Christine Gao

DS-GA 1015 Text as Data

Spring 2024

# DS-GA 1015: Text as Data Final Paper

## Topic Modeling Social Determinants of Health in Electronic Health Record Data

### Introduction

Extracting and interpreting useful information from electronic health records (EHRs) still remains one of the greatest challenges in NLP today. The nature of electronic clinical health data is complex and subjective to the note taker. Physicians are tasked with writing patient notes, documenting patient health history, and writing the discharge reports of the patient, making the task a tedious and complex one. Due to the inherently messy and unstructured nature of EHRs, subsequent analysis relies heavily on upstream tasks of preprocessing.

NLP poses a powerful tool to extract social determinants of health (SDoH) information from clinical notes, as they reveal the circumstances in which people are born, live, and work. SDoH are related closely to an individuals' health behaviors, lifestyle, and interpersonal relations. Previous studies have found associations between SDoH and various health outcomes, such as the effect of socioeconomic status, employment, and race on cancer or housing status on mental health (Patra, 2021).

This project will attempt to delve into the capabilities of NLP tasks for medical data, and seek to uncover whether there exists notable indicators of SDoH in physicians notes, better understand these trends, and capture social and behavioral concepts for leverage in clinical decision-making for improvement of health equity.

### Literature

Previous studies and models that have utilized NLP to detect various terms linked to risk factors for disease, mortality, and disabilities. A number of published studies have linked social

risk factors (such as smoking and substance abuse) and their impact on morbidity (Wang, 2015) The CDC has also named cancer, heart disease, and chronic lower respiratory disease as some of the leading causes of preventable premature deaths in the U.S from 2010-2022 (Garcia, 2024). It thus becomes imperative that efforts are focused on minimizing health disparities, identify populations with high prevalence of preventable mortality, and enact interventions against health risk factors. While there exists NLP advances that employ topic analysis to create interpretable text classification from EHRs, they have been used largely for predictive or classification tasks, such as predicting hospital readmission or disease classification. Furthermore, little analysis has been done on the MIMIC-III clinical database, making this task an interesting challenge for NLP in medical contexts.

**Theory and Hypothesis**

Clinical notes are difficult to parse and extracting information is inherently noisy, but NLP techniques can alleviate these tasks and reveal information about social risk factors that could help health care providers better attend to their patients. This project aims to construct indicators of an individuals health history within free-text clinical notes, as well as demonstrate the interpretive value of Latent Dirichlet Allocation (LDA) as an effective technique for finding keywords and hidden topics from large corpora using unsupervised learning. It is the hope that the topics constructed from this analysis will reflect those found in existing knowledge of social risk factors.

**Data and Methods**

The dataset utilized for this investigation is from the MIMIC-III clinical database. It comprises de-identified, health-related data linked to over forty thousand patients who were admitted to critical care units at the Beth Israel Deaconess Medical Center from 2001 and 2012

(Johnson, 2016). The tables contain patient stay data, critical care unit data, and hospital record data. For this analysis, only the 'NOTEEVENTS' and 'ADMISSIONS' tables were used, which contain the free text physician recorded notes and the demographics of patients admitted to the hospital. To reduce the data further, only the 'Discharge' notes were selected for the topic modeling task.

Because of the structure of the 'NOTEEVENTS' clinical notes, preprocessing was necessary by applying a standard text cleaning pipeline. The first step was removing newlines, any sensitive identifiers (i.e names, dates, and named entities), and removing special characters and excess punctuation. From this free text, four existing categories of health indicators were derived from the data: **history of present illness**, **past medical history**, **social history**, and **family history**. The 'ADMISSIONS' dataset contains additional categories that could be indicative of an individuals' health circumstances, and thus information regarding **insurance**, **religion**, **ethnicity**, and **marital status** were also considered in the final analysis.

To prepare the data for analysis, documents were tokenized and stopwords removed. The remaining data is stored as a document-term matrix, where each document represents a note and characteristics corresponding to one patient. To this LDA was applied to create topic-word probabilities and document-topic probabilities.

To determine the optimal number of topics, a manual gridsearch was performed to calculate the perplexity metric, which measures the predictive ability of the LDA model. Across all indicators, the optimal number of topics was 10. For each indicator, the LDA model calculates probabilistically the topics which contribute most to each document. From this distribution, the topic with the highest probability was assigned as the 'label' for that given document.

**Results**

The results of the Latent Dirichlet Allocation analysis (LDA) for each indicator, health category, and their corresponding topics are shown in **Figures 1-4.** For each topic, the top 10 words are presented, and can be viewed as a combination of keywords that compose the meaning of the topic. For example, in the 'social_history' topic, topic 8 contains terms such as 'heavy', 'past', 'smoker', and 'alcohol' - all of which point to an underlying concept of an individual with a history of substance abuse. Similarly, in the 'medical_history' topic, topic 1 contains the terms 'cardiac', 'stenosis', 'ventricular', and 'risk', which suggest a patient with a history of aortic stenosis, a medical condition involving the narrowing of the heart valve. These patterns can be identified by inspecting the LDA topics, and the assigned labels or overall concepts are included in **Table 1.** An interpretation of these topics is included in **Table 2**. Across the dominant topics, the demographic characteristics that occur most frequently are patients with Medicare or private insurance, married and single individuals, practice Catholicism and are of White or Black/African American ethnicity. However, we must consider that these demographic characteristics compose the majority of the patients in this subset of the database, and likely imply the population characteristics of the specific region.

The predominant topics in the **social history** indicator are 2, 4, and 6. These topics contain information pertaining to patients' history of smoking, alcohol, and tobacco use. The most prevalent topics from the **family history** indicator are 1, 5, and 9. These topics relate to a history of diabetes or cardiac issues in the family, family history of seizures, and family history of renal or kidney failure. The most common **medical history** topics are 2, 4, 7, and 8. These contain phrases that indicate a history of COPD linked to kidney failure, Hepatitis, diabetes linked to pulmonary hypertension, as well as heart disease. Finally**, history of present illness**

represents the patients' medical record and their condition upon admittance to the critical care unit. The dominant topics were 4, 6, 7, and 9, and the terms specify patients admittance to the emergency room as a result of fainting or headache, vomiting and abdominal pain, being transferred from a different hospital, or a surgical procedure.

The results of the topic models implies that the most common SDoH relate to previous substance use, existing history of diabetes or medical disorders amongst family members, and prevalence of heart disease. Analysis also found that across patients with Medicare or private insurance, previous tobacco use and diagnosis of diabetes were consistently high. This trend holds true for family history of cardiac disease. Interestingly, clinical notes pertaining to the present illness topic of pregnancy (Topic 10 in Table 1) were most prevalent in patients of White ethnicity with private or Medicaid insurance, and were insignificant in the marital status indicator. This could imply an imbalance in the type of maternity-related care being provided to White patients with insurance vs other ethnicities. In summary, this analysis demonstrates the commonness and uniqueness of topics around SDoH depicted across various diseases, conditions and health histories which afflict patients.

**Discussion**

The study demonstrates that unsupervised topic modeling with LDA can extract hidden themes from a large and unstructured corpus of electronic clinical notes and identify factors, namely pre-existing SDoH. By narrowing our search to various categories of health histories, namely social, family, and medical histories, common topics of diabetes, heart disease, smoking, alcohol and tobacco use are distributed across these categories. As one of the aforementioned examples, the analysis shows that social history of tobacco and alcohol use co-occur with medical history of diabetes, and family history of cardiac disease. This information can aid in

better understanding how SDoH are documented in EHRs and how clinical notes can yield insight into the previous health conditions of patients.

This analysis has its strengths and weaknesses. LDA provides comprehensive and interpretable topic modeling to a large corpus of notes that is mostly unexplored. It focuses on finding SDoH topics across various demographics and health indicators, and identifies common diseases and health histories that appear in conjunction with each other. Additionally, the topic probabilities provide logical clusters to describe and categorize documents as being indicative of various SDoH. However, LDA is not exhaustive, and can struggle with determining an optimal number of topics when generating the topic distribution. Additionally, it relies on a 'bag-of-words' representation, so order and context of words can result in lost meaning. Future work should assess the performance and interpretability of topic modeling across different subsets of notes within MIMIC (i.e Nursing, Radiology, or Social Work notes), and from some clustering techniques to further add interpretability to the topics generated. Additionally, a systematic method for data preprocessing could greatly improve future NLP tasks for the MIMIC datasets, and future EHR analysis.

**Bibliography**

Allen, Katie S, Hood, D, Cummins, J, Kasturi S, Mendonca E, Vest, J, *Natural language processing-driven state machines to extract social factors from unstructured clinical documentation*https://academic.oup.com/jamiaopen/article/6/2/ooad024/7128296?login=false

García MC, Rossen LM, Matthews K, et al. Preventable Premature Deaths from the Five Leading Causes of Death in Nonmetropolitan and Metropolitan Counties, United States, 2010–2022. https://www.cdc.gov/mmwr/volumes/73/ss/ss7302a1.htm

Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). *PhysioNet*. https://doi.org/10.13026/C2XW26.

Patra, Braja G et al. "Extracting social determinants of health from electronic health records using natural language processing: a systematic review." *Journal of the American Medical Informatics Association: JAMIA vol. 28,12 (2021): 2716-2727.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8633615/

Rijcken, Emil et al. "Topic Modeling for Interpretable Text Classification From EHRs." *Frontiers in Big Data. vol 5 (2022).* https://www.frontiersin.org/articles/10.3389/fdata.2022.846930/full

Wang, Yan et al. "Automated Extraction of Substance Use Information from Clinical Texts." *AMIA ... Annual Symposium proceedings. AMIA Symposium. 2015.* *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765598/*

## Appendix

### Table 1. LDA Topics and Terms from MIMIC-III Discharge Notes

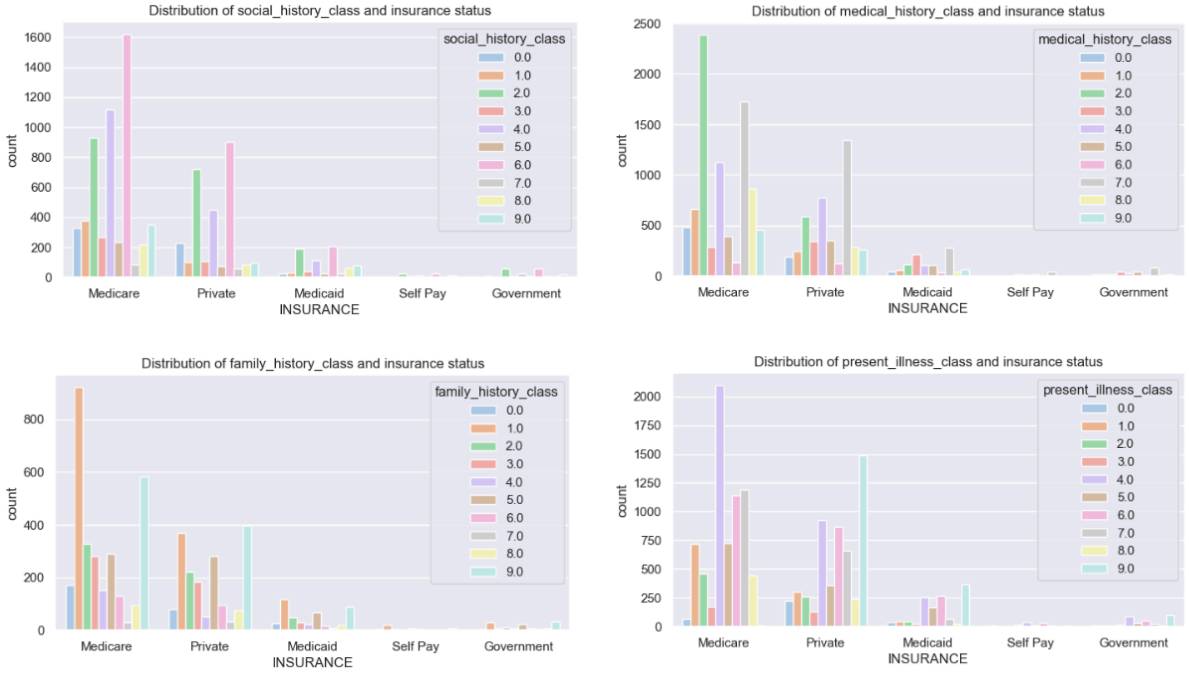
| Topic | Social History | Medical History | Family History | History of Present Illness |
|---|---|---|---|---|
| *1* | patient" + "drink" + "married" + "smoke" + "smoker" + "lives" + "former" + "alcohol" + "years" + "children" | "date" + "cardiac" + "left" + "stenosis" + "cabg" + "risk" + "mild" + "ventricular" + "factors" + "showed" | "diabetes" + "mi" +"died" + "two" + "brother" + "brothers" + "cad" +"one" + "sisters" + "sister" | "right" + "femoral" + "patient" + "entity" + "patent" + "common" + "left" + "iliac" + "lower" + "accident" |
| *2* | "name" + "entity" + "patient" + "care" + "home" + "lives" + "proxy" + "history" + "health" +"nursing" | "sp" + "date" + "history" + "chronic" + "gerd" + "htn" + "kidney" + "copd" + "disease" + "l" | "noncontributory" +"history" + "family" +"patient" + "cardiac" + "death" + "sudden" + "disease" + "diabetes" + "significant" | "patient" + "history" + "name" + "pain" + "disease" + "renal" + "date" + "dyspnea" + "year" + "diabetes" |
| *3* | "day" + "per" +"years" + "one" + "two" + "pack" + "drinks" + "patient" + "lives" + "smoked" | "2" + "4" + "1" + "3" + "disease" + "5" + "6" + "post" + "≈status" + "7" | "myocardial" + "infarction" + "age" + "died" + "father" + "mother" + "brother" + "name" + "patients" + "history" | "aortic" + "date" + "patient" + "valve" + "heart" + "effusion" + "mitral" + "atrial" + "failure" + "fibrillation" |
| *4* | "patient" + "history" + "smoking" + "entity" + "lives" +"date" + "alcohol" + "name" + "use" + "wife" | "history" + "hepatitis" + "c" + "abuse" + "date" + "1" + "2" + "alcohol" + "b" + "3" | "died" + "father" + "mother" + "age" + "deceased" + "mi" +"sister" + "alive" + "well" + "diabetes" | "patient" + "emergency" + "head" + "headache" + "weakness" + "loss" + "entity" +"consciousness" + "history" + "left" |
| *5* | "history" + "years" + "quit" + "ago" + "use" + "alcohol" + "smoking" + "patient" + "tobacco" + "year" | "date" + "status" + "post" + "cancer" + "right" + "surgery" + "2" + "3" + "left" + "hernia" | "unknown" + "mother" + "father" + "seizures" + "76" + "history" + "died" +"known" + "hypertension" + "patients" | "patient" + "pain" + "chest" + "emergency" + "entity" + "admission" + "breath" + "history" + "shortness" + "prior" |
| *6* | "entity" + "patient" +"lives" + "family" +"involved" + "daughter" + "lived" + "alone" +"rehabilitation" + "mother" | "years" + "ago"≈+"patient" + "date" + "name" + "entity" + "history" + "per" + "2" + "dr" | "cancer" + "history" + "family" + "mother" + "father" + "breast" + "diabetes" + "died" + "lung" + "colon" | "pain" + "patient" + "abdominal" + "history" + "vomiting" + "nausea" + "blood" + "date" + "denies" +"bowel" |
| *7* | "use" + "alcohol" + "tobacco" + "patient" + "lives" + "denies" + "drug" + "entity" +"married" + "wife" | "hypertension" + "2" + "pulmonary" + "disease" + "1" + "date" + "lung" + "predicted" +"3" + "diabetes" | "unremarkable" + "patient" + "daughter" + "mother" + "carcinoma" + "son" + "living" + "name" + "per" + "breast" | "patient" + "date" + "entity" + "ct" + "right" + "name" + "left" + "hospital" + "admitted" + "transferred" |
| *8* | "patient" + "date" + "3" + "lives" + "heavy" + "drinks" + "past" + "active" + "smoker" + "alcohol" | "2" + "hypertension" + "diabetes" + "history" + "disease" + "1" + "3" + "mellitus" + "4" + "coronary" | "date" + "cancer" + "notable" + "positive" + "entity" + "works" + "denies" + "lung" + "hx" + "parents" | "artery" + "left" + "patient" +"coronary" +"cardiac" +"catheterization" + "entity" + "date" + "disease" + "showed" |
| *9* | "history" + "patient" + "alcohol" + "abuse" + "past" + "none" + "lives" | "artery" + "date" + "coronary" + "post" +"status" + "left" + | "mother" + "father" +"ca" + "cva" + "failure" + "renal" + "brother" + "history" + | "date" + "post" +"status" + "artery" +"history" + "disease" + "coronary" + |

| | | | | |
|---|---|---|---|---|
| | + "nonsmoker" + "cocaine" + "smoking" | "disease" + "right" + "history" + "graft" | "sister" + "cabg" | "graft" + "patient" + "male" |
| ***10*** | "lives" + "patient" + "name" + "alcohol" + "entity" + "tobacco" + "son" + "history" + "alone" + "date" | "date" + "ho" + "renal" + "chronic" + "htn" + "left" + "history" + "secondary" + "multiple" +"neuropathy" | "disease" + "coronary" + "artery" + "history" + "diabetes" +"family" + "mother" + "father" + "significant" + "mellitus" | "name" +"negative" + "delivery" + "mother" + "born" + "pregnancy" + "gestation" + "date" + "b" + "infant" |

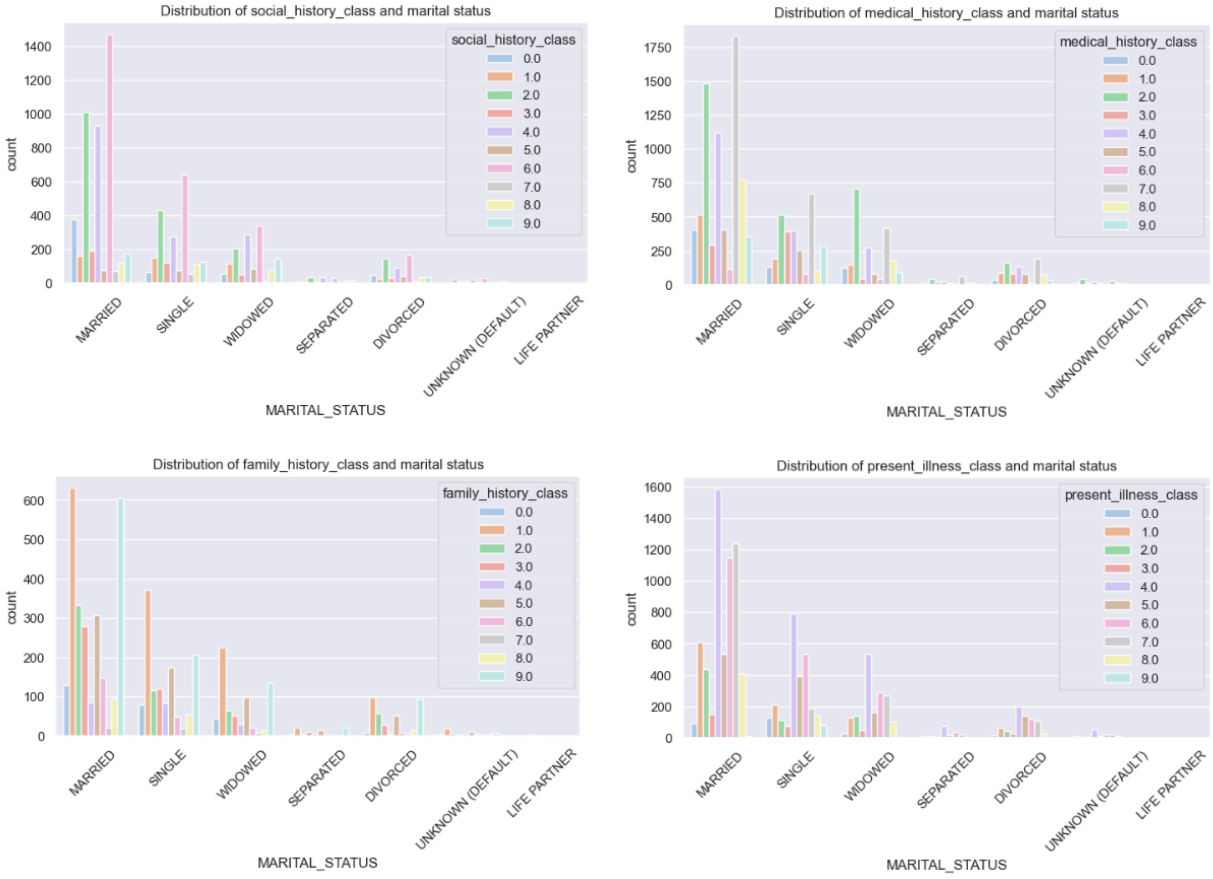
## Table 2. LDA Manually Defined Topics

| **Topic** | **Social Topics** | **Medical Topics** | **Family Topics** | **Illness Topics** |
|---|---|---|---|---|
| ***1*** | Smoking and alcohol history | Aortic stenosis - narrowing of heart valve | Family history of diabetes | Patient admitted due to accident |
| ***2*** | Patient residential conditions - nursing home | COPD - kidney failure | Family history of cardiac disease | Patient complains of dyspnea - kidney/renal disease |
| ***3*** | Smoking history | Disease - general | Family history of heart attack | Patient heart failure |
| ***4*** | Tobacco and alcohol use | Hepatitis | Family deceased | Patient emergency - fainting, headache |
| ***5*** | Tobacco and alcohol use | Previous surgery | Family history of seizures | Patient emergency - shortness of breath, heart attack |
| ***6*** | Patient family status | Patient info - general | Family history of lung and colon cancer | Patient emergency - vomiting, abdominal pain |
| ***7*** | Tobacco and alcohol use | Diabetes - Pulmonary hypertension | Family history of breast carcinoma | Patient transferred from different hospital |
| ***8*** | Severe alcohol and smoking history | Diabetes mellitus | No history of cancer | Patient procedure - cardiac catheterization |
| ***9*** | Substance abuse - cocaine | Coronary disease | Family history of renal (kidney failure) | Patient procedure - coronary artery graft |
| ***10*** | Patient lives alone | Renal insufficiency - neuropathy | Family history of diabetes (mellitus) | Patient admitted - pregnant woman in delivery |

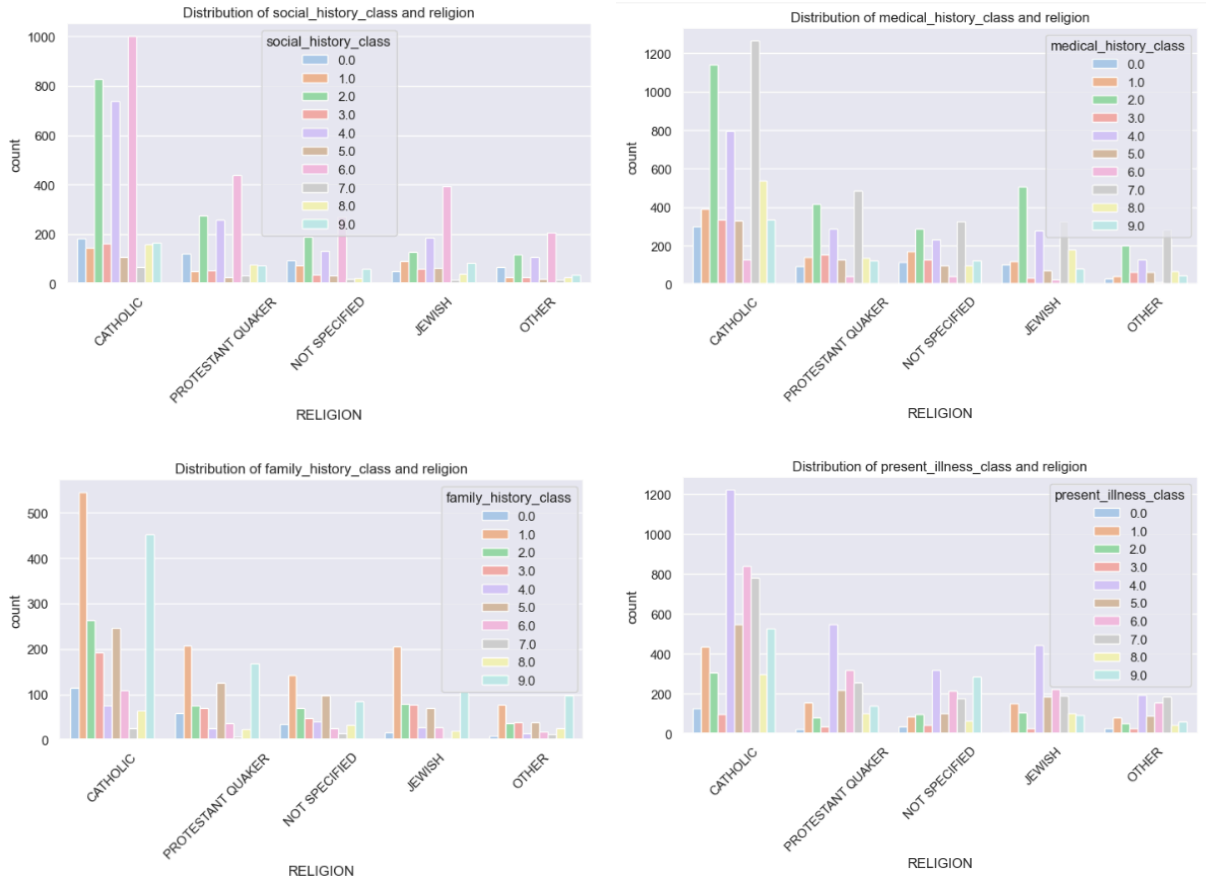
## Fig 1. Insurance Category Topic Distribution




## Fig 2. Marital Status Category Topic Distribution



## Fig 3. Religion Category Topic Distribution




## Fig 4. Ethnicity Category Topic Distribution