AWARD NUMBER: W81XWH-17-1-0556

TITLE: Utilizing Clinical Metadata to Predict High-Cost Complications and Treatment Response in IBD: Development of Clinical Decision Support Tools

PRINCIPAL INVESTIGATOR: David G. Binion

CONTRACTING ORGANIZATION: University of Pittsburgh Office of Research
123 University Place
Pittsburgh, PA 15213-2303

RECIPIENT: Catherine C. Henry

REPORT DATE: December 2021

TYPE OF REPORT: Final Report

PREPARED FOR: U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
DECEMBER 2021	Final	01 Sept 2017 - 31 Aug 2021

4. TITLE AND SUBTITLE

Utilizing Clinical Metadata to Predict High-Cost Complications and Treatment Response in IBD: Development of Clinical Decision Support Tools

5a. CONTRACT NUMBER
W81XWH-17-1-0556

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
David G. Binion and Claudia Ramos Rivers

E-Mail: binion@pitt.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Pittsburgh Office of Research
123 University Place
Pittsburgh, PA 15213-2303

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

IBD is a costly and debilitating disease, significantly affecting quality of life. Our research plans are to generate easy to use, internet based tools (similar to a calculator) to determine which patient will go on to have costly disease over the next several years, and/or is unlikely to respond to traditional biologic therapies with anti-TNF medications. We propose using an already available IBD patient registry database which has been developed by the P.I. and the research team at UPMC/University of Pittsburgh.

The short-term goal is to use accessible patient information and routinely collected prospective clinical data derived from the electronic medical record from over 3,000 IBD patients followed for >7 years, to generate personalized prediction models and tools to assess response to biologic therapy and risk of high costs complications, including enteric infection and disability for the care of patients with IBD. We will generate a publicly accessible computer-based risk prediction calculator that allows for risk stratification after entering routinely collected patient information. The goal of this web-based technology will be to use routine clinical information to facilitate a personalized clinical approach for treatment and stratification of IBD patients based on severity and phenotype.

Personalized approaches for IBD treatment will help to avoid unnecessary exposure to biologic therapies and their associated risks in patients likely to fail a standard biologic treatment (i.e., anti-TNF) approach. Similarly, identifying patients that are at risk for future high-cost complications will provide a window of opportunity for cost-saving outpatient care, proactive lifestyle modifications and dietary interventions to prevent hospitalization, surgery, infectious complications, or disability.

This personalized approach to IBD treatment will positively impact patients and their experience with disease, avoiding risks and given the opportunity for early interventions to avoid debilitating disease complications. Personalization of care will also benefit those taking care of IBD patients, as it will provide insight into disease subgroups and treatment choices, saving time and financial resources from the health system.

15. SUBJECT TERMS

Inflammatory Bowel Disease, anti-Tumor Necrosis Factor, Electronic Medical Records, Short Inflammatory Bowel Disease Questionnaire, Hemoglobin, Crohn's Disease, Ulcerative Colitis

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRDC
Unclassified	Unclassified	Unclassified	Unclassified	34	19b. TELEPHONE NUMBER *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

TABLE OF CONTENTS

1. **INTRODUCTION:** Narrative that briefly (one paragraph) describes the subject, purpose and scope of the research.

Using readily accessible patient demographics and routinely collected prospective clinical data harvested from the electronic medical record (EMR) from >3,000 consented IBD patients followed for >7 years, to generate personalized prediction models to assess response to anti-TNF biologic therapy and risk of high cost complications, including enteric infection and disability for the care of patients with inflammatory bowel disease (IBD). We will generate an accessible computer-based risk prediction platform that allows for risk stratification after entering routinely collected patient demographic and clinical information.

2. **KEYWORDS:** Provide a brief list of keywords (limit to 20 words).

Inflammatory Bowel Disease, anti-Tumor Necrosis Factor, Electronic Medical Records, Short Inflammatory Bowel Disease Questionnaire, Hemoglobin, Crohn's Disease, Ulcerative Colitis

3. **ACCOMPLISHMENTS:** The PI is reminded that the recipient organization is required to obtain prior written approval from the awarding agency grants official whenever there are significant changes in the project or its direction.

What were the major goals of the project?
List the major goals of the project as stated in the approved SOW. If the application listed milestones/target dates for important activities or phases of the project, identify these dates and show actual completion dates or the percentage of completion.

1. Develop a clinical decision support tool for identifying IBD patients at risk of complicated disease.
2. Develop a prediction model for anti-TNF treatment failure.

What was accomplished under these goals?
For this reporting period describe: 1) major activities; 2) specific objectives; 3) significant results or key outcomes, including major findings, developments, or conclusions (both positive and negative); and/or 4) other achievements. Include a discussion of stated goals not met. Description shall include pertinent data and graphs in sufficient detail to explain any significant results achieved. A succinct description of the methodology used shall be provided. As the project progresses to completion, the emphasis in reporting in this section should shift from reporting activities to reporting accomplishments.

The source code and documentation associated with this project are available for review in the project's GitHub repository at https://github.com/dbabichenko/ibd_dod_final

Major tasks 1 and 4:

Perform data queries of the IBD Registry database:

Developed pre-processing and data cleaning Python scripts for the following data sets:

a. Laboratory tests
b. Medications
c. Harvey-Bradshaw questionnaire
d. SIBDQ questionnaire

Developed a protocol for dealing with missing values for the following data sets:
a. Laboratory tests
b. Harvey-Bradshaw questionnaire
c. SIBDQ questionnaire

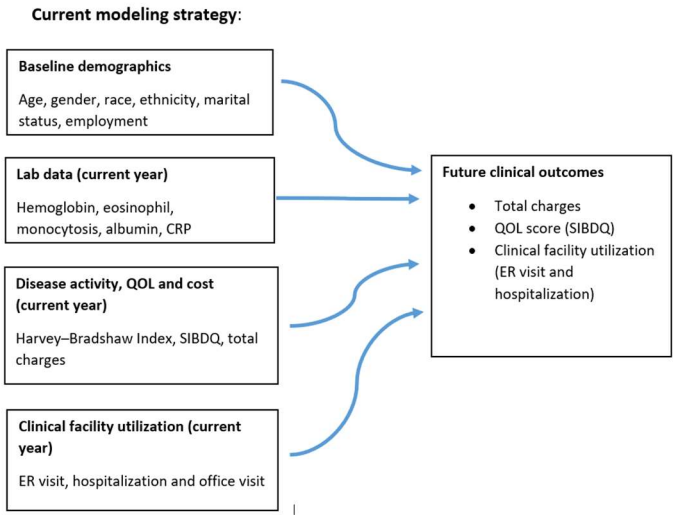
Develop novel factors associated with complicated/high-cost disease and Identification of IBD patients suffering from complex, high-cost disease trajectories using Bayesian networks. Identification and characterization of anti-TNF treated patients using registry data queries.

Creation of discovery and validation cohorts for independent confirmation of findings. Deployed updated MySQL schema for 2018 data import.

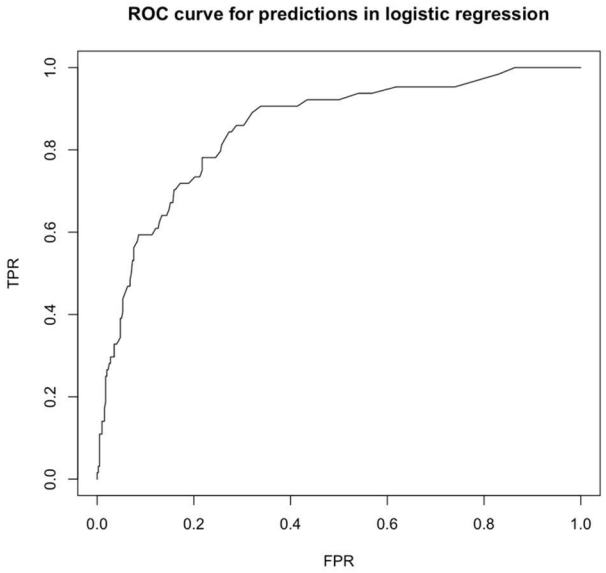
<u>Major tasks 2, 5 and 6</u>

Identify association of clinical/demographic/biochemical factors with complicated/high cost disease. Development of a prediction model for incurring high medical cost during the next years. Identify association of clinical/demographic/biochemical factors with anti-TNF response or anti-TNF treatment failure. Creation of anti-TNF failure in IBD prediction model.

Development of statistical models to predict medical charges, quality of life outcome SIBDQ scores, and use of medical facility (emergency room visit and hospitalization) with historical data from a pre-determined training set and assessed its performance in a separate testing set.



Results from an internal validation of a prediction model on the total charges (>$15K) of a future year:



Completed data de-identification automation scripts
Completed automation of training dataset generation
Completed initial predictor selection

Predictors (Independent variables):

a. **Demographic**: gender, marital status, employment status, age, distance to clinic, median income
b. **Medications**: 5-ASA, immunomodulators, systemic steroids, vitamin D, biologics
c. **Encounters**: procedure visit, contact with provider (office visit, email, phone)
d. **Labs**: Eosinophils, monocytes, inflammation markers, hemoglobin, vitamin D
e. **Quality of life (SIBDQ)**: SIBDQ total score, Question 1: feeling fatigue, question 8: feeling relaxed
f. **Psychiatric comorbidities** (including SIBDQ scores related to stress)

Response (Dependent variable):

a. Treatment cost

We have applied cluster analyses to screen patients into groups with different prognosis and differential response to standard IBD therapies and anti-TNF treatments.

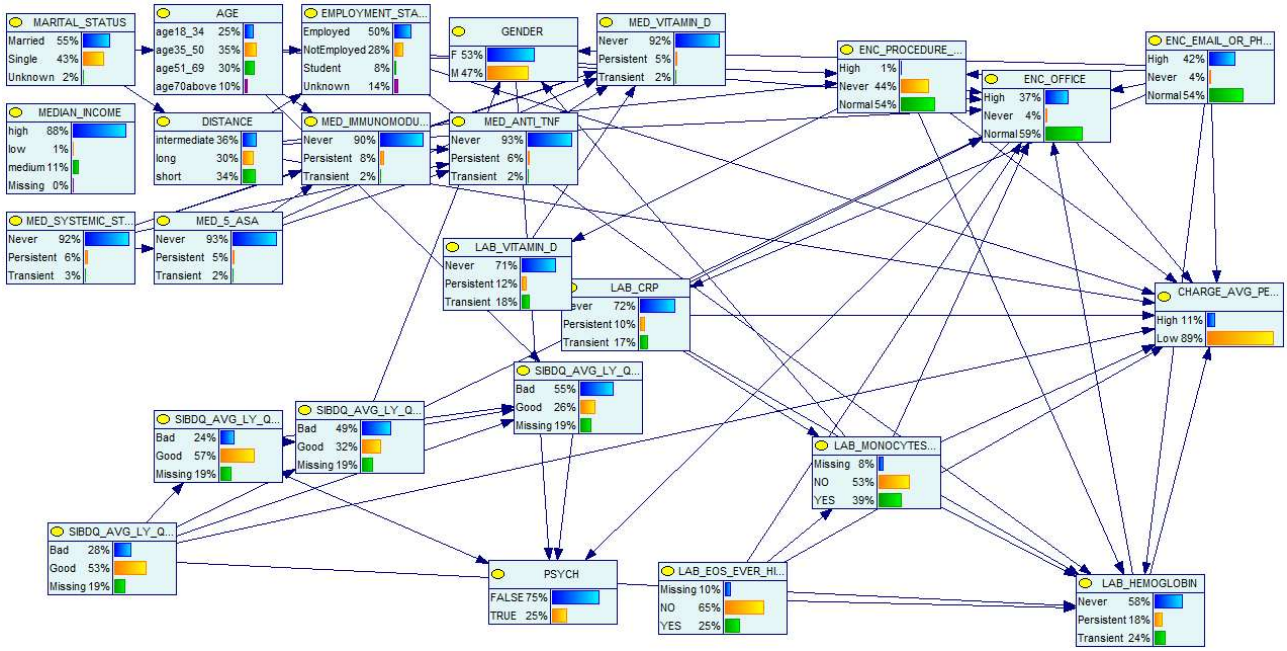
Created and evaluated a series of classification and probabilistic models trained from the IBD registry dataset

Table 1: Preliminary Classification Models

Classification Algorithm	Accuracy Score	Cross-validated Score (k = 10)
SVM with rbf kernel	0.908	0.910
Naïve Bayes	0.858	0.853
SVM with linear kernel	0.896	0.901
SVM with poly kernel	0.923	0.914
KNN with 3 neighbors	0.910	0.87
Random Forest	0.917	0.891
Decision Tree	0.903	0.898
Logistic Regression	0.929	0.911

Table 2: Preliminary Bayesian Network Model

Training Algorithm	Accuracy Score	Cross-validated Score (k = 10)
PC	0.892	0.877
Tree-augmented Naïve Bayes (TAN)	0.901	0.899



Using charges as a proxy for clinical outcomes we compared multiple classification models for predicting poor clinical outcomes. SVM + linear kernel models performed best (in some cases as good as 92% predictive accuracy) at predicting high-charge patients

By the end of year 2, we rewrote much of the computational logic for rating predictors using Random Forest regressor and classifier from Python's sklearn machine learning library. This approach allowed us to compare procedurally selected predictors with predictors identified with experts and through literature reviews.

We explored several non-modeling approaches for identifying poor clinical outcomes patients. A patient cohort and treatment pathways visualization system that integrates elements of exploratory search showed the most promise as a viable approach to developing a useful clinical decision support system. The two current approaches include using a Sankey chart to illustrate how different

clinical decisions result in different outcomes (Figure 1) and an exploratory search decision tree that illustrates immediate outcomes of every clinical decision (Figure 2).

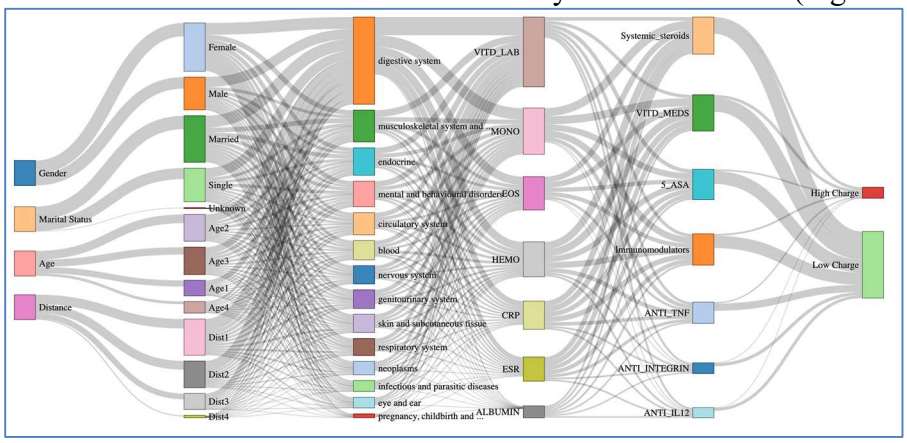


Figure 1: Sankey chart to illustrate how different clinical decisions result in different outcomes

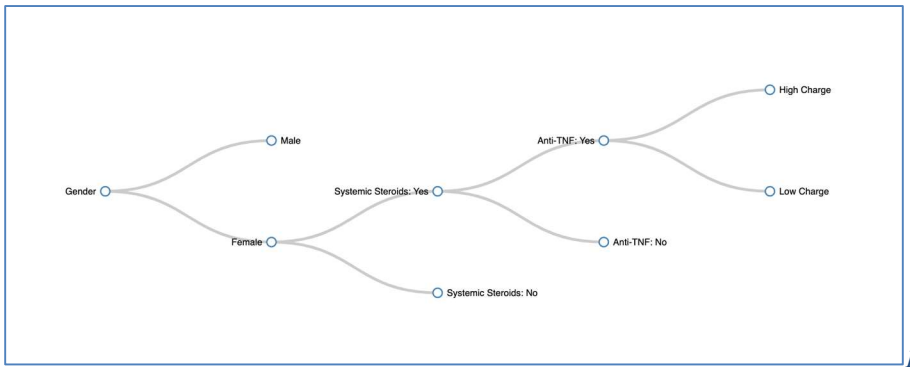


Figure 2: Decision tree-based user interface to support exploratory search of patient outcomes based on clinical decisions and demographic information

By examining the pattern of nodes and edges in Figure 3, we can quickly identify those patients 1 and 2 (P1 and P2) had encounters with the same doctor (D1) and that patient 2 (P2) was on two medications simultaneously and had an adverse reaction following these medication events. Representing patient clinical data as a graph affords numerous computational advantages over the relational model, including abilities to (1) quickly identify and extract patient cohorts based on similarity metrics, (2) visualize patient disease progression trajectories, (3) identify possible causal relationships in the data.

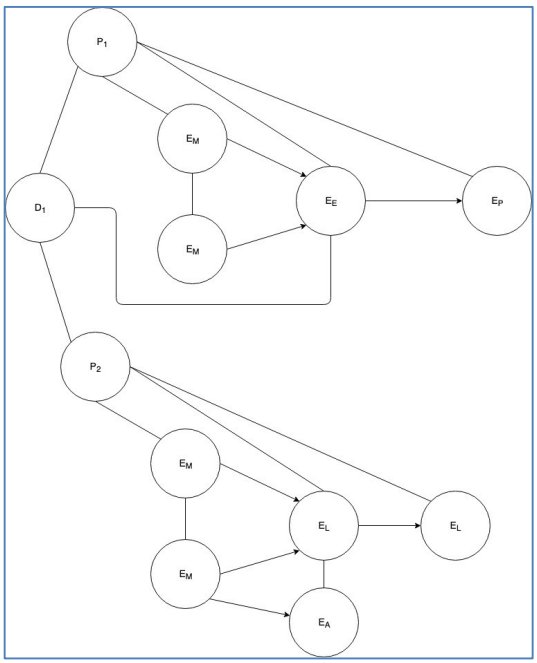


Figure 3: Example of patient data model represented as a graph

Figure 3 shows a possible representation of patient clinical data as a graph with three types of nodes: patient (P), event (E), and doctor (D). This example shows five types of events - medication (EM), encounter (EE), procedure (EP), laboratory test (EL), and adverse / allergic reaction (EA). Directed edges (arrows) indicate temporal precedence (i.e. event 1 occurred before event 2). Undirected edges (lines without arrows) indicate co-occurrence (i.e. event 1 and event 2 occurred at the same time).

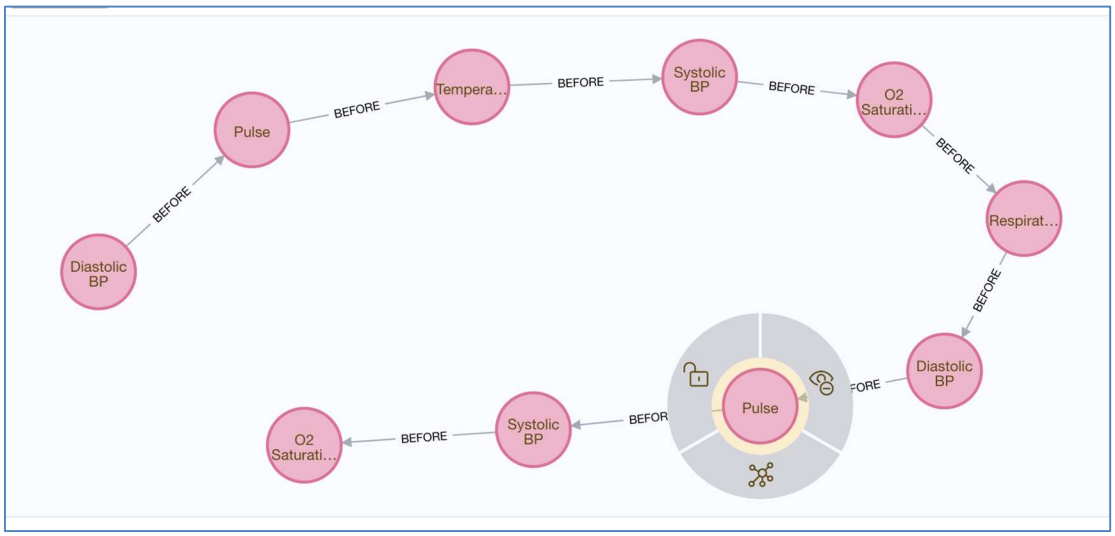


Figure 4: A temporal chain of events for a single patient represented by a graph. This data is modeled as a graph database using Neo4j graph DB engine

Major task 3:

Model development for web-based clinical decision support tool

We began developing a web-based decision support system that combines the visualization approaches described above with predictions (classifications) produced by XGBoost models. Figure 5.

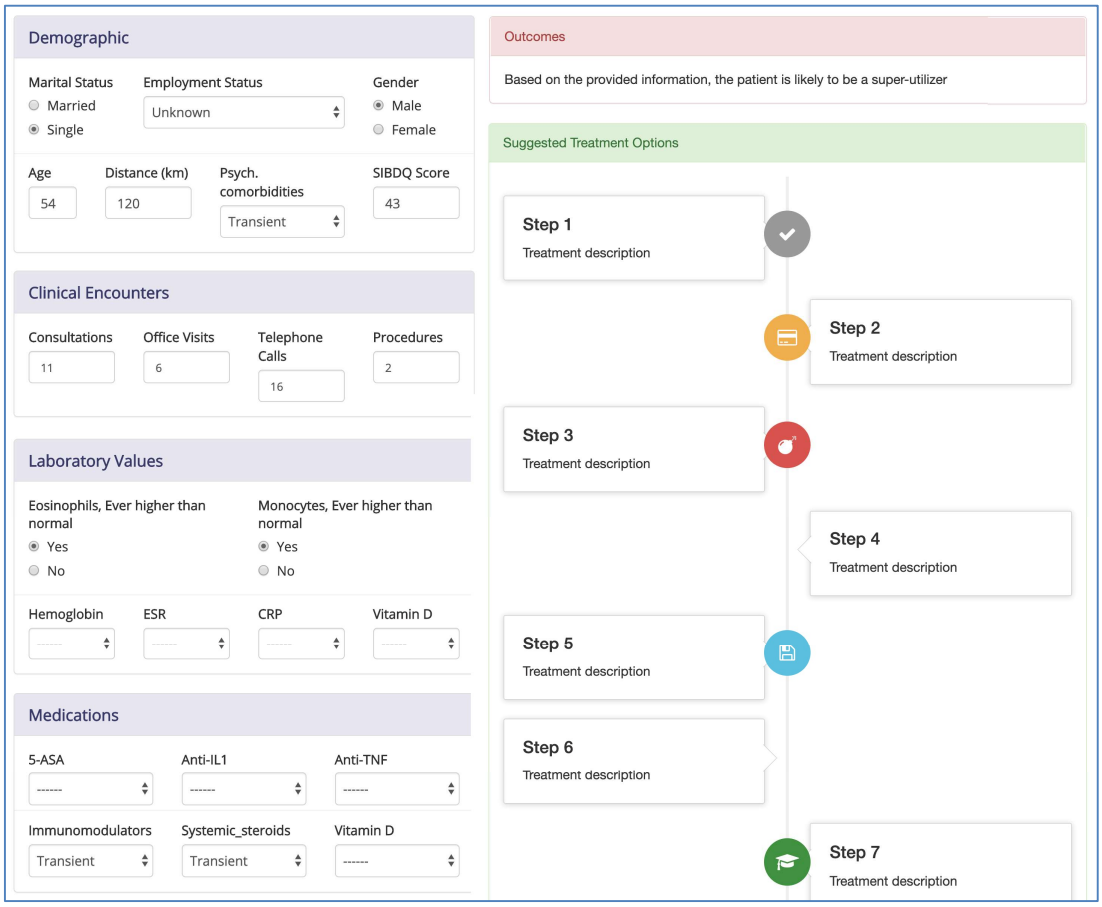


Figure 5: Proposed decision support system user interfacBy Year 3, we completed developing and validating Artificial Neural Network (ANN) - based approaches to IBD modeling. We have developed and validated a series of six machine learning models trained on a patient dataset consisting of 24-month treatment windows as predictors and the following 8 months as response variables. We have developed a data processing and model training pipeline in Python using the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) ANN architectures.

Table 3 presents the evaluation results of eight GRU-D models and two baseline Multilayer Perceptrons models. The four variations of multitask loss coefficients and both baselines are trained and tested on the Chronological Holdout (HO) and Chronological Overlap (OL) data splits.

Both baseline models bias their predictions to Low-Charge; Baseline-OL predicts 90% of the 2,192 High-Charge sequences as Low-Charge. In comparison, the best performing GRU-D-OL model has a recall of 34% on the High-Charge patients and predicts another 32% as Mid-Charge.

Among the GRU-D models, there is no clearly dominating performance in Table 3; from the metrics, it is unclear how multitask learning and target replication are contributing to model performance. However, it must be noted that the GRU-D models are trained only on 50 epochs for this experiment; these are probably not enough runs for the complex model to fit all its parameters to the data.

Unfortunately, due to high variance and high noise in the data, the models currently overfit and may not generalize to other IBD datasets. At the time of this writing, we believe that this approach shows the most potential out of all the modeling approaches explored in this project. To validate the full potential of this model architecture, this approach would have to be trained and evaluated on a larger and more homogeneous dataset (Aim 1).

Note: Exploration and evaluation of LSTM and GRU ANN models resulted in a masters' thesis for Mr. Suraj Subramanian, a graduate researcher from the University of Pittsburgh School of Computing and Information who worked on this project in 2019 and 2020. Complete description of the architecture, configuration, data cleaning and organization, and model training is available in Mr. Subramanian's thesis, found in this project's GitHub repository at https://github.com/dbabichenko/ibd_dod_final/blob/main/suraj_subramanian_thesis_final/suraj_subramanian_ms_thesis_final.pdf

Table 3: Evaluation results of eight GRU-D models and two baseline Multilayer Perceptrons models

Model	α_{aux}	α_{tr}	LogLoss	Kappa	MCC	AUPRCHC	R@P=0.4	Accuracy
Baseline-OL			0.442	0.101	0.200	0.272	0.221	0.872
GRU-D on OL	0	0	0.645	0.223	0.234	0.248	0.182	0.759
	0	0.5	0.584	0.228	0.231	0.274	0.228	0.796
	0.5	0	0.594	0.229	0.241	0.248	0.198	0.763
	1	1	0.589	0.236	0.243	0.267	0.211	0.778
Baseline-HO			0.369	0.014	0.079	0.213	0.119	0.904
GRU-D on HO	0	0	0.672	0.154	0.179	0.222	0.203	0.733
	0	0.5	0.814	0.099	0.136	0.174	0.119	0.606
	0.5	0	1.010	0.094	0.144	0.250	0.222	0.542
	1	1	0.806	0.141	0.180	0.268	0.241	0.677

We also completed a prototype of a visual interactive decision support system tool (Figures 6 and 7), have received an IRB approval to begin a user study to validate the design and the usefulness of this tool. We are currently in the process of evaluating the tool with IBD clinicians (Aim 1 and 2).

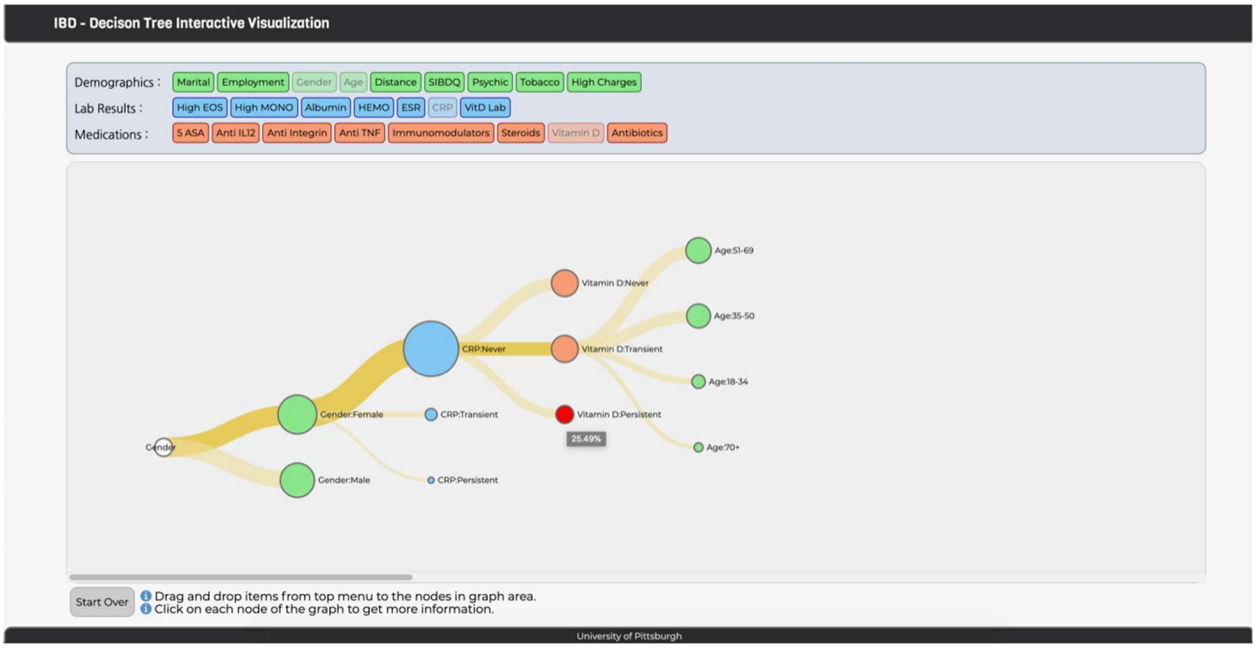


Figure 6: IBD Decision Tree Interactive Visualization

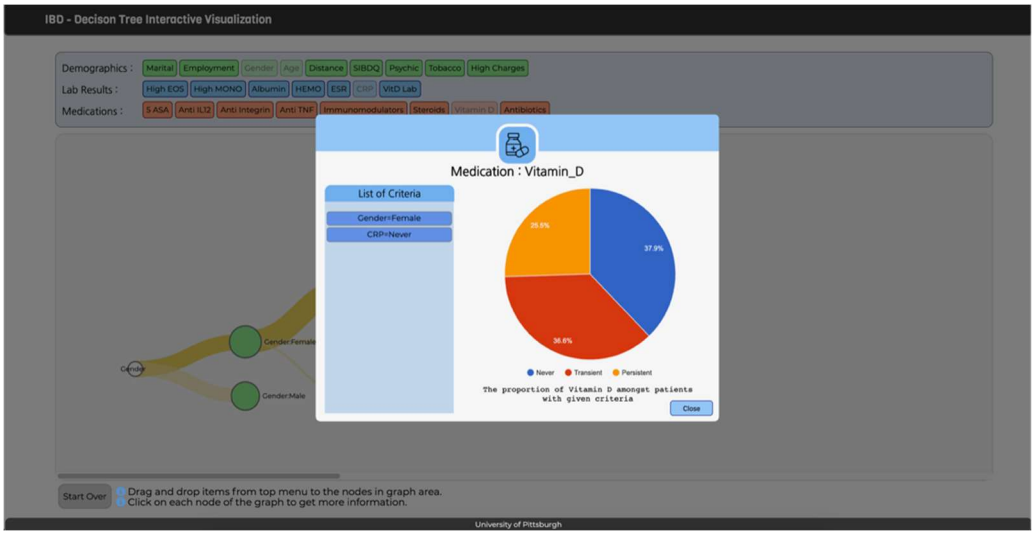


Figure 7: IBD Decision Tree Interactive Visualization – Details View

We implemented a web-based decision support system (DSS) based on a model that relies on high treatment charges as a proxy for clinical outcomes to identify and classify IBD patients at risk of complicated disease. The DSS user interface is shown in Figure 8. In addition to the DSS itself, we have also implemented a RESTful API framework to allow the system to interface with multiple models and to provide data to other user interfaces (Figures 6 & 7).

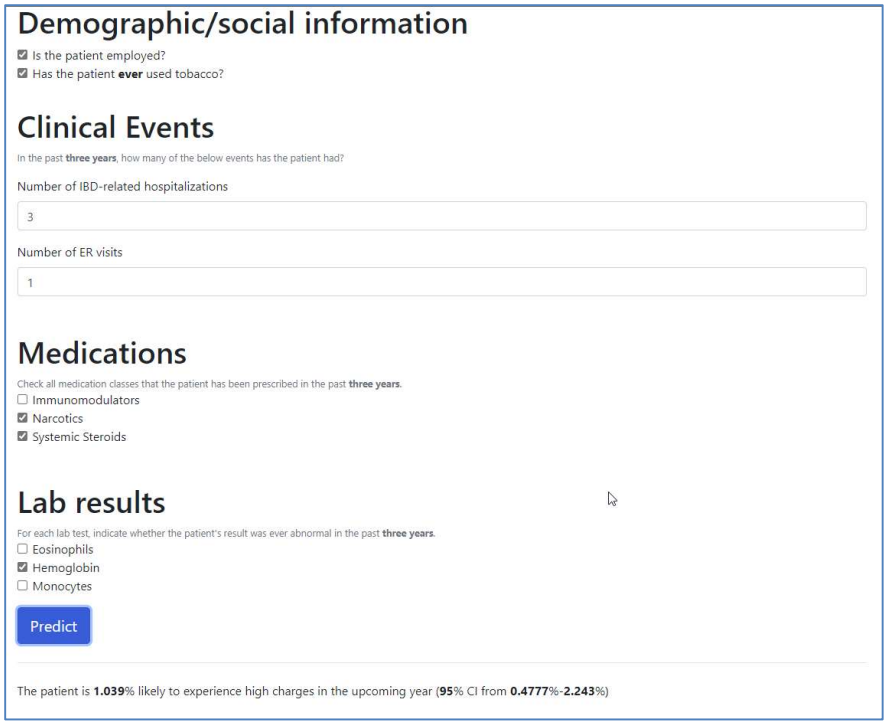


Figure 8: A web-based decision support system (DSS)

In recent years there have been some success with modeling patient trajectories using graph structures. To test the validity of this approach with IBD patient data, we have developed a temporal graph representation of treatment-related events. The current version of the graph representation is hosted within a Neo4J graph database. This graph (Figure 9) contains three major node types – *Patient*, *Diagnoses*, and *Event*. An *Event* node represents any type of event that took place during a patient's treatment – clinical encounter, procedure, change in vitals, medication prescriptions, etc. *Patient* is connected to *Diagnoses* via *DIAGNOSED-WITH* relationship; *Patient* is connected to *Event* with *HAS* relationship; *Event* nodes are connected to each other with *CO-OCCURS*, *PRECEEDS*, or *FOLLOWS* relationships. This design allows us to explore cohort relationships up to 15 times faster than with traditional relational databases. Moreover, a graph shows a natural representation of temporal relationships and allows us to explore patterns that are obscured by the traditional relational database design. At the time of this writing, we began to explore the use of graph embeddings to represent similar patients and to rapidly identify IBD patients who are at risk of complicated disease (**Aim 1**) and IBD patients who are at risk of poor response to anti-TNF biologic therapy (**Aim 2**).

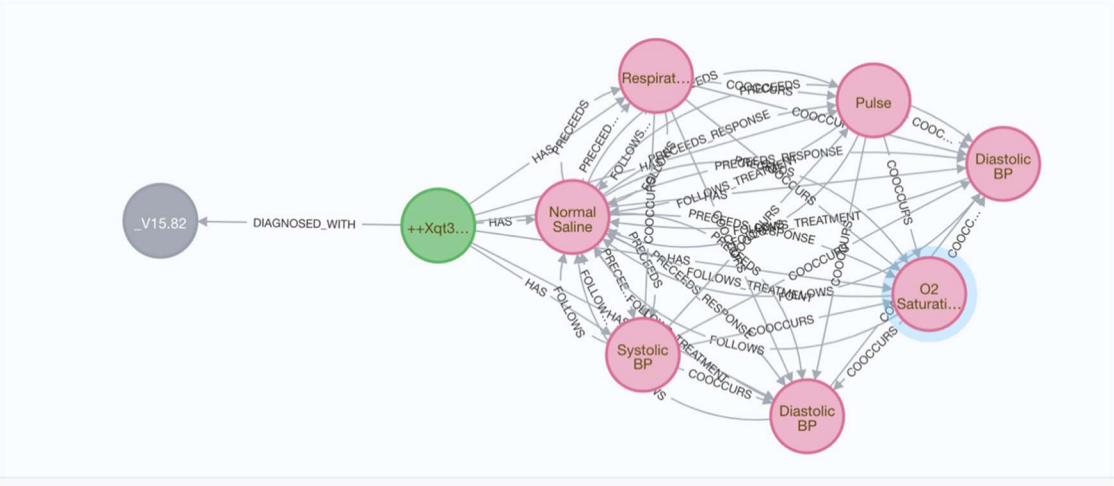


Figure 9: Partial graph representation of a single patient's treatment timeline

Steps completed towards building a graph representation of IBD data:

1. *Data preparation and Preprocessing:*
   The first step toward constructing the IBD graph is to ensure the quality of input data is consistent throughout the range of various datasets.
   The following dataset have been used to construct the graphs:
   - **Patients**: a complete set of demographic and geographic information about each patient in the dataset.
   - **Laboratories Test**: a sequence of lab test information and results for all the patients.
   - **Medications**: a sequence of prescribed medication for all the patients.
   - **Procedures**: a sequence of all the proceducedured that have been performed on a patient.
   - **Problems**: a sequence of all IBD related comorbidities that a patient has been diagnosed with during the course of treatment.
   - **Encounters**: a sequence of all patient's encounters with the medical center during the course of treatment.

2. *Creating the nodes:*

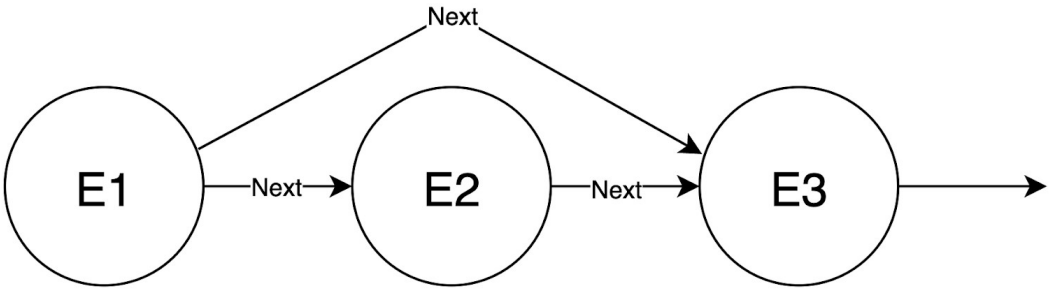


Figure 10: Example event nodes from the knowledge graph.

The first approach creates a significantly smaller graph but in return performing more complicated queries will become significantly more challenging. The second approach creates a denser graph which supports more complicated, easier and faster querying.
Based on the size of our dataset and the required queries we decided to choose the second approach.

Sample Queries:

Finding a sequence of low lab results in two categories of "hemoglobin" and "albumin" for patient 5:

```
Match(m:Lab {patientId:"5"})
where m.resultFlag <> "normal" and (m.group = 'hemoglobin' or
m.group = 'albumin')
return m
limit 10
```

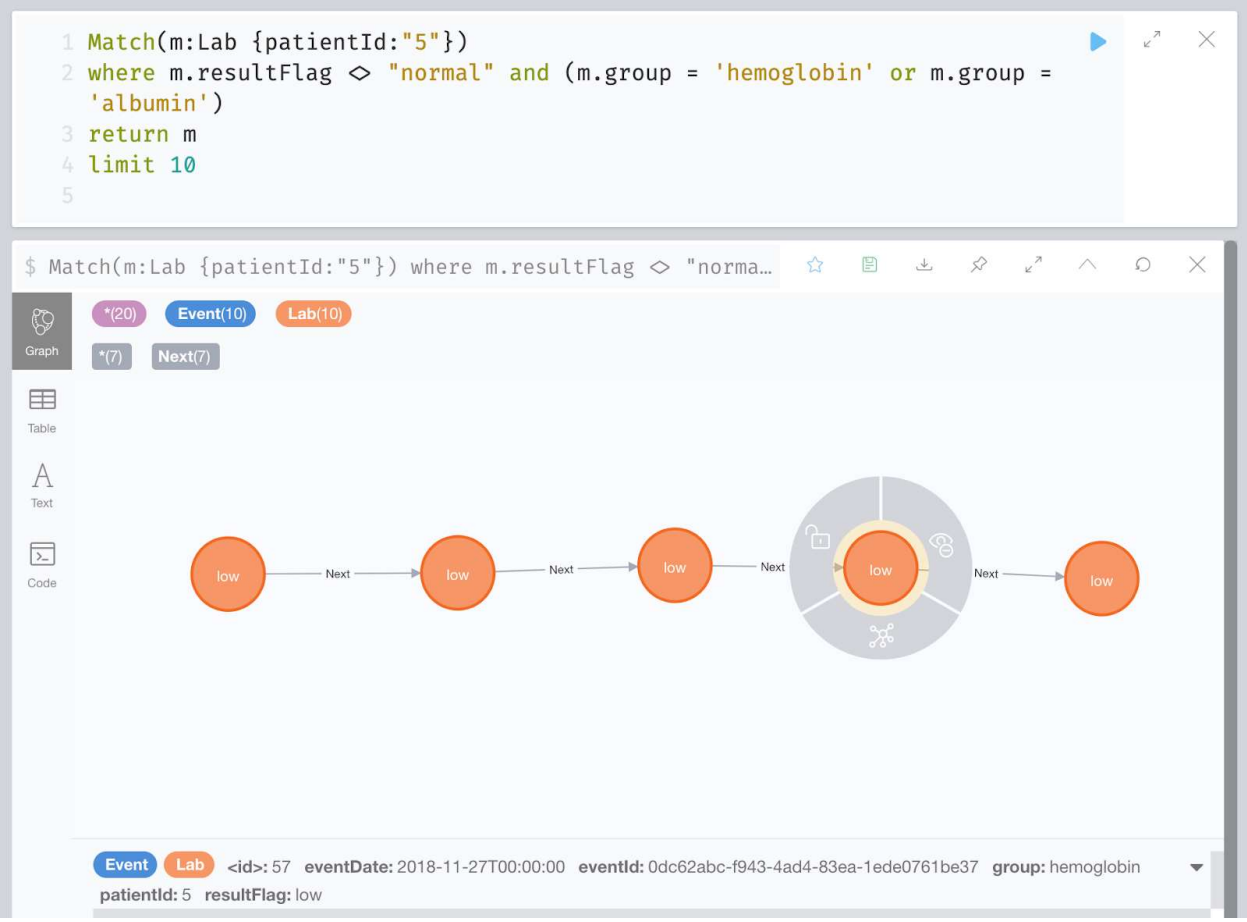


Figure 11: Sample query results from Neo4J representing a temporal event chain of clinical events related to a single patient.

Finding hemoglobin lab results and their corresponding distance for all patients

```
Match(e1:Lab)-[r:Next]->(e2:Lab)
where e1.group = 'hemoglobin' and e2.group = 'hemoglobin'
```

```
with e1.patientId as pid, duration.inDays(date(left(e1.eventDate,
10)), date(left(e2.eventDate, 10))).days as dis , e1.resultFlag
as f1, e2.resultFlag as f2
return pid, dis,f1,f2
order by dis
```

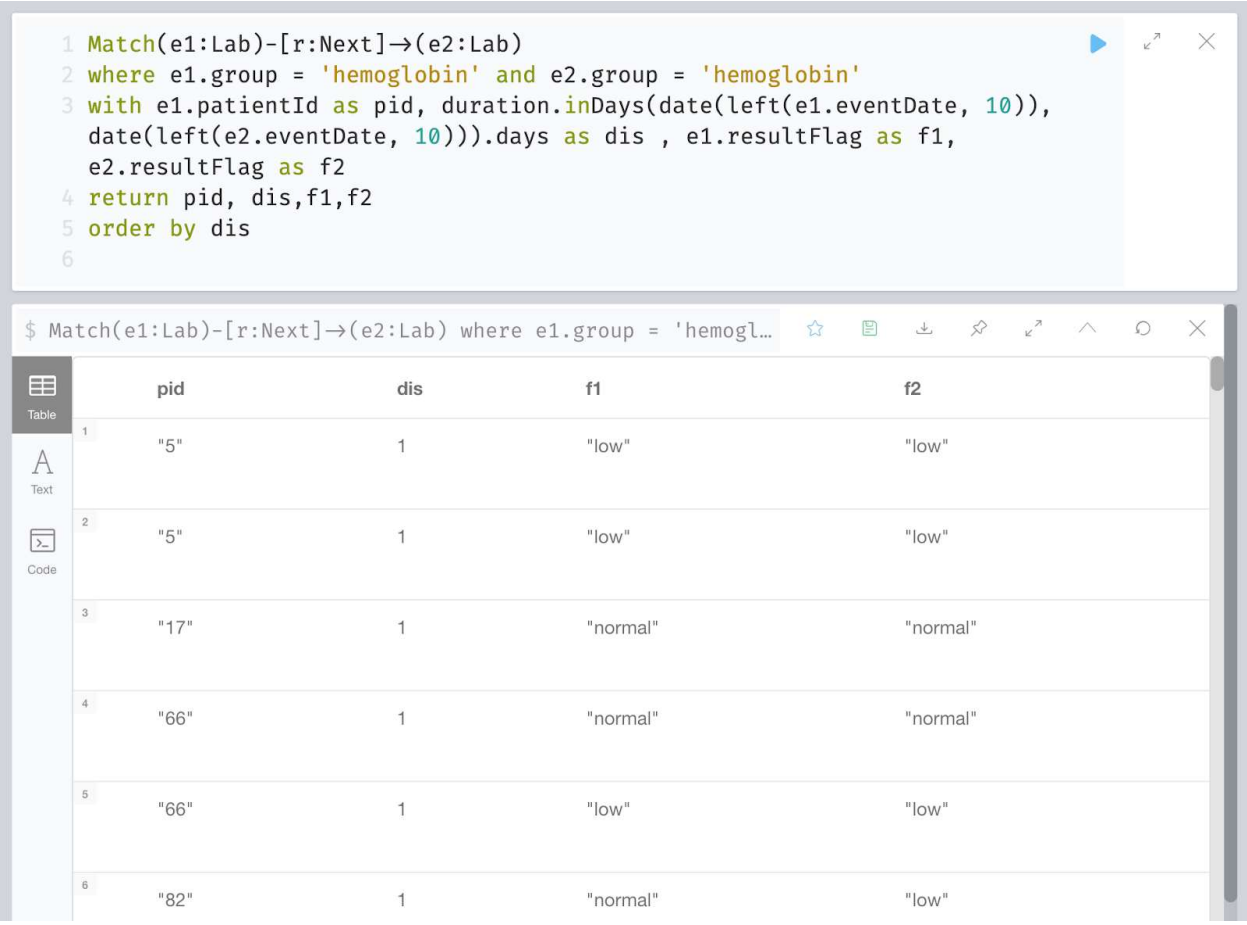


Figure 12: Sample query results from Neo4J representing hemoglobin lab results and their corresponding distance for all patients

Finding a complete list of a specific patient events and the distance between each event

```
Match(e1:Event)-[r:Next]->(e2:Event)
where e1.patientId = '5'
return LABELS(e1) as t1, LABELS(e2) as t2,  e1.eventDate as d1,
e2.eventDate as d2, duration.inDays(date(left(e1.eventDate, 10)),
date(left(e2.eventDate, 10))).days as dis
order by e1.eventDate, e2.eventDate
```

```
Match(e1:Event)-[r:Next]→(e2:Event)
where e1.patientId = '5'
return LABELS(e1) as t1, LABELS(e2) as t2,  e1.eventDate as d1,
e2.eventDate as d2, duration.inDays(date(left(e1.eventDate, 10)),
date(left(e2.eventDate, 10))).days as dis
order by e1.eventDate, e2.eventDate
```

$ Match(e1:Event)-[r:Next]→(e2:Event) where e1.patientId =…

Table | Text | Code

	t1	t2	d1	d2	dis
1	["Event", "Lab"]	["Event", "Med"]	"2011-08-15T00:00:00"	"2011-09-15T00:00:00"	31
2	["Event", "Lab"]	["Event", "Med"]	"2011-08-15T00:00:00"	"2011-09-15T00:00:00"	31
3	["Event", "Lab"]	["Event", "Med"]	"2011-08-15T00:00:00"	"2011-09-15T00:00:00"	31
4	["Event", "Lab"]	["Event", "Med"]	"2011-08-15T00:00:00"	"2011-09-15T00:00:00"	31
5	["Event", "Lab"]	["Event", "Med"]	"2011-08-15T00:00:00"	"2011-09-15T00:00:00"	31
6	["Event", "Lab"]	["Event", "Med"]	"2011-08-15T00:00:00"	"2011-09-15T00:00:00"	31

Figure 13: Sample query results representing a complete list of a specific patient events and the distance between each event

Last, but not least, we have trained and evaluated a series of machine learning models based on the aggregate data from the first three continuous years of each patient's treatment (three-year static dataset), as well as based on a rolling three-year window to predict outcomes in the fourth year. In other words, the predictors were aggregated from the first three years of each patient's data, and the outcome (the response variable) was from the fourth year (Figure 14).

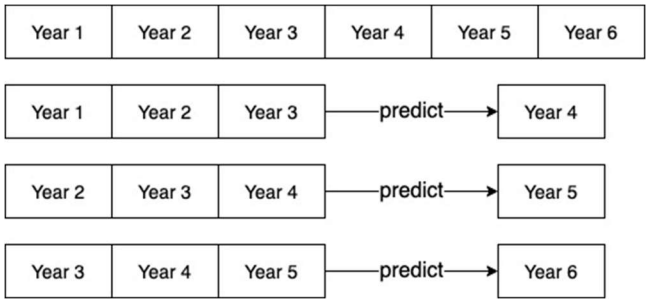


Figure 14: Illustration of generating a three-year rolling window dataset where aggregate data from three continuous years of treatment is used to predict outcomes in the fourth year of treatment.

A total of twelve ML models were trained and validated, with four models trained and validated using each of the datasets. These models were trained using Random Forest (RF), Support Vector

Machine (SVM) with a linear kernel, Gradient Boosted Trees (GBT), and a feedforward artificial neural network (ANN). Random Forest (RF), SVM, and GBT models were trained using the scikit-learn machine learning library in Python. The ANN was trained using the TensorFlow Keras framework's Sequential class and cross-validated using the scikit-learn library. All models' hyperparameters were tuned using grid search.

All models were validated using 10-fold cross-validation; cross-validated accuracy scores and AUC (area under receiver operating characteristic curve (ROC) curve) were used as metrics to identify and select best-performing classification models.

Table 4: Models' 10-fold Cross-Validated Accuracy and AUC Scores

	Master Dataset		**Three-Year Static Dataset**		**Three-Year Rolling Window Dataset**	
Algorithm	**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
Random Forest	0.876	0.733	0.823	0.741	0.729	0.689
Support Vector Machine (SVM)	0.867	0.671	0.799	0.602	0.626	0.591
Gradient-Boosted Trees (GBT)	0.891	0.748	0.847	0.724	0.732	0.699
Feedforward ANN	0.927	0.782	0.729	0.631	0.613	0.526

When trained on the largest (master) dataset, the ANN model outperformed all other models in terms of accuracy and AUC, with the GBT model coming in close second. However, as the size of the training data decreased in the static three-year dataset and decreased even further in the three-year rolling window dataset, the ANN's accuracy and AUC dropped in comparison with the respective GBT models.

Complete description of this experiment, including feature engineering, approaches for dealing with missing values, model training and model evaluation are available in the manuscript titled "Classification Models For Predicting Inflammatory Bowel Disease Healthcare Utilization." As of the time of this writing, this manuscript had been accepted for presentation and proceedings publication in the 15th International Conference on Healthcare Informatics (HEALTHINF2022). The manuscript can be previewed from this project's GitHub repository at https://github.com/dbabichenko/ibd_dod_final/blob/main/classification_models_healthinf_2022/HEALTHINF_2022_58_CR.pdf

Beyond the above effort, we strategically considered multiple statistical machine learning tools, including generalized estimating equations and classification and regression trees, to identify IBD patients with complicated disease courses using demographic and clinical predictors recorded under various time frames. Three proxy clinical outcomes were considered: 1) to incur over $100K medical charge during the next year; 2) to have at least two hospitalizations during the next year; 3) to require systemic steroids treatment during the next year. Logistic regression models and AUCs were used to identify candidate predictors in demographics (employment status, marital status, age), health behavior variables (history of tobacco use, alcohol use, psychological illness), clinical factors (duration of IBD diseases; hospitalization during past three years, number of ER visits and telephone call during the past year), drug prescriptions (use of immunomodulator and systemic steroids during the past year, narcotic usage and number of

systemic steroid prescriptions during past three years). Eventually generalized estimating equations were used to develop prediction models for the three proxy outcome variables of interest for patients with Crohn Disease and those with UC, respectively. The developed models can provide a patient or the treating physician the chance for the patient to have a complicated disease course during the future year in terms of those three clinical outcomes of interest and the corresponding 95% confidence intervals.

As an example, the GEE model for predicting a CD patient who would incur over $100K medical charge during the next year yields an AUC at 0.812 and a positive predictive probability at 0.351 in 10-fold cross validation among the training data. The prevalence of incurring over $100K medical charges during the future year was 0.083 among the training data and our predictive model had 4.2 times higher chance to identify a patient with a complicated disease course. External validation on the testing data yielded an AUC=0.825. The positive predictive value was 0.296, which was 3.5 times higher than the prevalence. This demonstrated that our GEE model was really powerful in predicting a rare disease outcome among the IBD patients.

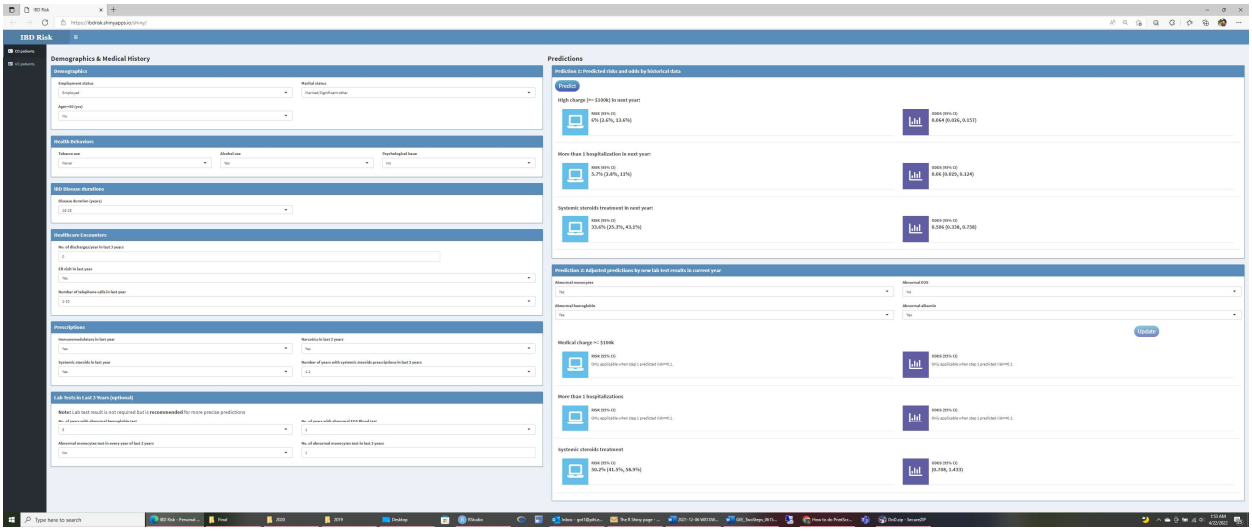


Figure 15: The convenient interface that provides these predictions has been programmed as an R Shiny tool and can be accessed by any IBD patients and physicians at: https://ibdrisk.shinyapps.io/shiny/.

To make it more flexible and improve its predictive power, we also considered the scenarios when lab test results (hemoglobin, EOS, monocytes) of past three years are available to the patients. Including past lab test results improved the performance of the prediction model.

To further provided informed decision on patient management, a two-stage method was also considered. In the first stage, the above-mentioned generalized estimating equation model was used to predict the risk of a clinical outcome of interest during the future year. For patients who were predicted to have a higher risk of complicated disease course during the future year, they would be followed by their treating gastroenterologist for more intense disease assessment. One common practice is to perform the usual blood tests for further screening. Thus at the second stage, a logistic regression model will be used to further assess the risk of the long-term outcome during the future year with observed data during past three years and the lab test results at the clinical visit among patients who were identified as higher risk from the first stage model. The

two-step method further improved the predictive power than the single-step GEEs. The two-stage model for predicting the risk of incurring over $100K medical charges during the future year reached a positive predictive value at 0.471 in 10-fold cross validation and 0.448 in the external validation on the testing data. The predicting power was further enhanced to 5.6 and 5.4 times higher, respectively.

Furthermore, we studied the development of optimal individualized treatment rule (ITR) on the use of the first biologics in minimizing the risk of having at least two hospitalizations during the next three years of staring biologics. Clinical data from 178 IBD patients who started their first biologics during the era of this patient registry. The R package "Personalized" was used to develop the optimal ITR for use of biologics among IBD patients. It was found that patient age and their usage of steroids in the past would be a good predictor for determining the optimal ITR for an individual patient. For example, for patients who never had systemic steroids in the past and of a median age at 36, Adalimumab would be a better biologic than Inflximab that it leads to less chance to have at least two hospitalizations during the next two years. However for a much older patient of 60 years old and who never had systemic steroids, Infliximab would be a better biologic than Adalimumab. For a patient who had systemic steroids in the past, Infliximab should always be favored. In the near future, we will expand our investigation on optimal ITRs in IBD patients with using cured data from hundreds of IBD patients.

What opportunities for training and professional development has the project provided?
If the project was not intended to provide training and professional development opportunities or there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe opportunities for training and professional development provided to anyone who worked on the project or anyone who was involved in the activities supported by the project. "Training" activities are those in which individuals with advanced professional skills and experience assist others in attaining greater proficiency. Training activities may include, for example, courses or one-on-one work with a mentor. "Professional development" activities result in increased knowledge or skill in one's area of expertise and may include workshops, conferences, seminars, study groups, and individual study. Include participation in conferences, workshops, and seminars not listed under major activities.

Nothing to report.

How were the results disseminated to communities of interest?
If there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe how the results were disseminated to communities of interest. Include any outreach activities that were undertaken to reach members of communities who are not usually aware of these project activities, for the purpose of enhancing public understanding and increasing interest in learning and careers in science, technology, and the humanities.

Nothing to report.

What do you plan to do during the next reporting period to accomplish the goals?
If this is the final report, state "Nothing to Report."

Describe briefly what you plan to do during the next reporting period to accomplish the goals and objectives.

Nothing to report.

4. **IMPACT:** Describe distinctive contributions, major accomplishments, innovations, successes, or any change in practice or behavior that has come about as a result of the project relative to:

What was the impact on the development of the principal discipline(s) of the project?
If there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe how findings, results, techniques that were developed or extended, or other products from the project made an impact or are likely to make an impact on the base of knowledge, theory, and research in the principal disciplinary field(s) of the project. Summarize using language that an intelligent lay audience can understand (Scientific American style).

Nothing to report.

What was the impact on other disciplines?
If there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe how the findings, results, or techniques that were developed or improved, or other products from the project made an impact or are likely to make an impact on other disciplines.

Nothing to report

What was the impact on technology transfer?
If there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe ways in which the project made an impact, or is likely to make an impact, on commercial technology or public use, including:

- *transfer of results to entities in government or industry;*
- *instances where the research has led to the initiation of a start-up company; or*

- *adoption of new practices.*

Nothing to report.

What was the impact on society beyond science and technology?
If there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe how results from the project made an impact, or are likely to make an impact, beyond the bounds of science, engineering, and the academic world on areas such as:
- *improving public knowledge, attitudes, skills, and abilities;*
- *changing behavior, practices, decision making, policies (including regulatory policies), or social actions; or*
- *improving social, economic, civic, or environmental conditions.*

Nothing to report.

5. **CHANGES/PROBLEMS:** The PD/PI is reminded that the recipient organization is required to obtain prior written approval from the awarding agency grants official whenever there are significant changes in the project or its direction. If not previously reported in writing, provide the following additional information or state, "Nothing to Report," if applicable:

Nothing to report.

Changes in approach and reasons for change
Describe any changes in approach during the reporting period and reasons for these changes. Remember that significant changes in objectives and scope require prior approval of the agency.

Through iterative testing of machine learning algorithm and through in-depth review of classification results we discovered that all tested ML classifiers were incorrectly classifying some high-charge patients as low-charge patients with high confidence. Through additional data review, we discovered that while charge data provides a good proxy for clinical outcomes, we have to take extra steps to identify patients whose charges are higher due to treatments unrelated to IBD. After removing patients with cancer-related diagnosis and organ transplant patients, the Random Forest and XGBoost models' accuracy improved to 92%, with misclassification rates significantly reduced.
While classification-modeling approach is producing reasonable results with predicting which patients will likely have poor response to anti-TNF biologic therapy (79% accuracy), we are

expanding our efforts to test causal probabilistic models to explore how treatment pathways affect response to anti-TNF biologic therapy. Initial Bayesian network (BN) models have accuracy comparable to classification models but may provide more insight into contributing factors.

Actual or anticipated problems or delays and actions or plans to resolve them
Describe problems or delays encountered during the reporting period and actions or plans to resolve them.

We began working exploring alternative approaches of representing temporal data and of rapidly generating / grouping patient cohorts. To achieve this, we need to make significant changes to how patient data is stored and structured.
Electronic Medical Record Systems (EMRS) such as Epic and Cerner represent patient data using relational data models. In such models, each object and event relevant to a patient, including the patient themselves, is represented by a table (an entity). Examples of such entities in EMRS include diagnosis, medications, visits, allergies, procedures, etc.. In turn, each table is described by a set of attributes (columns) that provide additional detail about the entity in question.
For example, a *Medication* table would contain columns such as a medication's unique identifier, brand name, generic name, therapeutic class, etc. Each row of data in such a table represents a single record (row of data) describing some part of the patient's treatment history. A row of data in the *Medication* table may look like the example shown in Table 1 and indicate that a patient with the medical record number (MRN) of 54321 was prescribed Tylenol on April 11, 2018.

In other words, storing patient data in relational data models facilitates state-based representations, where each record in relevant database tables represents the state of a patient at the time that the data was created.
While using relational models to store patient data is a common approach that is widely used by EMRS vendors such as Epic and Cerner, relational database management systems (RDBMS) have several shortcomings when it comes to helping answer questions about temporal events, especially when these questions concern exploring and understanding causality. Another key shortcoming is the difficulty of rapidly identifying patient cohorts based on temporal events. For example, retrieving a sub-cohort of CF patients based on criteria such as "CF patients ages 12-19 with cystic fibrosis-related diabetes (CFRD) that had multiple adverse reactions to Kalydeco" from a relational database, one would need perform computationally expensive connections between Patient, Diagnoses, Medication, and Allergy tables. Once the data were retrieved, further computations to extract temporal relationships would be required to extract the desired patient cohort.
To address these shortcomings, we propose representing patient data as a graph, a mathematical structure and a data model often used to represent, study, and model credit card fraud patterns, power consumption patterns, and information and influence propagation in social networks.

Changes that had a significant impact on expenditures
Describe changes during the reporting period that may have had a significant impact on expenditures, for example, delays in hiring staff or favorable developments that enable meeting objectives at less cost than anticipated.

Nothing to report.

Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents
Describe significant deviations, unexpected outcomes, or changes in approved protocols for the use or care of human subjects, vertebrate animals, biohazards, and/or select agents during the reporting period. If required, were these changes approved by the applicable institution committee (or equivalent) and reported to the agency? Also specify the applicable Institutional Review Board/Institutional Animal Care and Use Committee approval dates.

Significant changes in use or care of human subjects

Nothing to report.

Significant changes in use or care of vertebrate animals

Not applicable.

Significant changes in use of biohazards and/or select agents

Not applicable.

6. **PRODUCTS:** List any products resulting from the project during the reporting period. If there is nothing to report under a particular item, state "Nothing to Report."

- **Publications, conference papers, and presentations**
  Report only the major publication(s) resulting from the work under this award.

Journal publications. *List peer-reviewed articles or papers appearing in scientific, technical, or professional journals. Identify for each publication: Author(s); title; journal; volume: year; page numbers; status of publication (published; accepted, awaiting publication; submitted, under review; other); acknowledgement of federal support (yes/no).*

1. Ahsan M, et al., Interaction of sugar-sweetened beverage consumption, gender and socioeconomic status on Inflammatory Bowel Disease multi-year clinical trajectories. Accepted
2. Peripheral Blood Eosinophilia and Long-term Severity in Pediatric-Onset Inflammatory Bowel Disease. Prathapan KM, Ramos Rivers C, Anderson A, Koutroumpakis F, Koutroubakis IE, Babichenko D, Tan X, Tang G, Schwartz M, Proksell S, Johnston E, Hashash JG, Dunn M, Wilson A, Barrie A, Harrison J, Hartman D, Kim SC, **Binion DG.**Inflamm Bowel Dis. 2020 Nov 19;26(12):1890-1900. doi: 10.1093/ibd/izz323.PMID: 31960916
3. Disease Characteristics and Severity in Patients With Inflammatory Bowel Disease With Coexistent Diabetes Mellitus. Din H, Anderson AJ, Ramos Rivers C, Proksell S, Koutroumpakis F, Salim T, Babichenko D, Tang G, Koutroubakis IE, Schwartz M, Johnston E, Barrie A, Harrison J, Hashash J, Dunn MA, Hartman DJ, **Binion DG.**Inflamm Bowel Dis. 2020 Aug 20;26(9):1436-1442. doi: 10.1093/ibd/izz305.PMID: 31944255
4. Complete Resolution of Mucosal Neutrophils Associates With Improved Long-Term Clinical Outcomes of Patients With Ulcerative Colitis. Pai RK, Hartman DJ, Rivers CR, Regueiro M, Schwartz M, **Binion DG**, Pai RK.Clin Gastroenterol Hepatol. 2020 Oct;18(11):2510-2517.e5. doi: 10.1016/j.cgh.2019.12.011. Epub 2019 Dec 14.PMID: 31843598
5. Complete Resolution of Mucosal Neutrophils Associates With Improved Long-Term Clinical Outcomes of Patients With Ulcerative Colitis. Pai RK, Hartman DJ, Rivers CR, Regueiro M, Schwartz M, **Binion DG**, Pai RK.Clin Gastroenterol Hepatol. 2020 Oct;18(11):2510-2517.e5. doi: 10.1016/j.cgh.2019.12.011. Epub 2019 Dec 14.PMID: 31843598
6. Ustekinumab-Induced Remission of Two Cases of Refractory Cutaneous Crohn's Disease. Patel BM, Ramos Rivers C, Koutroumpakis F, Ahsan M, Dueker J, Hashash J, Johnston E, Barrie A, Harrison J, Schwartz M, Babichenko D, Tang G, **Binion D.**Inflamm Bowel Dis. 2021 Oct 18;27(10):e124. doi: 10.1093/ibd/izab115.PMID: 33999193
7. Monocytosis Is a Biomarker of Severity in Inflammatory Bowel Disease: Analysis of a 6-Year Prospective Natural History Registry. Anderson A, Cherfane C, Click B, Ramos-Rivers C, Koutroubakis IE, Hashash JG, Babichenko D, Tang G, Dunn M, Barrie A, Proksell S, Dueker J, Johnston E, Schwartz M, **Binion DG.**Inflamm Bowel Dis. 2021 Mar 9:izab031. doi: 10.1093/ibd/izab031. Online ahead of print.PMID: 33693659
8. Validated Indices for Histopathologic Activity Predict Development of Colorectal Neoplasia in Ulcerative Colitis. Pai RK, Hartman DJ, Leighton JA, Pasha SF, Rivers CR, Regueiro M, **Binion DG**, Pai RK.J Crohns Colitis. 2021 Sep 25;15(9):1481-1490. doi: 10.1093/ecco-jcc/jjab042.PMID: 33687061
9. Serum IgG4 Subclass Deficiency Defines a Distinct, Commonly Encountered, Severe Inflammatory Bowel Disease Subtype. Koutroumpakis F, Phillips AE, Yadav D, Machicado JD, Ahsan M, Ramos Rivers C, Tan X, Schwartz M, Proksell S, Johnston E, Dueker J, Hashash JG, Barrie A, Harrison J, Dunn MA, Konnikova L, Hartman DJ, Din H, Babichenko D, Tang G, **Binion DG.**Inflamm Bowel Dis. 2021 May 17;27(6):855-863. doi: 10.1093/ibd/izaa230.PMID: 32879976
10. The Impact of Cholecystectomy on Long-Term Disease Outcomes and Quality of Life in Patients with Crohn's Disease. Koutroumpakis F, Lodhi M, Ahsan M, Ramos Rivers C, Schwartz M, Hashash JG, Babichenko D, Tang G, Nagpal T, Dunn M, Keshavarzian A, **Binion DG.**Inflamm Bowel Dis. 2021 Feb 16;27(3):336-343. doi: 10.1093/ibd/izaa076.PMID: 32313925
11. Kurin M, et al. Clinical Characteristics of Inflammatory Bowel Disease Patients Requiring Long-Term Parenteral Support in the Present Era of Highly Effective Biologic Therapy. JPEN J Parenter Enteral Nutr. 2020;10.1002/jpen.1988. doi:10.1002/jpen.1988

Books or other non-periodical, one-time publications. *Report any book, monograph, dissertation, abstract, or the like published as or in a separate publication, rather than a periodical or series. Include any significant publication in the proceedings of a one-time conference or in the report of a one-time study, commission, or the like. Identify for each one-time publication: author(s); title; editor; title of collection, if applicable; bibliographic information; year; type of publication (e.g., book, thesis or dissertation); status of publication (published; accepted, awaiting publication; submitted, under review; other); acknowledgement of federal support (yes/no).*

Nothing to report.

Other publications, conference papers and presentations. *Identify any other publications, conference papers and/or presentations not reported above. Specify the status of the publication as noted above. List presentations made during the last year (international, national, local societies, military meetings, etc.). Use an asterisk (*) if presentation produced a manuscript.*

Digestive Disease Week 2018:

1. Does surgical anastamosis type impact rates of endoscopic recurrence in post-operative crohn's disease? An 8-year observational cohort study. *Furkan Ubeydullah Ertem, Claudia Ramos Rivers, Miguel D. Regueiro, Andrew R. Watson, Marc Schwartz, Ioannis Koutroubakis, Jana G. Hashash, Benjamin H. Click, Michael A Dunn, Dmitriy Babichenko, David G. Binion.*
2. Biomakers associated with early endoscopic recurrence of postoperative crohn's disease. *Furkan Ertem, Claudia Ramos Rivers, Miguel Regueiro, Benjamin Click, Ioannis Koutroubakis, Marc Schwartz, Andrew Watson, Jana Hashash, Michael Dunn, Dmitriy Babichenko, David Binion.*
3. Finding the sweet spot: the association between added dietary sugars and inflammatory bowel disease severity. *Maaz Ahsan, Alyce Anderson, Dmitriy Babichenko, Claudia Ramos Rivers, Stephen O'Keefe, Miguel Regueiro, Marc Schwartz, Jana Hashash, Benjamin Click, Ioannis Koutroubakis, Michael Dunn, David Binion*
4. Seven-year period prevalence and characteristics of hypertension in a large us inflammatory bowel disease cohort. *Kelly Gibbs, Alyce Anderson, Claudia Ramos Rivers, Benjamin Click, Ioannis Koutroubakis, Miguel Regueiro, Michael Dunn, Marc Schwartz, Jana Hashash, Dmitriy Babichenko, David Binion.*
5. The impact of serum cortisol on quality of life and dysautonomia in inflammatory bowel disease. *Benjamin Click, Alyce Anderson, Claudia Ramos Rivers, Marc Schwartz, Arthur Barrie, Michael Dunn, David Levinthal, Miguel Regueiro, David Binion.*

American College of Gastroenterology Annual meeting 2018:

6. Impact of cholecystectomy on clinical course of ibd, increased diarrhea, healthcare charges, narcotic use and decreased quality of life independent of inflammation. *Maham Lodhi, Claudia Ramos-Rivers, Marc Schwartz, Dmitriy Babichenko, Gong Tang, Tanvi Nagpal, Michael Dunn, David Binion.*

Digestive Diseases Week 2019:

1. Candidate And Exploratory Genetic Association Study Of Ibd Severity Using A Novel Phenotype (Multiyear Mean Healthcare Charges). *Tanvi Nagpal, David G. Binion, Claudia Ramos Rivers, Yan Lin.* DDW 2019
2. Peripheral Blood Eosinophilia Is A Biomarker Of Long-Term Severity In Pediatric-Onset Inflammatory Bowel Disease Patients. *Krishnapriya Marangattu Prathapan, Claudia Ramos Rivers, Sandra C. Kim, Ioannis Koutroubakis, Dmitriy Babichenko, Gong Tang, Marc Schwartz, Siobhan Proksell, Elyse Johnston, Jana G. Hashash, Michael A. Dunn, Annette Wilson, Alyce J. Anderson, Arthur Barrie, Janet Harrison, Douglas J. Hartman, David G. Binion.* DDW 2019*
3. Characterizing The Colonoscopic Features Of Clostridium Difficile Infection In Inflammatory Bowel Disease. *Vance Hartke, Siobhan Proksell, Alyce J. Anderson, Claudia Ramos Rivers, Marc Schwartz, Elyse Johnston, Jana G. Hashash, Arthur Barrie, Janet Harrison, Ioannis E. Koutroubakis, Douglas J. Hartman, Dmitriy Babichenko, Michael A. Dunn, David G. Binion.* DDW 2019
4. Peripheral Blood Eosinophilia Functions As A Candidate Biomarker Of Decreased Response To Anti-Tnf Therapy In Crohn's Disease. *Scott Friedberg, Weston Bettner, Xianling Wang, Claudia Ramos Rivers, Ioannis E. Koutroubakis, Gong Tang, Dmitriy Babichenko, Siobhan Proksell, Elyse Johnston, Marc Schwartz, Jana G. Hashash, Arthur Barrie, Janet Harrison, Douglas J. Hartman, Michael A. Dunn1, David G. Binion.* DDW 2019
5. Does Inflammation In Ibd "Burnout" Over Time? *Hassieb Din, Alyce J. Anderson, Claudia Ramos Rivers, Dmitriy Babichenko, Gong Tang, Ioannis E. Koutroubakis, Marc Schwartz, Siobhan Proksell, Elyse Johnston, Arthur Barrie, Janet Harrison, Jana G. Hashash, Michael A. Dunn, Douglas J. Hartman, Tariq Salim, Eva Szigethy, David G. Binion.* DDW 2019
6. Big Data Analytics Identifies Ulcerative Colitis Patients At Increased Risk For Incident Colorectal Neoplasia Using Multiyear Patterns Of Routine Clinical Lab Values. *Carlita Shen, Claudia Ramos Rivers, Dmitriy Babichenko, Douglas J. Hartman, Ioannis Koutroubakis, Marc Schwartz, Siobhan Proksell, Elyse Johnston, Jana G. Hashash, Arthur Barrie2, Janet Harrison, Gong Tang, Andrew R. Watson, David G. Binion.* DDW 2019
7. Natural History Of Diabetes Mellitus And Inflammatory Bowel Disease: Increased Disease Severity, Worse Quality Of Life, And Under-Treatment With Immunomodulator And/Or Biologic Agents. *Hassieb Din, Alyce J. Anderson, Claudia Ramos Rivers, Siobhan Proksell, Tariq Salim, Dmitriy Babichenko, Gong Tang, Ioannis E. Koutroubakis, Marc Schwartz, Elyse Johnston, Arthur Barrie, Janet Harrison, Michael A. Dunn, Douglas J. Hartman, David G. Binion.* DDW 2019*
8. *Endoscopic Patterns And Location Of* Post-Operative Recurrence In Crohn's Disease Patients With Side To Side Anastomosis Following Ileocecal Resection. *Furkan Ertem, Andrew R. Watson, Claudia Ramos Rivers, Dmitriy Babichenko, Gong Tang, Marc Schwartz, Siobhan Proksell, Elyse Johnston, Jana G. Hashash, Arthur Barrie, Janet Harrison, Ioannis E. Koutroubakis, Michael A. Dunn, Douglas J. Hartman, David G. Binion.* DDW 2019

American College of Gastroenterology Annual meeting 2019:

9. Low Serum Levels of IgG4 Antibodies May Function as a Biomarker of Severity in Inflammatory Bowel Disease. *Filippos Koutroumpakis, Anna Evans Phillips, Dhiraj Yadav, Jorge D Machicado, Claudia Ramos-Rivers, Marc Schwartz, Siobhan Proksell, Elyse Johnston, Jeffrey Dueker, Jana G Hashash, Arthur Barrie, Janet Harrison, Michael A Dunn, Liza Konnikova, Douglas J Hartman, Hasieb Din, Dmitriy Babichenko, Gong Tang, David G Binion.* ACG 2019*

American College of Gastroenterology Annual meeting 2020:

10. S0808 The biomarker peripheral blood eosinophilia predicts failure to achieve mucosal healing in inflammatory bowel disease. *Koutroumpakis, Filippos; Ahsan, Maaz; Rivers, Claudia Ramos; Schwartz, Marc; Johnston, Elyse; Dueker, Jeffrey; Proksell, Siobhan; Hashash, Jana; Barrie, Arthur; Tang, Gong; Babichenko, Dmitriy; Dunn, Michael; Binion, David. The American Journal of Gastroenterology. 115:S414-S415, October 2020*
11. S0810 DUAL BIOLOGIC Therapy for the treatment of severe/refractory inflammatory bowel disease: intravenous immunoglobulin in combination with standard biologic agents. *Koutroumpakis, Filippos; Ahsan, Maaz; Rivers, Claudia Ramos; Schwartz, Marc; Johnston, Elyse; Proksell, Siobhan; Dueker, Jeffrey; Hashash, Jana; Barrie, Arthur; Dunn, Michael; Tang, Gong; Babichenko, Dmitriy; Binion, David. The American Journal of Gastroenterology. 115:S416, October 2020.*
12. S0651 type of surgical anastomosis influences long-term clinical status, quality of life, and opioid requirement in post-operative crohn's disease: a 3-year comparative effectiveness study. *Alkaissy, Zaid; Koutroumpakis, Filippos; Watson, Andrew; Ahsan, Maaz; Rivers, Claudia Ramos; Proksell, Siobhan; Dueker, Jeffrey; Johnston, Elyse; Hashash, Jana; Schwartz, Marc; Barrie, Arthur; Babichenko, Dmitriy; Tang, Gong; Dunn, Michael; Binion, David. The American Journal of Gastroenterology. 115:S326, October 2020.*

13. S0890 Can ibd medications counteract the negative effects of a deleterious, high sugar diet? *Ahsan, Maaz; Koutroumpakis, Filippos; Rivers, Claudia Ramos; Wilson, Annette; Schwartz, Marc; Proksell, Siobhan; Johnston, Elyse; Dueker, Jeffrey; Hashash, Jana; Barrie, Arthur; o'keefe, Stephen; Dunn, Michael; Babichenko, Dmitriy; Tang, Gong; Binion, David. The American Journal of Gastroenterology .. 115:S459, October 2020.*
14. S0809 Is there a benefit of immunomodulator and anti-tnf combination therapy over monotherapy in the real world, long-term management of ibd? Assessing multiyear treatment persistence and mucosal healing. *Koutroumpakis, Filippos; Ahsan, Maaz; Rivers, Claudia Ramos; Alkaissy, Zaid; Proksell, Siobhan; Johnston, Elyse; Schwartz, Marc; Dueker, Jeffrey; Hashash, Jana; Barrie, Arthur; Tang, Gong; Babichenko, Dmitriy; Dunn, Michael; Binion, David. The American Journal of The American Journal of Gastroenterology October 2020.*

Digestive diseases week 2020:

15. 249 positive gluten sensitivity serologies and the impact of gluten free diet in patients with ibd. Dahar, maria m. Et al. Gastroenterology, volume 160, issue 6, s-55.
16. Su515 the association of oral vancomycin administration and faecal microbiome composition in patients with crohn's disease. Ghaffari, amir a. Et al. Gastroenterology, volume 160, issue 6, s-722
17. Sa487 dermatologic manifestations of ibd: association with natural history and biomarkers of severity patel, bansri m. Et al. Gastroenterology, volume 160, issue 6, s-518
18. Fr554 disparities in treatment and healthcare utilization beetween inflammatory bowel disease patients followed at a referral university center and community hospitals. Koutroumpakis, filippos et al. Gastroenterology, volume 160, issue 6, s-360 - s-361
19. Su524 peripheral blood monocytosis is a novel biomarker of long-term disease severity in pediatric-onset inflammatory bowel disease. Zhang, xiaoyi et al. Gastroenterology, volume 160, issue 6, s-726

ACG 2021:

20. P2660: disease severity and treatment response in esophageal crohn’s disease. Andy wu. Accepted. Acg 2021
21. P2629: does eating cheese worsen ibd? Prospective analysis of consumption, disease activity and quality of life. Susan jacobs. Acccepted. Acg 2021

American neurogastro-enterology and motility society annual meeting 2021:

22. Use of non-inflammatory anti-diarrheal medications and patient characteristics in an ibd natural history registry. R Sharma*, C Ramos Rivers*, D Levinthal*, E Szigethy*, F Koutroumpakis*, M Ahsan*, J Dueker*, E Johnston*, A Barrie*, J Harrison*, M Schwartz*, J Hashash*, D Babichenko†, G Tang†, D Binion*, A Tansel*. Accepted.
23. Treating ibs-c in ibd: prevalence of constipation and patient characteristics in inflammatory bowel disease (ibd) patients using a natural history registry. R Sharma*, C Ramos Rivers*, D Levinthal*, E Szigethy*, F Koutroumpakis*, M Ahsan*, J Dueker*, E Johnston*, A Barrie*, J Harrison*, M Schwartz*, J Hashash*, D Babichenko†, G Tang†, D Binion*, A Tansel*. Accepted .

- **Website(s) or other Internet site(s)**
  List the URL for any Internet site(s) that disseminates the results of the research activities. A short description of each site should be provided. It is not necessary to include the publications already specified above in this section.

Nothing to report.

- **Technologies or techniques**
  Identify technologies or techniques that resulted from the research activities. Describe the technologies or techniques were shared.

Nothing to report.

- **Inventions, patent applications, and/or licenses**
  Identify inventions, patent applications with date, and/or licenses that have resulted from the research. Submission of this information as part of an interim research performance progress report is not a substitute for any other invention reporting required under the terms and conditions of an award.

Nothing to report.

- **Other Products**
  Identify any other reportable outcomes that were developed under this project. Reportable outcomes are defined as a research result that is or relates to a product, scientific advance, or research tool that makes a meaningful contribution toward the understanding, prevention, diagnosis, prognosis, treatment and /or rehabilitation of a disease, injury or condition, or to improve the quality of life. Examples include:
  - *data or databases;*
  - *physical collections;*
  - *audio or video products;*
  - *software;*
  - *models;*
  - *educational aids or curricula;*
  - *instruments or equipment;*
  - *research material (e.g., Germplasm; cell lines, DNA probes, animal models);*
  - *clinical interventions;*
  - *new business creation; and*
  - *other.*

Nothing to report.

7. PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

What individuals have worked on the project?

Name: David G. Binion
Project Role: PI
Nearest person month(s) worked: 3
Contribution to Project: Dr. Binion oversaw all research in this project. Bi- weekly research meetings were held to disseminate progress. Dr. Binion has performed work providing strategies for extraction and preparation of clinically relevant variables.

Name: Gong Tang
Project Role: Co-Investigator
Nearest person month(s) worked: 2
Contribution to Project: Working with Ms. Wang, Dr. Tang has developed two-stage models to predict risk of high medical charges in the future using demographics, historical data on comorbidity, usage of medications, lab tests, surgery and healthcare utilization in both CD and UC patients. Medical charges during different future time frames were considered. Adjustment in medical charges was made according to the Consumer Price Index and generalized estimating equations were used in those models. The performance of those algorithms was assessed via implementation on a separate testing dataset.

Name: *Dmitriy Babichenko*
Project Role: Co- Investigator
Nearest person month(s) worked: 1
Contribution to Project: Dr. Babichenko completed data de-idenfication automation scripts
Created and evaluated a series of classification and probabilistic models trained from the IBD registry dataset. Completed data de-identification automation scripts. Worked on creating the initial decision support system user interface designs.

Name: *Marek Drudzel*
Project Role: *Co-investigator*
Nearest person month(s) worked: 2
Contribution to Project: Dr. Drudzel has performed extensive exploration of data to determine analysis to evaluate complicated/ high cost disease using Bayesian networks. Created and evaluated a series of classification and probabilistic models trained from the IBD registry dataset.

Name: *Mark Roberts*
Project Role: Co- Investigator
Nearest person month(s) worked: 1
Contribution to Project: Roberts has been providing advice about best strategies for extraction and preparation of clinical data. He has also provided advice on epidemiological relevance.

Name: *Michael Dunn*
Project Role: Co-Investigator
Nearest person month(s) worked: 1
Contribution to Project: Dr. Dunn has been providing advice and expertise about best strategies for extraction and preparation of clinically relevant variables.

Name: Claudia Ramos Rivers
Project Role: *Key personnel- Research Scientist*
Nearest person month(s) worked: 7
Contribution to Project: Dr. Ramos Rivers has overseen protocol submission for IRB approval as well as preparing progress reports. Dr. Ramos Rivers has also coordinated and attended to bi- weekly meetings to develop strategies on data extraction and preparation for analysis.

Name: *Annette Wilson*
Project Role: *Key personnel- Lab. Manager*
Nearest person month(s) worked: 2
Contribution to Project: Dr. Wilson has been responsible for the post award administrative work.

Name: *Yan Lin*
Project Role: *Key personnel - Faculty*
Nearest person month(s) worked: 1
Contribution to Project: *Dr. Lin has participated in developing prediction models for future clinical outcomes and worked with Dr. Binion and a research staff on bioinformatics analyses of genetic data from those IBD patients.*

Name: *Krauland, Mary G*
Project Role: *Graduate Student. Graduate School of Public Health*
Nearest person month(s) worked: 10
Contribution to Project: Under the supervision of Dr. Roberts, Mrs. Krauland has been providing advice about best strategies for extraction and preparation of clinical data. He has also provided advice on epidemiological relevance.

Name: *Marcin Kozniewski*
Project Role: *Graduate Student Researcher*
Nearest person month(s) worked: 11
Contribution to Project: *Under Marek Drudzel supervision, Marcin Kozniewski has performed exploration of data to determine analysis to evaluate complicated/ high cost disease using Bayesian networks.*

Name: *Xianling Wang*
Project Role: *Graduate Student Researcher*
Nearest person month(s) worked: 11
Contribution to Project: Under the supervision of Dr. Tang, Ms. Wang has performed extensive analyses to predict future clinical outcomes of IBD patients based on demographics, historical lab data and other medical records. Ms. Wang will explore more comprehensive modelling to include dimension reduction, regularization and causal pathway analysis.

Name: *Beata Pasek*
Project Role: Clinical Research coordinator
Nearest person month(s) worked: 2
Contribution to Project: Mrs. Pasek has consented patients currently in the study and has been responsible for regulatory activities.

Name: *Yongseok Park*
Project Role: *Key personnel - Faculty*
Nearest person month(s) worked: 1
Contribution to Project: *Dr. Park continued working with a resident MD (Dr. Thien-Bao Nguyen) and Xianling Wang on a project about using peripheral blood eosinophilia to predict earlier anti-TNF cessation and biologic witch in ulcerative colitis patients.*

Name: *Behnam Rahdari*
Project Role: *Graduate Student Researcher*
Nearest person month(s) worked: 9
Contribution to Project: *Under Dmitriy Babichenko supervision, Behnam Rahdari has continued to work on exploration of data to determine analysis to evaluate complicated/ high cost disease using Bayesian network and interface development.*

Name: Suraj Subramanian
Project Role: *Graduate Student Researcher*
Nearest person month(s) worked: 9
Contribution to Project: *Under Dmitriy Babichenko supervision, His primary responsibilities included developing and validating Artificial Neural Network (ANN) - based approaches to IBD modeling. Mr. Subramanian developed a pipeline for formatting patient data as a series of sliding temporal windows, using 24-month window to predict the clinical outcomes in the following 8 months. With larger training datasets the windows could be expanded to cover larger periods of time. Mr. Subramanian also developed and trained a series of ANN models using Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures.*

Name: Benjamin Stein
Project Role: *Graduate Student Researcher*
Nearest person month(s) worked: 9
Contribution to Project: *Under Dmitriy Babichenko supervision, Graduate student of* Computing and Information who began working on this project in May 2020. Mr. Stein replaced Mr. Subramanian (who graduated from the Master's in Information Sciences program). Mr. Stein is responsible for continuing Mr. Subramanian's work on ANNs, as well as using the same sliding temporal window approach to train conventional gradient boosted trees models.

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?

Nothing to Report.

What other organizations were involved as partners?

Nothing to Report.

8. SPECIAL REPORTING REQUIREMENTS

COLLABORATIVE AWARDS: N/A

QUAD CHARTS: N/A

9. APPENDICES: N/A