**Survival Analytics Approaches for Adherence**

Erin Rebholz, CSCI 82, Fall 2023: Final Project

**Problem**

Understanding the drivers behind medication adherence helps pharmacies, payors and drug manufacturers develop interventions that may improve patient adherence. Diabetes and statin medications, in particular, help treat chronic conditions that require daily medications and are interesting examples with large patient populations. The goal of this study is to examine how survival analytics methodologies may be applied to predict patient adherence. For many years, state-of-the-art analytics for medication adherence has been cohort analytics studies using Kaplan Meiers and Cox Proportional Hazard curves. This study evaluates tree-based methods such as Random Forest and Gradient Boosting to see how well they might predict patient adherence behavior, in conjunction with other measures, including proportion of days covered of (PDC).

**Data Set**

Initially, I’d hoped to use the SyntheaMass[[1]](#footnote-1) data set, which is a synthetic patient data set that has been developed by the US federal government to support research and development use. However, the medication data set was not suitable for pharmacy adherence analysis, as it is structured such that patients are 100% adherent to chronic care medications, from the start of the first fill to when the patient dies. That said, because there is information on patient death, I was able to structure the censor & duration fields to reflect when the patient died and how long the patient had been taking statins prior to death. Therefore, I was able to provide a working sample code for statins using my selected survival prediction methodologies.

Given the Synthea limitations, I acquired a pharmacy deidentified data set under license to my personal LLC, Merilytic Health. While I can share analytic outputs for this data, there are limitations in ability to share the data set more broadly due to potential risk of patient reidentification and limitations for sharing the data outside the United States. Therefore, for this data set I’m only able to provide the python workbook in pdf form.

**Approach**

***Treatment Area and Study Window Selection***

The RxNorm data set[[2]](#footnote-2) was used to identify the NDCs for three different treatment categories, anti-tnf drugs, GLP-1’s, a newer class of diabetes and weight control medications, and DPP-4’s, an older class of daily diabetes medications. GLP1-s and **Wegovy and Ozempic,** in particular, were selected as they had good data coverage. The analysis period was set at **Jan 1, 2023 to Dec, 1, 2023**, given there was good data completeness and patient availability, with **3,162** patientsand **9,102** fills.

From the SyntheaMass data set, I found that the dataset consisted mainly of generics products, so I selected the statin class, given it would have good longitudinal data (first statin was approved and released in 1991). That said, the sample data set contained missing fill data for a large number of the medication records, it also contained examples of statin use prior to the commercial availability of the drug category. For the **statin** study, **660** patients remained, with records between **Jan 1991 to July 2021**.

***Exploratory Data Analysis and Censor Duration Calculations.***

Exploratory analytics was conducted on the data set to better understand the ranges of values for features. The distributions of days supply and fill dates, which were used to conduct the censor and duration analytics, were also profiled.

For the censor and duration analysis, a patient is considered adherent if they refilled their medication within 30 days of the expected fill date, with the expected fill data calculated as the previous fill plus the days of supply. If the patient didn’t fill in time, they were considered censored, even if there were subsequent fills of the medication available. This is a forgiving design choice that could be refined to narrow, or expand the window based on the use case being considered.

Finally, the proportion of days covered, or PDC, was also calculated which looks at the ‘days covered’, or days supply of the medication within the study window , divided by the total days the patient was in the study. The begin date in the study varies by patient as new patients come into the study. The end date is either the last day of the study for non-censored patients, or the last end date at the time of the censored fill. This way PDC score gives only information to the model about known past behavior at the time of censor.

***Adherence Prediction Analytics***

Data was divided into train and test data sets, and then three methods were used to predict patient adherence: Cox Proportional Hazards, Random Forest Survival, and Gradient Boosting. In addition to class materials and Lifelines package used in class, the Scikit-Survival package in Python was employed for the ensemble models and result evaluation, and previous research from Humboldt University was helpful in developing my approach.[[3]](#footnote-3)[[4]](#footnote-4) I was able to generate time-dependent AUC curves for each method, which I found useful to compliment, Concordance Index and Integrated Breier Score in evaluating model performance.

**Results**

Initial Kaplan Meier and Cox Proportional hazard curves show a sharp drop in probability for survival after the first and second fills (approximately 30 and 60 days for a product that is dispensed in 28 and 30-days supply). Based on the time-dependent AUC curves for GLP-1s, we see our predicted adherence curves improve as the duration increases and we have more knowledge of patient behavior. In the future, introducing average patient compliance or structuring PDC scores in windows of previous compliance (3, 6, 12 month views) could also be used to demonstrate history prior to the study period and could help improve predictive capabilities earlier in the duration window. Finally, additional factors such as prior authorization status and loyalty program use could be interesting to see how or if they improve adherence, given that patient pay amount is a significant factor in predicting adherence.

**Reflections | Next Steps**

I look forward to continuing to refining this model and expanding both methodologies and feature engineering to support the analysis. DeepSurv is an interesting approach that I began but requires additional work given the age of the code available (last updated to support python 3.8)[[5]](#footnote-5).

**Note:** Several other studies and web-based resources were helpful in developing my approach and are referenced within the code based. The reference list below were the most important sources.

**References:**

1. “RxNorm.” Product, Program, and Project Descriptions. U.S. National Library of Medicine. Accessed December 19, 2023. https://www.nlm.nih.gov/research/umls/rxnorm/index.html.
2. Walonoski, Jason, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. “Synthea: An Approach, Method, and Software Mechanism for Generating Synthetic Patients and the Synthetic Electronic Health Care Record.” *Journal of the American Medical Informatics Association* 25, no. 3 (March 1, 2018): 230–38. https://doi.org/10.1093/jamia/ocx079.
3. “Ensemble Models — Scikit-Survival 0.22.2.” Accessed December 19, 2023. https://scikit-survival.readthedocs.io/en/latest/api/ensemble.html.
4. “Deep Learning for Survival Analysis.” Accessed December 13, 2023. https://humboldt-wi.github.io/blog/research/information\_systems\_1920/group2\_survivalanalysis//.
5. Katzman, Jared L., Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. “DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network.” *BMC Medical Research Methodology* 18, no. 1 (February 26, 2018): 24. https://doi.org/10.1186/s12874-018-0482-1.

1. Walonoski et al., “Synthea.” [↑](#footnote-ref-1)
2. “RxNorm.” [↑](#footnote-ref-2)
3. “Deep Learning for Survival Analysis.” [↑](#footnote-ref-3)
4. “Ensemble Models — Scikit-Survival 0.22.2.” [↑](#footnote-ref-4)
5. Katzman et al., “DeepSurv.” [↑](#footnote-ref-5)