Data Sci 223 Final Project

Kevin Shi

Semi-Quantitative Analysis of Ambulatory Nephrology Notes: Application to Quality of Telehealth Encounters

**Note: Generative AI (ChatGPT) was used in this project for coding and debugging assistance. No protected health information was included in generative AI prompts. No generative AI was used to assist in results analysis or discussion. Other than coding assistance, all work in this project reflects the work of Kevin Shi (no other students).**

Introduction:

The modern healthcare landscape has been recently and irrevocably changed by the widespread adoption of telehealth services that have changed the nature of patient and provider interactions. Telehealth has been generally regarded favorably both by providers and patients, offering higher flexibility and access to care. This is especially impactful in specialties where there is a relative provider shortage and thorough physical exams, or at least physical exams that require elements that cannot be transmitted remotely, are not critical to clinical decision making, such as psychiatry or nephrology. However, there are no consensus guidelines on how to best implement telehealth clinically, and the fundamental question of quality of care during telehealth encounters remains unanswered: can telehealth be used as a substitute for in-person care? Currently, COVID-19 era emergency waivers for telehealth, after being extended numerous times, are set to expire September 30, 2025. The performance of telehealth encounters can help answer this urgent question.

For this project, I will apply data science concepts learned in Data Sci 223 to analyze clinical information of both telehealth and in-person encounters. Fundamentally, my hypothesis is that, if there are differences between telehealth and in-person encounters with respect to documentation, orders, or actions done in the encounters, then perhaps telehealth and in-person encounters are not substitutes and should be considered differently for future policy, education, and implementation reasons. My dataset is 37728 encounters from January 1, 2020 to December 31, 2024 at the at the University of California San Francisco adult general nephrology clinic. These encounters were completed (i.e. not canceled or no-show), served patients 18 years or older, and were done with MD/DO physicians. There were 55 features in this dataset (though some of the features were redundant), yielding a total dataset size of 37728 x 55.

I will be using multiple outcomes of interest. For my first outcome, I will utilize sentiment analysis. Sentiment analysis offers a window into healthcare quality, providing a semi-quantitative way to secondarily analyze patient-provider interactions from clinical documentation. This has numerous applications and has been studied in the realm of predicting mental illness, characterizing provider bias, forecasting clinical outcomes (Denecke 2023). For example, at UCSF, there is ongoing research using natural language processing to analyze notes to see if stigmatizing language is associated with different rates of kidney transplant referral.

I initially used a transformer-based model for sentiment analysis, "distilbert-base-uncased-finetuned-sst-2-english". This is a default Hugging Face model. In order to facilitate run time, I used a random sample of 1000 notes.

The Distilbert model returns one of two labels for a piece of text, either “positive” or “negative” and a corresponding score from 0-1 that represents how confident the model is in its score. Surprisingly, when run on the clinical notes of 1000 nephrology encounters, the vast majority were scored as having “negative” sentiment (988:8; 4 missing/NA).

A blue rectangular bar graph

AI-generated content may be incorrect.  
 **Figure 1 – DistilBERT sentiment analysis labels for 1000 clinical notes**

To see if this performance was affected by miscellaneous content in clinical notes, such as lab values, medication names, etc., I attempted to isolate the “subjective” portion of clinical notes, following the logic that the subjective portion of notes would have the most variation in provider diction and word choice, and thus be the most expressive in terms of sentiment. A variety of methods were trialed to best extract the subjective portion of notes. Unfortunately, most of these, including using pre-trained clinical NLP tools like medSpaCy, did not perform well. I believe that this is due to the high heterogeneity of clinical note text, which varied by provider, clinical context (e.g. an encounter for glomerulonephritis vs. polycystic kidney disease vs. kidney transplant vs. stone disease etc.), encounter context (follow-up vs. new patient visit), and institutional context (use of different templates for billing, coding, or compliance reasons). I ultimately ended up using a regular expression-based strategy. This required a significant amount of fine-tuning, and ultimately ended up only extracting subjective portions of 865/1000 notes, likely due to fundamental limitations of a regular expression-based strategy, which is subject to very small variation in text.

I re-ran the sentiment analysis on just the “subjective” component of the notes. The results were similar, with the vast majority of notes having a “negative” sentiment.

A graph showing negative and negative

AI-generated content may be incorrect.

**Figure 2 – DistilBERT sentiment analysis labels for the subjective sections of 1000 clinical notes**

The distribution of positive versus negative sentiment was not substantially different for telehealth versus in-person encounters.

A graph of a positive result

AI-generated content may be incorrect.

**Figure 3 – Distribution of sentiment labels across encounter types**

Given the somewhat unexpected output of sentiment analysis in this case, I pursued a slightly different way of sentiment analysis using the VADER (Valence Aware Dictionary and sEntiment Reasoner) tool. It seemed to have a slightly different output compared to the transformer method, with slightly more “positive” sentiment notes.  
  
A blue and white rectangular box with numbers

AI-generated content may be incorrect.

**Figure 4 – Confusion Matrix of VADER vs. DistilBERT for sentiment analysis on the subjective sections of 865 clinical notes**

I manually evaluated a handful of instances of disagreement where VADER and DistilBERT assigned different sentiment labels. I evaluated instances where VADER assigned a positive label and DistilBERT assigned a negative label. Subjectively, I felt that some of the diction used here was potentially more negative. Some snippets are included below:

*She could not tolerate a full dose of cellcept due to night sweats/palpitations. She had worsening proteinuria in Sept through Dec 2021 and* ***refused a repeat biopsy****. She was reinitiated on pulse steroids which worked slowly to improve her proteinuria.* ***She then finally agreed to a biopsy*** *which was performed in February 2021 and showed mild chronicity with class II lupus nephritis only.*

***Complaining*** *of daytime somnolence and poor sleep onset with seroquel. Stating that he functions better without seroquel Because he takes all of his medications at once (at bedtime),* ***he misses his medications*** *such as lisinopril because* ***he wishes to skip his seroquel*** *dose Prescribed lisinopril 10mg daily.* ***Does not regularly check his BP at home****. Now back to work at a church, parttime in the office, states it is low stress and a positive sign in his life. But he is lo longer under care at LPPI. Ongoing issues with disordered sleep Ongoing B12 injections Hasn't been able to exercise regularly. Been gaining weight, but knows the importance of weight loss.* ***Lacks motivation to exercise.***

Subjectively, I would lean towards the DistilBERT label of these encounters, though this is not a rigorous analysis. Using the VADER labels did not really seem to show a relationship between sentiment label and encounter modality either.

A comparison of a graph

AI-generated content may be incorrect.

**Figure 5** – **Sentiment label distributions across encounter type for VADER and DistilBERT**

I then refactored sentiment labels, defining a combined label that was positive when both modalities labeled the note as positive, negative when both labeled negative, and unclear when the methods disagreed. With this adjustment, there still was not a significant difference in distributions of sentiments across encounter type. In addition, there were so few instances where both models assigned a positive label that association was difficult to assess. Using a Fisher exact test and combining unclear and positive labels into a “not-negative” class returns a p-value of 0.43, suggesting that we do not have evidence to reject a null hypothesis that encounter modality and sentiment label are independent.   
  
A graph of a diagram

AI-generated content may be incorrect.

**Figure 6** **– Distribution of combined sentiment labels by visit type**

Overall, I’m not entirely sure if the conclusion above is valid. I’m suspicious if these general sentiment analysis methods work all that well on medical text.

I then pivoted to another outcome of interest, medication prescription changes, defined as the number of new prescriptions, prescription cancellations, or dose changes made during an encounter. Reordering of medications was not counted as a medication prescription change. Similarly, changes to historical or patient-reported medications, such as over-the-counter medications, or medications prescribed from other institutions, did not count as a medication prescription change. Medication changes were defined and tabulated in a process not included in this final project (i.e. cannot be found in the code I submit).

In the spirit of this class, I tried to model medication prescription changes using Python. Understanding that there are repeated observations per patient and provider in this data set, I leveraged a hierarchical mixed effects model. However, this seemed excruciatingly difficult to do in Python. After numerous issues with packages, upgrading and downgrading versions, I gave up. While this functionality seems to be directly supported by the PyMC package, this seems to require probabilistic simulations that, even for my relatively reasonable data set (37728 observations), ended up requiring over 25 hours of computation time to run. I ended up doing this in R, which may admittedly be against the spirit of this class, but there is some formal interoperability between the two languages.

I ended up modeling the number of medication prescription changes as an outcome variable predicted by encounter modality, patient factors (age, sex, and race), and encounter factors [estimated glomerular filtration rate (eGFR), baseline number of medications, time since last visit, and encounter type (new patient or follow-up), history of diabetes, and history of hypertension] that were hypothesized to affect medication change rate, as well as patient and provider random effects. Using ANOVA tests, models with random slopes and random intercepts for patient-level effects, but only random intercepts for provider level effects, worked best. This model found that telehealth encounter modality was strongly associated with fewer medication prescription changes than in-person encounters.

A graph with red and white dots

AI-generated content may be incorrect.

**Figure 7 – Forest plot of medication changes.** For the categorical variables of race, sex, first visit, and telehealth, IRRs are given in comparison to reference values of "White," "Female," "Follow-up", and "In-person," respectively.

When looking at this more clearcut, “hard” outcome, there do seem to be systematic differences between telehealth and in-person encounters.

Overall, I was able to apply natural language processing and traditional modeling methods to this data set. The general natural language processing processing methods did not detect any significant difference between the sentiment of telehealth versus in-person notes. A mixed effects model did detect differences between telehealth versus in-person notes. In the future, I would like to see how specialized, medical natural language processing technologies do in this context.

**Quick answers**

Who worked on the project, if not just yourself – Only Kevin Shi

Overview of the problem – See introduction

Description of the dataset you used (input features, outcome,dimensions, etc) – See introduction

Tools/methods used – Hugging Face distilBERT, VADER from the nltk package, glmmTMB from the R glmmTMB package

Decisions made along the way, including trade-offs e.g., cut X for time so our solution may lack Y) – Cutting the dataset down to 1000 for natural language processing methods – the transformer-based approach seemed to take significant computational power

Issues overcome along the way – Problems with packages, environments, versions, regular expressions, and mixed effects models in Python

How to run the code (dependencies, etc.) - Straightforward

Example output (what does it do?) – Creates the plots in this write-up.

Citations (data, code, papers) – Below and attached in git repository.

References

1. Denecke K, Reichenpfader D. Sentiment analysis of clinical narratives: A scoping review. J Biomed Inform. 2023 Apr;140:104336. doi: 10.1016/j.jbi.2023.104336. Epub 2023 Mar 22. PMID: 36958461.