Subtyping Gastrointestinal Surgical Outcomes from Real World Data: A Comprehensive Analysis of UK Biobank

Uri Kartoun PhD FAMIA[1], Kingsley Njoku MD[2], Tesfaye Yadete MD[3], Sivan Ravid MSc[4], Eileen Koski MPhil FAMIA[5], William Ogallo RPh PhD[6], Joao Bettencourt-Silva PhD[7], Natasha Mulligan MSc[7], Jianying Hu PhD[5], Julia Liu MD[2], Thaddeus Stappenbeck MD PhD[3], Vibha Anand PhD FAMIA[1]

[1]IBM Research, Cambridge, MA, USA; [2]Department of Internal Medicine, Morehouse School of Medicine, Atlanta, GA, USA; [3]Department of Inflammation and Immunity, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA; [4]Healthcare Informatics, IBM Research-Haifa, Mount Carmel Haifa, Israel; [5]IBM Research, Yorktown Heights, NY, USA; [6]IBM Research Africa, Kenya; [7]IBM Research Europe, Dublin, Ireland.

Abstract

Chronic gastrointestinal (GI) conditions, such as inflammatory bowel diseases (IBD), offer a promising opportunity to create classification systems that can enhance the accuracy of predicting the most effective therapies and prognosis for each patient. Here, we present a novel methodology to explore disease subtypes using our open-sourced BiomedSciAI toolkit. Applying methods available in this toolkit on the UK Biobank, including subpopulation-based feature selection and multi-dimensional subset scanning, we aimed to discover unique subgroups from GI surgery cohorts. Of a 12,073-patient cohort, a subgroup of 440 IBD patients was discovered with an increased risk of a subsequent GI surgery (OR: 2.21, 95% CI [1.81–2.69]). We iteratively demonstrate the discovery process using an additional cohort (with a narrower definition of GI surgery). Our results show that the iterative process can refine the subgroup discovery process and generate novel hypotheses to investigate determinants of treatment response.

Introduction

Gastrointestinal conditions such as Crohn's disease (CD) and ulcerative colitis (UC) are the two major forms of Inflammatory Bowel Disease (IBD) where there is no cure.[1] IBD is characterized by recurrent, chronic inflammation of the gastrointestinal tract, with frequent involvement of other body systems which are called extra-intestinal manifestations. Although CD and UC have some similar features, there are some key differences: inflammation in CD is typically transmural and can occur anywhere in the gastrointestinal tract from the mouth to anus; inflammation in UC is continuous and occurs predominantly within the inner lining (the mucosa) of the colon only. IBD presentations may differ by race and ethnicity, and age of onset.[2] The current disease classification system is based on disease extent and characteristics to guide treatment and prognosis. Despite advances in understanding of disease pathogenesis and new therapeutic options, IBD patients still require surgeries and hospitalizations for management. Thus, a more accurate patient stratification system may facilitate a targeted approach for existing therapies and a guide to the development of new therapies.[1] The heterogeneity of IBD is complex due to the significant contribution of environmental and microbiome factors that interplay with complex genetic predisposition. Disease subtyping that incorporate a multitude of factors such as sociodemographics[2], lifestyle, clinical and relevant immunological profiles[3] is expected to be helpful in development of effective interventions.

Biobanks and Electronic Medical Records (EMRs) are rich sources of real-world data (RWD) to study IBD heterogeneity, from the perspective of both disease subtyping and treatment and/or therapeutic response perspective. They contain comprehensive disease-related information such as comorbidities, procedures, and medication data, as well as sociodemographic data such as age, gender, anthropometrics (weight, height, body mass index), social and behavioral history such as smoking status, sleep behavior, pain, depression, and details on duration, location of disease activity and severity. They also contain pertinent laboratory test results (e.g., c-reactive protein, fecal calprotectin, hemoglobin, neutrophil-lymphocyte ratio, albumin) relevant to response to administered drug therapies and adverse events.[3] Analytic approaches such as feature selection and clustering utilizing modern machine learning and artificial intelligence (AI) methods are designed to incorporate such diverse data in big datasets, such as those from the UK Biobank (UKBB) to subclassify IBD.[4–7] We propose to develop a candidate classification system for GI surgeries using this comprehensive dataset to provide the experimental basis to investigate determinants of treatment response. In this study, we examined disease subtypes with a primary outcome of repeat major GI surgery within 5 years of initial GI surgery using the UKBB and our proposed pipeline which leverages the open-sourced BiomedSciAI toolkit (https://github.com/BiomedSciAI).

Methods

The discovery methodology (Fig. 1) is applied as an iterative process to refine both cohort selection and outcome(s) of interest.

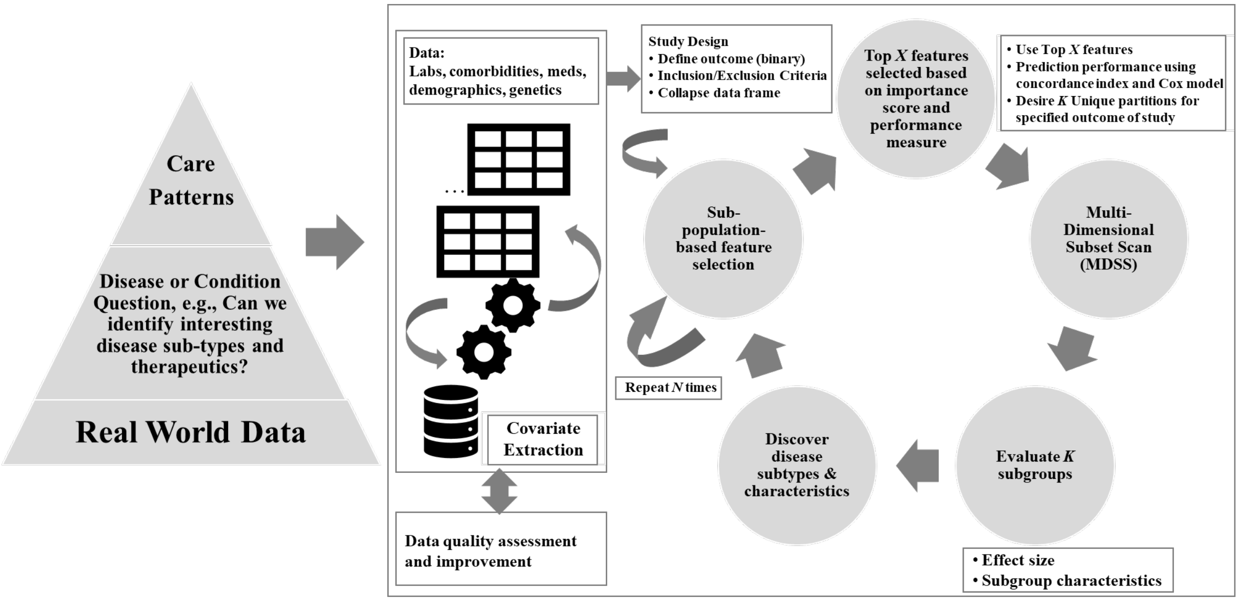

Care Patterns
Disease or Condition Question, e.g., Can we identify interesting disease sub-types and therapeutics?
Real World Data
Data: Labs, comorbidities, meds, demographics, genetics
Covariate Extraction
Data quality assessment and improvement
Study Design
• Define outcome (binary)
• Inclusion/Exclusion Criteria
• Collapse data frame
Top X features selected based on importance score and performance measure
• Use Top X features
• Prediction performance using concordance index and Cox model
• Desire K Unique partitions for specified outcome of study
Multi-Dimensional Subset Scan (MDSS)
Evaluate K subgroups
• Effect size
• Subgroup characteristics
Discover disease subtypes & characteristics
Sub-population-based feature selection
Repeat N times

Figure 1. Disease subtype identification. RWD = Real-world evidence.

Cohort Construction and Variable Extraction

a. Data Source

We use the UK Biobank, prospective repository that contains healthcare data for over 500,000 individuals.[8] Our analysis focused on a subset of patients with general practice (GP) observational data, comprising approximately 230,000 individuals who received care between January 1, 2000, and December 31, 2015. We curated a list of Office of Population Censuses and Surveys-4 (OPCS-4) codes that indicate major surgeries related to IBD while excluding codes for routine evaluations such as colonoscopies. This list comprised 146 OPCS-4 codes from the full list of 927 surgery types. We included surgeries pertaining to total and partial colectomy, ostomy-related procedures (formation, revision, and takedown), and bypass with anastomosis after resection. We then identified all patients whose first major GI surgery was performed during the study period. To ensure adequate data ascertainment and follow-up, we only included individuals with at least two years of care prior to their first major GI surgery, which we defined as the "baseline" period. Inclusion filters were applied to patients who were 18 years or older at baseline and showed no indication of pregnancy up to 9 months before the baseline period. Data access was provided under UK Biobank application #95318.

b. Extracting Covariates

We used RWD data entries to define age, sex, race, and smoking status. We identified the laboratory types that represented 95% of all available numerical raw observations in the GP data, resulting in 135 covariates. For each laboratory covariate, we used the values that were available at baseline, and if none were available, we used the most recent value found during the preceding 6 months. We imputed missing values by using the mean value in each laboratory covariate (all had missing values with high prevalences of at least 50%). We also included 24 variables indicating prior prescription of GI-related medications, focusing on IBD. A list of known IBD medications was extracted by mining the unstructured data corpus of IBD clinical trials data from ClinicalTrials.gov using the open-source IBM Deep Search platform (https://ds4sd.github.io). A Unified Medical Language System (UMLS)-based annotator (https://lhncbc.nlm.nih.gov/ii/tools/MetaMap.html) was used to identify biomedical concepts in trial protocols and subsequently the UMLS Semantic Network was used to identify those concepts that are represented in medication vocabularies (RxNorm and DrugBank). In addition, we used a RWD extraction tool[9], which enables rapid extraction of covariates, to obtain over 400 Clinical Classifications Software Refined (CCSR)[10] comorbidities (at any time on or before baseline). We mapped GP and inpatient diagnosis codes to CCSR categories - diagnosis codes are encoded in ICD10 in the hospitalization data, and Read versions 2/3[11] in the GP data. Each of these taxonomies has thousands of different codes. We first mapped GP diagnosis from Read2/3 to ICD10 (inpatient data is already in ICD10). Then we mapped ICD10 to CCSR categories to form the 400 categories. Additionally, we obtained UKBB-specific sociodemographic status indicators such as "Average total household income before tax", "Job involves heavy manual or physical work," and "Attendance/disability/mobility allowance."[8,9] We also included two specific covariates for CD and UC (identified using ICD-10 codes K50* and K51*, respectively).

c. Outcome Definition

We defined the primary outcome as the occurrence of a subsequent major GI surgery after the initial surgery (baseline) within the subsequent 5 years. All individuals were censored either at their last follow-up or at the end of the 5-year prediction window. Furthermore, to be considered an outcome, surgeries had to occur after 30 days or more after baseline. We applied this filter to exclude surgeries that represented a continuation or complication of the one observed at baseline (these surgeries were often observed during the same admission). The last follow-up date was determined by the date of the individual's last interaction with the healthcare system, such as an office visit, hospital encounter, laboratory test, or medication. Additionally, we report the incidence rate for the 5-year subsequent GI surgery per 1,000 patient years (1K PY).

Subpopulation-based Feature Selection (SPBFS) and Evaluation of Prediction Performance

We used the open-sourced version of the tool developed by IBM Research and Massachusetts Institute of Technology (https://github.com/IBM/spbfs) which will be part of BiomedSciAI toolkit. We divided the 12,073-patient cohort (Fig. 2A) into two sets: a derivation set (2/3) and a validation set (1/3), consisting of 8,052 and 4,021 individuals, respectively. Prior to applying sub-population-based feature selection[12,13] to the derivation set, we removed features that were uninformative, i.e., contained only a single value for all patients. Table 1 highlights characteristics of the cohort including incident rate and comorbidity prevalences (e.g., CD and UC). Out of the 600 features, we randomly selected 6 (1% of 600) to match cases and controls in each iteration, with a caliper value of 0.5 (a statistical standard upper bound threshold indicating the highest allowed standard deviation for each covariate)[14] and a case-control ratio of 1 (an integer representing the total number of controls that are matched per case). We repeated this process 1,000 times, resulting in a subset of the 600 features that were selected and ranked by importance as 0 (not selected) or in the range of 1 to 100 (higher value indicating higher importance). We then used the validation set to assess the prediction performance of all possible feature combinations from the selected features. Importance scores (as shown in Fig. 3) indicate the percentage of runs in which each feature was selected as informative, with a score of 99% indicating that a particular feature was selected in 990 out of the 1,000 runs. Although all selected features were used for further analyses, the smaller combination set that yielded the highest prediction performance, as measured by the concordance index (c-index), was given priority for further investigation. Additional thresholds may be considered as well (e.g., top 1% of the selected features).

Discovering Data Partitions using Multi-Dimensional Subset Scan (MDSS)

RWD-based subgroup analysis can become complicated and time-consuming as the number of features increases. Furthermore, prior knowledge based on limited exposure to relevant subgroups and a targeted clustering or prediction model may not reveal hidden differentiating (or anomalous) groups in patterns of care. To address these subgroup analysis challenges, we extended the Multi-dimensional Subset Scanning (MDSS) algorithm from the anomalous pattern detection literature[7,15,16] to automatically detect and rank differentiating subgroups from a set of discretized features and their feature values. MDSS (to be open sourced as part of BiomedSciAI toolkit) largely focuses on discovering subsets of data by maximizing a scoring function over all subsets in a tabular dataset *in linear time.* Our MDSS extensions can be used to: (1) discover and rank multiple anomalous subgroups with mean outcomes in the subgroups are the most divergent from the global mean or the predicted probability of a binary classifier; and (2) discover and rank subgroups with the most significant heterogeneous treatment effects. The latter functionality is not used in this work. As the number of selected features could be large, we defined importance score >1% as a threshold selection score for the selected features. Feature values were either categorical (True/False) or discretized in quartiles if not categorical, e.g., laboratory values such as albumin, creatinine, etc., for input to the MDSS tool. Detected subgroups (which can be mutually exclusive or overlapping as requested via Application Programming Interface calls to the tool) were discovered with statistical validity and here we chose to discover four mutually exclusive groups as the number of desired subgroups for purposes of demonstration and readability.

Exploring Disease Subtypes and their Characteristics

Here we analyze the discovered data partitions; the MDSS tool identifies unique partitions of data and segregates these groups from the larger cohort based on unique subject IDs, i.e., it can label individuals to the data partitions. The data partitioning functionality helps in identifying the most important features and their values for a group (or discovered subtype) at a higher level and the tool's labeling functionality helps to discover covariate patterns (and group characteristics) at a deeper level (e.g., polygenic risk scores for bowel cancer). For purposes of illustration, here we only describe selected important (largely Present/Absent) features if their value in the discovered subtype is found to be "True", e.g., male = True for the entire subgroup. Note that while the feature value "False" may also be as important, but in our context, it largely means a covariate was absent, e.g., as in is_azathioprine = False. To assess differences between discovered subgroups we generated density distributions of selected variables, e.g., time to outcome, and time to outcome considering smokers vs. nonsmokers (Fig. 4A–F).

Defining Iterations

The above steps can be iterated over to fine tune subgroup discovery, for example using input from clinical collaborators. This in turn could help to refine the definition of cohorts, and/or the specific outcome of interest. In the work presented here, we refined the definition of the outcome for each iteration. Essentially, we modified what is considered as a "major" GI surgery at both the index and/or the subsequent surgical time (outcome). We performed two iterations, listed below as two scenarios to demonstrate the discovery process -

Scenario 1: An inclusive scenario whereby the definition of *"major" GI* surgery included 146 OPCS-4 codes (as the broad definition mentioned in the Data Source subsection) yielding a larger analytic cohort, and,

Scenario 2: An exclusive scenario whereby *"Open Endoscopic operations on colon, OPCS code H18*"* and *"Endoscopic cauterisation of lesion of lower bowel using fibreoptic sigmoidoscope, OPCS code H23.2"* were excluded as major GI surgeries yielding a much smaller cohort (144 OPCS-4 codes were used in total).

Results

We first describe the characteristics of the extracted GI surgery cohort in both scenarios from UKBB. We then describe importance of extracted covariates (features) for risk of major subsequent GI surgery in each scenario, followed by discovered subgroups and their characteristics.

Cohort characteristics: Using the methodology depicted in Fig. 1, we identified cohorts (N = 12,073 in Scenario 1) and (N = 4,534 in Scenario 2) of patients who underwent at least one major GI surgery (Fig. 2A–B). In Table 1, the characteristics of these cohorts are presented, including a comparison of key variables in between groups of patients who did or did not experience a subsequent GI surgery within the 5 years of initial surgery. The baseline surgical incidence was 20.7% and 23.1% in the two scenarios (1 and 2, respectively). On average, participants were 61.9 vs. 61.0 years old at baseline and 29.2% vs. 23.3% had a history of smoking (scenarios 1 and 2, respectively). The median follow-up period was 2.4 years (Q1: 0.9, Q3: 4.8) and 2.7 years (Q1: 0.8, Q3: 5.0) in two scenarios (1 and 2, respectively). CD prevalences were 3.0% and 4.2% in scenarios 1 and 2, respectively. UC prevalences were 4.7% and 5.4% in scenarios 1 and 2, respectively.

Comorbid histories focused on GI-related pre-existing conditions are presented as well as based on CCSRs, with "Other gastrointestinal disorders," "Diverticulosis and diverticulitis," and "Gastrointestinal hemorrhage" as the most prevalent conditions with 40.5%, 31.4%, and 30.4%, respectively in scenario 1. The most prescribed medications were hydrocortisone, prednisolone, and metronidazole, with a prevalence of 36.0%, 20.5%, and 15.4%, respectively in scenario 1. The populations with and without the outcome were associated with statistically significant differences in values for most of the covariates as shown in Table 1 in both scenarios. Most covariates found differentiating the scenarios – for example, "Diverticulosis and diverticulitis" prevalence was higher and "Anal and rectal conditions" prevalence was lower in scenario 1 vs. 2. Use of prescription drugs including hydrocortisone and azathioprine, as highlighted in further detail in Table 1 differentiates the two scenarios.

Feature importance: Of the 600 candidate features, 89 were found informative by SPBFS (i.e., had an importance > 0). An evaluation to assess performance using all combinations of the top 89 selected features determined that using the 7 top features (Fig. 3) yielded the highest prediction performance with c-index of 0.607 [95% CI, 0.586–0.627]. Colorectal cancer was found as the top predictor for a subsequent GI surgery with an importance score of 99.0%. Note that using only the 3 top features yielded a similar performance as using the top 7 with a c-index of 0.594 [95% CI, 0.575–0.613]. From a clinical perspective it is valuable to analyze additional selected features as they could initiate investigations to understand patient outcomes specific to unique subgroups. Medications such as azathioprine (importance score = 21.5%) and hydrocortisone (importance score = 7.2%) also made it to the list of selected features, out of 24 medication candidates. Since our data is from a general population (patients who received care between 2000 and 2015) who had a GI surgery in the UKBB, it is expected that most would have moderate to severe disease, as failure of medical management is the general indication for surgery and past induction therapy, which aligns with our feature ranking (using importance score) of these medications. Although laboratory observations in patients with GI surgeries such as total bilirubin, albumin, creatinine, c-reactive protein, and white blood cell count were included in the 89 selected features, their importance scores were small (i.e., below 0.2%). Of the 89 features, 35 (in Scenario 1) and 21 (in Scenario 2) were selected with importance score > 1% and were used for subgroup discovery using MDSS tool. Polygenic risk score for standard bowel cancer was indicated as a top feature (importance score = 68.9% in Scenario 1), highlighting the importance of incorporating genetic biomarkers into prediction models.

Disease subtypes: The MDSS tool can detect both large and small subgroups of differentiating patterns with statistical validity (Table 2). In both scenarios, the MDSS tool discovered and ranked four groups of patients where the outcome of subsequent major surgery within 5 years was most significantly different from the mean of the cohort. Furthermore, the MDSS tool ranks the top K mutually exclusive anomalous subgroups, also referred to here as subtypes. These are labeled as P1–P4 in both scenarios and detailed in Table 2.

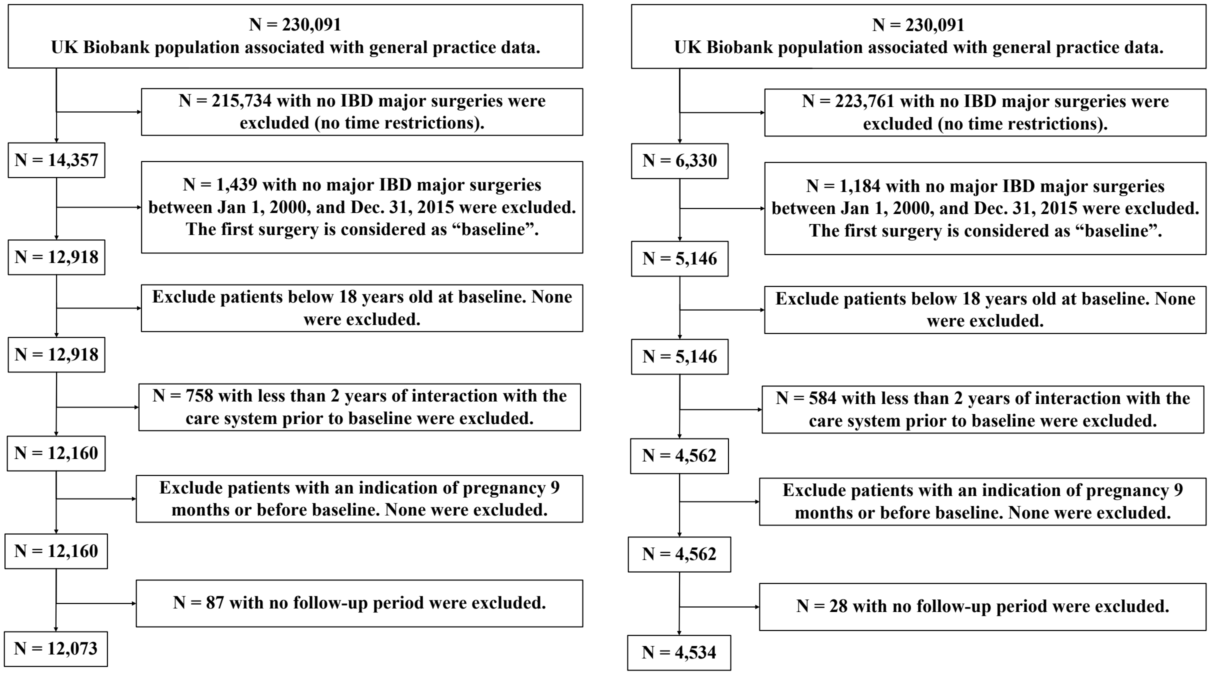
Scenario 1: Of the 12,073 patients, 2,462 (20.4%) had higher observed subsequent GI major surgery. MDSS discovered 4 unique subgroups, namely those based on present (or true) features, i.e., with *Gastrointestinal colorectal cancer* (P1), *Gastrointestinal and biliary perforation* (P2), *Regional enteritis and ulcerative colitis* (P3), and *Benign neoplasms* (P4) in male patients (Table 2). Among patients with gastrointestinal and biliary perforation (P2) rate of GI subsequent surgeries (55%) was highest and diagnosis of UC was absent in this group. This subgroup P2 may warrant further investigation. In the subgroup regional enteritis and ulcerative colitis (P3), the use of balsalazide was noted but was absent from other subgroups. Interestingly, our data-driven method was able to discover a disease subtype, a group of 440 patients (3.6%) with either CD or UC (P3), with a higher rate of subsequent major GI surgery (36%).

Scenario 2: Of the 4,534 patients, 562 (12.4%) had higher observed subsequent GI major surgery. However, this cohort was younger in age compared to the larger scenario 1 cohort suggesting the subsequent surgeries were very specific. MDSS discovered 4 unique subgroups, namely those with *Gastrointestinal and biliary perforation* (P1), *Peritonitis and intra-abdominal abscess, skin disorders in male patients with a median age of 55y* (P2), *Regional enteritis and ulcerative colitis with an established diagnosis of UC* (P3), and "*other" group of male patients with median age 56y* (P4) as specified in Table 2. Among the group of male patients (P2) with Peritonitis and intra-abdominal abscess and with skin disorders (N = 15), the outcome of GI subsequent surgery (93%) was highest followed by *Gastrointestinal and biliary perforation* (P1) group (57%), *Regional enteritis and ulcerative colitis with an established diagnosis of UC* (P3) (50%), and *"other" group of male patients with median age 56y* (P4) (37%). Therefore, all these subgroups warrant further investigation. The use of the medication azathiopirine was particularly absent in *Gastrointestinal and biliary perforation* (P1) and the *"other" group of male patients* (P4). Note that UC was absent in groups with gastrointestinal and biliary perforation (in both scenarios). This validates our methodology as perforation is pathognomonic to distinguish CD from UC. Note also the higher observed outcome rate comparing P3 (S2 vs. S1), which further validates our method as re-surgeries are more preferred in UC, which is potentially curative, than in CD, where disease is more widespread through the GI tract and surgeries are only used to treat complications.

Disease Subtype Characteristics: Distributions of selected variables for each scenario are shown in Fig. 4. Overall, the trend in both scenarios for peak time to outcome (of subsequent surgery) from index surgery varied by the subgroup but was generally within 1–2 years (Figs. 4A and 4B) except for benign neoplasm group (P4) in scenario 1 where a second peak was observed in 3–4 years post index surgery. This intrigued our clinical collaborators and initiated another iteration (Scenario 2) investigation and is an example of our iterative methodology. Furthermore, comparing Fig. 4C with 4E (smokers vs. nonsmokers), we hypothesize that smoking status plays a role in timing of subsequent major GI surgery (groups P1–P3 smokers had shorter time to the outcome). Figs. 4D and 4F (Scenario 2) indicate that smokers were associated with earlier time to events in all subgroups.

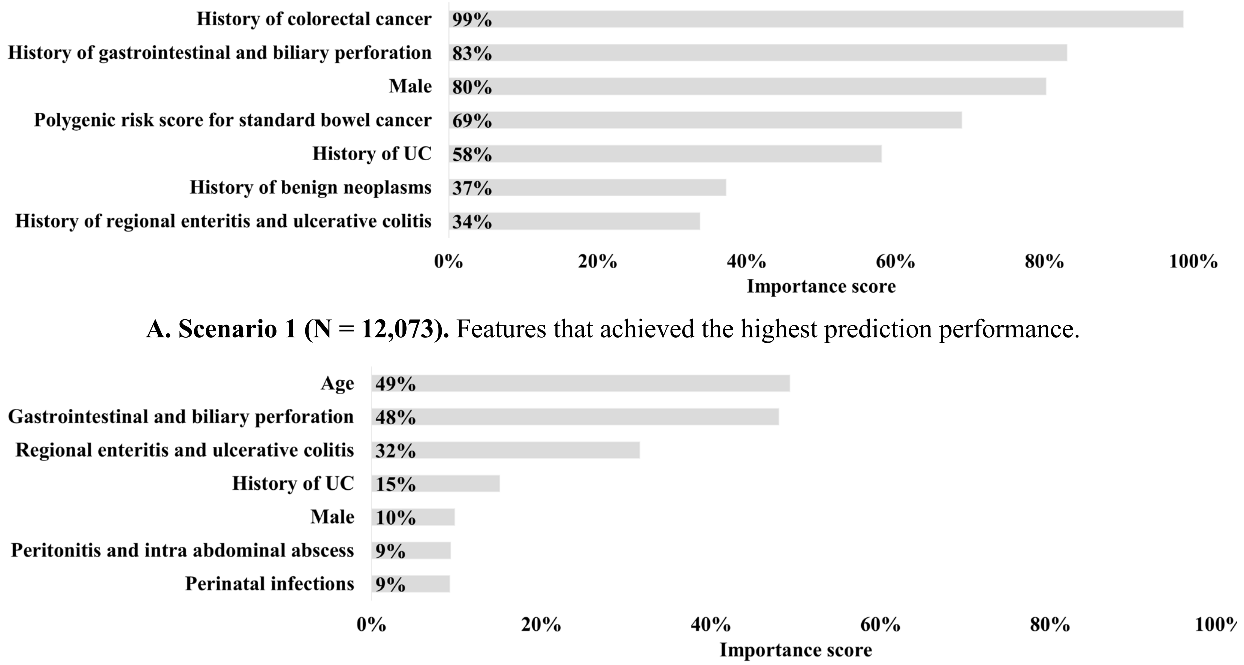
Table 1. Characteristics of GI major surgery cohort. Given the limitation of space, only a few comorbidities (those related to the digestive system) are presented, those focused on the digestive system (DIG). The most prevalent medications are presented (>1% of the population).

		Scenario 1		Scenario 2	
		All	**Subsequent surgery?**	**All**	**Subsequent surgery?**
Variable and category (N)		12,073	Yes (2,495) No (9,578) (*P*-value)	4,534	Yes (1,048) No (3,486) (*P*-value)
Outcome	**Number of patients with a subsequent surgery**	2,495 (20.7%)	-	1,048 (23.1%)	-
	Median time to outcome, years (Q1, Q3)	0.9 (0.3, 2.1)		0.6 (0.3, 1.1)	
	Subsequent major GI surgeries per 1,000 person-years	78.7		83.4	
Median follow-up, years (Q1, Q3)		2.4 (0.9, 4.8)		2.7 (0.8, 5.0)	
Age (years); Mean (Standard deviation)		61.9 (8.0)	61.6 (7.7); 61.9 (8.1) (0.06)	61.0 (8.3)	59.8 (8.1); 61.4 (8.3) (<0.001)
Male (%)		47.1	53.3; 45.5 (<0.001)	49.7	53.7; 48.5 (0.004)
Non-Black (%)		99.2	99.2; 99.3 (0.7)	99.2	99.1; 99.2 (1.0)
Current or past smoker (%)		29.2	28.1; 29.5 (0.2)	23.3	22.6; 23.5 (0.6)
Comorbid history; Prevalence (%)					
Other gastrointestinal disorders (DIG025)		40.5	37.5; 41.3 (0.001)	36.6	35.3; 37.0 (0.3)
Diverticulosis and diverticulitis (DIG013)		31.4	27.9; 32.3 (<0.001)	23.2	25.4; 22.5 (0.05)
Gastrointestinal hemorrhage (DIG021)		30.4	32.1; 30.0 (0.05)	28.4	28.8; 28.3 (0.7)
Esophageal disorders (DIG004)		25.1	22.7; 25.7 (0.002)	20.5	20.5; 20.5 (1.0)
Anal and rectal conditions (DIG015)		23.7	26.5; 23.0 (<0.001)	29.8	30.2; 29.7 (0.8)
Crohn’s (ICD10)		3.0	3.9; 2.7 (0.003)	4.2	5.8; 3.8 (0.005)
Ulcerative Colitis (ICD10)		4.7	6.7; 4.2 (<0.001)	5.4	8.1; 4.6 (<0.001)
Medication history; Prevalence (%)					
Hydrocortisone		36.0	31.6; 37.1 (<0.001)	28.5	24.8; 29.6 (0.003)
Prednisolone		20.5	19.7; 20.1 (0.3)	16.5	17.1; 16.4 (0.6)
Metronidazole		15.4	14.3; 15.6 (0.1)	13.4	12.9; 13.5 (0.6)
Budesonide		3.5	3.2; 3.6 (0.4)	2.7	2.4; 2.8 (0.5)
Dexamethasone		1.6	1.6; 1.6 (0.9)	1.4	1.0; 1.4 (0.3)
Azathioprine		1.5	2.5; 1.2 (<0.001)	2.1	3.3; 1.7 (0.001)


N = 230,091
UK Biobank population associated with general practice data.
N = 215,734 with no IBD major surgeries were excluded (no time restrictions).
N = 14,357
N = 1,439 with no major IBD major surgeries between Jan 1, 2000, and Dec. 31, 2015 were excluded. The first surgery is considered as "baseline".
N = 12,918
Exclude patients below 18 years old at baseline. None were excluded.
N = 12,918
N = 758 with less than 2 years of interaction with the care system prior to baseline were excluded.
N = 12,160
Exclude patients with an indication of pregnancy 9 months or before baseline. None were excluded.
N = 12,160
N = 87 with no follow-up period were excluded.
N = 12,073
N = 230,091
UK Biobank population associated with general practice data.
N = 223,761 with no IBD major surgeries were excluded (no time restrictions).
N = 6,330
N = 1,184 with no major IBD major surgeries between Jan 1, 2000, and Dec. 31, 2015 were excluded. The first surgery is considered as "baseline".
N = 5,146
Exclude patients below 18 years old at baseline. None were excluded.
N = 5,146
N = 584 with less than 2 years of interaction with the care system prior to baseline were excluded.
N = 4,562
Exclude patients with an indication of pregnancy 9 months or before baseline. None were excluded.
N = 4,562
N = 28 with no follow-up period were excluded.
N = 4,534

A. Scenario 1 (N = 12,073) **B. Scenario 2 (N = 4,534)**

Figure 2. Cohort selection.



A. Scenario 1 (N = 12,073). Features that achieved the highest prediction performance.

B. Scenario 2 (N = 4,534). Top 7 features.

Figure 3. Selected features identified by the subpopulation-based feature selection. Comorbidity history is based on CCSR definitions for all condition, but CD and UC in which ICD10 was used.

Table 2. Disease Subtypes from GI Major Surgery Cohort.

ID	N	Median age in years (Q1, Q3)	Observed outcome (%)	Odds Ratio [95% CI]	Differentiating Features [P = Present, A = Absent]
Scenario 1: Surgeries include 146 OPCS-4 codes as “major” surgery (at index and as subsequent surgery outcome); N = 12,073 (incidence rate = 20.7%)					
P1	1,570	64 (59, 68)	570 (36%)	2.54 [2.27–2.85] *P* < 0.001	**Colorectal cancer [P],** Regional enteritis and ulcerative colitis [A], Calculus of urinary tract [A], Ulcerative colitis [A], Foodborne intoxications [A], Balsalazide [A], Azathioprine [A]
P2	89	59.0 (55, 67)	49 (55%)	4.78 [3.14–7.27] *P* < 0.001	**Gastrointestinal and biliary perforation** [P], Intestinal infection [A], Maternal outcome – delivery [A], Uncomplicated pregnancy [A], Other pregnancy complications [A], Sinusitis [A], Colorectal cancer [A], Calculus of urinary tract [A], **Ulcerative colitis [A],** Foodborne intoxications [A], Balsalazide [A]
P3	440	59 (51, 65)	157 (36%)	2.21 [1.81–2.69] *P* < 0.001	**Regional enteritis and ulcerative colitis [P],** Other unspecified injury [A], Colorectal cancer [A], Maternal outcome – delivery [A], Diverticulosis and diverticulitis [A], Gastrointestinal and biliary perforation [A], Calculus of urinary tract [A]
P4	363	64 (60, 68)	134 (37%)	2.32 [1.86–2.88] *P* < 0.001	**Benign neoplasms [P], Male [P],** Abdominal pain and other digestive abdomen sign and symptoms [A], Neoplasms of uncertain nature or behavior [A], Uncomplicated pregnancy [A], Maternal outcome – delivery [A], Intestinal infection [A], Diverticulosis and diverticulitis [A], Gastrointestinal and biliary perforation [A], Foodborne intoxications [A], Disability live allowance [A], Azathioprine [A], Balsalazide [A], Hydrocortisone [A]
Scenario 2: Surgeries include 144 OPCS-4 codes (excluding H18* and H23.2) as “major” surgery (at index and as subsequent surgery outcome); N = 4,534 (incidence rate = 23.1%)					
P1	86	56 (51, 59)	49 (57%)	4.57 [2.95–7.05] *P* < 0.001	**Gastrointestinal and biliary perforation [P],** Nervous system signs and symptoms [A], Hypertension [A], Nutritional Anemia [A], Anxiety and fear-related disorders [A], Azathioprine [A]
P2	15	55 (48, 67)	14 (93%)	47.19 [6.2–359.26] *P* < 0.001	**Peritonitis and intra-abdominal abscess [P], Age group [33–56y], Male [P], Skin disorders [P],** lipid metabolism disorders [A], Perinatal infections [A], Hypertension [A], Viral infection [A], Secondary malignancies [A], Anxiety and fear-related disorders [A], Nervous system signs and symptoms [A], Nutritional Anemia [A], **Ulcerative colitis [A]**
P3	107	52 (47, 58)	54 (50%)	3.52 [2.39–5.17] *P* < 0.001	**Ulcerative colitis [P], Regional enteritis and ulcerative colitis [P],** lipid metabolism disorders [A], Hypertension [A], **Age group [33–56y, 56–62y],** Nervous system signs and symptoms [A], Anxiety and fear-related disorders [A], Gastrointestinal and biliary perforation [A],
P4	354	56 (50, 60)	132 (37%)	2.12 [1.69 –2.66] *P* < 0.001	**Male [P], Age group [33–56y, 56–62y],** Viral infection [A], Skin disorders [A], Upper respiratory infections [A], lipid metabolism disorders [A]

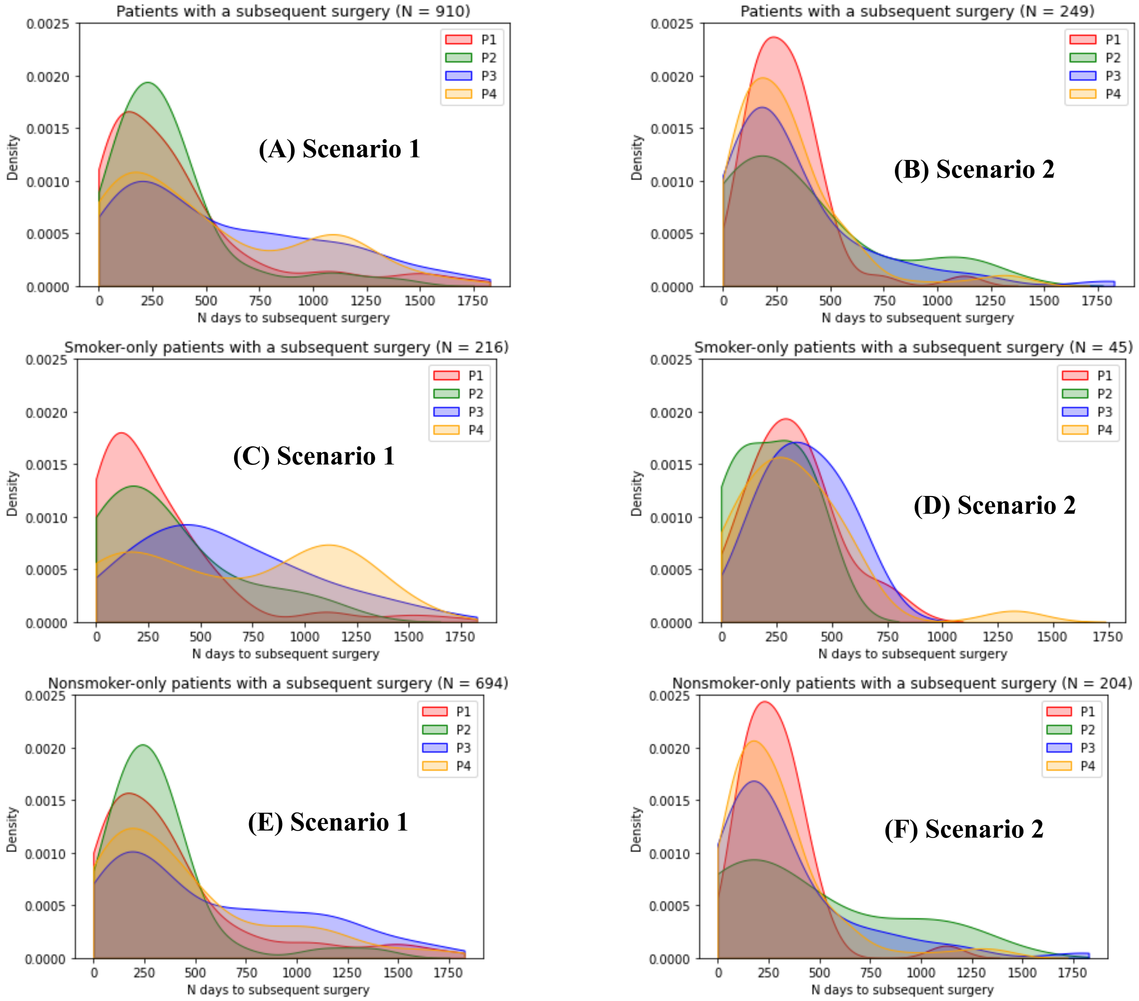


Figure 4. Density distributions.

Discussion and Conclusions

We present an iterative methodology to leverage RWD to discover important features and subgroups in relation to subsequent GI surgery outcomes. Using open-sourced data-driven methods from our toolkit, we identified a set of 3 features with highest ability to predict subsequent GI surgery, with the top predictor being colorectal cancer.[12,13] The high importance score of colorectal cancer is consistent with previous research highlighting the linkage between the conditions.[17] We observed the presence of colorectal cancer and benign neoplasms in large patient clusters of Scenario 1, a finding consistent with our identification method to include diagnostic procedures (e.g., colonoscopies) in Scenario 1 and exclude them in Scenario 2. The poor discrimination of predicting a subsequent major GI surgery (c-index of approximately 0.6) highlights the complexity of treating and following-up on patients post their first surgeries. We also highlight the importance of a variety of GI conditions including esophageal disorders, and particularly lower GI conditions such as diverticulosis/diverticulitis and anorectal disorders for higher risk of subsequent surgery. Inflammatory etiologies of diverticulosis/diverticulitis have gained significant acceptance only recently, with mesalamine treatment offered to patients with recurrent diverticulitis.

The use of cortisone is an indicator of inadequate disease control, thus expected to be a marker for repeat surgery. Azathioprine has been used as maintenance therapy in areas where penetration of biologic agents is not as high as in the United States. It is interesting that biologic use was not a predictor, but azathioprine was, suggesting inadequate control of azathioprine as a single agent to manage post-operative disease.[18,19] The smaller cohort in Scenario 2 particularly teases out GI subgroups. The subgroup identification of young male UC patients with a triad of biliary perforation, intra-abdominal abscess and skin disorders that close to universally predicts repeat surgery is a unique finding that has not been previously reported.

Our study has limitations. The UKBB may not be representative of the general population. We only included patients who underwent major GI surgery which may limit generalizability to patients with less severe disease. The prevalence of Whites was over 99.0% forming a potential for misrepresentation (our finding should be evaluated in more diverse data sources, e.g., https://allofus.nih.gov/). Our study may benefit from evaluating additional feature selection and clustering approaches. Finally, our study only considered a limited set of predictors, and may benefit from, for example, incorporating information extracted from narrative notes. In conclusion, following an iterative pipeline confirms existing knowledge and generates novel hypotheses to investigate further determinants of treatment response.

References

1. Selin KA, Hedin CRH, Villablanca EJ. Immunological Networks Defining the Heterogeneity of Inflammatory Bowel Diseases. *Journal of Crohn's and Colitis*. 2021;15(11):1959-1973. doi:10.1093/ecco-jcc/jjab085
2. Liu JJ, Abraham BP, Adamson P, et al. The Current State of Care for Black and Hispanic Inflammatory Bowel Disease Patients. *Inflamm Bowel Dis*. Published online July 11, 2022:izac124. doi:10.1093/ibd/izac124
3. Atreya R, Neurath MF, Siegmund B. Personalizing Treatment in IBD: Hype or Reality in 2020? Can We Predict Response to Anti-TNF? *Frontiers in Medicine*. 2020;7. Accessed August 24, 2022. https://www.frontiersin.org/articles/10.3389/fmed.2020.00517
4. Denson LA, Curran M, McGovern DPB, et al. Challenges in IBD Research: Precision Medicine. *Inflamm Bowel Dis*. 2019;25(Suppl 2):S31-S39. doi:10.1093/ibd/izz078
5. Li J, Qian JM. Artificial intelligence in inflammatory bowel disease: current status and opportunities. *Chin Med J (Engl)*. 2020;133(7):757-759. doi:10.1097/CM9.0000000000000714
6. Olivera P, Danese S, Jay N, Natoli G, Peyrin-Biroulet L. Big data in IBD: a look into the future. *Nat Rev Gastroenterol Hepatol*. 2019;16(5):312-321. doi:10.1038/s41575-019-0102-5
7. Ogallo W, Tadesse GA, Speakman S, Walcott-Bryant A. Detection of Anomalous Patterns Associated with the Impact of Medications on 30-Day Hospital Readmission Rates in Diabetes Care. *AMIA Jt Summits Transl Sci Proc*. 2021;2021:495-504.
8. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*. 2015;12(3):e1001779. doi:10.1371/journal.pmed.1001779
9. Ozery-Flato M, Yanover C, Gottlieb A, Weissbrod O, Parush Shear-Yashuv N, Goldschmidt Y. Fast and Efficient Feature Engineering for Multi-Cohort Analysis of EHR Data. *Stud Health Technol Inform*. 2017;235:181-185.
10. Clinical Classifications Software Refined (CCSR). Accessed March 2, 2023. https://hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp
11. Read Codes. NHS Digital. Accessed March 21, 2023. https://digital.nhs.uk/services/terminology-and-classifications/read-codes
12. Sub-population-based feature selection (SBPFS) | Computational Cardiovascular Research Group. Accessed February 24, 2023. https://www.rle.mit.edu/cb/sub-population-based-feature-selection-sbpfs/
13. Feature Selection Based on Subpopulations and Propensity Score Matching: A Coronary Artery Disease Use Case using the UK Biobank. Accessed February 27, 2023. https://knowledge.amia.org/76677-amia-1.4637602/f007-1.4641746/f007-1.4641747/580-1.4642000/269-1.4641997?qr=1
14. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat*. 2011;10(2):150-161. doi:10.1002/pst.433
15. Idrees I, Speakman S, Ogallo W, Akinwande V. Successes and Misses of Global Health Development: Detecting Temporal Concept Drift of Under-5 Mortality Prediction Models with Bias Scan. *AMIA Jt Summits Transl Sci Proc*. 2021;2021:286-295.
16. Trivedi A, Ogallo W, Tadesse GA. Hierarchical Representation of Complex Intervention Sequences for Automated Subgroup Analysis in Critical Care Settings. *Stud Health Technol Inform*. 2022;290:789-793. doi:10.3233/SHTI220187
17. Olén O, Erichsen R, Sachs MC, et al. Colorectal cancer in Crohn's disease: a Scandinavian population-based cohort study. *Lancet Gastroenterol Hepatol*. 2020;5(5):475-484. doi:10.1016/S2468-1253(20)30005-4
18. Honig G, Larkin PB, Heller C, Hurtado-Lorenzo A. Research-Based Product Innovation to Address Critical Unmet Needs of Patients with Inflammatory Bowel Diseases. *Inflamm Bowel Dis*. 2021;27(Suppl 2):S1-S16. doi:10.1093/ibd/izab230
19. Kurowski JA, Milinovich A, Ji X, et al. Differences in Biologic Utilization and Surgery Rates in Pediatric and Adult Crohn's Disease: Results From a Large Electronic Medical Record-derived Cohort. *Inflammatory Bowel Diseases*. 2021;27(7):1035-1044. doi:10.1093/ibd/izaa239