STAT 598

NLP based Named Entity method of Clinical Text Extraction from Medical Transcripts and K-means Clustering on Electronic Medical Records based on Patient’s chief complaint with Triage Severity Score

By: Vidyashree Ramesh

The production of healthcare data has increased significantly in recent years, with healthcare data accounting for 30% of the world's data production in 2018. Electronic medical records (EMRs) are detailed patient medical records that contain a wealth of information about demographics, medical history, lab results, and clinical notes. NLP can help extract vital information from these unstructured clinical notes, which can lead to more proactive and personalized healthcare for patients. By identifying patterns and trends in large datasets, healthcare providers can make more informed clinical decisions and provide evidence-based care.

Clinical notes are an important part of EMRs as they contain critical patient information, including triage chief complaints. However, clinical notes are unstructured and require extraction of medical information and updating to data tables for further analysis. In my research analysis I didn’t have access to clinical notes, hence I come across medical transcripts from Kaggle scraped from mtsamples.com by taking this idea I manually created 30 samples of clinical notes based on the emergency department triage assessment information and used pre-trained SciSpacy model to extract drug and disease information of ED. Once this information is extracted, it can be converted to a database in a structure format. However, clinical note sample is not large enough for clustering. I come across larger dataset of emergency department from Beth Israel Deaconess Center. Using k-means clusters found respiratory, cardiovascular, diabetics, and neurological are the common disease patient visiting ED. Finally, evaluation of k-means clustering performed using Silhouette score of 0.7 shows cluster are clearly distinguishable.

In conclusion, the use of EMRs and NLP has significant potential to revolutionize healthcare by improving the quality of care and enabling healthcare providers to make more informed clinical decisions.

Table of Contents

1. Introduction

Healthcare data production has witnessed an enormous growth in recent years, and healthcare data has become a significant proportion of the world's data production in 2018, according to Forbes. One major source of healthcare data is Electronic Medical Records (EMRs), which are comprehensive patient medical records that contain a vast amount of information, including demographics, medical history, lab results, and clinical notes. However, the clinical narratives within free-text EMRs can be challenging to access, and the information contained in them is critical for patient triage assessments, diagnosis, treatments, and outcomes. This is where Natural Language Processing (NLP) comes into play. NLP is a technique that can extract valuable information from clinical narratives within EMRs, such as symptoms, diagnoses, and treatment plans, which can then be used to improve patient care and outcomes. By analyzing the data with NLP, patterns and trends within healthcare data that might have been missed can be identified, providing insights that can inform clinical decision-making and improve population health management. Therefore, NLP is an essential tool for making sense of the vast amount of unstructured data generated by EMRs and other healthcare data sources.

Table 1. Describes the different clinical data used in healthcare

Clinical Data	Description
EHR / EMR	Detailed patient medical records that contain information about demographics, medical history, lab, imaging, and progress notes
Clinical Notes / Medical Transcripts	Notes taken by healthcare providers during patient encounters, including assessments, diagnosis, treatments, and outcomes
Discharge Summaries	Summaries of a patient’s hospital stay, including admission and discharge diagnosis, treatments, and outcomes
Radiology & Pathology reports	Reports generated by radiologists to describe imaging results, such as X-rays, CT scans and MRIs, Reports generated by pathologists to describe tissue samples and biopsy results, Reports generated by lab technicians to describe results of test performed on biological samples

Clinical Trials / Drug reviews	Detailed reports of clinical trials including study design, participant characteristics, outcome, and adverse events, various drug reviews
Medical Research Docs / guidelines	Medical research documents, guidelines, and many more unstructured data are generated daily contributing to this massive medical corpus

The motivation to choose this topic because emergency department (ED) is a vital component of healthcare that provides urgent medical care to patients in need. However, the ED operates in a resource-limited environment, where human attention is the most important resource. Patients can present with a wide range of conditions, making it essential to prioritize patients based on the severity of their condition.

In the ED, syndromic surveillance can help detect disease outbreaks by tracking symptoms such as fever, cough, and diarrhea. By analyzing the clinical notes by healthcare professionals can identify trends and patterns that may indicate the presence of a disease outbreak, allowing for timely public health interventions.

When patient visited emergency department, a triage assessment is done based on patient temperature, oxygen saturation, blood pressure and provided acuity (severity) score in the range of high to low severity in the range of 1 – 4 and diagnosis and treatment outcomes is written in the form of clinical notes by physicians and nurse. All these key information is unstructured in the form of clinical notes. By extracting key terms like drug, disease and diagnosis from clinical notes helps the healthcare to understand trends or any key information of outbreaks.

My goal of this project to learn and apply NLP techniques to clinical notes. I come across a medical transcripts dataset from Kaggle. A Kaggle is an online community of sharing and using the dataset. The medical transcript data is scrapped from mtsamples. A mtsamples is designed to give you access to a big collection of transcribed medical reports. These samples can be used by learning, as well as working medical transcriptionists for their daily transcription needs for research purposes. However, the transcripts don’t have information of emergency department. Hence, I created 30 samples of data based of emergency department disease, drug and diagnosis information.

For clinical note a pre-trained SciSpacy model is used to extract drug and disease information using name entity method which is trained by biomedical data with drug and disease information. To perform clustering analysis, I didn’t have enough data and wanted a larger dataset. Hence I found a structured dataset of emergency department of Beth Israel Deaconess Center Boston to know common disease patient’s visiting emergency department

2. Literature Review

Natural language processing (NLP) is a field of computer-based methods aiming to facilitate the processing of the human (natural) language (Spyns, 1996). NLP is an invaluable tool to extract and process information residing in the natural language format into a more structured layout for research (Thomas, 2014). Using pre-defined computer algorithms, NLP systems can be automated to parse textual information and search for key words and phrases to extract pertinent clinical data. NLP systems can rapidly analyze large clinical datasets and deliver real-time outcomes for various research applications and healthcare outcomes analyses.

Another reason clinical text needs different NLP models is that it contains a large amount of data spread across different sources, such as EHRs, clinical notes, and radiology reports, which need to be integrated. This requires models that can process and understand the text and link and integrate the data across different sources and establish clinically acceptable relationships (Maiti, 2023)

Lastly, clinical text often contains sensitive patient information and needs to be protected by strict regulations such as HIPAA. NLP models used to process clinical text must be able to identify and protect sensitive patient information while still providing useful insights.

The textual data within medicine requires a specialized Natural Language Processing (NLP) system capable of extracting medical information from various sources such as clinical texts and other medical documents (Thomas, 2014)

We need to use specific NLP libraries in biomedical domain like spacy and Scispacy. It is an open-source NLP library that provides out-of-the-box models for various domains, including the medical domain. A specialized version of spaCy that is trained specifically on scientific and biomedical text, which makes it ideal for processing medical text.

There are unique challenges with processing the text that is found in clinical notes as opposed to other types of text or biomedical literature typically processed using NLP. Clinical text written by medical providers includes telegraphic language and semi-structured text. Nursing notes and surveys are replete with check boxes and structured question and answer templates. These structures are the source of most causes for false positive errors (Guy, 2016)

NLP pipeline utilizing SciSpacy to perform custom Named Entity Recognition on clinical texts. The outcome will be extracting information regarding diseases, drugs, and drug doses from clinical text, which can then be utilized in various NLP downstream applications. Use of ScispaCy pre-trained NER model en_ner_bc5cdr_md-0.5.1 to extract disease and drugs. Drugs are extracted as Chemicals. en_ner_bc5cdr_md-0.5.1 is a spaCy model for named entity recognition (NER) in the biomedical domain. The “bc5cdr” refers to the BC5CDR corpus, a biomedical text corpus used to train the model. The “md” in the name refers to the biomedical domain. The “0.5.1” in the name refers to the version of the model (Maiti, 2023)

The rule-based matching resembles the usage of regular expressions, but spaCy provides additional capabilities. Using the tokens and relationships within a document enables you to identify patterns that include entities with the help of NER models. The goal is to locate drug names and their dosages from the text, which could help detect medication errors by comparing them with standards and guidelines.

A major goal of Natural Language Processing (NLP) in the medical domain is the automatic extraction and encoding of data stored in free text patient records. This extracted data can then be utilized by Information Technology systems to perform syndromic surveillance. In particular, the chief complaint a short string that describes a patient’s symptoms before even a preliminary diagnosis has been made has come to be a vital resource for syndromic surveillance in the North American context. Despite the acknowledged importance of chief complaints for syndromic surveillance, considerable variation exists in how they are used in practice by system builders, both in terms of algorithms chosen to map chief complaint strings to syndromes, and the syndrome definitions themselves (Conway, 2013)

Free-text triage chief complaints are the earliest clinical data available on most hospital information systems. A TCC is short phrase entered by a triage nurse describing the reason for a patient’s visit to an emergency department. Some examples of common TCCs include ‘‘cough,’’ ‘‘n/v/d,’’ and ‘‘luq abd pain.’’ The purpose of a TCC is to describe a patient’s condition in as short a space as possible, therefore, TCCs contain abbreviations and punctuation that can often confuse even experienced emergency room personnel. Researchers are investigating auto- mated, knowledgebased methods for expanding TCC strings from abbreviated form into a more complete form (Chapman, 2005)

Classifying free-text phrases into the syndromic categories raised questions about which chief complaints belong in which syndromes. As we developed the classification method described below, we generated heuristics for classification of specific complaints into syndromes.

These heuristics were sometimes easily justifiable and other times quite arbitrary. Among other examples, decisions about appropriate classifications arose when a chief complaint represented a symptom that is not a very specific indicator of a public health outbreak or when a complaint could be due to more than one (Chapman, 2005)

Gastrointestinal includes pain or cramps anywhere in the abdomen, nausea vomiting, diarrhea and abdominal distension or swelling.

Constitutional is made up of non-localized, systemic problems including fever, chills, body aches, flu symptoms (viral syndrome), weakness, fatigue, anorexia, malaise, lethargy, sweating (diaphoresis), light headedness, faintness and fussiness. Shaking (not chills) is not constitutional but is other.

Respiratory includes the nose (coryza) and throat (pharyngitis), as well as the lungs. Examples of respiratory include congestion, sore throat, tonsillitis, sinusitis, cold symptoms, bronchitis, cough, shortness of breath, asthma, chronic obstructive pulmonary disease (COPD) and pneumonia. If both cold symptoms and flu symptoms are present, the syndrome is respiratory.

Rash includes any description of a rash, such as macular, papular, vesicular, petechial, purpuric or hives. Ulcerations are not normally considered a rash unless consistent with cutaneous anthrax (an ulcer with a black eschar).

Hemorrhagic is bleeding from any site, e.g., vomiting blood (hematemesis), nose bleed (epistaxis), hematuria, gastrointestinal bleeding (site unspecified), rectal bleeding and vaginal bleeding. Bleeding from a site for which we have a syndrome should be classified as hemorrhagic and as the relevant syndrome (e.g., Hematochesia is gastrointestinal and hemorrhagic; hemoptysis is respiratory and hemorrhagic).

Botulinic includes ocular abnormalities (diplopia, blurred vision, photophobia), difficulty speaking (dysphonia, dysarthria, slurred speech) and difficulty swallowing (dysphagia).

Neurological covers non-psychiatric complaints which relate to brain function. Included are headache, head pain, migraine, facial pain or numbness, seizure, tremor, convulsion, loss of consciousness, syncope, fainting, ataxia, confusion, disorientation, altered mental status, vertigo, concussion, meningitis, stiff neck, tingling and numbness. (Dizziness is constitutional and neurological.)

Other is a pain or process in a system or area RODS is not monitoring. For example, flank pain most likely arises from the genitourinary system, which RODS does not model, and would be considered other. Chest pain with no mention of the source of the pain is considered other (e.g., chest pain (other) versus pleuritic chest pain (respiratory). Earache or ear pain is other. Trauma is other.

COCO (Complaint Coder) uses a naive bayesian classifier to sort chief complaints into one of eight syndromic categories (constitutional, respiratory, gastrointestinal, hemorrhagic, botulism, neurological, respiratory and other). The probability of each word belonging to a syndromic category is learned from a manually created data set. According to (Olszewsk, 2003)i (a direct precursor of the COCO system) chief complaints were converted to lower case and all punctuation removed. Additionally, compared chief complaints to ICD-9 codes for determining syndromic categories and found that chief complaints yielded better results (using a data set of 28,990 reports) (Olszewsk, 2003).

COCO can be used as an off-the-shelf product (that is, using the syndromic probabilities and syndromic categories developed by RODS). Alternatively, bespoke syndromic categories and training data can be developed for new sites or public health contexts. Note that COCO has been used as a component within other surveillance systems (Conway,2013)

Clustering is a data mining exploratory method to form object groups and identify structures of unlabeled data set. It is one of the most famous unsupervised learning methods. These methods form groups of objects or individuals with maximizing their within-group similarity and their between- group dissimilarity. Time-series clustering is an application of clustering to non-static data, i.e. data depending on time. Most of clustering algorithms used for times series are algorithms derived from those used in static data like partitioning methods or hierarchical methods (Gregory, 2018). The main step of a clustering algorithm is to define the distance between objects or time-series to quantify their similarity during the algorithm.

K-means: Given the set of n time-series {xi |i = 1, ..., n}, with xi = (xi1, ..., xip) each of length p, and {vk|k = 1, ..., K } center clusters with K, number of clusters, known, we assign each time series xi to one, and only one cluster vk. For short, after initialization of center clusters, we assign each time-series to the cluster whose its center cluster is the nearest with respect to the Euclidean distance. Next, center clusters are updated, being average time-series of all time-series' cluster. Then, the processes is repeated until stabilization (Warren, 2005). In our scenario, n refers to the number of ICD-9 or ICD-10 codes in our data and p to the acuity score.

Each time series, which corresponds to either one ICD-10 code or the influenza test series, is standardized, by subtracting from it its mean and by dividing by the classical L2- norm. This standardization is necessary because we need to make clusters of diagnosis having similarities in terms of variations and not in terms of occurrence (Gavrilov, 2000). The main disadvantage of K-means is that the number of clusters must be specified. As we do not know the numbers of clusters K, this is chosen with criteria such as elbow variance or silhouette score.

For each number of clusters possible (from 1 to n), after computing the K-means algorithm, the elbow variance and the silhouette score are calculated. The elbow method consists in plotting the ratio of variance explained (variance between center clusters divided by total variance) in function of the number of clusters chosen. The location of a bend (or the beginning of an asymptote) only give an indication of the appropriate number of clusters K (D.J Ketchen, 1996). The silhouette score is calculated using the mean intra-cluster distance (ai) and the mean nearest-cluster distance (bi) for the time series.

3. Methods

The project framework is divided in two parts such as clinical note extraction using NLP named entity method and K-means clustering using Elbow Method.

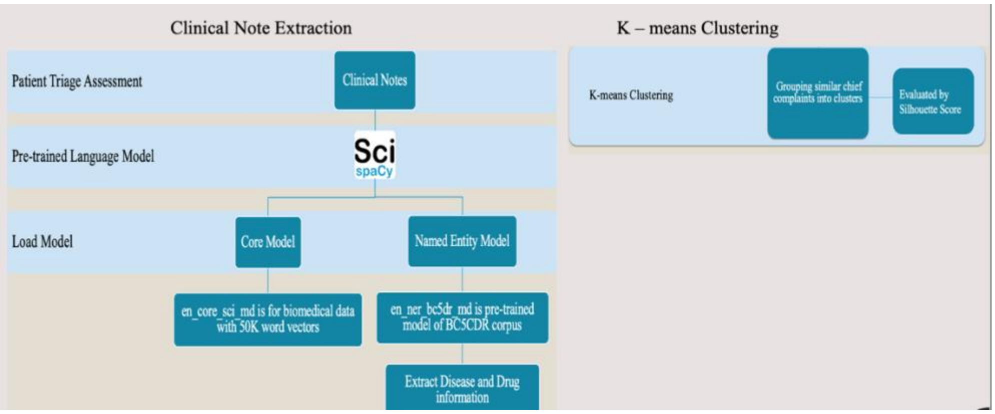

Clinical Note Extraction
K – means Clustering
Patient Triage Assessment
Clinical Notes
Pre-trained Language Model
Sci
spaCy
Load Model
Core Model
Named Entity Model
en_core_sci_md is for biomedical data with 50K word vectors
en_ner_bc5dr_md is pre-trained model of BC5CDR corpus
Extract Disease and Drug information
K-means Clustering
Grouping similar chief complaints into clusters
Evaluated by Silhouette Score

Figure 1. A flow chart that depicts steps for clinical note extraction and K-means clustering

3.1.1. Objective 1 - Clinical Note Extraction -Steps to extract disease and drug from Medical Transcripts-NLP Name Entity Method

1. Clinical Notes or Medical Transcript refers to comprehensive information about an individual's healthcare history, including diagnoses, treatments, lab results, imaging studies, and other relevant health information

2. The Test Corpus used for this project was collected from mtsamples.csv, which was scraped from mtsample

3. Using this similar mtsample data, 30 free text information was manually created for further analysis.

4. ScispaCy is a pre-trained name entity model that can be used to extract valuable information from biomedical text

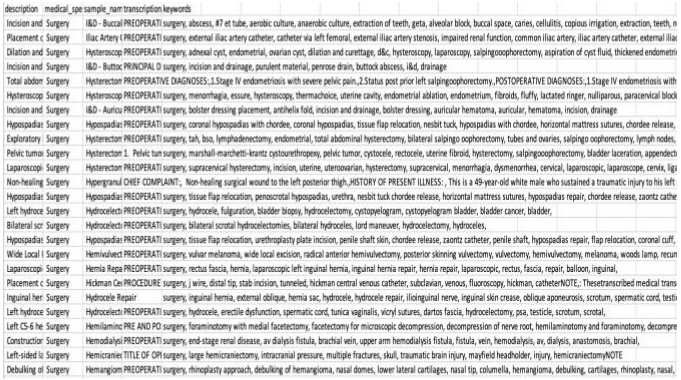
5. To analyze biomedical data, it is essential to load specific models that are designed for this purpose.

6. The en_core_sci_sm and en_core_sci_md are core models that are specifically created for biomedical data analysis

7. Additionally, the en_ner_bc5cdr_md model can be used to extract disease and drug-related information from clinical notes.

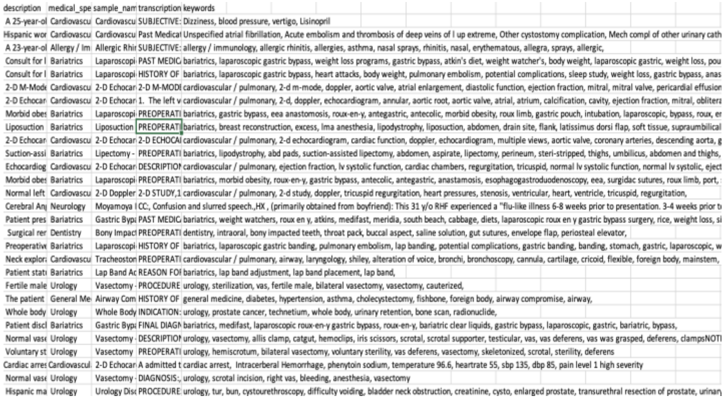
7. Additionally, the en_ner_bc5cdr_md model can be used to extract disease and drug-related information from clinical notes.

Clinical Notes or Medical Transcripts refer to comprehensive information on an individual's healthcare history, including diagnoses, treatments, lab results, imaging studies, and other relevant health information. However, obtaining clinical data can be difficult due to HIPAA/credential access from the hospital. Therefore, I come across a medical transcript's dataset on Kaggle, an online community that provides datasets for building AI models. As Figure 2 shows the medical transcripts used in Kaggle were obtained from MTsamples.com, which provides transcribed medical transcription sample reports and examples for reference purposes only.



description	medical_spe	sample_nam	transcription	keywords
Incision and	Surgery	I&D - Buccal	PREOPERATI	surgery, abscess, #7 et tube, aerobic culture, anaerobic culture, extraction of teeth, geta, alveolar block, buccal space, caries, cellulitis, copious irrigation, extraction, teeth, n
Placement c	Surgery	Iliac Artery (	PREOPERATI	surgery, external iliac artery catheter, catheter via left femoral, external iliac artery stenosis, impaired renal function, common iliac artery, iliac artery catheter, external iliac
Dilation and	Surgery	Hysteroscop	PREOPERATI	surgery, adnexal cyst, endometrial, ovarian cyst, dilation and curettage, d&c, hysteroscopy, laparoscopy, salpingooophorectomy, aspiration of cyst fluid, thickened endometri
Incision and	Surgery	I&D - Buttoc	PRINCIPAL D	surgery, incision and drainage, purulent material, penrose drain, buttock abscess, i&d, drainage
Total abdom	Surgery	Hysterectom	PREOPERATIVE DIAGNOSES:,1.Stage IV endometriosis with severe pelvic pain.,2.Status post prior left salpingoophorectomy.,POSTOPERATIVE DIAGNOSES:,1.Stage IV endometriosis with	
Hysteroscop	Surgery	Hysteroscop	PREOPERATI	surgery, menorrhagia, essure, hysteroscopy, thermachoice, uterine cavity, endometrial ablation, endometrium, fibroids, fluffy, lactated ringer, nulliparous, paracervical block
Incision and	Surgery	I&D - Auricu	PREOPERATI	surgery, bolster dressing placement, antihelix fold, incision and drainage, bolster dressing, auricular hematoma, auricular, hematoma, incision, drainage
Hypospadias	Surgery	Hypospadias	PREOPERATI	surgery, coronal hypospadias with chordee, coronal hypospadias, tissue flap relocation, nesbit tuck, hypospadias with chordee, horizontal mattress sutures, chordee release,
Exploratory	Surgery	Hysterectom	PREOPERATI	surgery, tah, bso, lymphadenectomy, endometrial, total abdominal hysterectomy, bilateral salpingo oophorectomy, tubes and ovaries, salpingo oophorectomy, lymph nodes,
Pelvic tumor	Surgery	Hysterectom	1. Pelvic tun	surgery, marshall-marchetti-krantz cystourethropexy, pelvic tumor, cystocele, rectocele, uterine fibroid, hysterectomy, salpingooophorectomy, bladder laceration, appendect
Laparoscopi	Surgery	Hysterectom	PREOPERATI	surgery, supracervical hysterectomy, incision, uterine, uteroovarian, hysterectomy, supracervical, menorrhagia, dysmenorrhea, cervical, laparoscopic, laparoscope, cervix, liga
Non-healing	Surgery	Hypergranul	CHIEF COMPLAINT:, Non-healing surgical wound to the left posterior thigh.,HISTORY OF PRESENT ILLNESS: , This is a 49-year-old white male who sustained a traumatic injury to his left	
Hypospadias	Surgery	Hypospadias	PREOPERATI	surgery, tissue flap relocation, penoscrotal hypospadias, urethra, nesbit tuck chordee release, horizontal mattress sutures, hypospadias repair, chordee release, zaontz cathe
Left hydroce	Surgery	Hydrocelect	PREOPERATI	surgery, hydrocele, fulguration, bladder biopsy, hydrocelectomy, cystopyelogram, cystopyelogram bladder, bladder cancer, bladder,
Bilateral scr	Surgery	Hydrocelect	PREOPERATI	surgery, bilateral scrotal hydrocelectomies, bilateral hydroceles, lord maneuver, hydrocelectomy, hydroceles,
Hypospadias	Surgery	Hypospadias	PREOPERATI	surgery, tissue flap relocation, urethroplasty plate incision, penile shaft skin, chordee release, zaontz catheter, penile shaft, hypospadias repair, flap relocation, coronal cuff,
Wide Local I	Surgery	Hemivulvect	PREOPERATI	surgery, vulvar melanoma, wide local excision, radical anterior hemivulvectomy, posterior skinning vulvectomy, vulvectomy, hemivulvectomy, melanoma, woods lamp, recun
Laparoscopi	Surgery	Hernia Repa	PREOPERATI	surgery, rectus fascia, hernia, laparoscopic left inguinal hernia, inguinal hernia repair, hernia repair, laparoscopic, rectus, fascia, repair, balloon, inguinal,
Placement c	Surgery	Hickman Cei	PROCEDURE	surgery, j wire, distal tip, stab incision, tunneled, hickman central venous catheter, subclavian, venous, fluoroscopy, hickman, catheterNOTE,: Thesetranscribed medical trans
Inguinal her	Surgery	Hydrocele Repair		surgery, inguinal hernia, external oblique, hernia sac, hydrocele, hydrocele repair, ilioinguinal nerve, inguinal skin crease, oblique aponeurosis, scrotum, spermatic cord, testi
Left hydroce	Surgery	Hydrocelect	PREOPERATI	surgery, hydrocele, erectile dysfunction, spermatic cord, tunica vaginalis, vicryl sutures, dartos fascia, hydrocelectomy, psa, testicle, scrotum, scrotal,
Left C5-6 he	Surgery	Hemilamino	PRE AND PO	surgery, foraminotomy with medial facetectomy, facetectomy for microscopic decompression, decompression of nerve root, hemilaminotomy and foraminotomy, decompre
Constructior	Surgery	Hemodialysi	PREOPERATI	surgery, end-stage renal disease, av dialysis fistula, brachial vein, upper arm hemodialysis fistula, fistula, vein, hemodialysis, av, dialysis, anastomosis, brachial,
Left-sided la	Surgery	Hemicraniec	TITLE OF OPI	surgery, large hemicraniectomy, intracranial pressure, multiple fractures, skull, traumatic brain injury, mayfield headholder, injury, hemicraniectomyNOTE
Debulking ol	Surgery	Hemangiom	PREOPERATI	surgery, rhinoplasty approach, debulking of hemangioma, nasal domes, lower lateral cartilages, nasal tip, columella, hemangioma, debulking, cartilages, rhinoplasty, nasal,

Figure 2. Sample medical transcripts from mtsample dataset



description	medical_spe	sample_nam	transcription	keywords
A 25-year-ol	Cardiovascu	Cardiovascu	SUBJECTIVE:	Dizziness, blood pressure, vertigo, Lisinopril
Hispanic wor	Cardiovascu	Cardiovascu	Past Medical	Unspecified atrial fibrillation, Acute embolism and thrombosis of deep veins of l up extreme, Other cystostomy complication, Mech compl of other urinary cath
A 23-year-ol	Allergy / Im	Allergic Rhir	SUBJECTIVE:	allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal, erythematous, allegra, sprays, allergic,
Consult for l	Bariatrics	Laparoscopi	PAST MEDIC	bariatrics, laparoscopic gastric bypass, weight loss programs, gastric bypass, atkin's diet, weight watcher's, body weight, laparoscopic gastric, weight loss, pou
Consult for l	Bariatrics	Laparoscopi	HISTORY OF	bariatrics, laparoscopic gastric bypass, heart attacks, body weight, pulmonary embolism, potential complications, sleep study, weight loss, gastric bypass, anas
2-D M-Mode	Cardiovascu	2-D Echocar	2-D M-MODI	cardiovascular / pulmonary, 2-d m-mode, doppler, aortic valve, atrial enlargement, diastolic function, ejection fraction, mitral, mitral valve, pericardial effusion
2-D Echocar	Cardiovascu	2-D Echocar	1. The left v	cardiovascular / pulmonary, 2-d, doppler, echocardiogram, annular, aortic root, aortic valve, atrial, atrium, calcification, cavity, ejection fraction, mitral, oblitera
Morbid obes	Bariatrics	Laparoscopi	PREOPERATI	bariatrics, gastric bypass, eea anastomosis, roux-en-y, antegastric, antecolic, morbid obesity, roux limb, gastric pouch, intubation, laparoscopic, bypass, roux, en
Liposuction	Bariatrics	Liposuction	PREOPERATI	bariatrics, breast reconstruction, excess, lma anesthesia, lipodystrophy, liposuction, abdomen, drain site, flank, latissimus dorsi flap, soft tissue, supraumbilical
2-D Echocar	Cardiovascu	2-D Echocar	2-D ECHOCAI	cardiovascular / pulmonary, 2-d echocardiogram, cardiac function, doppler, echocardiogram, multiple views, aortic valve, coronary arteries, descending aorta, g
Suction-assi	Bariatrics	Lipectomy -	PREOPERATI	bariatrics, lipodystrophy, abd pads, suction-assisted lipectomy, abdomen, aspirate, lipectomy, perineum, steri-stripped, thighs, umbilicus, abdomen and thighs,
Echocardiog	Cardiovascu	2-D Echocar	DESCRIPTION	cardiovascular / pulmonary, ejection fraction, lv systolic function, cardiac chambers, regurgitation, tricuspid, normal lv systolic function, normal lv systolic, eject
Morbid obes	Bariatrics	Laparoscopi	PREOPERATI	bariatrics, morbid obesity, roux-en-y, gastric bypass, antecolic, antegastric, anastamosis, esophagogastroduodenoscopy, eea, surgidac sutures, roux limb, port,
Normal left	Cardiovascu	2-D Doppler	2-D STUDY,1	cardiovascular / pulmonary, 2-d study, doppler, tricuspid regurgitation, heart pressures, stenosis, ventricular, heart, ventricle, tricuspid, regurgitation,
Cerebral Ang	Neurology	Moyamoya I	CC:, Confusion and slurred speech.,HX , (primarily obtained from boyfriend): This 31 y/o RHF experienced a "flu-like illness 6-8 weeks prior to presentation. 3-4 weeks prior t	
Patient pres	Bariatrics	Gastric Bypa	PAST MEDIC	bariatrics, weight watchers, roux en y, atkins, medifast, meridia, south beach, cabbage, diets, laparoscopic roux en y gastric bypass surgery, rice, weight loss, si
Surgical rer	Dentistry	Bony Impact	PREOPERATI	dentistry, intraoral, bony impacted teeth, throat pack, buccal aspect, saline solution, gut sutures, envelope flap, periosteal elevator,
Preoperative	Bariatrics	Laparoscopi	HISTORY OF	bariatrics, laparoscopic gastric banding, pulmonary embolism, lap banding, potential complications, gastric banding, banding, stomach, gastric, laparoscopic, w
Neck explor	Cardiovascu	Tracheoston	PREOPERATI	cardiovascular / pulmonary, airway, laryngology, shiley, alteration of voice, bronchi, bronchoscopy, cannula, cartilage, cricoid, flexible, foreign body, mainstem,
Patient stati	Bariatrics	Lap Band Ac	REASON FOI	bariatrics, lap band adjustment, lap band placement, lap band,
Fertile male	Urology	Vasectomy -	PROCEDURE	urology, sterilization, vas, fertile male, bilateral vasectomy, vasectomy, cauterized,
The patient	General Me	Airway Com	HISTORY OF	general medicine, diabetes, hypertension, asthma, cholecystectomy, fishbone, foreign body, airway compromise, airway,
Whole body	Urology	Whole Body	INDICATION:	urology, prostate cancer, technetium, whole body, urinary retention, bone scan, radionuclide,
Patient discl	Bariatrics	Gastric Bypa	FINAL DIAGN	bariatrics, medifast, laparoscopic roux-en-y gastric bypass, roux-en-y, bariatric clear liquids, gastric bypass, laparoscopic, gastric, bariatric, bypass,
Normal vase	Urology	Vasectomy -	DESCRIPTION	urology, vasectomy, allis clamp, catgut, hemoclips, iris scissors, scrotal, scrotal supporter, testicular, vas, vas deferens, vas was grasped, deferens, clampsNOTI
Voluntary st	Urology	Vasectomy	PREOPERATI	urology, hemiscrotum, bilateral vasectomy, voluntary sterility, vas deferens, vasectomy, skeletonized, scrotal, sterility, deferens
Cardiac arres	Cardiovascul	2-D Echocar	A admitted t	cardiac arrest, Intracerberal Hemorrhage, phenytoin sodium, temperature 96.6, heartrate 55, sbp 135, dbp 85, pain level 1 high severity
Normal vase	Urology	Vasectomy -	DIAGNOSIS:,	urology, scrotal incision, right vas, bleeding, anesthesia, vasectomy
Hispanic ma	Urology	Urology Disc	PROCEDURE	urology, tur, bun, cystourethroscopy, difficulty voiding, bladder neck obstruction, creatinine, cysto, enlarged prostate, transurethral resection of prostate, urinary

Figure 3. Manually Created Medical Transcripts for the project– Dataset's

My research goal was to extract disease and drug-related information from emergency department clinical notes. In Figure 3 I manually created 30 dataset samples based on patient triage assessment clinical notes from the emergency department at Beth Israel Deaconess Medical Center (BIDMC), Boston. I utilized ScispaCy, a pre-trained named entity model designed for extracting valuable information from biomedical text. To analyze biomedical data effectively, it is essential to load specific models created for this purpose, such as the en_core_sci_sm and en_core_sci_md core models. In addition, I used the en_ner_bc5cdr_md model to extract disease and drug-related information from clinical notes, which I updated in the "chief complaint" and "medication name" columns, respectively. Finally, I added the extracted information to a data table for further analysis.

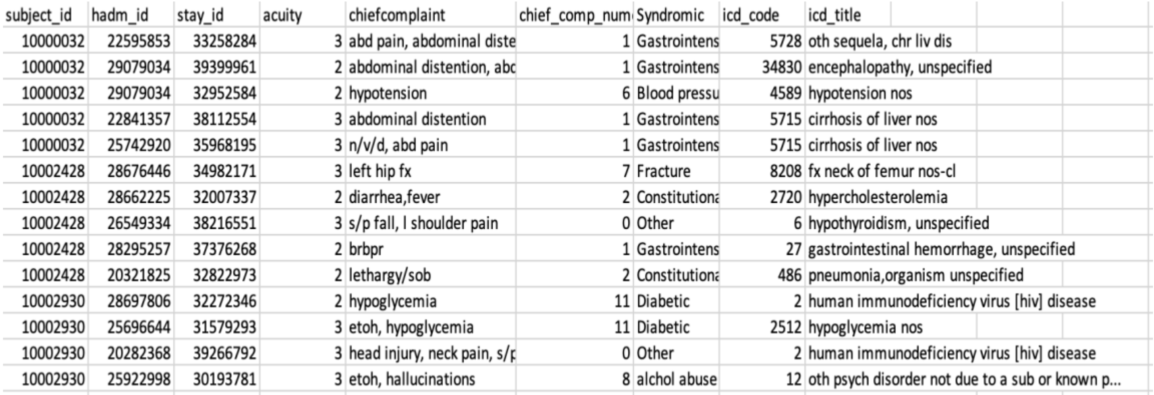
3.1.2. K – means Clustering Dataset

The 30 clinical note extraction samples on drug and disease information were insufficient for clustering analysis. As a result, I came across the MIMIC-ED dataset, which is a large, freely available database of emergency department (ED) admissions at the Beth Israel Deaconess Medical Center between 2011 and 2019, containing 425,000 ED stays. For research purposes, a subset of 100 patient EMR information was provided. The dataset comprised 5874 observations, 33 variables, and repeated observations of the same patient.

Before joining the tables, duplicate observations were removed from all the tables, including EDSTAY, DIAGNOSIS, and TRIAGE, which were joined using the stay_id column. The final dataset contained 8 variables, including subject_id, hadm_id, stay_id, chief_complaint, medication name, icd_code, and icd_title. The icd_title column was converted to lower case, and missing information was removed from the table.

Since my research goal was to group similar chief complaints into clusters, I opted for k-means clustering, as clustering only works for numeric variables. To facilitate this analysis, I manually added syndromes to my dataset based on a literature review of "Classifying free-text triage chief complaints into syndromic categories with natural language processing" by Chapman (2004). Later, I created a chief complaint numeric column and assigned respective numbers to each syndrome (e.g., 1 for gastrointestinal).

The data sample includes the MIMIC-ED dataset columns, as well as additional variables such as chief_comp_numeric and syndromic. The use of chief_comp_numeric suggests that the chief complaint data may have been converted into a numerical format for analysis purposes. These added variables aim to facilitate the analysis of the MIMIC-ED dataset columns and help answer the research question.



subject_id	hadm_id	stay_id	acuity	chiefcomplaint	chief_comp_num	Syndromic	icd_code	icd_title
10000032	22595853	33258284	3	abd pain, abdominal diste	1	Gastrointens	5728	oth sequela, chr liv dis
10000032	29079034	39399961	2	abdominal distention, abc	1	Gastrointens	34830	encephalopathy, unspecified
10000032	29079034	32952584	2	hypotension	6	Blood pressu	4589	hypotension nos
10000032	22841357	38112554	3	abdominal distention	1	Gastrointens	5715	cirrhosis of liver nos
10000032	25742920	35968195	3	n/v/d, abd pain	1	Gastrointens	5715	cirrhosis of liver nos
10002428	28676446	34982171	3	left hip fx	7	Fracture	8208	fx neck of femur nos-cl
10002428	28662225	32007337	2	diarrhea,fever	2	Constitutiona	2720	hypercholesterolemia
10002428	26549334	38216551	3	s/p fall, l shoulder pain	0	Other	6	hypothyroidism, unspecified
10002428	28295257	37376268	2	brbpr	1	Gastrointens	27	gastrointestinal hemorrhage, unspecified
10002428	20321825	32822973	2	lethargy/sob	2	Constitutiona	486	pneumonia,organism unspecified
10002930	28697806	32272346	2	hypoglycemia	11	Diabetic	2	human immunodeficiency virus [hiv] disease
10002930	25696644	31579293	3	etoh, hypoglycemia	11	Diabetic	2512	hypoglycemia nos
10002930	20282368	39266792	3	head injury, neck pain, s/p	0	Other	2	human immunodeficiency virus [hiv] disease
10002930	25922998	30193781	3	etoh, hallucinations	8	alchol abuse	12	oth psych disorder not due to a sub or known p...

Figure 4. Data table after merging and adding syndrome and chief complaint numeric for k-means clustering.

To address research question, I wanted to explore the relationship between the chief complaint and acuity level in emergency department patients. To do this, use of K-means clustering is a powerful tool that can be used to group large datasets into distinct clusters based on similar features or attributes. In the context of emergency medicine, k-means clustering can be applied to the chief complaint data of patients visiting the emergency department (ED) to identify the most common diseases or syndromes that patients present with. For example, if we have data on the chief complaints of patients visiting the ED, such as "chest pain," "shortness of breath," "dizziness," and "fatigue," we can use k-means clustering to group these complaints into clusters based on similarity.

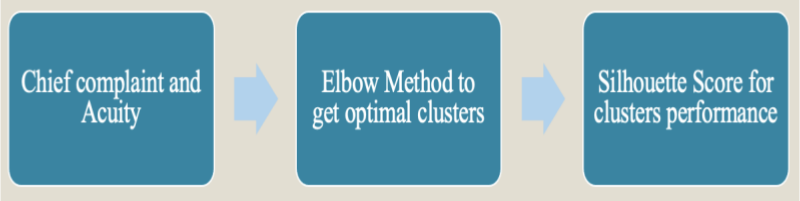

Chief complaint and Acuity
Elbow Method to get optimal clusters
Silhouette Score for clusters performance

Figure 5. Steps to perform k-means clustering.

To determine the optimal number of clusters, I used the elbow method. This involved calculating the sum of squared errors for different numbers of clusters and plotting them on a graph. The elbow point on the graph represents the point where the addition of another cluster doesn't significantly decrease the sum of squared errors. In my case, the graph showed that two clusters were the optimal number for the 125 data frames I was analyzing.

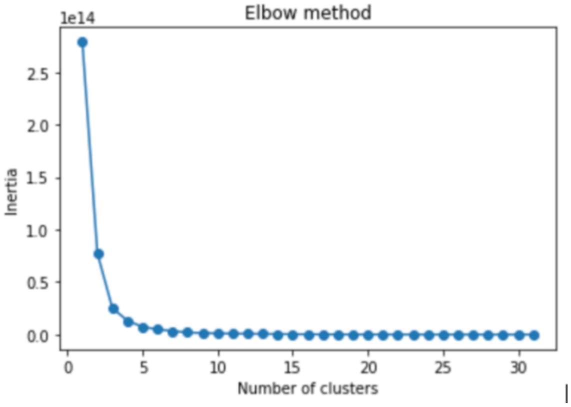


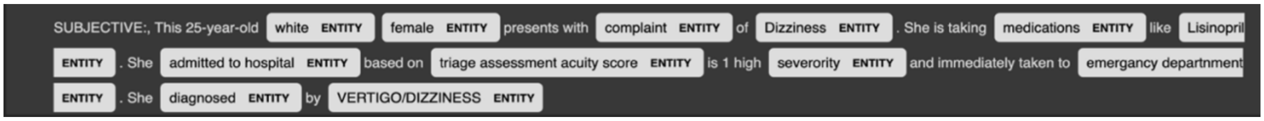
Figure 6. Graphical representation of elbow method to get optimal clusters

4. Results

4.1. NLP Chief complaint extraction

Based on the implementation of the NLP pipeline using SciSpacy, able to successfully perform custom Named Entity Recognition on clinical texts.

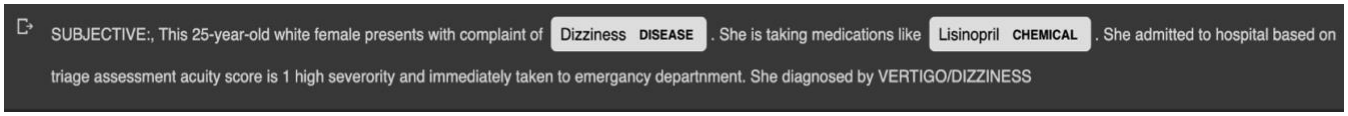
The core model en_core_sci_md is specifically designed for biomedical vocabulary and contains over 50K words. The model extracted medical terms like dizziness and lisinopril along with generic information such as complaint, severity terms as entity



SUBJECTIVE:, This 25-year-old white ENTITY female ENTITY presents with complaint ENTITY of Dizziness ENTITY . She is taking medications ENTITY like Lisinopril ENTITY . She admitted to hospital ENTITY based on triage assessment acuity score ENTITY is 1 high severority ENTITY and immediately taken to emergancy departnment ENTITY . She diagnosed ENTITY by VERTIGO/DIZZINESS ENTITY

Figure 7. en_core_sci_md : Extracted medical terms like disease and drug information along with generic terms like complaint as entity

The en_ner_bc5cdr_md model is a pre-trained model specifically designed for extracting disease and medication information from text. This model has been trained on a large corpus of biomedical literature and is capable of accurately identifying disease and medication entities in clinical texts.



SUBJECTIVE:, This 25-year-old white female presents with complaint of Dizziness DISEASE . She is taking medications like Lisinopril CHEMICAL . She admitted to hospital based on triage assessment acuity score is 1 high severority and immediately taken to emergancy department. She diagnosed by VERTIGO/DIZZINESS

Figure 8. en_ner_bc5cdr_md – Name entity model extracted only disease and drug information.

4.2. Visualizing each variable in the dataset

4.2.1. Subject_id

The subject_id column in the dataset represents a unique patient ID that remains the same regardless of hospital stay. The data shows extreme values in subject_id, with patients 10014, 10015, and 10040 being admitted to the hospital more than 10 times.

The graph provides a clear visualization of these extreme values, allowing for easy identification and analysis. On average, the data shows that patients are admitted to the hospital three times, indicating a recurring need for medical care among patients.

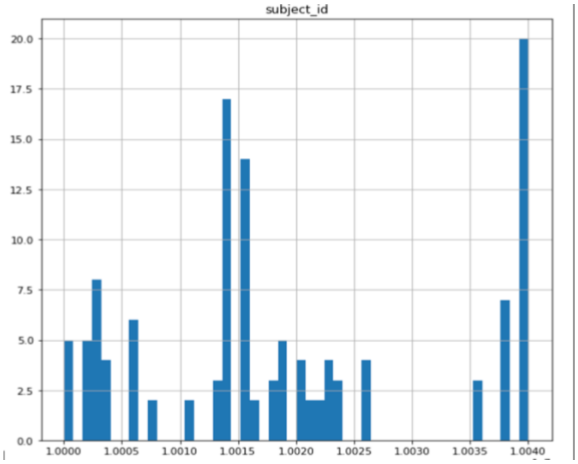


Figure 9. Histogram of subject_id showcasing extreme value observations.

4.2.2. Hadm_id

The bar graph shows that some patients have been admitted to the hospital more times than others. On average, each patient has been admitted three times.

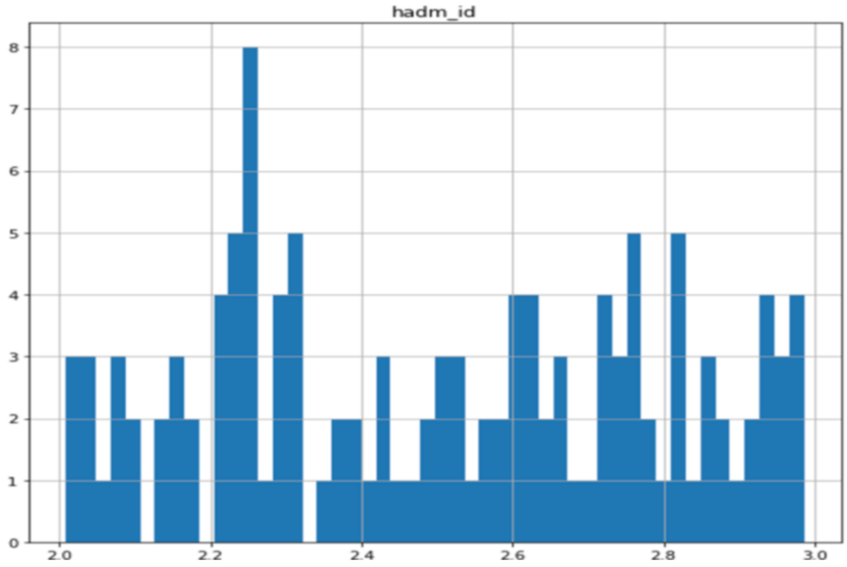


Figure 10. Histogram of hadm_id showcasing only patients transferred to hospital after ED.

4.2.3. Stay_id

The same patient was given a different stay ID each time they were admitted to the hospital for a different stay. On average, each patient stayed in the emergency department two times. But some patients with complicated diagnoses came to the emergency department up to five times.

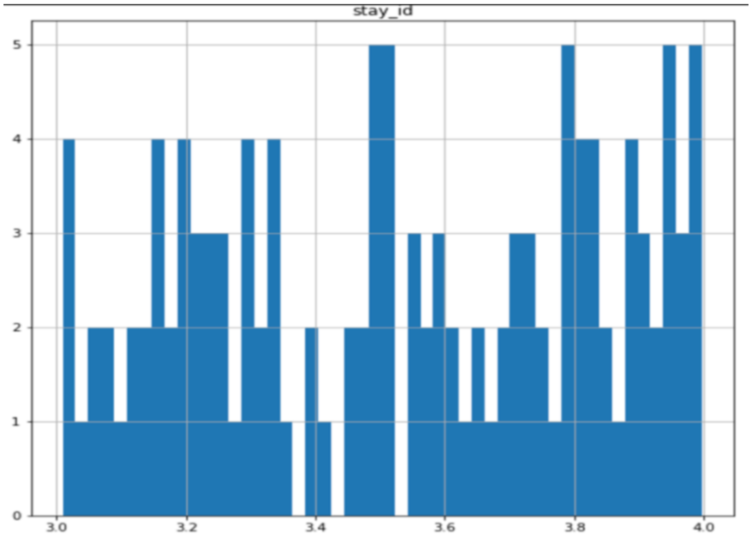


Figure 11. Histogram of patient stay information

4.2.4. Acuity

During the triage assessment process in the emergency department, patient's temperature, oxygen saturation, and blood pressure are measured to determine their acuity score. The acuity score determines the level of urgency and the appropriate level of care needed for the patient.

The data shows that Level 1 patients require immediate physician intervention and are taken directly to a room. Level 2 patients are high risk and require appropriate placement, while Level 3 patients require two or more resources with stable vital signs. Level 4 patients require only one resource.

The graph shows that most patients visited the ED at level 2 or level 3 acuity. This information is valuable for understanding patient acuity and can help healthcare providers allocate resources and prioritize care based on the patient's needs. By identifying the most common acuity levels, healthcare providers can anticipate and prepare for the most common patient needs in the emergency department.

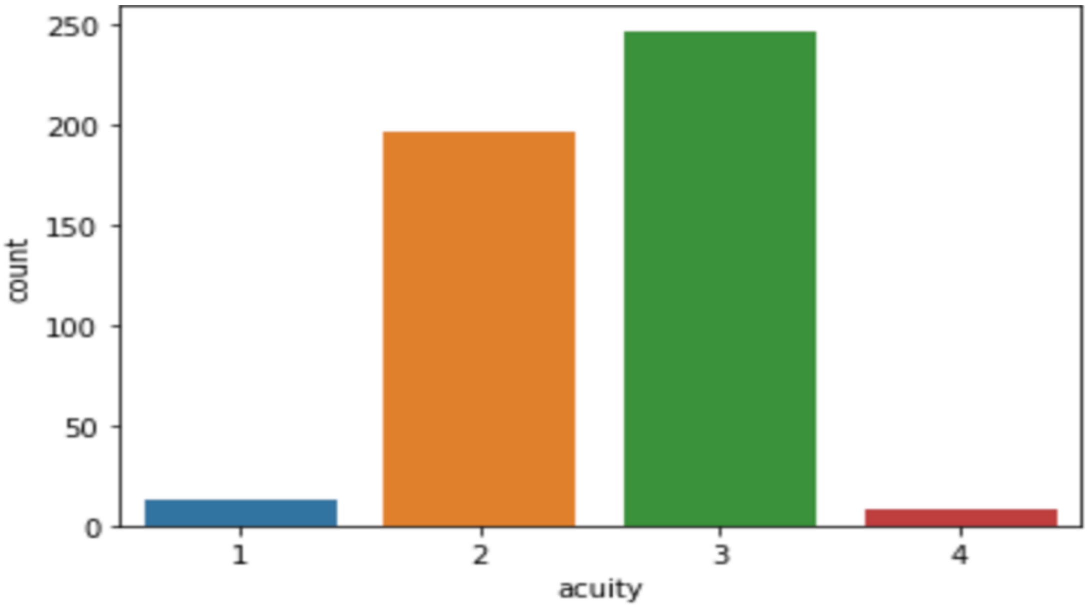


Figure 12. Bar graph depicting acuity score level in triage assessment

4.2.5. Chief complaint

The plot displays the most common chief complaints in order of highest to lowest ratings. The rating system assigns a numerical value to each chief complaint, with 1 indicating gastrointestinal issues such as abdominal pain or cramps, 0 indicating other pain types, 3 indicating respiratory issues such as shortness of breath or pneumonia, and 8 indicating infections such as cellulitis or skin redness.

Based on the plot, gastrointestinal issues are the most common chief complaint, followed by respiratory issues, and then infections. Other pain types are the least common chief complaint. This information is valuable for understanding the most common reasons patients seek care in the emergency department, allowing healthcare providers to allocate resources and prioritize care accordingly.

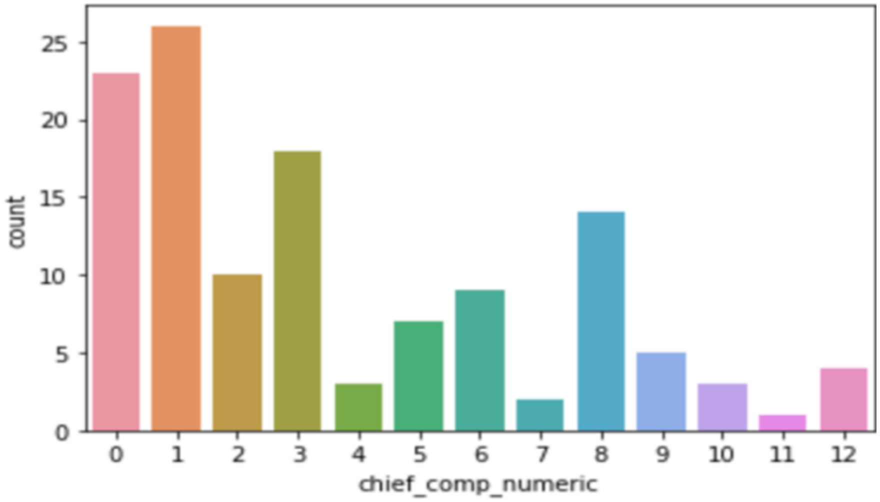


Figure 13. Bar plot of specific syndromes in the dataset

4.3. K – means clustering of syndrome and acuity

The clustering analysis conducted on the dataset identified respiratory, neurological, cardiovascular, and diabetic diseases as the most common reasons for patients visiting the emergency department at Beth Israel Deaconess Medical Center in Boston. These findings are consistent with previous research, which has highlighted the high incidence of respiratory and cardiovascular diseases in emergency departments.

This information is valuable for healthcare providers and policymakers, as it can be used to allocate resources and prioritize care for patients with these conditions. By understanding the most common reasons for emergency department visits, healthcare providers can tailor their services to meet the needs of their patients, ultimately improving patient outcomes and satisfaction.

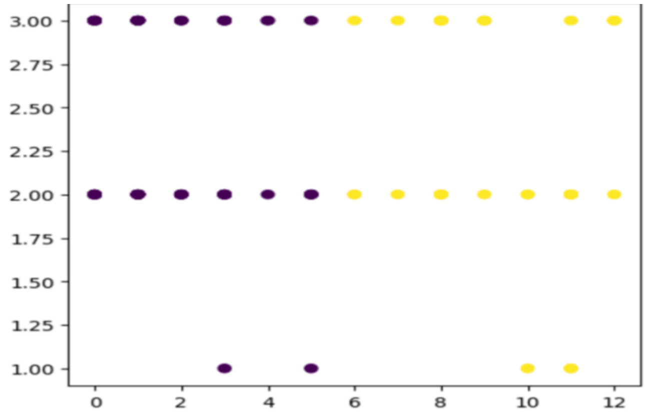
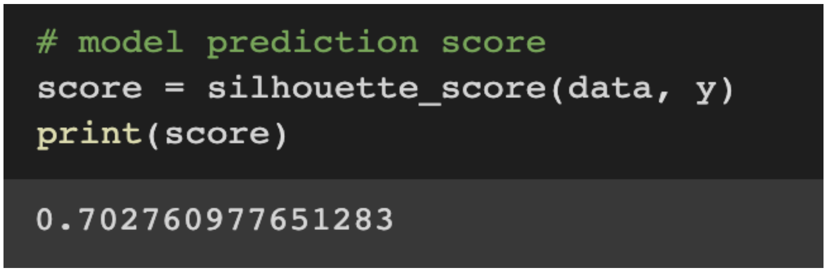


Figure 14: Clusters based on syndrome and acuity score.

4.3.1. Evaluation of K – means Clustering model

The Silhouette Coefficient is a commonly used metric for evaluating the quality of clustering results. The score ranges from -1 to 1, with a value of 1 indicating that the clusters are well separated and distinct, and a value of -1 indicating that the clustering is inappropriate or incorrect.

A Silhouette Coefficient of 0.7 is considered a very good score and indicates that the clusters are well separated from each other and clearly distinguished. This suggests that the clustering technique used in the analysis was effective in grouping the data points into distinct clusters, and that the clusters themselves are meaningful and useful for further analysis.



```
# model prediction score
score = silhouette_score(data, y)
print(score)
```

```
0.702760977651283
```

Figure 12: Code snippet of k-means clustering evaluation metric

5. Discussion of Results

5.1.1. Clinical Text Extraction by NLP

The use of the en_core_sci_md core model is highly advantageous in the context of clinical texts, given its specialized biomedical vocabulary that contains numerous medical terms and drug names. This feature allows the pipeline to accurately identify and categorize a wide range of named entities that are pertinent to clinical practice.

The en_ner_bc5cdr_md model's extensive training on a vast corpus of biomedical literature is critical as it enables the model to handle the specific language and vocabulary used in clinical texts with ease. Moreover, pre-trained models can significantly reduce the time and resources required to train custom NLP models for clinical text analysis.

However, since I didn't have access to hospital clinical notes, I manually created 30 sample datasets for clinical text extraction. Hence, I couldn't evaluate the model's performance. In future studies, it is essential to assess the model's accuracy and minimize errors in extracting only disease and drug information. Therefore, healthcare providers must manually review the model's results and verify its accuracy before making any clinical decisions based on the extracted information.

5.1.2. Limitations of K-means clustering

K-means clustering analysis is a machine learning technique that groups similar data points based on their characteristics. In this study, it was used to identify patterns in the data related to the chief complaints of patients visiting the emergency department. To facilitate clustering, I created a column called "syndrome" based on a literature review that classified diseases into syndromes. However, since I had other diseases not present in the literature, I had to Google and manually add all 125 syndromes.

If I had access to ED surveillance information, I could have classified chief complaints automatically using NLP techniques such as Coco Classifier, leading to a much better clustering model. It's crucial to consult with domain experts like physicians to understand which disease leads to a specific syndrome. The surveillance data could detect outbreaks, and using these models every day and continually training them for any new diseases is essential. If a specific disease like respiratory increases by a certain threshold value of ED stays, hospitals could detect early outbreaks, helping physicians and medical staff prepare well by allocating a lot of ventilators and medical resources, leading to saving millions of lives.

In addition to K-means clustering, it's essential to use other models like k-mode clustering and hierarchical clustering for better accuracy. K-means clustering with two clusters revealed that the most common diseases among patients visiting Beth Israel Deaconess Medical Center in Boston were respiratory, neurological, cardiovascular, and diabetic, with higher acuity. The consistency of these findings with previous research highlights the robustness and generalizability of the results. This consistency across studies suggests that respiratory and cardiovascular diseases are major health concerns across different populations and healthcare settings. Identifying these diseases as the most common reasons for emergency department visits has significant implications for healthcare resource allocation and planning since healthcare providers may need to prioritize resources for treating these conditions in emergency departments.

By identifying the most common conditions that lead patients to visit the emergency department, healthcare providers and policymakers can better allocate resources to provide the necessary care for these conditions. This information can also help healthcare providers tailor their services to their patients' needs, ultimately improving patient outcomes and satisfaction. For example, if respiratory and cardiovascular diseases are identified as the most common reasons for emergency department visits, healthcare providers can focus on providing appropriate diagnostic tests, treatment, and preventive care for these conditions. Policymakers can also prioritize funding for research and programs that aim to reduce the incidence and prevalence of these conditions and improve access to care for affected patients.

Furthermore, identifying the most common reasons for emergency department visits can inform the development of public health policies and interventions. For instance, policies can be developed to promote healthy lifestyles, prevent the onset of chronic diseases, and reduce the burden of these conditions on emergency departments.

6. Conclusion and Recommendations

The results of analyzing patient behavior data suggest that this information is crucial for healthcare providers to better understand patient health and develop more effective healthcare interventions. By analyzing patient behavior data, healthcare providers can anticipate future hospitalizations for certain patients and identify patients who may benefit from more proactive healthcare interventions to prevent the need for frequent hospitalizations.

However, the results also suggest that further work is necessary to fully leverage the potential of patient behavior data in healthcare. For example, healthcare providers may need to develop more sophisticated data analytics tools to better understand patient behavior patterns and identify patients at higher risk for hospitalization. Additionally, healthcare providers may need to work more closely with patients to better understand their health behaviors and develop targeted interventions that address their specific needs.

To improve or advance the data, follow-on work could include:

1. Further research into patient behavior data and its potential applications in healthcare, including identifying new patterns and trends that may be relevant for predicting future hospitalizations.
2. Development of more advanced analytics tools and algorithms that can process large amounts of patient behavior data and identify patterns and trends in real-time.
3. Collaboration between healthcare providers and patients to better understand patient behavior and develop targeted interventions that address specific needs and preferences.
4. Integration of patient behavior data with other types of health data, such as electronic health records and genetic data, to develop a more comprehensive picture of patient health and potential future health outcomes.

The integration of NLP and K-means clustering techniques has proven to be effective in identifying common diseases among patients visiting the Emergency Department. By extracting vital information from unstructured clinical notes and analyzing large datasets, healthcare providers can make more informed clinical decisions and provide evidence-based care. This can ultimately lead to more proactive and personalized healthcare for patients.

Moving forward, the use of NLP and K-means clustering in healthcare should be further explored and implemented in a larger scale. Hospitals and healthcare organizations can benefit from investing in technologies that utilize these techniques to improve patient care and reduce healthcare costs. It is also important to continue improving the accuracy and efficiency of these methods, to ensure that healthcare providers can make the best use of the data available to them.

Finally, as the use of technology in healthcare continues to grow, it is important to ensure that patient privacy and security are protected. Healthcare organizations must adhere to strict data protection regulations and ensure that patient data is kept confidential and secure. Furthermore, it is important to involve healthcare professionals in the development and implementation of these technologies, to ensure that they are effective in improving patient care and workflow, while being practical for use in a clinical setting.

References

1. Wendy W. Chapman, Lee M. Christensen, Michael M. Wagner, Peter J. Haug, Oleg Ivanov, John N. Dowling, Robert T. Olszewski, Classifying free-text triage chief complaints into syndromic categories with natural language processing, Artificial Intelligence in Medicine, Volume 33, Issue 1, 2005, Pages 31-40, ISSN 0933-3657, https://doi.org/10.1016/j.artmed.2004.04.001
2. Wagner, M. M., Hogan, W. R., Chapman, W. W., & Gesteland, P. H. (2005). Chief Complaints and ICD Codes. *Handbook of Biosurveillance*, 333. https://doi.org/10.1016/B978-012369378-5/50025-9
3. Thomas, A. A., Zheng, C., Jung, H., Chang, A., Kim, B., Gelfond, J., Slezak, J., Porter, K., Jacobsen, S. J., & Chien, G. W. (2014). Extracting data from electronic medical records: validation of a natural language processing program to assess prostate biopsy results. *World journal of urology*, *32*(1), 99–103. https://doi.org/10.1007/s00345-013-1040-4
4. Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., and Mark, R.G. (2020). MIMIC-IV (version 0.4). PhysioNet. doi: 10.13026/C2HM2Q.
5. Sugerman, D., & Nordenberg, D. (2004). Syndromic surveillance: a useful tool for detecting emerging disease threats. The Journal of Urban Health, 81(suppl 1), i24-i34. doi: 10.1093/jurban/jth136
6. Weiss, S. J., Ernst, A. A., & Nick, T. G. (2000). Comparison of the National Emergency Department Overcrowding Scale and the Emergency Department Work Index for quantifying emergency department crowding. Annals of Emergency Medicine, 35(6), 547-554. doi: 10.1016/s0196-0644(00)70014-0
7. Lombardo J, Burkom H, Elbert E, et al. Using chief complaints for syndromic

surveillance: a review of chief complaint-based classifiers in North America. J Public Health Manag Pract. 2003;9(3)

8. Wong W, Moore AW, Cooper G, Wagner M. Rule-based anomaly pattern detection for detecting disease outbreaks. In: Proceedings of the 18th National Conference on Artificial Intelligence (AAAI-02); 2002.
9. Travers DA, Haas SW. Using nurse's natural language entries to build a concept-oriented terminology for patient's chief complaints in the emergency department. J Biomed Inform 2003;36:260—70.
10. Travers DA, Waller A, Haas SW, Lober WB, Beard C. Emergency department data for bioterrorism surveillance: electronic data availability, timeliness, sources and stan- dards. In: Proceedings of the AMIA Symposium; 2003. p. 664—8.
11. Gregory Soler, Guillaume Bouleux, Eric Marcon, Aymeric Cantais, Sylvie Pillet, et al.. Emergency Department Admissions Overflow Modeling by a Clustering of Time Evolving Clinical Diagnoses. 14th IEEE International Conference on Automation Science and Engineering (CASE 2018), Aug 2018, Munich, Germany. hal-01885975
12. Extracting Medical Information From Clinical Text With NLP - https://www.analyticsvidhya.com/blog/2023/02/extracting-medical-information-from-clinical-text-with-nlp/?utm_source=related_WP
13. Medical transcripts from mtsamples - https://mtsamples.com
14. Kaggle medical transcripts dataset from mtsamples https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions
15. scispaCy package - https://spacy.io/universe/project/scispacy

Appendix

Data Source

- Medical Transcripts: https://mtsamples.com
- Kaggle Dataset: https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions
- Emergency Department Dataset: https://mimic.mit.edu/docs/iv/modules/ed/

Clinical Note Extraction

Install necessary packages

```
!pip install scispacy
!pip install https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.5.1/en_core_sci_sm-0.5.1.tar.gz
!pip install https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.5.1/en_core_sci_md-0.5.1.tar.gz
!pip install https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.5.1/en_ner_bc5cdr_md-0.5.1.tar.gz
```

Loading packages

```
import pandas as pd
import spacy
import scispacy
import en_core_sci_sm
import en_core_sci_md
#NER specific models
import en_ner_bc5cdr_md
#Tools for extracting & displaying data
from spacy import displacy
```

Upload csv files

```
from google.colab import files
uploaded = files.upload()

df = pd.read_csv(r'mimic_clinical_note.csv', encoding='ISO-8859-1')
df

text = df.loc[0, 'transcription']
text
```

```
# Load specific model: en_core_sci_sm and pass text through
nlp_sm = en_core_sci_sm.load()
doc = nlp_sm(text)

# display results by entity extraction
displacy_image = displacy.render(doc, jupyter = True, style = 'ent')

# Load specific model: en_core_sci_md and pass text through
nlp_md = en_core_sci_md.load()
doc = nlp_md(text)
displacy_image = displacy.render(doc, jupyter = True, style = 'ent')

# Now Load specific model: import en_ner_bc5cdr_md and pass text through
nlp_bc = en_ner_bc5cdr_md.load()
doc = nlp_bc(text)
#Display resulting entity extraction
displacy_image = displacy.render(doc, jupyter=True,style='ent')

print("TEXT", "START", "END", "ENTITY TYPE")
for ent in doc.ents:
    print(ent.text, ent.start_char, ent.end_char, ent.label_)

df.dropna(subset=['transcription'], inplace=True)
df_subset = df.sample(n=2, replace=False, random_state=42)
df_subset.info()
df_subset.head()

from spacy.matcher import Matcher
pattern = [{'ENT_TYPE':'CHEMICAL'}, {'LIKE_NUM': True}, {'IS_ASCII': True}]
matcher = Matcher(nlp_bc.vocab)
matcher.add("DRUG_DOSE", [pattern])
for transcription in df_subset['transcription']:
    doc = nlp_bc(transcription)
    matches = matcher(doc)
    for match_id, start, end in matches:
        string_id = nlp_bc.vocab.strings[match_id]  # get string representation
        span = doc[start:end]  # the matched span adding drugs doses
        print(span.text, start, end, string_id,)
#Add disease and drugs
        for ent in doc.ents:
            print(ent.text, ent.start_char, ent.end_char, ent.label_)
```

Emergency Department data pre-processing

```
import pandas as pd
from google.colab import files
uploaded = files.upload()
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")
from sklearn.cluster import KMeans

# patient stay
edstay = pd.read_csv(r'edstays.csv', encoding = "ISO-8859-1")
edstay = edstay.dropna()
df_ed = edstays.drop(columns = ['intime', 'outtime', 'gender', 'race', 'arrival_transport',
'disposition'])
df_ed = df_ed.astype({'hadm_id': 'int'})

# triage
triage = pd.read_csv(r'triage.csv', encoding = "ISO-8859-1")
triage = triage.dropna()
df_tri = triage.drop(columns = [ 'temperature', 'heartrate', 'resprate', 'o2sat', 'sbp', 'dbp', 'pain'])
df_tri = df_tri.astype({'acuity':'int'})
df_tri['chiefcomplaint'] = df_tri['chiefcomplaint'].str.lower()
df_tri

# removing duplicates
df_tri.drop_duplicates(subset=['stay_id'])

combined = df_ed.merge(df_tri.drop_duplicates(subset=['stay_id']), how='left')
df = combined.dropna()
df = df.astype({'acuity': int})

# medication reconillation
medrecon = pd.read_csv(r'medrecon.csv', encoding = "ISO-8859-1")
df_medrecon = medrecon.dropna()
df_medrecon.head(10)
df_medrecon = df_medrecon.drop(columns = ['charttime', 'gsn', 'ndc', 'etc_rn', 'etccode'])

df_medrecon.drop_duplicates(subset=['stay_id'])

combined_med = df.merge(df_medrecon.drop_duplicates(subset=['stay_id']), how='left')
df_med = combined_med
df_med

# diagnosis
diagnosis = pd.read_csv(r'diagnosis.csv', encoding = "ISO-8859-1")
df_diagnosis = diagnosis.drop(columns = ['seq_num','icd_version', 'Unnamed: 6', 'Unnamed: 7',
'Unnamed: 8'])
df_diagnosis['icd_title'] = df_diagnosis['icd_title'].str.lower()
df_diagnosis

df_dia = df_diagnosis.drop_duplicates(subset = ['stay_id'])

combined_dia = pd.merge(df_med, df_dia, how = 'left')
# joined all necessary tables, cleaned dataset
```

```
df_final = combined_dia

df = pd.read_csv(r'df_final (2).csv', encoding = "ISO-8859-1")
df

# acuity score distribution
import seaborn as sns
ax = sns.countplot(x="acuity", data=df_final)

# visualizing each variable
fig, ax = plt.subplots(figsize=(20, 20))
df_final.hist(bins=50, ax=ax)
```

K – means clustering

```
df = pd.read_csv(r'df_final_chiefcomp.csv', encoding="ISO-8859-1")
df = df.rename(columns={'ï»¿subject_id': 'subject_id'})

# clustering with acuity and chief complaint
df = df.loc[:125]
data = list(zip(df.chief_comp_numeric, df.acuity))
print(data)

inertias = []
for i in range(1,126):
    kmeans = KMeans(n_clusters=i, init='k-means++', random_state=42)
    kmeans.fit(data)
    inertias.append(kmeans.inertia_)

plt.figure(figsize=(10,10))
plt.plot(range(1,126), inertias, marker='o')
plt.title('Elbow method')
plt.xlabel('Number of clusters')
plt.ylabel('Inertia')
plt.show()

kmeans = KMeans(n_clusters=2)
kmeans.fit(data)
k_means_optimum = KMeans(n_clusters = 2, init = 'k-means++',  random_state=42)
y = k_means_optimum.fit_predict(data)
print(y)

plt.figure(figsize=(5,5))
plt.scatter(df.chief_comp_numeric, df.acuity,   c=kmeans.labels_)
plt.show()

from sklearn.metrics import silhouette_score

# model prediction score
```

```
score = silhouette_score(data, y)
print(score)


# elbow method - clustering icd_code and acuity

df = df.loc[:125]
data = list(zip(df.icd_code, df.acuity))
print(data)

inertias = []
for i in range(1,126):
    kmeans = KMeans(n_clusters=i)
    kmeans.fit(data)
    inertias.append(kmeans.inertia_)

# plt.figure(figsize=(10,10))
plt.plot(range(1,126), inertias, marker='o')
plt.title('Elbow method')
plt.xlabel('Number of clusters')
plt.ylabel('Inertia')
plt.show()

kmeans = KMeans(n_clusters=3)
kmeans.fit(data)
plt.figure(figsize=(10,5))
# plt.scatter(df.stay_id, df.chief_comp_numeric,  c=kmeans.labels_)
plt.scatter(df.icd_code, df.acuity,  c=kmeans.labels_)
plt.show()

k_means_optimum = KMeans(n_clusters = 3, init = 'k-means++',  random_state=42)
y = k_means_optimum.fit_predict(data)
print(y)

# model prediction score
score = silhouette_score(data, y)
print(score)
```