donors were deceased, had hepatitis B, had hepatitis C, and were considered high risk by Centers for Disease Control guidelines. For transplant factors, the posttransplant length of stay was higher among those readmitted within 30 days compared to those not readmitted (5 days; interquartile range: 4−7) versus 4 days (interquartile range: 4−6); and other factors, including higher number of HLA B and HLA DR mismatches, ABO incompatibility, lower white blood cell count at the time of transplant, higher Hemoglobin A1C at transplant were significantly associated with 30DR.

NLP Confirms the Relevance of Other Variables

In addition to elucidating potentially novel variables, TF-IDF can also confirm the importance of other variables in clinical notes that would not otherwise be captured with structured data alone. Bigrams (i.e., word pairs) can identify common forms of support that patients require. For example, among all available notes for all patients, we searched for the most common words that preceded the word "support." The top 5 most common forms of support identified by notes were "care support" ($n = 1756$), "transportation support" ($n = 823$), "emotional support" ($n = 536$), "social support" ($n = 325$), and "family support" ($n = 149$).

Predictive Model Performance of Structured and Unstructured Data

There were 80 structured variables included in the predictive model (Supplementary Table S1). The ROC for the structured model alone was estimated as 0.6821 (95% CI: 0.6644, 0.6998) for 30DR. In contrast, the AUPRC for the structured model alone was 0.4975 CI (95% CI: 0.4706, 0.5244). The results of the TF-IDF and logistic regression models to examine the predictive accuracy of individual clinical notes are shown in Table 2. The ROC model with discharge summary notes had the best performance (c-statistic 0.6262; 95% CI: 0.5836, 0.6688). We then added individual clinical note types to structured data (Table 3). These models improved results, with the best performing model including structured data plus progress notes data (c-statistic 0.6902, [95% CI: 0.6699, 0.7105]), followed by the model with structured data and consultation notes (c-statistic 0.684, [95% CI: 0.664, 0.704]). The results of the predictive accuracy of combining clinical notes together along with the structured data model are reported in Table 4. For ROC curves, the models with different combinations of clinical notes did not perform better than the model with structured data and progress notes only. The top performing AUPRC was 0.5060

Table 3. AUROC and AUPRC model prediction of 30 day hospital readmission adding individual clinical notes to structured data

Data sets (structured + singular clinical note)[a]	n	ROC	95% CI	PRC	95% CI
Progress	1415	0.6902	0.6699, 0.7105	0.5038	0.4805, 0.5270
Consultations	1354	0.684	0.664, 0.704	0.5002	0.4807, 0.5196
Discharge summary	517	0.6832	0.6606, 0.7058	0.4991	0.4673, 0.5308
Selection conference	2033	0.6796	0.6602, 0.699	0.5009	0.4737, 0.5280
History and physical	1422	0.6789	0.6558, 0.702	0.5001	0.4906, 0.5095
Social worker	1118	0.6687	0.6509, 0.6865	0.4808	0.4500, 0.5116
Echo	1110	0.6671	0.6507, 0.6835	0.4768	0.4584, 0.4953
Operative	1472	0.6602	0.6391, 0.6813	0.4760	0.4473, 0.5048
Structured data only	2060	0.6821	0.6644, 0.6998	0.4975	0.4706, 0.5244

AUROC, x; AUPRC, x; CI, confidence interval; PRC, precision-recall curve; ROC, receiver operating curve.
[a]On the structured data, we used boosting trees. For unstructured notes we used TF-IDF followed by logistic regression. We assessed combining structured data with different types of unstructured data.

(95% CI: 0.4757, 0.5362) and layered the following notes on top of structured data: Consultation, Discharge Summary, history and physical, progress, and selection conference notes.

The top predictive terms from the best predictive model in Table 4 can be seen in Supplementary Table S2. The top 15 predictors were all structured variables. For unstructured variables, the beta-coefficient and associated CI was used to rank importance. In a sensitivity analysis, multiple different NLP techniques, such as Word2Vec and Doc2Vec, were evaluated, as well as classifiers other than logistic regression (i.e., Random Forrest). Description of these techniques is outside the scope of this paper, however, NLP techniques employed here had the highest predictive performance.

DISCUSSION

Our study confirms that predictive features described in prior literature and, using NLP, generate several new hypotheses about factors throughout the transplant process that may be predictive of 30DR (from end stage kidney disease diagnosis to posttransplant discharge). To date, studies on readmission among kidney transplant recipients have mostly focused on risk factor identification. Important risk factors for hospitalization previously identified include demographic factors (e.g., older age), socioeconomic factors (lower education and Medicaid insurance), clinical factors (high body mass index and various comorbidities), surgical factors (longer length of stay, receipt of a deceased [vs. Living] donor, older donor age, and surgical complications), utilization factors (pretransplant hospitalization), and adherence to medication.[1,2,5-7,11] Many of these risk factors are reproduced in our study in structured data. Our study lends further support to these previously identified predictive features. In addition, our study offers important insight into the limitations and potential opportunities of using clinical notes to enhance prediction of transplant outcomes.

To our knowledge, there have been no studies that have used clinical notes and NLP to predict 30DR among kidney transplant recipients. Taber *et al.*[12] designed a model for 30DR including fixed transplant predictors (area under the curve 0.63; 95% CI: 0.58–0.69) where the predictive accuracy significantly improved to 0.73 (95% CI: 0.67–0.79) after including posttransplant (and predischarge) dynamic factors such as systolic blood pressure slope during transplant admission, but did not examine clinical notes or use NLP methods to enhance prediction. Other studies have examined the inclusion of unstructured data in the prediction of other outcomes in transplant, including graft failure.[18] In a sensitivity analysis using

Table 4. AUROC and AUPRC from models after adding multiple clinical note types to predictive models for hospital readmission after kidney transplants

Data sets (multiple notes added to structured data model)	ROC	95% CI	PRC	95% CI
Structured data alone	0.6821	0.6644, 0.6998	0.4975	0.4706, 0.5244
Structured data + progress[a]	0.6902	0.6699, 0.7105	0.5038	0.4805, 0.5270
Structured data + history and physical + progress	0.6834	0.6613, 0.7055	0.5023	0.4874, 0.5171
Structured data + consultations + discharge summary + history and physical	0.6827	0.6587, 0.7067	0.5033	0.4782, 0.5285
Structured data + consultations + discharge summary + history and physical + progress + selection conference	0.679	0.6555, 0.7025	0.5060	0.4757, 0.5362
Structured data + all clinical notes	0.6639	0.6409, 0.6869	0.4884	0.4644, 0.5124

AUROC, x; AUPRC, x; CI, confidence interval; PRC, precision-recall curve; ROC, receiver operating curve.
[a]The best performing model was structured data + progress note only. We examined adding multiple types of notes on top of structured data. As we added more notes, the ROC declined and PRC remained relatively stable.

AUPR, the best overall performing model changed from "Structured Data + Progress Notes" to "Structured Data + Consultations + Discharge Summary + history and physical + Progress + Selection Conference." The latter combination, with the higher AUPR, may be better at distinguishing readmission under a skewed class distribution and therefore may be a better option in practice. Our model shows a range between 0.4884 and 0.5060. A high PRC is a more difficult metric to acquire than ROC. There is no agreement about what constitutes a good or adequate precision. But given the imbalanced nature of our data, our expert coauthors categorized PRC ranges above 83% as good precision, 50% to 83% as moderate precision, and below 50% as low precision.

NLP employs computational techniques to learn, understand, and produce human language content and can be used to analyze and learn from the enormous quantity of human language content that is now available in the EMR and health care systems.[27] Srinivas *et al*. successfully applied this type of approach to the prediction of graft loss and mortality after kidney transplantation and reported a high accuracy of their models of 0.87; 95% CI: 0.81 to −0.94 for 1-year graft loss and 0.84; 95% CI: 0.80–0.89 for 3-year mortality.[18] Of note, the variables they used were based on previously reported risk factors and clinical input of transplant experts. This process can be time-consuming and complex, which makes it difficult to generalize outside of a single institution because of different reporting techniques. In other words, identifying novel predictive features in this way requires a supervised approach. However, using an unsupervised approach as described above identified potentially novel text variables that would otherwise be less likely to be identified as a predictor. DuBay *et al*.[28] also used data from clinical notes from the transplant period through 1 year posttransplant to predict 5-year graft failure, finding that EMR and NLP methods substantially increased the predictive validity compared to United Network for Organ Sharing registry data alone; however, differences between structured and unstructured data were not reported.[28]

We leveraged machine learning techniques to generate predictive features in an unsupervised manner. This approach requires minimal input from clinical experts, instead relying on computer algorithms to identify important predictive features. We report that layering of data sources does not augment predictive accuracy much more than adding single notes data to the baseline structured model. Although many potential predictors were identified, the approach did not yield an overall higher predictive accuracy and perhaps reinforces the need for a more balanced machine-human partnership and/or the need for more data to fully take advantage of these big data methods. In other words, using a machine learning NLP approach to generate previously unrecognized important words or topics but employing a human-driven decision regarding which predictive features to include in predictive models. Given the success in improving predictive accuracy that Srinivas *et al*. demonstrates, this more time-intensive approach rather than a machine learning based approach alone may be more impactful.

After characterizing the clinical notes, the question remains why all these new potential variables resulted in minimal to no improvement in predictive accuracy. One reason may be prefilled note templates (e.g., commonly used Epic "dot" phrases) that employ prepopulated text, which limits the predictive utility of NLP-derived variables. Although exploring term frequency on its own can provide insight into how language is used in a collection of natural language, TF-IDF is an unsupervised process that identifies words that are unique to each note out of the collection of different types of notes. Each of these terms are theoretically the words most unique to that note and were considered as predictive features in addition to other structured features. For example, there are very few structured data that indicate a patient's mental well-being or literacy despite knowing that it is likely an important predictor of 30DR. However, in the social worker note, the words "affording," "literacy," and "motivations" pertain to patients' social history and mental health and might be used as predictive features. Clinically, however, given the extent of template notes that are prefilled, the utility of the predictive power of these words was questioned by our clinical experts.

A limitation of this study is the size of our cohort and the unbalanced number of different types of notes. These were all the notes available to us within our local EMR, however some may have been missing for various reasons. Some patients may have been cared for at other transplant hospitals and were transferred to our center, some processes have changed clinical care and the location where notes are stored over time. In addition, the nomenclature of notes may have changed. However, we expect these are similar issues across other large transplant centers. The transplant center providing the data for our cohort is one of the largest transplant centers in the US, and yet n=2060 is quite small when attempting to use machine learning techniques. Increasing the sample size for the purposes of analyzing free-text clinical notes is a challenge, however, because doing so would require multiple transplant centers to create a repository of notes. Such a database for kidney transplant patients does not

currently exist, and it is often difficult to completely de-identify clinical notes for sharing. A long-term goal for the field, as others have suggested, should be an informatics approach, using a "common data model" to integrate disparate data sources, data elements (including clinical notes), and temporal data points.[29] It would require further research into how free-text notes are written and organized in EMRs at transplant centers around the country.

Another limitation is that 7.9% of our cohort had a prior transplant according to a sensitivity analysis linking United Network for Organ Sharing to our data set. In addition, if a patient was readmitted at another hospital, it would not be reflected in our analysis. The reason for only selecting patients using local institution's data, however, is because a predictive model that can be integrated into clinical care must be able to use data in real time. In addition, United States Renal Data System data have a lag time of 2 years.

Missing data is a major concern in big data research. Because academic medical centers have relatively recently transitioned from paper to electronic medical charts, the available free-text data is limited. Furthermore, as the databases for EMR have developed, the scaffolding has been built in a patchwork approach. Therefore, some clinical note types were written in the EMR before others or were transferred from one database storage type to another. At each stage, clinical notes loose information within the document, go missing after transfer, or were never transferred to the EMR in the first place. This is a missing data problem that can introduce bias in multiple forms. For example, patients receiving a transplant many years ago may not have as much free-text data. To address this, we looked at patients receiving transplant in 2005 or later, when there was a clear change in the availability of free-text notes in the EMR database. The difference in the number of missingness in the 8 notes analyzed for this study indicate the challenge of analyzing EMR data.

Finally, though this study is the first to explore NLP in kidney transplant patients, there remains room for improvement and refinement. Looking at word frequency, for example, does not account for the context in which those words appear. For example, a word can often be preceded by negating words like, "not" or "never." If a social worker writes, "the patient denies alcohol use," and we assume that alcohol use is a risk factor for 30DR, the analysis employed in this study would possibly erroneously associate the word "alcohol" as predictive of 30DR, even though the patient denied using it. Increasing the sophistication of NLP techniques may be worthwhile.

In conclusion, we have characterized 8 clinical notes and mined them for possible new predictive features that might be useful to improve predictive accuracy of 30DR. Predictive models using unstructured, free-text clinical notes were built using unsupervised approaches to machine learning. These predictive models did not meaningfully improve predictive accuracy above structured data alone. However, the results generated several new hypotheses regarding potentially novel predictors to be examined in future research applying more human-driven approaches. The vast amount of free-text data in the form of clinical notes that exist in the EMR has been untapped in the field of kidney transplant. As we become more reliant on big data and machine learning methods, NLP is possibly the key to leveraging these notes for research that will ultimately help the patient, the physicians, and the hospital improve outcomes. To do this, a "common data model" should be built to integrate data elements (e.g., clinical notes) from disparate data sources (e.g., from different transplant centers).

DISCLOSURE

All the authors declared no competing interests.

ACKNOWLEDGMENTS

We would like to acknowledge Bonggun Shin, PhD for his analysis during early stages of this project. This research was funded by the National Institutes on Minority Health and Health Disparities (R01MD011682), and was also supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR002378 and TL1TR002382.

SUPPLEMENTARY MATERIAL

Supplementary File (PDF)

Table S1. List of all structured data variables and all unstructured data sources by time collected in the kidney transplant process.

Table S2. Ranking top 10 positive and negative predictive features for higher 30-day readmission of kidney transplant recipients (2005-2015) in highest performing predictive model from Table 4 (structured data + progress note).

STROBE Statement.

REFERENCES

1. Lynch RJ, Zhang R, Patzer RE, et al. First-year waitlist hospitalization and subsequent waitlist and transplant outcome. *Am J Transplant*. 2017;17:1031–1041. https://doi.org/10.1111/ajt.14061
2. Lynch RJ, Zhang R, Patzer RE, et al. Waitlist hospital admissions predict resource utilization and survival after renal transplantation. *Ann Surg*. 2016;264:1168–1173. https://doi.org/10.1097/sla.0000000000001574
3. Johnson CD, Wicks MN, Milstead J, et al. Racial and gender differences in quality of life following kidney transplantation.

Image J Nurs Sch. 1998;30:125–130. https://doi.org/10.1111/j.1547-5069.1998.tb01266.x

4. Kripalani S, Theobald CN, Anctil B, Vasilevskis EE. Reducing hospital readmission rates: current strategies and future directions. *Annu Rev Med*. 2014;65:471–485. https://doi.org/10.1146/annurev-med-022613-090415
5. McAdams-Demarco MA, Grams ME, Hall EC, et al. Early hospital readmission after kidney transplantation: patient and center-level associations. *Am J Transpl*. 2012;12:3283–3288. https://doi.org/10.1111/j.1600-6143.2012.04285.x
6. Axelrod DA, Dzebisashvili N, Schnitzler MA, et al. The interplay of socioeconomic status, distance to center, and interdonor service area travel on kidney transplant access and outcomes. *Clin J Am Soc Nephrol*. 2010;5:2276–2288. https://doi.org/10.2215/CJN.04940610
7. Schold JD, Buccini LD, Kattan MW, et al. The association of community health indicators with outcomes for kidney transplant recipients in the United States. *Arch Surg*. 2012;147:520–526. https://doi.org/10.1001/archsurg.2011.2220
8. Tsai TC, Joynt KE, Orav EJ, et al. Variation in surgical-readmission rates and quality of hospital care. *N Engl J Med*. 2013;369:1134–1142. https://doi.org/10.1056/NEJMsa1303118
9. Patzer RE, Serper M, Reese PP, et al. Medication understanding, non-adherence, and clinical outcomes among adult kidney transplant recipients. *Clin Transpl*. 2016;30:1294–1305. https://doi.org/10.1111/ctr.12821
10. Hogan J, Arenson MD, Adhikary S, et al. Timing matters: improving prediction of hospital readmission post kidney transplantation. *Transplantation*. 2019;19:1096–1097.
11. Harhay M, Lin E, Pai A, et al. Early rehospitalization after kidney transplantation: assessing preventability and prognosis. *Am J Transpl*. 2013;13:3164–3172. https://doi.org/10.1111/ajt.12513
12. Taber DJ, Palanisamy AP, Srinivas TR, et al. Inclusion of dynamic clinical data improves the predictive performance of a 30-day readmission risk model in kidney transplantation. *Transplantation*. 2015;99:324–330. https://doi.org/10.1097/TP.0000000000000565
13. Goldfield NI, McCullough EC, Hughes JS, et al. Identifying potentially preventable readmissions. *Health Care Financ Rev*. 2008;30:75–91.
14. Molnar MZ, Nguyen DV, Chen Y, et al. Predictive score for posttransplantation outcomes. *Transplantation*. 2017;101:1353–1364. https://doi.org/10.1097/tp.0000000000001326
15. Amarasingham R, Patzer RE, Huesch M, et al. Implementing electronic health care predictive analytics: considerations and challenges. *Health Aff (Millwood)*. 2014;33:1148–1154. https://doi.org/10.1377/hlthaff.2014.0352
16. Parreco J, Hidalgo A, Kozol R, et al. Predicting mortality in the Surgical Intensive Care Unit using artificial intelligence and natural language processing of physician documentation. *Ann M Surg*. 2018;84:1190–1194. https://doi.org/10.1177/000313481808400736
17. Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA*. 2011;306:848–855. https://doi.org/10.1001/jama.2011.1204
18. Srinivas TR, Taber DJ, Su Z, et al. Big data, predictive analytics, and quality improvement in kidney transplantation: a proof of concept. *Am J Transplant*. 2017;17:671–681. https://doi.org/10.1111/ajt.14099
19. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Accessed April 18, 2022. https://www.R-project.org/.
20. Silge J, Robinson D. Text Mining With R. Accessed June 21, 2018. https://www.tidytextmining.com
21. Tokunaga T, Iwayama M. Text categorization based on weighted inverse document frequency. In: *Special Interest Groups and Information Process*. Society of Japan; 1994. Computer Science.
22. Shin B, Hogan J, Adams AB, et al. Multimodal ensemble approach to incorporate various types of clinical notes for predicting readmission. Proc IEEE-EMBS Int Conf Biomed Heal Informatics. Accessed April 18, 2022 https://www.bhi-bsn-2019.org/bhi/
23. Xu L, Hogan J, Patzer RE, Choi JD. Noise pollution in hospital readmission prediction: long document classification with reinforcement learning. In: *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. Association for Computational Linguistics; 2020:95–104.
24. Seni G, Elder J. Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions (Synthesis Lectures on Data Mining and Knowledge Discovery). Accessed February 23, 2019. https://doc.lagout.org/Others/Data Mining/Ensemble *Methods in Data Mining*_ Improving Accuracy through Combining Predictions %5BSeni %26 Elder 2010-02-24%5D.pdf
25. Boyd K, Eng KH, Page CD. Area under the precision-recall curve: point estimates and confidence intervals. In: Blockeel H, Kersting K, Nijssen S, Železný F, eds. *Machine Learning and Knowledge Discovery in Databases*. Springer; 2013:451–466.
26. Python Software Foundation. Python Language Reference, version 2.7. Available at http://www.python.org
27. Névéol A, Zweigenbaum P. Clinical natural language processing in 2014: foundational methods supporting efficient healthcare. *Yearb Med Inform*. 2015;10:194–198. https://doi.org/10.15265/IY-2015-035
28. DuBay DA, Su Z, Morinelli TA, et al. Development and future deployment of a 5 years allograft survival model for kidney transplantation. *Nephrol (Carlton)*. 2019;24:855–862. https://doi.org/10.1111/nep.13488
29. Cho S, Mohan S, Husain SA, Natarajan K. Expanding transplant outcomes research opportunities through the use of a common data model. *Am J Transplant*. 2018;18:1321–1327. https://doi.org/10.1111/ajt.14892