UC San Diego
SCHOOL OF MEDICINE
Department of BioMedical Informatics

Electronic Health Record-Based Cohort of Patients with Inflammatory Bowel Diseases for Patient-Centered Outcomes Research

Siddharth Singh, MD, MS; Paulina Paul, MS; Chun-Nan Hsu, PhD; Lucila Ohno-Machado, MD, PhD

Abstract

We are developing an electronic health record (EHR)-based cohort of patients with inflammatory bowel diseases (IBD) seen at any of the five University of California medical campuses, using a novel informatics approach combining structured data and narrative data, through natural language processing (NLP), for observational comparative effectiveness and patient-centered outcomes research.

Background/Introduction

Observational research in inflammatory bowel diseases (IBD) has traditionally been limited by small sample size and event rates (in single-center studies) or lack of detailed phenotype analysis (in administrative health claims databases). We propose to overcome these barriers by creating a large contemporary electronic health record (EHR)-based cohort of patients with IBD seen across any of the five University of California (UC) medical campuses, which combines structured (and codified) data with NLP-derived phenotype data, utilizing the platform developed for the patient-centered SCAlable National Network for Effectiveness Research (pSCANNER) a clinical data research network that is part of PCORnet. A combination of structured and narrative data has been shown to have superior performance to identify and characterize patients in "phenotyping" activities for numerous diseases.

Methods and Materials

Cohort identification:
The cohort for this study will be identified using a sequential three-step process as shown in Figure.

1. In the first step, we'll screen all patients (seen across any of the five UC health sciences campuses), using ICD-9 codes for Crohn's disease (ICD 9 555.x) and ulcerative colitis (ICD 9 556.x), to identify a large cohort of patients with a potential diagnosis of IBD.
2. From this cohort, in the next step, we'll develop an algorithm based on structured data elements, listed below, to more accurately identify patients with CD and UC.

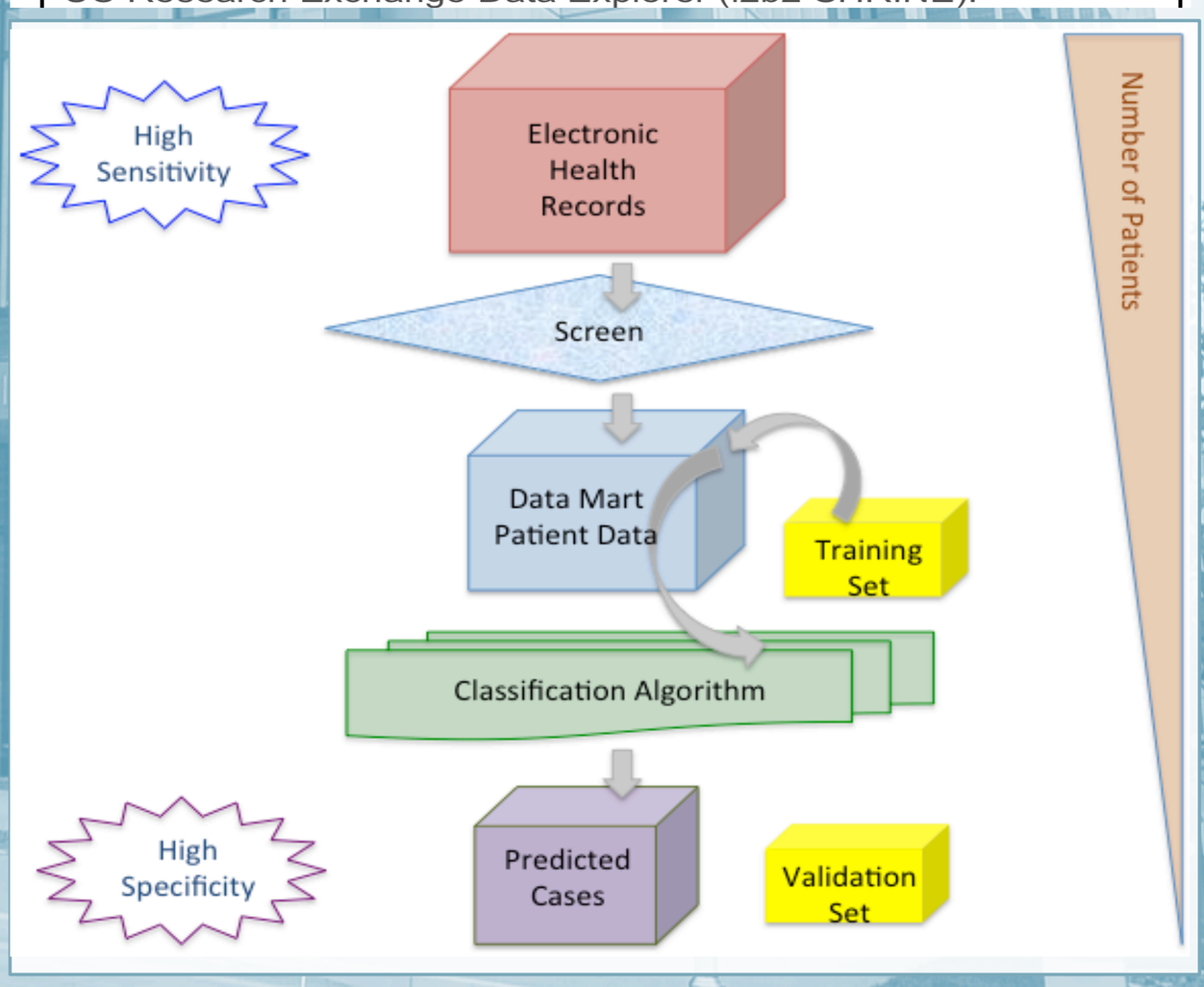
- Number of ICD-9 codes for CD and/or UC
- Diagnostic codes for competing diagnosis (irritable bowel syndrome, diverticulitis),
- Codes for inpatient hospitalization, gastroenterologist visit or endoscopic procedure,
- Diagnostic codes for CD or UC-related complications (intestinal fistula, stricture, perianal fistula or abscess),
- Procedural codes for abdominal or perianal surgery (using Current Procedural Technology [CPT] codes),
- Laboratory values for elevated inflammatory markers (such as C-reactive protein), and
- Prescription of IBD-related medications in the EMR prescription program (5-ASA, corticosteroids, immunomodulators, anti-TNF agents, anti-integrin agent).

Methods/Materials

3. To analyze narrative text, we'll apply natural language processing (NLP) techniques to mine hospital notes and endoscopy reports separately, to evaluate IBD-disease extent, phenotype and severity at time of cohort entry. We'll map expert-defined terms to SNOMED [Systemized Nomenclature of Medicine-Clinical Terms (SNOMED-CT), a hierarchically organized clinical healthcare terminology index with over 300,000 concepts, to allow for variations in language use] or RxNorm. We'll process clinical notes using clinical notes for cohort identification (CICT).
4. Next, we'll combine structured data elements with NLP-identified narrative text to create a combined model to identify patients with IBD. The accuracy of the models at various specificity levels will be calculated using hierarchical generalized linear model, and the overall prediction performance evaluated based on the area under the receiver operating characteristic curve (AUC), using a training set of randomly selected 500 patients with CD and UC (who will be classified as having CD or UC or neither, through manual chart review).

BMI Class	Crohn's Disease		Ulcerative Colitis	
	Overall	On anti-TNF	Overall	On anti-TNF
Class II/III obesity (BMI≥35.0 kg/m^2)	683	197	713	113
Class I obesity (BMI 30.0-34.9 kg/m^2)	1305	406	1543	270
Overweight (BMI 25.0-29.9 kg/m^2)	3139	979	3589	609
Normal weight (BMI 18.5-24.9 kg/m^2)	5089	1670	5176	901
Underweight (BMI ≤18.5 kg/m^2)	1141	387	918	194
TOTAL	**11,357**	**3,639**	**11,939**	**2,087**

Table 1. Total number of patients (and patients treated with anti-TNF) with IBD seen at any of the five University of California Medical Centers, stratified by body mass index, identified using UC Research Exchange Data Explorer (i2b2 SHRINE).



Conclusions

Our approach combining structured and narrative data, using NLP algorithms, will help to identify and characterize a large, EHR-based contemporary cohort of IBD patients across the UC system for use in observational comparative effectiveness and patient-centered outcomes research.

This work is funded by ACG Junior Faculty Development Award (SS), CCFA Career Development Award (SS), PCORI CDRN-1306-04819 (LOM) and NIH CTSA grant UL1TR001442 (LOM).

sis040@ucsd.edu