**Project Introduction:** Using Text-Based Procedural Data to Identify Patients with Active Inflammatory Bowel Disease

**Is it possible to develop NLP algorithms to automate the identification of patients with active IBD (which is currently entirely manual)?**

Summary:

A Massachusetts-based AI company is using videos and electronic health record data to develop machine learning algorithms to improve gastroenterologists' ability to detect, diagnose, and treat abnormalities in the colon, with a focus on identifying patients with active Inflammatory Bowel Disease (IBD) for clinical trials. The company relies on analyzing images, videos, and text-based data, including exam notes and image metadata, to surface patients. **They are exploring the possibility of developing Natural Language Processing (NLP) algorithms to automate the identification of patients with active IBD, which is currently a manual process**.

The team has analyzed text-based data and identified keyword-centric patterns among patients with Inflammatory Bowel Disease (IBD). They found that certain keywords, such as "ileitis," have a high positive predictive value for IBD patients. However, the challenge is that these keywords also appear in patients with a range of disease severities. The team also identified combinations of words and phrases that are correlated with higher conversion rates for surfacing patients, such as:

"ileitis" + "Crohn's - Simple Endo Score" + "Inflammatory Bowel Disease"

**Additional Information**: MUST READ !!!

1. **What is Crohn’s Disease**? Crohn's disease is a chronic inflammatory bowel disease (IBD) that can affect any part of the gastrointestinal (GI) tract, from the mouth to the anus. It is characterized by inflammation, ulceration, and thickening of the bowel wall, which can lead to abdominal pain, diarrhea, weight loss, and other symptoms.
2. **What is Endoscopy**? Endoscopy is a procedure that uses a flexible tube with a camera to examine the inside of the GI tract.
3. **What is a Simple Endo score?** The *Simple Endoscopic Score for Crohn's Disease* (SES-CD) is a tool used to assess the severity of Crohn's disease based on endoscopic findings. The SES-CD is a scoring system that evaluates four different aspects of the bowel lining:

* Severity of **ulceration**(Ulceration is a process in which a portion of tissue or skin breaks down, forming an open sore or lesion)
* Extent of affected bowel
* Size of ulcers
* Presence of **narrowing or strictures**(narrowing or strictures refer to a condition in which the lumen (the opening) of a hollow organ or structure in the body becomes narrower or more constricted than it should be.)

Each of these aspects is given a score from 0 to 3, and the total score ranges from 0 to 12. Higher scores indicate more severe disease.

1. **What does ileitis mean**? Ileitis is a term used **to describe inflammation of the ileum**, which is the last section of the small intestine. In the context of Crohn's disease, ileitis is a common form of the condition, which is characterized by chronic inflammation and ulceration in the gastrointestinal tract.

An algorithm can be used to more easily and consistently identify words and phrases that are indicative of Inflammatory Bowel Disease (IBD) and predict its severity. This contrasts with the **current approach** of using crafted queries based on assumptions about how healthcare providers document patient data, which can **lead to biases and overlook important terms**.

The proposed methodology involves using unstructured textual processing of Electronic Health Record (EHR) data to **identify multi-word sequences associated with active IBD**. If certain text can reliably predict active IBD, this information can be leveraged to automate the patient surfacing process for trial referral, which currently requires manual processing by an expert physician. Additionally, **the findings can be used to predict which patient exams are IBD in an upcoming EHR Pilot(**explained below**)**.

This methodology can help improve the accuracy and efficiency of IBD diagnosis and treatment, while minimizing biases and human error in the process.

Additional Information: MUST READ!!!

1. What is **EHR pilot**? An Electronic Health Record (EHR) pilot is a trial program that tests the implementation of an EHR system on a small scale before it is rolled out across an entire healthcare organization. The **purpose** of an EHR pilot **is to identify and address any issues or challenges that may arise during the implementation process**, and to evaluate the effectiveness of the EHR system in achieving the desired goals.

During an EHR pilot, a selected group of healthcare providers and staff may use the EHR system to manage patient data, track medical histories, and schedule appointments. This pilot group can provide feedback on the system's usability, functionality, and overall effectiveness. This feedback can be used to refine the system and improve the user experience before it is deployed more widely.

An EHR pilot can help healthcare organizations to minimize disruptions to patient care, reduce costs, and improve the quality of care by ensuring that the EHR system is properly implemented and integrated into existing workflows.

**Current Manual Process**:

1. **ICD Filtering for IBD**: Exams with ICD K50 & K51 identifying IBD patients(5.3K Exams with comments)

ICD filtering for IBD refers to the use of International Classification of Diseases (ICD) codes to identify patients with Inflammatory Bowel Disease (IBD) from electronic health records (EHRs).

ICD codes are standardized medical codes that are used to classify diseases and other health conditions. By applying ICD codes for IBD to EHR data, healthcare providers can quickly and easily identify patients who have been diagnosed with IBD or who have symptoms of the disease.

ICD filtering can be used in combination with other data processing methods, such as natural language processing (NLP), **to extract additional information from EHR data that may not be captured by ICD codes alone**. For example, **NLP can be used to identify key terms and phrases in unstructured text that are associated with active IBD, which can help to refine the diagnosis and treatment of the disease**.

1. **ICD filtering for I/E**: Using exclusionary criteria from study protocol, eliminate patients with ICD codes representing any exclusionary criteria(approximately 800 Exams)

ICD filtering for I/E refers to the process of using International Classification of Diseases (ICD) codes to filter and analyze healthcare data related to hospital inpatient (I) and outpatient (E) encounters.

ICD filtering can be used to identify specific conditions or procedures associated with hospital admissions or outpatient visits. For example, a hospital administrator may use ICD filtering to identify the **most common reasons for hospital admissions** or **to track trends in specific medical conditions over time**.

**To perform ICD filtering,** healthcare data is typically collected and coded using ICD codes. These codes are then used to sort and filter the data according to specific criteria, such as the type of medical condition, the type of procedure performed, or the demographic characteristics of patients.

1. **Keyword Queries**: Query exam notes and image comments for keywords and phrases indicating presence of active disease.

NLP algorithms are used to analyze unstructured data, such as text notes and image comments, and identify relevant keywords and phrases that indicate the presence of active disease. These algorithms use various techniques, **including part-of-speech tagging**, **named entity recognition**, and **sentiment analysis**, to extract meaningful information from the text.

Machine learning algorithms are then used to analyze the extracted data and identify patterns and trends that are associated with active disease. These algorithms use a combination of supervised and unsupervised learning techniques to train models that can accurately classify exam notes and image comments based on the presence or absence of active disease.

The resulting models can be used to automatically identify and flag exam notes and image comments that contain relevant information about active disease, which can then be reviewed and analyzed by healthcare professionals to make informed decisions about patient care.

1. **Internal image review**: Review images for signs of clearly active disease
2. **GI(Gastro-Intestinal) image review**: Review images for disease severity and assign threshold score(350+ classified as active by GIs)

**Requirements:**

The project requires the use of natural language processing (NLP) and machine learning techniques to develop models that can detect active Inflammatory Bowel Disease (IBD) from expert physician's inputs. This involves the **vectorization of words through neural networks** using different techniques such as **bag of words** and **n-grams** available in packages like **word2vec**, **doc2vec**, and **fasttext**.

The outputs from this process will be helpful to find associations between words, particularly relevant to find words that the team at the company might omit due to our internal bias. Additionally, **the inputs provided by expert physicians to detect active IBD on an exam can be considered tags**. The vectorized text could be used as an input to predict the **tag -1 if the exam presents active disease**, **0 otherwise**- using machine learning techniques.

The **vectorized text will help find associations between words to overcome internal biases that may cause important words to be omitted**. **The vectorized text can also be used as input to predict the tag that indicates the presence of active IBD on an exam using machine learning techniques**.

The developed models' predicted likelihood of finding a patient with active IBD could then be used to modify internal processes, reducing the need for manual review of images for some exams.

**Deliverables**:

1. A report showing:

1. Methodology
2. Descriptive statistics on the frequency of words from the exams and images comments with active IBD vs the rest, by type of IBD (Chron’s disease and Ulcerative Colitis)
3. A results section that should include:
4. Individual words and multi-word sequences associated with exams showing active IBD.
5. A comparison of the accuracy of different ML models with different parameters for the prediction of active disease (for example, in the form of a confusion matrix)

2. The code(Python/R).

**Data**

 ~5K+ exams with ICD codes for IBD (K50 for Crohn's Disease and K51 for Ulcerative Colitis).

~800 of these exams satisfied the criteria for a patient to be a good trial candidate and were manually reviewed by expert physicians.

From these cases, **350+ exams** were classified as having active disease. The exam and image comments for each of these exams is in **JSON** format.