Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Michael D. Arenson Date

Identification of Kidney Transplant Recipients at High-Risk for Post-Transplant Hospitalization using Natural Language Processing

By

Michael D. Arenson
Master of Science

Clinical Research

_____

Rachel E. Patzer, PhD, MPH
Mentor

_____

Amita Manatunga, PhD
Committee Member

_____

Lindsay Collin, PhD Candidate
Committee Member

Accepted:

_____

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

_____

Date

Identification of Kidney Transplant Recipients at High-Risk for Post-Transplant Hospitalization using Natural Language Processing

By

Michael D. Arenson
B.S., University of Minnesota, 2010
M.A., Emory University, 2012

Advisor: Rachel E. Patzer, PhD, MPH

An abstract of
A thesis submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science
in Clinical Research
2019

Abstract

Identification of Kidney Transplant Recipients at High-Risk for Post-Transplant Hospitalization using Natural Language Processing
By Michael D. Arenson

Post-discharge rehospitalization after kidney transplant is a common and preventable problem that is both costly to patients and healthcare systems and is associated with poor outcomes. There is epidemiological evidence that up to 50% of surgical readmissions may be preventable (e.g. through discharge planning, patient education, and/or follow-up communication). Predictive analytics have previously been used to identify patients at risk of rehospitalization with limited success.

The vast amount of free-text data in the form of clinical notes that exist in the electronic medical record (EMR) has been untapped in the field of kidney-transplant. To date EMR free-text clinical notes have not been included in predictive models of 30-day rehospitalization (30DR) post-kidney transplant. Unstructured data describes any source of data that is not easily placed in a traditional numeric dataset. Analyzing free-text requires Natural language processing (NLP), which is a subfield of Artificial Intelligence that uses computer algorithms to analyze human language. Here, NLP was used to analyze EMR free-text documentation of kidney transplant recipients with the ultimate goal of reducing readmission post-kidney transplant.

This was a retrospective observational analysis of first-time recipients of kidney transplant at a large institution in the Southeast between January 2005 and December 2015. Both structured and unstructured data in the form of clinical notes written in the EMR were analyzed. Eight clinical notes were characterized and mined for possible new predictive features that might be useful to improve predictive accuracy of 30DR post-kidney transplant. Predictive models using unstructured, free-text clinical notes were built using machine-learning, unsupervised approaches. These predictive models did not meaningfully improve predictive accuracy above structured data alone. However, the results generated a number of new hypotheses regarding potentially novel predictors to be examined in future research applying more human-driven approaches.

Identification of Kidney Transplant Recipients at High-Risk for Post-Transplant Hospitalization using Natural Language Processing

By

Michael D. Arenson
B.S., University of Minnesota, 2010
M.A., Emory University, 2012

Advisor: Rachel E. Patzer, PhD, MPH

A thesis submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science
in Clinical Research
2019

Acknowledgments

I would like to thank my mentors throughout my time at Emory, in particular Rachel E. Patzer, PhD, MPH who taught me how to be a researcher and a leader. Julian Hogan, MD, PhD and Bonggun Shin, PhD provided daily mentorship and this research would not have been possible without their help. I would also like to thank my other thesis readers, Amita Manatunga, PhD and Lindsay Collin, PhD Candidate.

I must also thank the entire MS Clinical Research program leadership and administration. In particular I have to thank Dr. Ziegler who connected me with Dr. Patzer.

As always, I have to thank my family for their love and support.

Table of Contents

INTRODUCTION

In the field of transplantation, organs are a scarce resource. According to the National Kidney Foundation, currently more than 100,000 individuals are waitlisted each year for kidney transplant.[1] The average time on the waitlist is 3.5 years. Each year 7,500 patients on the waitlist will either become too sick or pass away before making it to the top of the list. As such, once a patient receives a kidney transplant there is a moral obligation to set both the transplant and the patient up for a healthy future. Rehospitalization after kidney transplant, however, is common, costly to both the patient and the healthcare system, and associated with worse outcomes and racial disparities.[2,3]

Predictive analytics, or the use of electronic algorithms to forecast future events in real time, makes it possible to harness the power of big data to improve the health of patients and lower the cost of health care.[4] However, this opportunity raises policy, ethical, and legal challenges. New technologies have become available that can harness the power of large data sets to help identify which medical interventions will benefit which patients. The power of predictive analytics, however, comes with great responsibilities; Responsibilities that we are only now beginning to understand.[5]

Throughout time, technology has served those that wielded it. We have learned over time through many mistakes and pitfalls in research – from the Nuremberg Code (1947) to the U.S. Common Rule (1991) – how to more ethically engage in the research and development of these technologies. Predictive analytics has already fallen into pitfalls of its own.[6,7] With predictive analytics, however, some of those ethical challenges are different from past technologies in scale, if not entirely in scope. Predictive analytics does push the art of what is possible. A wide range of data from many sources is now available, and insights can be made dynamically, practically in real-time.

While this thesis aims to move the field of predictive analytics in kidney transplant forward, it does not touch on the ethical dimensions of the research. The transplant field has historically had to grapple with ethical challenges given scarcity of organs and high cost of transplant surgeries.[8] As the field of transplant continues to experiment with artificial intelligence and predictive analytics, it will have to grapple with new ethical challenges that pertain to, for example, predicting rehospitalization after kidney transplant. Efforts have begun and should continue.[9]

BACKGROUND

Post-discharge rehospitalization after kidney transplant is a common and preventable problem that is both costly to patients and healthcare systems and is associated with poor outcomes. More than 50% of patients are hospitalized in the year following kidney-transplantation, and post-transplant hospitalization is associated with higher rates of graft loss[3,10], lower patient survival, and poor quality of life.[11] These poor outcomes are significantly more pronounced in disadvantaged groups. The causes of post-transplant hospitalization are multifactorial and include post- transplant complications such as infections, and non-transplant related factors such as patients' comorbidities.[12]

There is epidemiological evidence that up to 50% of surgical readmissions may be preventable (e.g. through discharge planning, patient education, and/or follow-up communication).[13] In the field of kidney transplant, the number of preventable readmissions may be lower, and leading reasons for rehospitalization are surgical complications (15%), rejection (14%), intravascular volume overload or depletion (11%), and systemic and surgical wound infections (11% and 2.5%).[14] This reflects the broader surgical literature, which has shown that most readmissions are related to new post-discharge complications.[15]

Predictive analytics have previously been used to identify patients at risk of rehospitalization with limited success due to reliance on static, structured data from national registries. Important risk factors for hospitalization following kidney transplantation include demographic, socioeconomic, clinical, transplant surgery, utilization factors, and timing of readmission.[12,16–20] Research on post-transplant hospitalization risk prediction models are limited in their utility by static data that do not reflect dynamic aspects of the transplant process and typically rely on administrative data that do not capture important patient-centered risk factors known to impact all levels of the End Stage Renal Disease (ESRD) care

trajectory.[12,14,21–23] Thus, transplant risk prediction models do not capture the salient information that could identify those at highest risk for post-transplant hospitalization. Given the medical implications for patients and reimbursement considerations for transplant centers for rehospitalization, improvements in predictive accuracy would allow for improved patient outcomes and less waste of medical resources.

Unstructured data describes any source of data that is not easily placed in a traditional numeric dataset. Examples include images, audio or video files, and free-text (such as those found in electronic medical record (EMR) notes or news website comments sections). In the transplant field, the latter is one such untapped data source.[24] Clinical free-text notes authored by transplant team providers and stored in the EMR are a large, untapped data source that could provide novel, high-yield predictor variables. Analyzing free-text requires Natural language processing (NLP), which is a subfield of Artificial Intelligence that uses computer algorithms to analyze human language.

NLP can be used to analyze EMR free-text documentation and has been experimented with in surgery.[25,26] For example, analyzing physician documentation to predict mortality in patients admitted to the surgical intensive care unit[25], predicting graft failure18, or automating identification of post-operative complications[26]. There are many unstructured data sources that serve as potential high-yield targets for NLP in kidney transplant such as detailed social worker notes prior to discharge from transplant or physician-authored notes during the pre-transplant evaluation. As such, we hypothesized that incorporating unstructured data from clinical notes through the use of Natural Language Processing (NLP) into predictive models would improve predictive accuracy for post-discharge hospitalization of kidney transplant recipients.

Ultimately, predictive models could be incorporated into a point-of-care clinical dashboard used by the transplant team to assess in real-time a patient's risk of 30DR. An example of one possible dashboard design can be found in **Figure 1**. In support of this overarching vision, our study had three aims. First, since NLP has been minimally experimented with in kidney transplant, we aimed to characterize kidney transplant-related clinical notes using NLP. For example, NLP might be used to identify the most common forms of support kidney transplant patients require – whether it be transportation or emotional support – which has not been included in prior predictive models. Second, we aimed to predict 30-day readmission (30DR) using NLP on individual clinical notes. And third, we aimed to maximize predictive accuracy from Aim 2 by combining multiple clinical notes.

METHODS

Study Population and Data Sources

This was a retrospective observational analysis of first-time recipients of kidney transplant at a large institution in the Southeast between January 2005 and December 2015. Patients were included if they were adults (>18 years at time of transplant), solitary kidney transplant (KTx) recipients, transplanted once between 1/1/2005 – 12/30/2015 at the institution, and received follow-up care at the study institution during that time. The only exclusion criteria were whether patients were missing substantial amounts of structured data, but there were not any patients excluded subsequent to inclusion.

Both structured and unstructured data in the form of clinical notes written in the EMR were analyzed. Baseline models were constructed using all available structured variables. These included features such as age, race, dialysis vintage, comorbidities and other variables known to be associated with 30DR.[14,21,27] In total, 80 variables were included in the structured (i.e. baseline) model.

Study Outcome

The outcome of interest was 30-day rehospitalization within thirty days of discharge from transplant (30DR). Rehospitalization was defined as the first unplanned hospital admission post-discharge 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 **Figure 2**. The data were collected from local electronic medical records and included data at the time of ESRD diagnosis, date of waitlisting prior to transplant, and transplant. Comprehensive baseline recipient and donor socio-demographics, transplant characteristics, laboratory, and other transplant-related data were collected up to the time of discharge post-transplant.

Unstructured Data Characteristics

Unstructured data comprised free-text clinical notes written in the EMR. We pulled eight note types available from the local institution's EMR database. Only notes written prior to discharge after transplant were included since post-discharge information would not be available for a patient prospectively. All documented clinical pre-kidney transplant discharge events were captured, including hospitalizations. Due to their file size, only those progress notes written in between the times of admission for transplant surgery and subsequent discharge were analyzed. With the exception of Selection Committee and Progress notes, all other note types were analyzed if they were written at the earliest a year prior to the date of the patient's transplant surgery and at the latest up to the time of discharge. Analysis relied on combining data sourced from both structured and unstructured data.

Statistical Analysis

Description of demographics and other structured variables was performed using SAS. Characterization of transplant-related clinical notes (Aim 1) was performed using R and methods described in Silge and Robinson's book, "Text Mining with R".[28] Free-text from multiple clinical notes of the same type were merged into one long free-text file for each patient. For example, if a patient had multiple social work notes, they were combined into

one large composite social work text file. Next, text was preprocessed by cleaning the text of words that occur one time, numbers, punctuation, and "stop words". Stop words are usually the most common words in a language or words that do not provide much meaning. For example, "the", "is", "because", or "about". There is not a single universal list of stop words, and the stop words may change depending on the analysis.

The process of turning free-text into structured data (i.e. turning text into a format that can be analyzed using tabular spreadsheet-like datasets with individual cells) is called tokenization. A token is a meaningful unit of text, such as a word, that we are interested in using for analysis, and tokenization is the process of splitting text into tokens. The token that is stored in each row is most often a single word but can also be an n-gram (i.e. strings of length n), sentence, or paragraph. When text is organized in a format with one token per row, tasks like removing stop words or calculating word frequencies are natural applications of familiar operations performed with structured data. Exploring term frequency on its own can give us insight into how language is used. In addition to considering words as individual units, many interesting text analyses are based on the relationships between words. For example, whether certain words tend to follow others immediately, or if they tend to co-occur within the same documents. The one-token-per-row framework can be extended from single words to n-grams and other meaningful units of text, as well as to many other analysis priorities.

In addition to term frequency, there are many approaches that have been used to analyze tokenized data. In this analysis we used two: Term Frequency-Inverse Document Frequency (TF-IDF)[29], and Latent Dirichlet Allocation (LDA)[28]. TF-IDF is intended to measure how important a token is to a document in a collection (or "corpus") of documents. For example, the importance of one token in a novel in a corpus of novels or, for the

purposes of this research, to one type of clinical note in a collection of different types of notes. TF-IDF is calculated using the equation below and is composed of two terms: Term Frequency (TF) and Inverse Document Frequency (IDF).

$$TF - IDF = TF\ (t, d) x\ IDF(t)$$

$$TF - IDF = TF\ (t, d)\ x\ log \frac{n}{df(t)}$$

TF is a normalized measure of how frequently a token occurs in a document. It is the number of times a token (e.g. a single word) appears in a document divided by the total number of words in the document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the total number of tokens in the document (i.e. normalized).

The second part of the equation describes the IDF, which describes how unique a token is to a specific document compared to all of the other documents in the collection. If only term frequency was used to characterize documents, however, commonly occurring words such as "is", "the", or even "kidney" or "transplant" would be identified but may not provide as much insight into how NLP can be leveraged. Thus, to put more weight on the rarer words in a collection of documents, IDF is calculated by taking the log of the total number of documents in the collection divided by the number of documents with the token of interest in it.

In addition to TF-IDF, *topic modeling* is a method for unsupervised classification of a collection of documents. Topic modelling refers to the task of identifying topics that best describes a set of documents. These topics will only emerge during the topic modelling process (therefore called latent). And one popular topic modelling technique is known as Latent Dirichlet Allocation (LDA). It treats each document as a mixture of topics, and each topic as a mixture of words. This allows documents to "overlap" each other in terms of content, rather than being separated into discrete groups, in a way that mirrors typical use of natural language.

Each clinical note was vectorized using both methods and subsequently analyzed using logistic regression. Predictive models were developed using combined structured and unstructured data beginning at ESRD diagnosis and ending at discharge from transplant event. Prior to establishing the predictive models, free-text data were pre-processed using natural language processing (NLP).

To account for missing data, logistic regressions was used in an Ensemble logistic regression method by averaging the probabilities of 30DR from each data source for each patient (**Figure 3**). Creating a predictive model often entails putting all data into one dataset and running a Concatenated logistic regression model. In contrast, the Ensemble approach uses multiple models by applying logistic regression (or any other classifier) to each unique source of data separately and then takes the average of the all of the logistic regression outputs.[30] Ensemble methods, with respect to classification algorithms are relatively new techniques. Ensemble is a more sophisticated approach for increasing model accuracy as compared to the traditional practice of parameter tuning on a single model.

Ensembling has been shown to address three challenges in the traditional Concatenation approach.[31] First, many tasks in medical domains inherently consist of small

sample data with lengthy documents. This is particularly true in the field of kidney transplant where often the total number of patients receiving a transplant numbers in the hundreds. Thus only a few thousand patients with EMR clinical notes are available, often sporadically missing note types, and consisting of millions of words. Given the length of these notes, complex deep learning methods cannot be applied to these kinds of domains, making the Ensemble method a good alternative.

In constructing our cohort, only patients with structured data were included. Patients who were missing clinical notes did not exclude a patient from the cohort since we used the Ensemble method. In total, 61 predictive risk models were built for the outcome of interest (30DR). These included one baseline model comprised of structured-only data. For each of the eight note-types, we created a (1) model with TF-IDF words as input, a (2) model with LDA topic vectors as input, and (3) a model with both TF-IDF + LDA features. Thus each note-type yielded 3 separate models (Aim 2). Finally, in attempting to maximize predictive accuracy (Aim 3) we combined features from and analyzed twelve separate combinations of notes (e.g. combining Progress + Consultation note features) and created three models (TF-IDF, LDA, and TF-IDF + LDA) for each of the combinations as previously described for individual notes.

Once the separate structured and unstructured models were selected, internal validation and area under the curve (AUC) was used to determine and compare model accuracy. We used *5-fold cross-validation* to test the performance of each model. The area under the receiver operating curve (ROC) was recorded for the testing set. The average ROC curve was calculated for all 5 cross-validation testing set. In order to identify the most heavily weighted predictor variables, we used a neural network-based feature selection

algorithm called Wx, which has also been used in next-generation sequencing data.[32] Statistical analyses were performed using SAS (Cary, NC), R, and Python.

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 4**). Of the final cohort, 633 (30.7%) experienced 30DR. Although 18 were missing all clinical notes, they were still included in the cohort because they had structured data. We pulled eight note types available from the local institution's EMR database. These notes were written at various timepoints throughout the transplant process. Only notes recorded prior to post-KTx discharge were analyzed: Social Work (n=1118), Selection Committee (n=2033), Echo Report (n=1110), History and Physical (H&P) (n=1422), Consultants (n=1354), Progress (n=1415), Operative (n=1472), and Discharge Summaries (n=517).

All structured variables, including demographic and clinical characteristics of the patient population, can be found in **Table 1**. Variables were chosen based on statistical significance of $p \leq 0.05$ in the analysis or previous identification in the literature of contributing to rehospitalization.

Using two NLP approaches (TFIDF and LDA) we identified three main themes. First, that NLP can be used to identify potentially novel predictive variables not previously identified in structured data. Second, that NLP confirms the relevance of other predictive variables. And third, that broad KTx-related topics are shared amongst many clinical note types, making some note types high-yield and others less so for predicting outcomes.

Recipient, donor, and transplant characteristics

The study population was predominately African American or black (47%) and male (58%). The most common cause of ESRD amongst subjects associated with 30DR in the

study population was diabetes (223 [35.2%] vs. 321 [22.5%]), hypertension (162 [26%] vs. 415 [25%]), and primary glomerulonephritis (116 [18%] vs. 307 [22%]). Also significantly associated with 30DR were: ever carrying a diagnosis prior to transplant of Infectious or parasitic diseases (n=177 [27.5%]), endocrine, nutritional and metabolic disease, or immunity disorders (n=583 [90.7%]), mental disorders (n=215 [33.4%]), and diseases of the nervous system and sense organs (n=285 [44.3%]). In addition, 30DR was associated with the number of days from referral end to evaluation end (314.4 days [S.D. 378.1], n=449 vs. 250.8 days [S.D. 393.7], n=1,130), number of days from evaluation start to evaluation end (326.4 days [S.D. 353.0], n=460 vs. 262.8 days [S.D. 305.8], n=1,186), and number of days from evaluation end to waitlist start (-104.9 days [S.D. 300.7], n=460 vs. -68.1 days [259.3], n=1,186).

For donor factors, the recipients with deceased donors, positive hepatitis C and B status, and those that were considered high risk by CDC Guidelines were more likely to be readmitted within 30 days. For transplant factors, the post-KTx length of stay (5 [4-7] vs. 4 [4-6]), HLA B and HLA DR mismatches, ABO incompatibility, White Blood Cell count at time of transplant, and Hemoglobin A1C at transplant were significantly associated with 30DR.

NLP used to identify novel variables not previously identified in structured data

Frequency of words amongst all notes and within each note could possibly indicate novel predictive features overall and for each note type. For example, **Figure 5** illustrates three example graphs for three different note types: (A) Consultation, (B) Selection Conference, and (C) Discharge Summary notes. Terms are plotted by frequency and according to those that appear in patients who were readmitted within 30-days and not. In

the Consultation note, words such as, "disseminated", "dialyvite", "chemotherapy", "juven", "bela" (i.e. belatacept), and "ablation" are associated with patients who experienced 30DR. These terms indicate infection, dialysis, cancer, supplemental nutrition, immunosuppression, and arrhythmia or other heart conditions, respectively, which are largely related to health and well-being post-kidney transplant.[33] In contrast, words such as "bupropion" (an anti-depressant), "bmt" (i.e. bone marrow transplant), and "esbl" (i.e. Extended spectrum beta-lactamases), can be found associated with patients not experiencing 30DR. The literature has not previously identified these as predictive features. Similarly, in (B) Selection Conference notes, other terms such as, "hbv" (Hepatitis B infection), "adenocarcinoma", and "fna" (fine needle aspiration, oftentimes performed for a biopsy of abnormal tissue) are reasonably associated with 30DR.

Patterns emerge comparing **Figure 5(B)** Selection Conference notes, which is one of the first notes recorded for many patients in the transplant process and **Figure 5(C)** Discharge Summary notes, which is the last note recorded in a patient's transplant odyssey. For example, early in the process while the patient is being evaluated for transplant suitability **5(B)**, 'abscesses' are protective of 30DR. At discharge **5(C)**, only days post-transplant surgery, 'abscess' flips to the opposite side, becoming associated with 30DR.

Focusing only on **Figure 5(C)**, one sees words that suggest similar risk factors for 30DR. The words 'lymphoma', "amikacin", "pna" (shorthand for pneumonia) which, as has been seen in Figure **5(A)** and **5(B)**, indicate cancer and infection are associated with 30DR. Also found is the word, "aphasia", which means the patient has language troubles (either speaking or comprehending). This is usually a sign of stroke. In **5(C)**, "fluoxetine" can be seen, which is another antidepressant that seems to be associated with patients not

readmitted within 30-days, as well as "amputation" for reasons that are not immediately apparent.

NLP confirms the relevance of other variables

In addition to elucidating potentially novel variables, TF-IDF can also confirm the importance of other variables that would not otherwise be captured with structured data, alone. For example, bigrams (i.e. word pairs) can identify common forms of support that patients require. Amongst all available notes for all patients, we searched for bigrams where the second word is "support". We found that the three words most frequently preceding the word "support" are "care", "transportation", and "emotional" (**Table 2**).

While 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. The top thirty words in **Figure 6** are, as measured by TF-IDF, the most important to each note. For example, out of the top 30 words identified in the Selection Committee note, "adenosine", "thallium", "cardio", "dobutamine", "peak", "velocity", "transthoracic", and "scintigraphic" can be found which pertain to heart health and evaluation. In the same note, "nutrition" and "dietary" indicate the importance of nutrition in the evaluation process. Each of these terms are the variables most unique to that note and thus can possibly be used as predictive features in addition to other structured features that indicate a patient's cardiac and dietary status.

Also, there are very few structured data that indicate a patient's mental well-being despite knowing that it is likely an important predictor of 30DR. But, in the Social Work note, the words "educated", "activities", "mood", and "mental" pertain to patients' social

history and mental health and might be used as predictive features. Also found in the Social Work TF-IDF terms are the words, "budd" and "terrace". Separately, these do not make sense, but Budd Terrace is a nursing facility that patients are often discharged to if they are not able to care for themselves upon discharge. Therefore, it not only makes sense that they would have a similar TF-IDF rank, but that this would be a term unique to social work notes.

Kidney transplant topics are shared amongst many of the clinical note types

While TF-IDF identifies words that are unique to each note, LDA identifies overarching topics. This is done by extracting probabilities that each word is generated from each topic, called β ("beta"), from the model (note this is different than betas used in regression modeling). For example, the term "nutrition" might have a large probability of being generated from one topic but a very low probability of being generated from a different topic. This allows one to identify the terms that are most common within each topic. **Figure 7** shows the 20 terms with the highest beta for each topic. Since each topic is generated in an unsupervised manner, the concept or title of each topic requires interpretation and as such will be examined further in the Discussion section below.

Some words in **Figure 7**, such as "patient", "transplant", "start", or other dose measurements such as "mg" or "ml" are common within multiple topics. This is an advantage of topic modeling as opposed to looking only at individual note types, because topics used in natural language could have some overlap in terms of words. While this should not be surprising, it is useful to know to what degree clinical notes overlap. For

example, Selection Committee notes are more correlated to Operative notes than they are to Social Work notes. This makes sense, since Selection Committee and Operative notes are both authored by a transplant surgeon. However, in the quest to reduce readmissions, social issues may not be adequately addressed in the evaluation stage, decreasing chances of identifying social problems and preventing readmission downstream.

In addition to estimating each topic as a mixture of words, LDA also models each document as a mixture of topics (**Figure 8**). We can examine the probabilities that each document is generated from each topic, called γ ("gamma"). Each of these values is an estimated proportion of words from that document that are generated from that topic. For example, the model estimates that only about 37.2% of the words in Progress notes were generated from Topic 2. In contrast, 100% of the words in the Echo note were generated by Topic 3. Many of the notes were drawn from a mix of topics, such as Progress, Consultation, and H&P notes.

Predictive model performance of structured and unstructured data

There were 80 structured variables included in the model (**Table 1**). When added to structured data the AUC did not improve statistically significantly after including data from individual clinical note types (**Table 3**). For the structured model, the AUC was estimated as 0.6523 (95%CI 0.6218, 0.6829) for 30DR.

Layering of data sources does not augment predictive accuracy much more than adding single notes data to the baseline model, alone (Table 4). The best performing model included Structured data and the following clinical notes: Consultations, H&P, Progress, and Selection Conference notes (AUC 0.6744, 95%CI: 0.6587, 0.6900). This is neither a

statistically significant nor a meaningful improvement over structured data alone, as it only improves predictive accuracy by 2.21%.

The top predictive terms from the best predictive model in **Table 4** can be seen in **Table 5**. All of the top 20 predictors with the exception of one are structured variables. The highest-ranking unstructured variable contributing to the highest performing model was the term "mg". In a sensitivity analysis, multiple different NLP techniques were evaluated, such as Word2Vec and Doc2Vec, as well as classifiers other than logistic regression (i.e. Random Forrest). Description of these techniques is outside the scope of this paper, however, the NLP techniques employed here as well as using logistic regression had the highest predictive performance.

DISCUSSION

Our study both confirms predictive features described in prior literature and, using natural language processing, generates a number of new hypotheses about factors throughout the transplant process that may be predictive of 30DR (from ESRD diagnosis to post-transplant 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 (older age and AA race), socioeconomic factors (lower education and Medicaid insurance), clinical factors (high BMI and various comorbidities), transplant surgery factors (longer length of stay, receipt of a deceased (vs. living) donor, older donor age, and surgical complications), utilization factors (pre-transplant hospitalization), and adherence to medication.[3,10,12,14,16,17] Many of these risk factors are reproduced in our study in structured data. Our study lends further support to these previously identified predictive features.

To our knowledge, there has only been one other study that has published a post-transplant specific predictive model.[21] In this study, Taber et al. first designed a model including fixed transplant predictors that remained modestly predictive (AUC 0.63; 95% CI: 0.58–0.69). The predictive accuracy significantly improved to 0.73; 95% CI, 0.67–0.79 after including post-transplant but pre-discharge dynamic factors such as the systolic blood pressure slope during transplant admission. Thus, the development of more accurate predictive models of readmission after kidney transplantation will require the collection of more granular data than those usually available in transplant registries. These data include socio-economic data (e.g. familial support, transportation issues) and clinical data such as labs values or vitals.

Natural language processing (NLP) employs computational techniques to learn, understand, and produce human language content. NLP can be used to analyze and learning from the enormous quantity of human language content that is now available in the EMR and healthcare systems.[34] Recently, 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–-0.94 for 1-year graft loss and 0.84; 95% CI, 0.80–0.89 for 3-year mortality.[27] Using NLP algorithms, Srinivas et al. parsed Banff lesion scores from pathology reports in text form. Lesion scores transcribed as g0, t0, i2, t2, v0 were extracted and transferred to analytic databases.

However, the extraction of these predictors was based on previously reported risk factors and clinical input of transplant experts. This process is time-consuming and complex which makes it difficult to generalize outside of a single institution due to 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 such as "juven" (risk factor) or "fluoxetine" (protective). These would otherwise be less likely to be identified as a predictor.

For these reasons, 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. While 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. Using a machine-learning NLP approach to generate previously unrecognized important words or topics but using a more human-driven decision regarding which predictive features to extract from clinical notes and include in predictive models. Given the

success in improving predictive accuracy that Srinivas et al. demonstrates (albeit with a much more finite outcome like graft failure), this seems like the next important step for this research.

In characterizing the clinical notes as we have done for our first aim, we have generated a number of new hypotheses about potential predictive features to extract using NLP. For example, terms and their synonyms in **Figure 5** that are both found in higher frequency in patients with and without 30DR should be extracted and used as predictive features. Specifically, terms that are associated with infection, dialysis, cancer, supplemental nutrition, immunosuppression, and arrhythmia or other heart conditions. Some terms require further exploration and understanding. For example, it is unclear why the word, "amputation" in Consultation notes might be found more often in patients who did not experience 30DR. Also, multiple notes identify antidepressant medications as being protective or 30DR. Identifying patients discharged on antidepressants may is another example of a potentially novel variable to include in future models (identified either by using structured or unstructured data). Novel variables also arise from exploring the TF-IDF for each note separately. For example, identifying which patients are being discharged to "budd terrace" or another acute care facility could be a predictive feature of 30DR.

In the case of topic models, novel predictive features take the form of vectors of words. These vectors are included in predictive models, but their usefulness might lie more in identifying the highest yield clinical notes to include. The notes that had the largest predictive accuracy were those that incorporated multiple topics. That is, the consultation, H&P, progress, and selection-conference notes. In contrast, the notes that were described by only one topic had less predictive ability. This demonstrates that notes that incorporate multiple sources of data (e.g. physical exam, labs, imaging) are more useful for predicting

readmission. When attempting to calculate using NLP the real-time risk of a patient prior to discharge, these "multifaceted" notes should be prioritized.

The question remains why all of these new potential variables led to minimal to no improvement in predictive accuracy. The preventability of readmission post-kidney transplant is unknown, but studies have shown variability (from only 8% of readmissions being preventable to 50%).5,6 This contradicts literature in general surgery, which indicates major differences between general surgery and transplant. As seen in Table 5, the unstructured data did not rank high on predictive importance. The only variable that did was the word, "mg", which could be a proxy for patients who are prescribed a lot of medications either postoperatively or at time of discharge, both of which would indicate severity of disease.

Limitations

Our study has some limitations. First, we restricted our cohort to patients receiving only one transplant from our transplant center between 2005–2015. Our cohort was meant to include only first-time kidney transplant recipients, however, 163 (7.9%) of patients were found to have had a prior transplant after linking to national-level United States Renal Database System (USRDS) registry. Given that patients with multiple transplants are at higher risk of 30DR than patients with only one, this could bias the results of our study. 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. Conversely, USRDS data have a lag time of two years. Although linking local data to national-level registries such as USRDS would avoid the inclusion of subjects with multiple

transplants, the predictive model would not be built to use the data available in real-time. Thus, any clinical dashboard using this predictive model would be clinically irrelevant.

This speaks to a larger limitation of missing data. As academic medical centers have transitioned from paper to electronic medical charts, the available free text data is limited. While software exists that can translate scanned paper documents into electronic free-text data, the data are almost certain to be too messy to analyze and contain many errors and typos. As the databases for EMR have developed, the scaffolding has been built in a patchwork approach. Thus, 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, limiting the availability of free text data for that patient. In order to address this, we initiated our cohort 2005 when there was a clear change in the availability of free-text notes in the EMR database. However, the difference in the number of missingness in the eight notes analyzed for this study indicate the challenge of analyzing EMR data.

Another limitation is the size of our cohort. The transplant field treats a relatively small number of patients. The transplant center providing the data for our cohort is one of the largest transplant centers in the U.S., and yet an n of 2060 is quite small when attempting to use machine learning techniques. Increasing the n 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. Given the vast amount of free-text data in the transplant field, however,

this should perhaps be a long-term goal. It would require a further research into how free-text notes are written and organized in EMR's at transplant centers around the country.

Future Directions

Many of the NLP techniques used in this study are also used by industry leaders in machine learning. For example, the Google search employs a TF-IDF technique, albeit much more advanced. As 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 addition, looking at other forms of basic NLP analysis such as sentiment of words (i.e. words that are positive versus negative, or that connote joy). For example, a patient with more negative words in their post-transplant notes may have a higher likelihood of being readmitted. Furthermore, using industry developed tools such as Amazon Comprehend Medical or ClarityNLP developed by Georgia Tech to search for specific words/phrases identified might provide a more systematic, scalable search algorithm that can be used at multiple transplant centers.

CONCLUSIONS

In this work, I have characterized eight 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 machine-learning, unsupervised approaches. These predictive models did not meaningfully improve predictive accuracy above structured data alone. However, the results generated a number of 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, Natural language processing is possibly the key to leveraging these notes for research that will ultimately help the patient, the physicians, and the hospital improve outcomes.

REFERENCES

1. Foundation NK. Organ Donation and Transplantation Statistics.
2. Patzer RE, McClellan WM. Influence of race, ethnicity and socioeconomic status on kidney disease. *Nat Rev Nephrol*. 2012;8(9):533-541. doi:10.1038/nrneph.2012.117.
3. Lynch RJ, Zhang R, Patzer RE, Larsen CP, Adams AB. First-Year Waitlist Hospitalization and Subsequent Waitlist and Transplant Outcome. *Am J Transplant*. 2017;17(4):1031-1041. doi:10.1111/ajt.14061.
4. Cohen IG, Amarasingham R, Shah A, Xie B, Lo B. The Legal And Ethical Concerns That Arise From Using Complex Predictive Analytics In Health Care. *Health Aff*. 2014;33(7):1139-1147. doi:10.1377/hlthaff.2014.0048.
5. Ito J, Zittrain J. The Ethics and Governance of Artificial Intelligence. MIT Media Lab.
6. Angwin J, Larson J, Mattu S, Kirchner L. Machine Bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Published 2016. Accessed April 26, 2018.
7. Emerging Technologies. Neural Network Learns to Identify Criminals by Their Faces. MIT Technology Review. https://www.technologyreview.com/s/602955/neural-network-learns-to-identify-criminals-by-their-faces/. Published 2016. Accessed April 26, 2018.
8. UNOS. Ethical Principles in the Allocation of Human Organs. https://optn.transplant.hrsa.gov/resources/ethics/ethical-principles-in-the-allocation-of-human-organs/. Published 2015. Accessed April 26, 2018.
9. Amarasingham R, Audet A-MJ, Bates DW, et al. Consensus Statement on Electronic Health Predictive Analytics: A Guiding Framework to Address Challenges. *EGEMS (Washington, DC)*. 2016;4(1):1163. doi:10.13063/2327-9214.1163.
10. Lynch RJ, Zhang R, Patzer RE, Larsen CP, Adams AB. Waitlist Hospital Admissions Predict Resource Utilization and Survival After Renal Transplantation. *Ann Surg*. 2016;264(6):1168-1173. doi:10.1097/sla.0000000000001574.
11. Johnson CD, Wicks MN, Milstead J, Hartwig M, Hathaway DK. Racial and gender differences in quality of life following kidney transplantation. *Image J Nurs Sch*. 1998;30(2):125-130. http://www.ncbi.nlm.nih.gov/pubmed/9775552.
12. McAdams-Demarco MA, Grams ME, Hall EC, Coresh J, Segev DL. Early hospital readmission after kidney transplantation: patient and center-level associations. *Am J Transpl*. 2012;12(12):3283-3288. doi:10.1111/j.1600-6143.2012.04285.x.
13. Jones CE, Hollis RH, Wahl TS, et al. Transitional care interventions and hospital readmissions in surgical populations: a systematic review. *Am J Surg*. 2016;212(2):327-335. doi:10.1016/j.amjsurg.2016.04.004.
14. Harhay M, Lin E, Pai A, et al. Early rehospitalization after kidney transplantation: assessing preventability and prognosis. *Am J Transpl*. 2013;13(12):3164-3172. doi:10.1111/ajt.12513.
15. Merkow RP, Ju MH, Chung JW, et al. Underlying reasons associated with hospital readmission following surgery in the United States. *JAMA*. 2015;313(5):483-495. doi:10.1001/jama.2014.18614.
16. 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(12):2276-2288. doi:10.2215/CJN.04940610.
17. 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(6):520-526. doi:10.1001/archsurg.2011.2220.
18. Tsai TC, Joynt KE, Orav EJ, Gawande AA, Jha AK. Variation in surgical-readmission rates and quality of hospital care. *N Engl J Med*. 2013;369(12):1134-1142. doi:10.1056/NEJMsa1303118.
19. Patzer RE, Serper M, Reese PP, et al. Medication understanding, non-adherence, and clinical outcomes among adult kidney transplant recipients. *Clin Transplant*. 2016;30(10):1294-1305.

doi:10.1111/ctr.12821.
20. Hogan J, Arenson MD, Adhikary S, et al. Timing matters: Improving Prediction of Hospital Readmission Following Kidney Transplantation. 2019.
21. 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(2):324-330. doi:10.1097/TP.0000000000000565.
22. Goldfield NI, McCullough EC, Hughes JS, et al. Identifying potentially preventable readmissions. *Health Care Financ Rev*. 2008;30(1):75-91. http://www.ncbi.nlm.nih.gov/pubmed/19040175.
23. Molnar MZ, Nguyen D V, Chen Y, et al. Predictive Score for Posttransplantation Outcomes. *Transplantation*. 2017;101(6):1353-1364. doi:10.1097/tp.0000000000001326.
24. Amarasingham R, Patzer RE, Huesch M, Nguyen NQ, Xie B. Implementing Electronic Health Care Predictive Analytics: Considerations And Challenges. *Health Aff*. 2014;33(7):1148-1154. doi:10.1377/hlthaff.2014.0352.
25. Parreco J, Hidalgo A, Kozol R, Namias N, Rattan R. Predicting Mortality in the Surgical Intensive Care Unit Using Artificial Intelligence and Natural Language Processing of Physician Documentation. *Am Surg*. 2018;84(7):1190-1194. http://www.ncbi.nlm.nih.gov/pubmed/30064586.
26. 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(8):848-855. doi:10.1001/jama.2011.1204.
27. 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(3):671-681. doi:10.1111/ajt.14099.
28. Silge J, Robinson D. *Text Mining with R*. 2019th-02-10th ed. O'Reilly; 2019. https://www.tidytextmining.com.
29. Tokunaga T, Iwayama M. Text Categorization Based on Weighted Inverse Document Frequency. In: *Special Interest Groups and Information Process Society of Japan*. ; 1994. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.7015.
30. Seni G, Elder J. *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions (Synthesis Lectures on Data Mining and Knowledge Discovery)*. 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. Accessed February 23, 2019.
31. Shin B, Hogan J, Adams AB, Lynch RJ, Patzer RE, Choi JD. Multimodal Ensemble Approach to Incorporate Various Types of Clinical Notes for Predicting Readmission. *Proc IEEE-EMBS Int Conf Biomed Heal Informatics*. 2019. https://www.bhi-bsn-2019.org/bhi/.
32. Park S, Shin B, Choi Y, Kang K, Kang K. Wx: a neural network-based feature selection algorithm for next-generation sequencing data. *bioRxiv*. November 2017:221911. doi:10.1101/221911.
33. National Kidney Foundation. Care After Kidney Transplant. https://www.kidney.org/atoz/content/immunosuppression#page16. Published 2015. Accessed March 27, 2019.
34. Névéol A, Zweigenbaum P. Clinical Natural Language Processing in 2014: Foundational Methods Supporting Efficient Healthcare. *Yearb Med Inform*. 2015;10(1):194-198. doi:10.15265/IY-2015-035.

TABLES / FIGURES

Figure 1: Conceptual mock-up of clinical dashboard identifying kidney transplant patients at high-risk of 30-day readmission

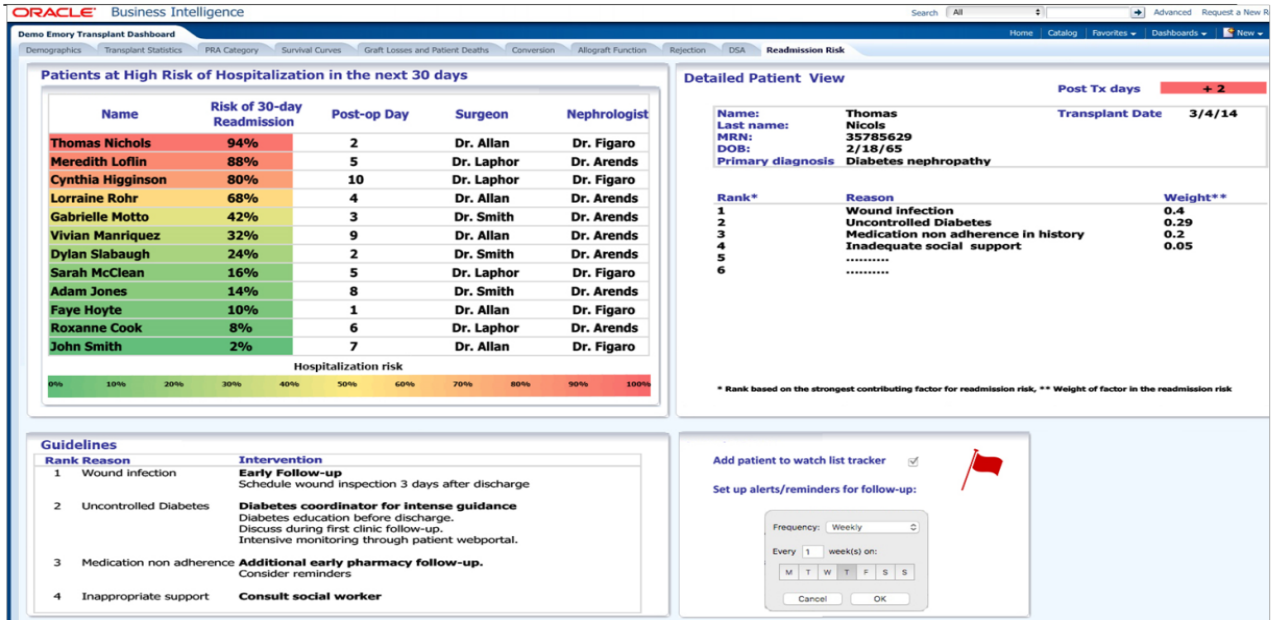


Figure 2: List of all Structured Data Variables and all Unstructured Data Sources by Time Collected in Transplant Process

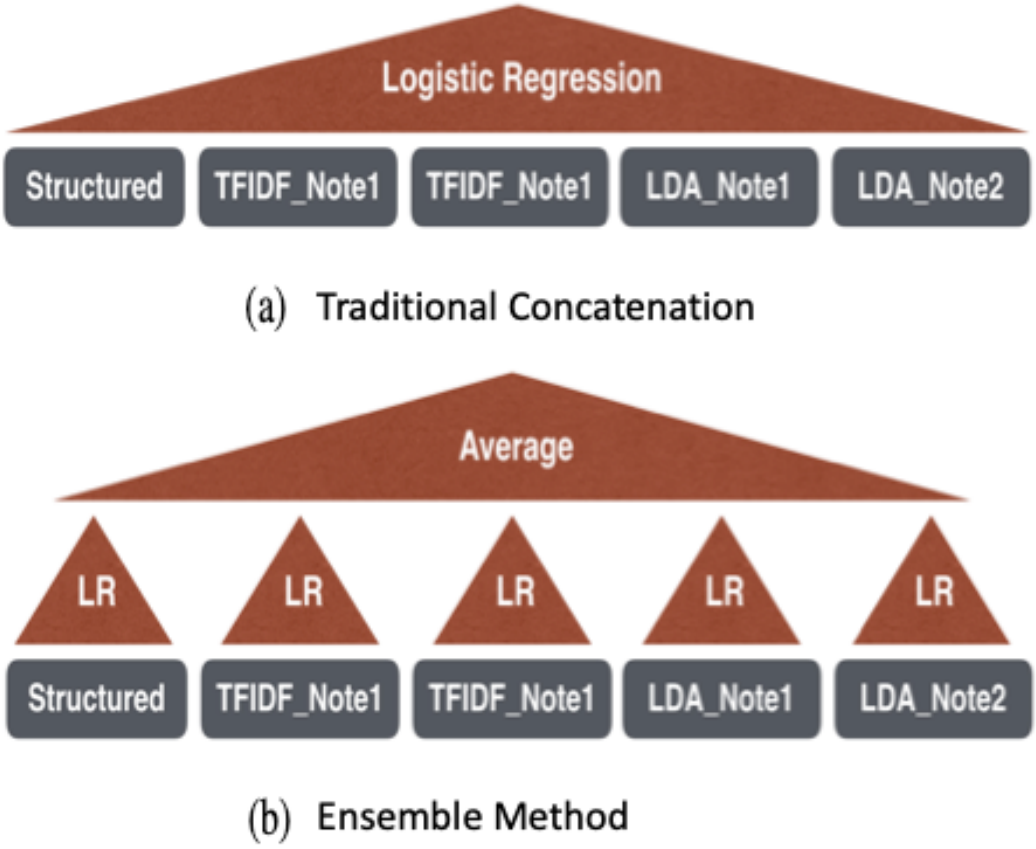
	Recipient			Donor		Transplant Process			
Variables included in Predictive Model	Demo-graphics	Clinical	Social	Demo-graphics	Clinical	ESRD Diagnosis, Dialysis, & Referral	Evaluation & Waitlisting	Transplant Operation	Post-Transplant
Structured Data									
Age at Transplant	•			•					
Race	•								
Ethnicity	•								
Gender	•								
Religion	•								
Language	•								
Primary Cause of ESRD		•							
Karnofsky Status		•							
Prior Transplant		•							
EBV Status		•			•				
CMV Status		•			•				
Toxoplasmosis Status		•			•				
Hepatitis C Status		•			•				
Blood Type (ABO)		•			•				
Alcohol Use			•						
Smoking Status			•						
Whether Patients' address zipcode change within 1 year prior to transplant			•						
Donor Type (Living vs. Deceased)				•					
Hepatitis B Core Status					•				
Hepatitis B Surface Antigen Status					•				
Hepatitis B Surface Antibody Status					•				
HCC High Risk					•				
CDC High Risk					•				
# of days from referral start to evaluation start						•			
# of days from referral end to evaluation end						•			
# of days from evaluation start to evaluation end							•		
# of days from evaluation end to waitlist start							•		
T Cell cross match								•	
B Cell cross match								•	
HLA A mismatches								•	
HLA B mismatches								•	
HLA DR mismatches								•	
Kidney Received On (Ice vs. Pump)								•	
EBV Infection Risk								•	
CMV Infection Risk								•	
Blood Type Compatibility								•	
Albumin at Transplant*								•	
Albumin at Discharge Post Transplant*									•
Change in Albumin from Transplant to Discharge									•
Creatinine at Transplant*								•	
Creatinine at Discharge Post Transplant*									•
Blood Nitrogen Urea at Transplant*								•	
Blood Nitrogen Urea at Discharge Post Transplant*									•

White Blood Cell at Transplant*			•	
White Blood Cell at Discharge Post Transplant*				•
Hemoglobin A1C at Transplant*			•	
Hemoglobin at Transplant*			•	
Hemoglobin at Discharge Post Transplant*				•
Change in Hemoglobin from Transplant to Discharge				•
Prograf level at Discharge Post Transplant*				•
Unstructured Data				
Social Work	•	•	•	•
Selection Committee		•		
Echo Report		•		
H & P		•	•	
Consultants		•	•	•
Progress		•	•	•
Operative		•	•	
Discharge Summary		•	•	•

*Also included are minimum and maximum values of each laboratory value at specified timepoint.

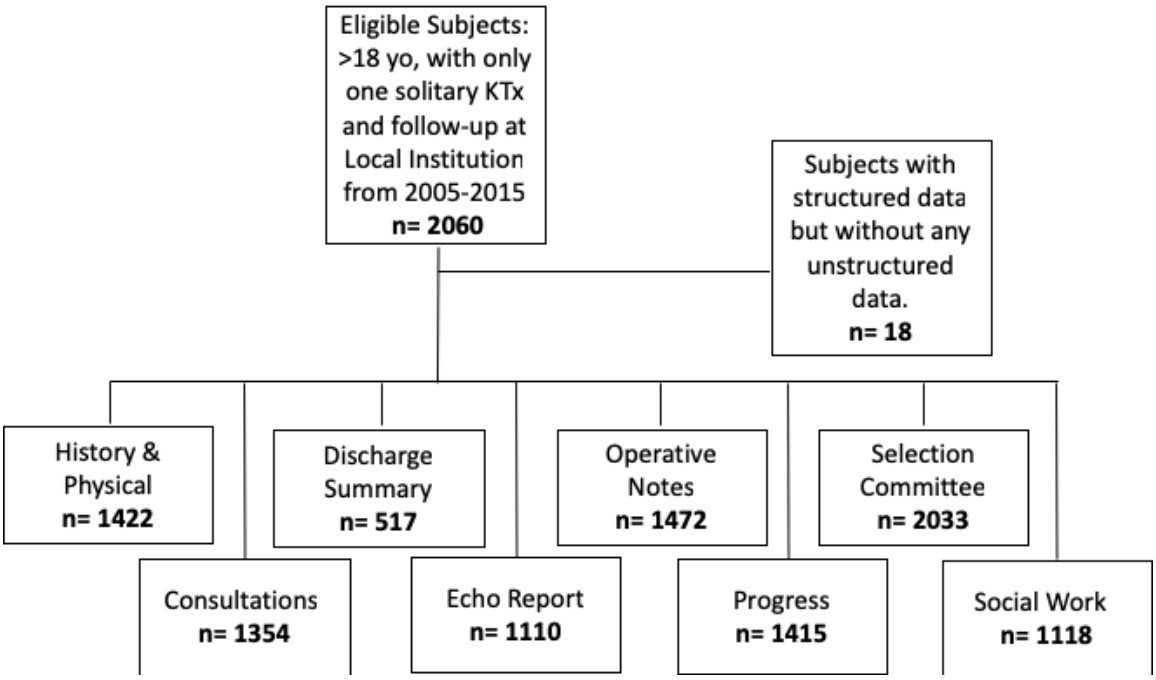
Structured data variables (n=80) are highlighted in blue. Unstructured data sources, highlighted in yellow, are identified as opposed to all unstructured data variables due to space constraints. Variables in the Recipient and Donor columns were administrative data collected early in the transplant process or at time of transplant. HCC coding is a payment model designated by the Centers for Medicare and Medicaid Services (HCC, Hierarchical Condition Category). CDC high risk guidelines developed by Centers for Disease Control in 1994 to notify and protect candidates (CDC, Centers for Disease Control). Length of Hospital stay from the day of transplant to the day of discharge. Risk describes risk of active infection; Comprised of High (donor +, recipient -), Intermediate (donor -, recipient +), and Low Risk (donor -, recipient -).

Figure 3: Traditional Concatenation vs. Ensemble logistic regression method.



Ensemble methods combine multiple models into one usually more accurate than the best of its components. Ensembles are useful with all modeling algorithms, not just logistic regression. Building an ensemble consists of two steps:(1) constructing varied models and (2) combining their estimates. Ensembling depends heavily on the quality of the individual models (e.g. without overfitting).

Figure 4: Inclusion and Exclusion Flowchart



Only patients with structured data were included in the analysis (n=2060, years 2005-2015). Thus, both the number of patients with structured data and our n=2060. Patients did not have to have any clinical notes to be included in the analysis. Number of patients without any clinical notes were n=18. The clinical note with the most patients represented was the Selection Committee note (n=2033) and the least being the Discharge Summary (n=517).

Table 1: Baseline characteristics of kidney transplant recipients from Emory Transplant center, stratified by readmission within 30 days post -transplant, 2005–2015

Characters	Study Population n=2060	Within 30 Days Readmission n=633 (30.7%)	Not Within 30 Days Readmission n=1427 (69.3%)	P-Value
RECIPIENT FACTORS				
Demographic				
Age, years, 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)	
African American or Black	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 ESRD, n (%)				<.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 (%)				<.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 (%)				<.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 diagnoses prior to transplant) **(If a diagnosis is not present, it is assumed that the subject did not have the condition. Therefore, there is no missing value)**				
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	1,795 (87.1%)	583 (90.7%)	1,212 (85.5%)	0.001
Diseases of the blood and blood-forming organs	1,412 (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%)	<.0001
Diseases of the circulatory system	2,055 (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	1,261 (61.2%)	414 (64.4%)	847 (59.8%)	0.05
Diseases of the genitourinary system	2,059 (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 (%) (1,624 subjects with non-missing value)				0.0002
Deny	1,103 (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 (%) (1,227 subjects with non-missing value)				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, years, Median (IQR)	39 (25-49)	39 (23-49)	39 (27-49)	0.41

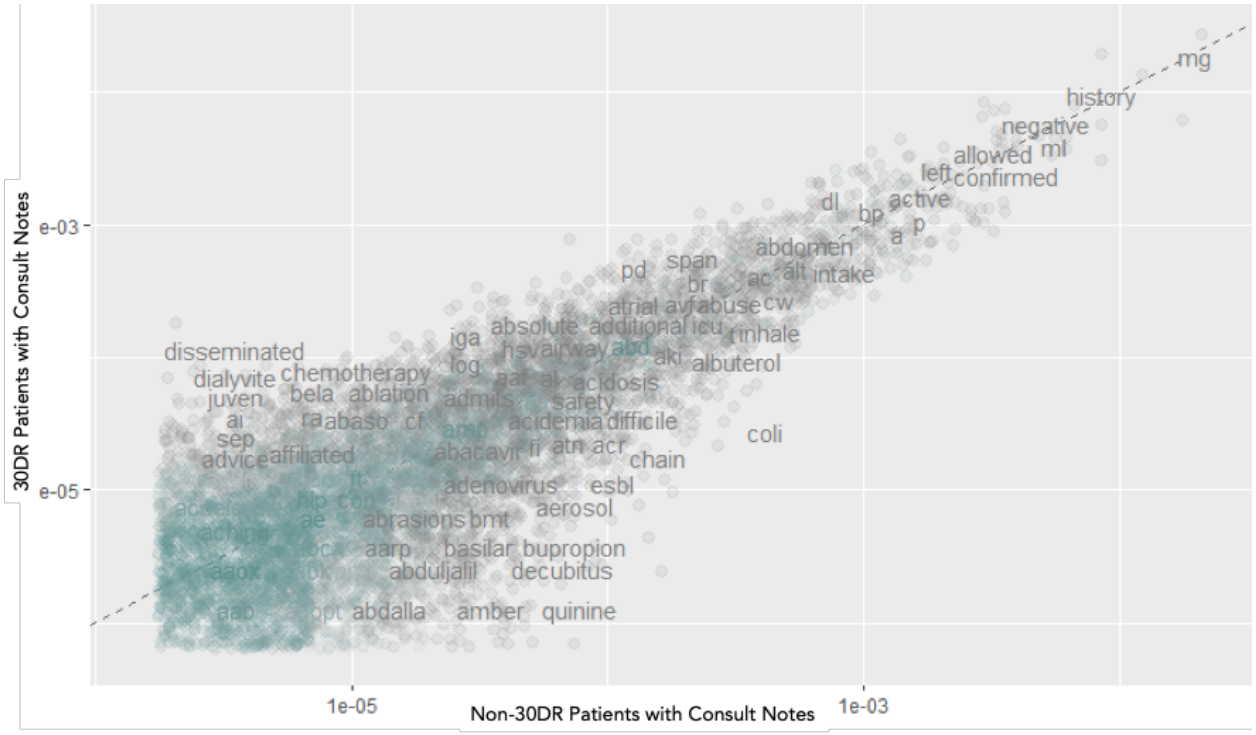
Type of Transplant Donor				0.002
Deceased	1386 (67.3)	458 (72.4)	928 (65.0)	
Living	643 (31.2)	164 (25.9)	479 (33.6)	
Pediatric	31 (1.5)	11 (1.7)	20 (1.4)	
TRANSPLANT FACTORS				
Length of Hospital Stay, Days, Median (IQR)	4 (4-6)	5 (4-7)	4 (4-6)	<.0001
ABO compatible, n (%)				<.0001
Yes	1,930 (93.7)	570 (90.1)	1,360 (95.3)	
No	24 (1.2)	4 (0.6)	20 (1.4)	
unknown	106 (5.2)	59 (9.3)	47 (3.3)	
Labs (peri-transplant)				
Creatinine at Discharge Post Transplant (g/dL)				0.09
Missing n (%)	17 (0.83)	5 (0.79)	12 (0.84)	
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]	
White Blood Cell at Transplant (10E3MCL)				0.001
Missing n (%)	54 (2.6)	18 (2.8)	36 (2.5)	
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]	
Hemoglobin A1C at Transplant (Percent)				<.0001
Missing n (%)	1037 (50.3)	194 (46.5)	743 (52.1)	
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]	
Hemoglobin at Discharge Post Transplant (GMDL)				0.05
Missing n (%)	14 (0.7)	4 (0.6)	10 (0.7)	
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]	

Transplant Milestones				
# of days from referral start to evaluation start, mean (S.D.) (n=number of subjects with non-missing data)	113.5 (953.4)	89.6 (246.4) (n=449)	123.1 (1117.3) (n=1,130)	0.34
# of days from referral end to evaluation end, mean (S.D.) (n=number of subjects with non-missing data)	268.7 (389.9)	314.4 (378.1) (n=449)	250.8 (393.7) (n=1,130)	0.005
# of days from evaluation start to evaluation end, mean (S.D.) (n=number of subjects with non-missing data)	280.3 (320.6)	326.4 (353.0) (n=460)	262.8 (305.8) (n=1,186)	0.001
# of days from evaluation end to waitlist start, mean (S.D.) (n=number of subjects with non-missing data)	-78.2 (271.6)	-104.9 (300.7) (n=460)	-68.1 (259.3) (n=1,186)	0.03
# of days from waitlist start to waitlist end, mean (S.D.) (n=number of subjects with non-missing data)	798.9 (676.8)	817.2 (719.9) (n=521)	788.6 (655.3) (n=1,300)	0.58
# of days from waitlist start to transplant, mean (S.D.) (n=number of subjects with non-missing data)	799.1 (677.5)	819.8 (721.9) (n=521)	787.7 (655.4) (n=1,300)	0.52

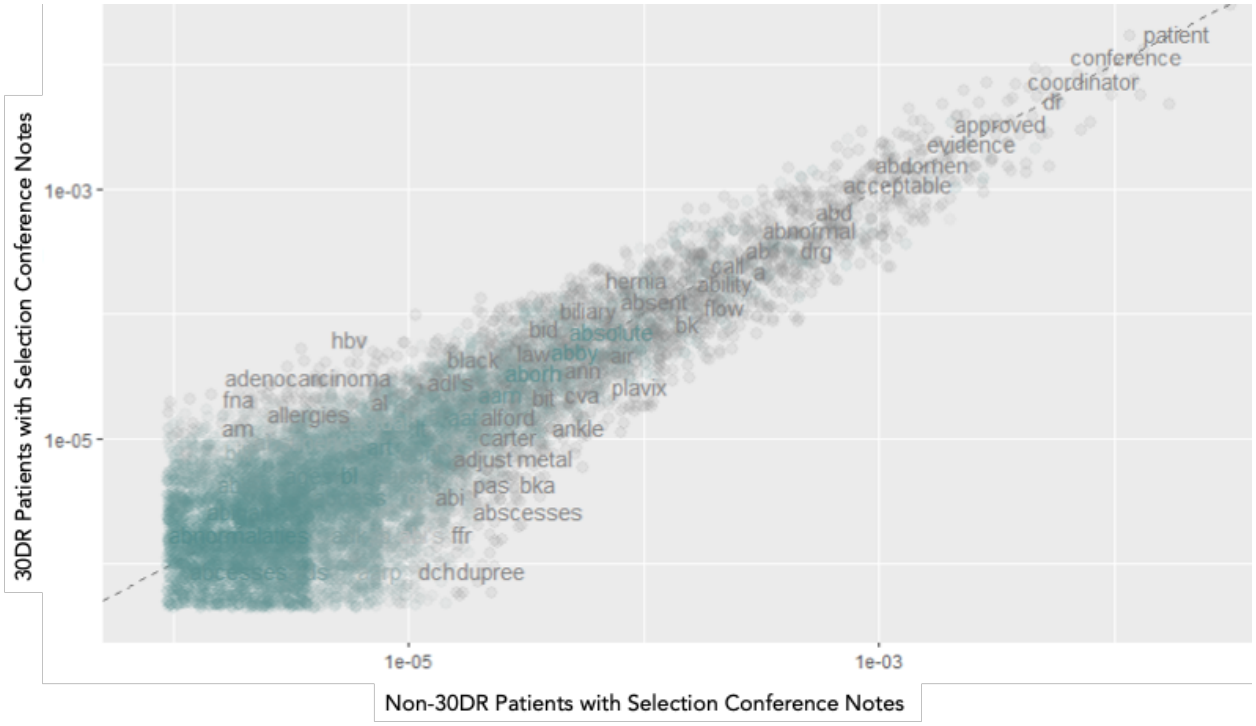
HCC coding is a payment model designated by the Centers for Medicare and Medicaid Services (HCC, Hierarchical Condition Category). CDC high risk guidelines developed by Centers for Disease Control in 1994 to notify and protect candidates (CDC, Centers for Disease Control). Length of Hospital stay from the day of transplant to the day of discharge. Risk describes risk of active infection; Comprised of High (donor +, recipient -), Intermediate (donor -, recipient +), and Low Risk (donor -, recipient -). (n=2060, years 2005-2015)

Figure 5: Frequency of words in three types of notes graphed by 30DR vs. non-30DR patients.

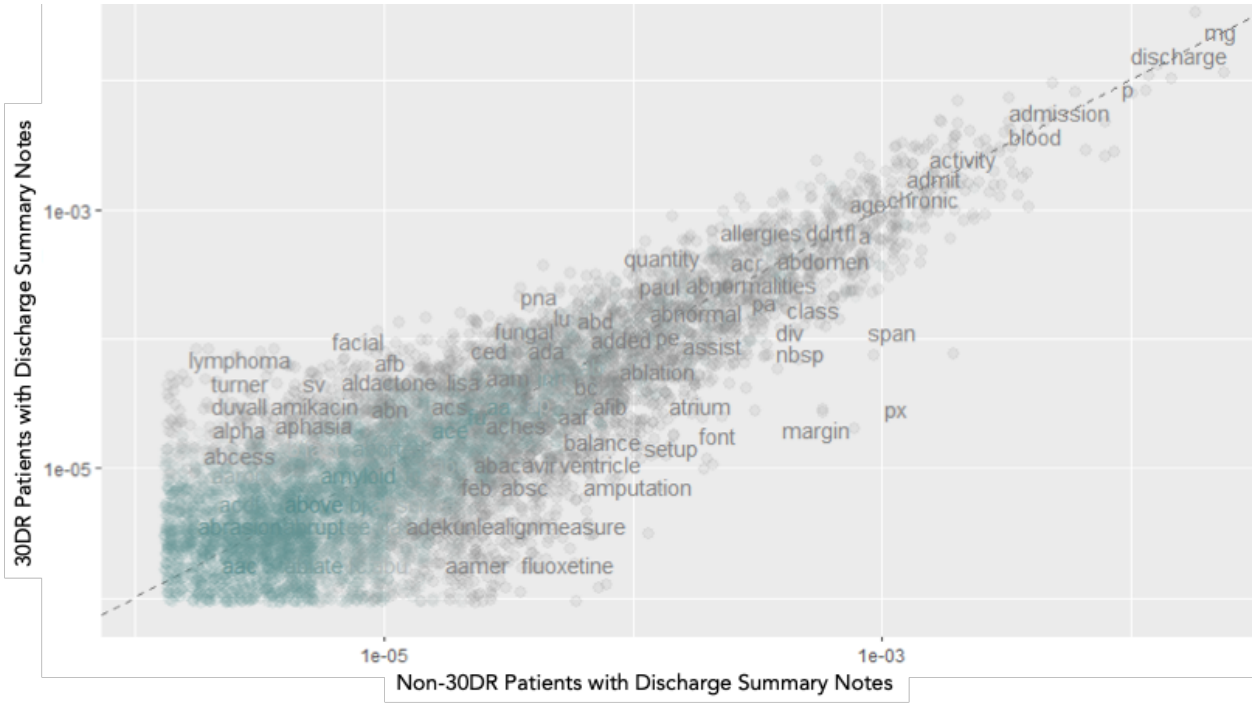
(A) Words Frequently Associated with 30DR in Consultation Notes



(B) Words Frequently Associated with 30DR in Selection Conference Notes



(C) Words Frequently Associated with 30DR in Discharge Summaries

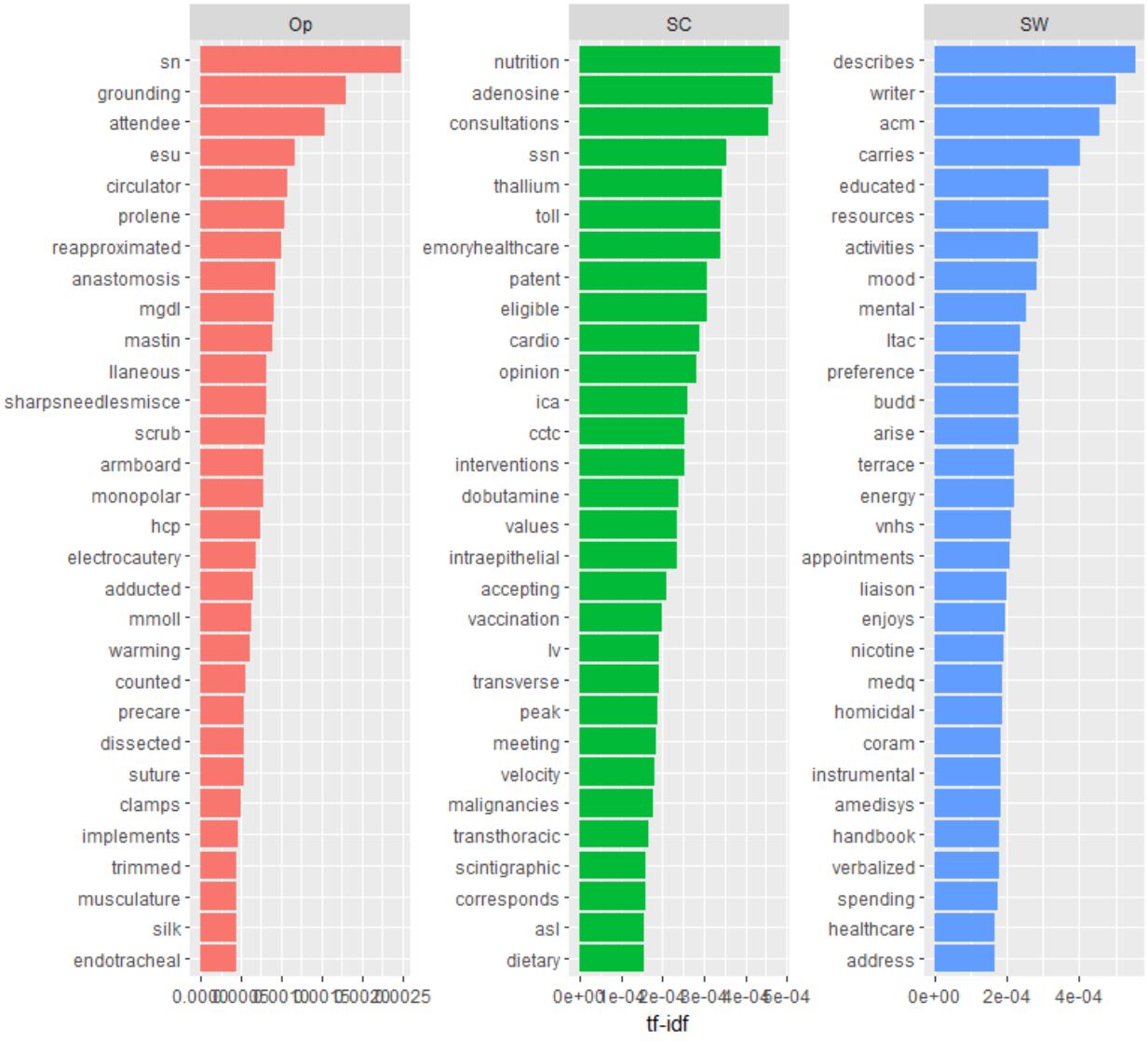


Graphs of three different note type examples plotting words according to term frequency. The x-axis are patients who did not experience 30DR. The y-axis are patients that did experience 30DR. Words on the grey dashed are terms that appear equally as frequent in patients notes that experienced 30DR as those that did not. The words appearing closest to the top-right occur more frequently. If a word appears closer to the top left of the graph, the word appears more often in patients notes who were readmitted within 30-days. (n=2060, years 2005-2015)

Table 2: Most Common Words that Precede the Word "Support" Amongst All Notes as Identified by Term Frequency

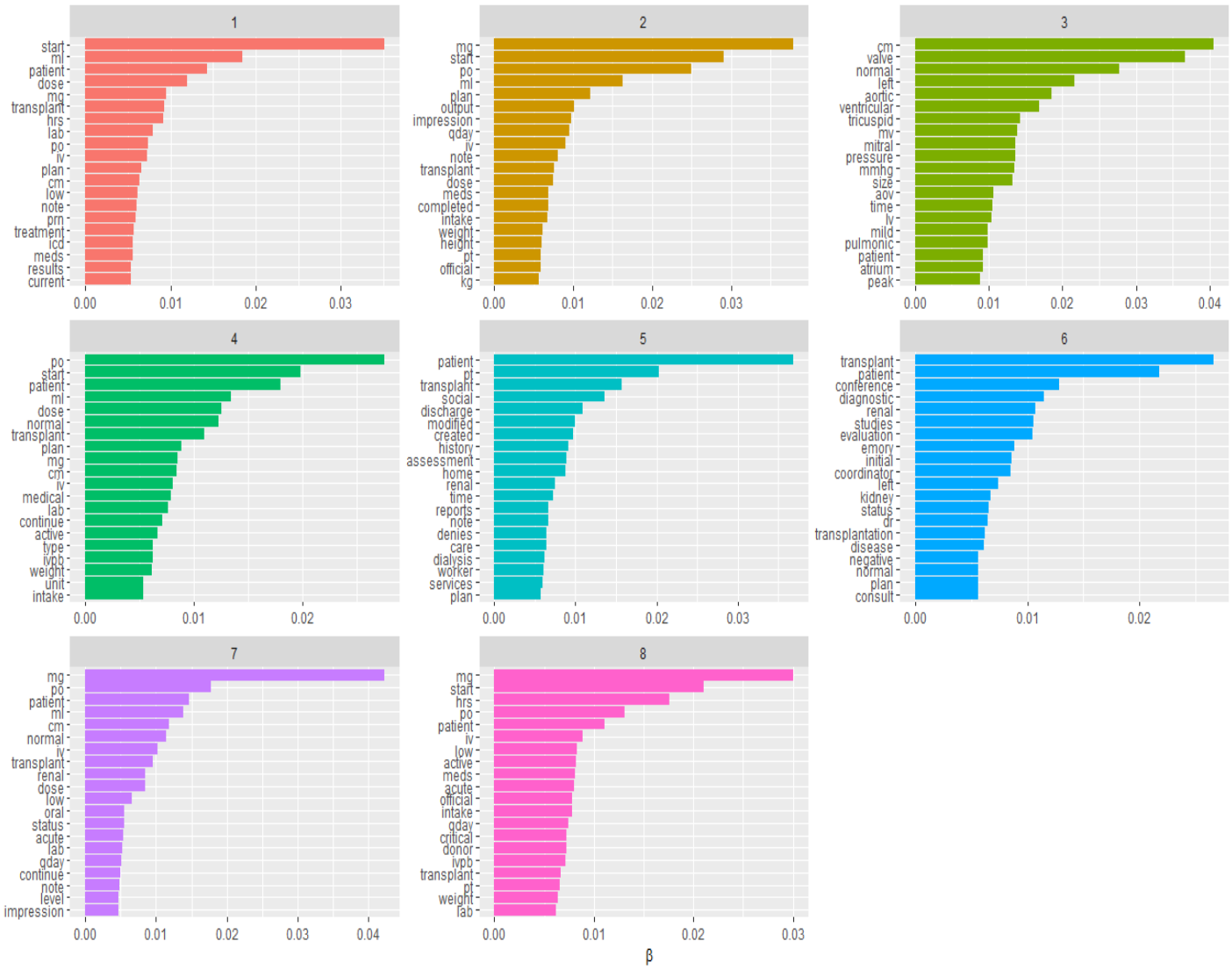
First Word	Frequency
care	1756
transportation	823
emotional	536
social	325
family	149
offer	99
strong	79

Figure 6: Top 30 TF-IDF features for Operative, Selection Conference, and Social Work notes



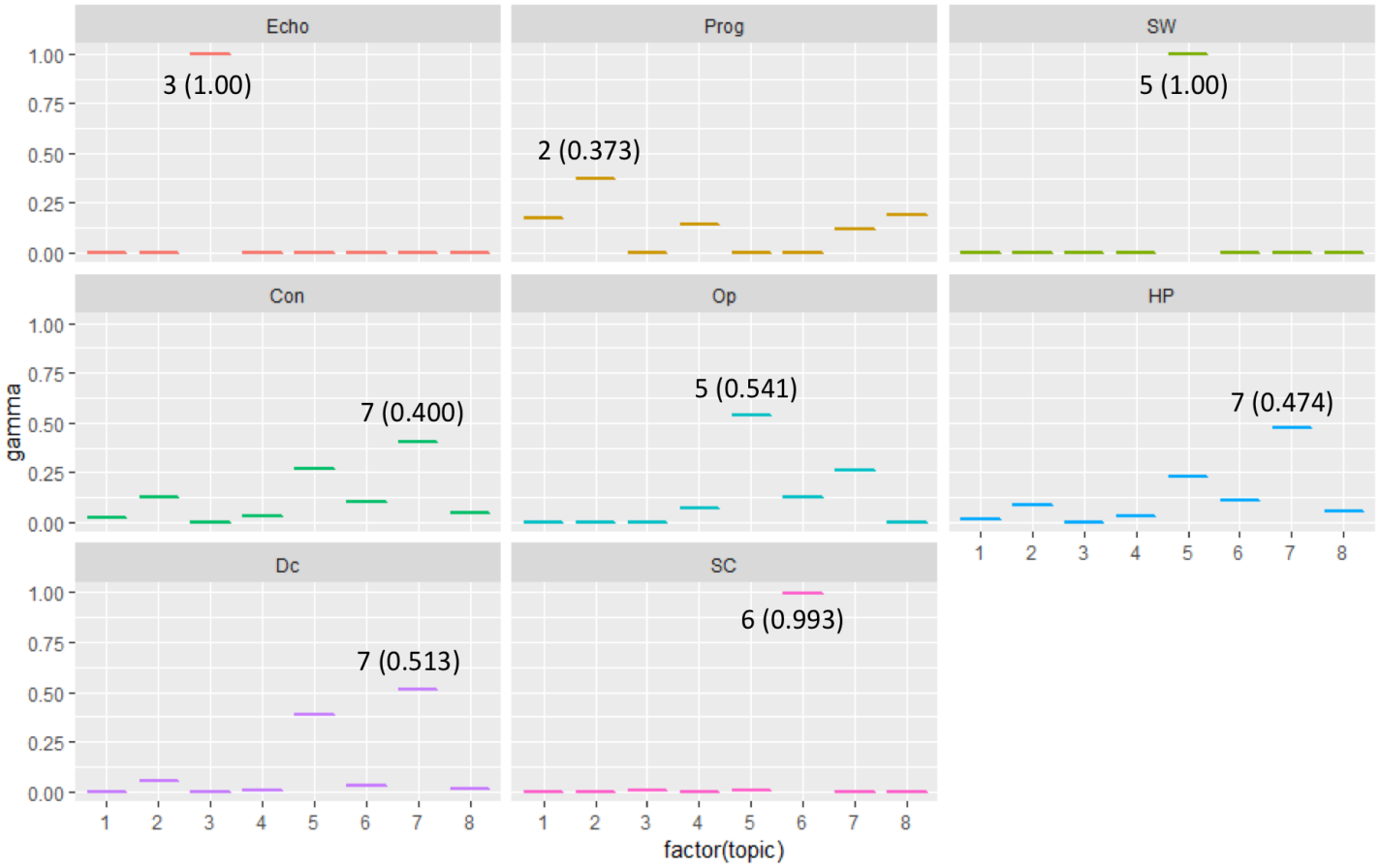
These words are, as measured by TF-DIF, the most important to each of the three example notes: Operative ("Op"), Selection Conference ("SC"), and Social Work ("SW") notes. The transplant team uses different language across note types. The words above are those that most frequently appear in one document but least frequently appear in all other documents. Thus, they are characteristic of one document more than any other and denote the importance of that word to that particular document. (n=2060, years 2005-2015)

Figure 7: Top 20 Terms in Topic Model using LDA with k=8 Topics



The researcher chooses how many topics, k, they want to create, but the topics are generated in an unsupervised manner. Beta is the probability of a word appearing in any given topic. The 20 words with the highest beta for each topic (k=8) is above. The central theme of each topic is up to the researcher's interpretation. See Discussion section for interpretation.

Figure 8: Gamma for Each Topic by Note Type



Each note broken down by the gamma of each topic. Notes include Echocardiography (Echo), Progress (Prog), Social Work (SW), Consultation (Con), Operative (Op), History & Physical (HP), Discharge (Dc), and Selection Committee (SC) notes. The topic with the highest gamma is identified for each document. The gammas for some documents are entirely comprised of one topic, whereas the gamma for other documents are spread out across many topics. Gamma is the words in a topic generated from a given document.

Table 3: Individual Clinical Notes Added to Structured Variables to Create Predictive Model

	LDA Ensemble		TFIDF Ensemble		LDA + TFIDF Ensemble	
Datasets (Added to Structured)	**AUC**	**95% CI**	**AUC**	**95% CI**	**AUC**	**95% CI**
Structured Only	0.6523	(0.6218, 0.6829)	0.6523	(0.6218, 0.6829)	0.6523	(0.6218, 0.6829)
Consultations	0.6597	(0.6367, 0.6826)	0.6609	(0.6398, 0.6819)	0.6617	(0.6428, 0.6807)
Discharge Summary	0.6535	(0.6238, 0.6831)	0.6551	(0.6249, 0.6853)	0.6543	(0.6252, 0.6834)
Echo	0.65	(0.6202, 0.6798)	0.6504	(0.6215, 0.6794)	0.6469	(0.6178, 0.6761)
H & P	0.6552	(0.6290, 0.6815)	0.6622	(0.6369, 0.6876)	0.6585	(0.6347, 0.6822)
Operative	0.6494	(0.6218, 0.6771)	0.6421	(0.6155, 0.6686)	0.6362	(0.6104, 0.6620)
Progress	0.6633	(0.6386, 0.6880)	0.6635	(0.6406, 0.6865)	0.6668	(0.6463, 0.6873)
Selection Conference	0.6575	(0.6282, 0.6868)	0.6617	(0.6350, 0.6883)	0.6587	(0.6323, 0.6850)
Social Worker	0.6482	(0.6193, 0.6770)	0.6459	(0.6175, 0.6743)	0.6411	(0.6137, 0.6684)

Table 4: Adding Multiple Note Types to Predictive Models for Hospital Readmission after Kidney Transplants

	LDA Ensemble		TFIDF Ensemble		LDA + TFIDF Ensemble	
Datasets (Added to Structured)	**AUC**	**95% CI**	**AUC**	**95% CI**	**AUC**	**95% CI**
Structured Only	0.6523	(0.6218, 0.6829)	0.6523	(0.6218, 0.6829)	0.6523	(0.6218, 0.6829)
Consultations + H&P	0.6586	(0.6366, 0.6807)	0.6642	(0.6427, 0.6856)	0.6601	(0.6388, 0.6814)
Progress + Consultations	0.6664	(0.6446, 0.6882)	0.6661	(0.6473, 0.6850)	0.668	(0.6489, 0.6871)
Progress + Discharge Summary	0.663	(0.6385, 0.6875)	0.6649	(0.6413, 0.6884)	0.6667	(0.6464, 0.6870)
Progress + Echo	0.6595	(0.6326, 0.6863)	0.6595	(0.6356, 0.6834)	0.6598	(0.6361, 0.6835)
Progress + H&P	0.662	(0.6377, 0.6863)	0.6679	(0.6465, 0.6893)	0.6657	(0.6442, 0.6871)
Progress + Operative	0.6571	(0.6325, 0.6817)	0.6498	(0.6279, 0.6717)	0.6475	(0.6259, 0.6691)
Progress + Selection Conference	0.6676	(0.6454, 0.6899)	0.6713	(0.6513, 0.6913)	0.6714	(0.6553, 0.6875)
Progress + Social Worker	0.6574	(0.6324, 0.6823)	0.6553	(0.6331, 0.6775)	0.6544	(0.6337, 0.6751)
Progress + Consultations + Selection Conference	0.6707	(0.6526, 0.6888)	0.6738	(0.6583, 0.6892)	0.6734	(0.6635, 0.6834)
Consultations + H&P + Progress + Selection Conference	0.6683	(0.6519, 0.6847)	0.6744	(0.6587, 0.6900)	0.6704	(0.6593, 0.6815)
Consultations + Discharge_Summary + Echo + Progress + Selection Conference	0.6672	(0.6486, 0.6859)	0.6724	(0.6563, 0.6884)	0.6699	(0.6589, 0.6810)
Consultations + Discharge Summary + Echo + H&P + Operative + Progress + Selection Conference + Social Worker	0.6574	(0.6400, 0.6749)	0.6608	(0.6454, 0.6763)	0.6535	(0.6397, 0.6674)
Best Score	0.6707		0.6744		0.6734	
Improvement	1.84%		2.21%		2.11%	

Table 5: Ranking top predictive features for higher readmission of kidney transplant recipients (2005-2015) in highest performing predictive model from Table 4

Rank	Data Source	Feature
1	Structured	Albumin minimum at Discharge
2	Structured	Creatinine maximum prior to transplant
3	Structured	Recipient CMV infection risk: High
4	Structured	Albumin minimum at time of transplant
5	Structured	Hepatitis C Status of Recipient
6	Structured	Creatinine prior to transplant
7	Structured	Race (Recipient)
8	Structured	Creatinine maximum at Discharge
9	Structured	Prograf maximum at Discharge
10	Structured	Recipient EBV infection risk: High
11	Structured	Donor Type (Living vs. Deceased)
12	Structured	Hemoglobin minimum at Discharge
13	Structured	# of days from referral start to evaluation start
14	Structured	Recipient HCC Risk
15	Structured	Albumin maximum at Discharge
16	Structured	Albumin at Discharge
17	Structured	Donor Blood Type (ABO)
18	Structured	Recipient White Blood Cell count at time of Transplant
19	TFIDF: Progress Notes	"mg"
20	Structured	Change in Albumin from Transplant to Discharge

HCC coding is a payment model designated by the Centers for Medicare and Medicaid Services (HCC, Hierarchical Condition Category). CDC high risk guidelines developed by Centers for Disease Control in 1994 to notify and protect candidates (CDC, Centers for Disease Control). Length of Hospital stay from the day of transplant to the day of discharge. Risk describes risk of active infection; Comprised of High (donor +, recipient -), Intermediate (donor -, recipient +), and Low Risk (donor -, recipient -). (n=2060, years 2005-2015)