**INTRODUCTION**

- Inflammatory bowel disease (IBD), encompassing ulcerative colitis (UC) and Crohn’s disease, is a chronic, relapsing-remitting gastrointestinal disease that affects ∼1.5 million people in the United States.\(^2\)
- The disease course of IBD is unpredictable, and methods to predict disease flares could help tailor therapy to high-risk patients, reducing morbidity and high medical costs associated with flare.\(^3\)
- Wajie et al.\(^4\) developed a novel machine learning model to predict IBD-related hospitalization and outpatient corticosteroid use in a cohort of US veterans with IBD, but the study population is likely not representative of the general IBD population.
- It is important to assess the predictive value of this machine learning model in a larger, more diverse group of patients with IBD.

**METHODS**

- **Data Source and Patient Selection.**
  - Data for this retrospective cohort study came from Optum Electronic Health Records (EHRs), which include ∼197 million patients treated in 38 US states.
  - Patients (≥ 18 years) were selected based on ≥ 2 International Classification of Diseases Clinical Modification (9th or 10th revision [ICD-9-CM or ICD-10-CM]) diagnosis codes for UC, Crohn’s disease, or indeterminate colitis between January 1, 2007, and December 31, 2017, with ≥ 1 diagnosis code from an outpatient visit.
  - Patients were classified as having UC if all diagnosis codes were 505.xx for ICD-9-CM or K50.xx for ICD-10-CM, and indeterminate colitis otherwise. For Crohn’s disease, ICD-9-CM diagnosis codes 555.xx were used for IBD, or K51.xx for ICD-10-CM, Crohn’s disease if all diagnosis codes were 555.xx for ICD-9-CM or K51.xx for ICD-10-CM, and indeterminate colitis otherwise.
  - The date of first recorded IBD diagnosis was defined as the index date, and patient records must have been available for ≥ 12 months after the index date.

- **Outcome Measures.**
  - The primary outcome of interest was IBD flare, defined as a composite measure capturing either a hospitalization with a diagnosis of IBD or an outpatient corticosteroid prescription indicated for IBD.
  - We determined the indication for corticosteroid prescription by searching for diagnosis codes for a variety of common inflammatory comorbid conditions within 7 days before the corticosteroid prescription date.

- **Predictor Variables.**
  - Variables included age, sex, number of previous IBD flare visits, use of immunosuppressive medication (immunomodulators and/or anti–tumor necrosis factor agents).
  - Data were derived in the current visit. Focal calcification was not included because it was performed in ≤ 1% of the study population.
  - Mixing laboratory values were imputed based on the median laboratory value from all available visits.

**RESULTS**

- **Patient Characteristics.**
  - We identified a total of 59,678 patients with IBD across 780,559 visits for this study. Of these, 22,245 (29.2%) patients had at least 1 IBD flare (Table 1). Patients were predominantly female (57.1%) and white (87.7%), with a median age of 48.2 years (16.8 years) at index.

- **Imported Predictor Variables.**
  - The 10 most important predictor variables were number of previous IBD flares, age at visit, mean potassium level before visit, and mean and median white blood cell count before visit.

- **Model Performance.**
  - Area under the ROC curve (AUC) for the benchmark logistic regression, logistic regression with L1 regularization, and random forest models were 0.641 (95% confidence interval [CI]: 0.632-0.650), 0.668 (95% CI: 0.656-0.680), and 0.675 (95% CI: 0.660-0.690), respectively (Figure 1 and Table 2).

**CONCLUSIONS**

- Our random forest model outperforms logistic regression models in predicting flare in a nationally representative sample of patients with IBD and has similar predictive accuracy to that of Wajie et al.\(^4\) Random forest models are able to capture nonlinear relationships and interactions implicitly, whereas linearity and interactions in logistic models need to be modeled explicitly.
- Addition of laboratory values to clinical features improves predictive performance over clinical features alone.
- SHAP values are useful in interpreting important random forest model predictor variables, which align with clinical expertise on known risk factors for future IBD flare.
- Limitations of this model include limited EHR data availability (particularly for corticosteroid prescriptions), the possibility of patients receiving care that would not have been captured in Optum EHRs, and the potentially restrictive definition of flare used herein.
- Predictive models could help enable personalized care in IBD and could be incorporated into routine practice as a tool for monitoring and identifying patients at high risk of flare.

**REFERENCES**


**ACKNOWLEDGMENTS**

- Ryan Gan, Diana Sun, Amanda Tatro, Shirley Cohen-Mekelburg, Wyndy L. Wiitala, Akbar K. Wajie
- Genentech, Inc., South San Francisco, CA, USA; Roche Pharms A. G, Basel, Switzerland; University of Michigan Health System, Ann Arbor, MI, USA; Veterans Affairs Health Care System, Center for Clinical Management Research, Ann Arbor, MI, USA

---

**Table 1. Patient Characteristics**

<table>
<thead>
<tr>
<th>Age (years), mean (SD)</th>
<th>UC (11,500 visits)</th>
<th>Crohn’s disease (11,998 visits)</th>
<th>Indeterminate colitis (11,773 visits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male, n (%)</td>
<td>48.1 (16.8)</td>
<td>48.6 (16.6)</td>
<td>47.1 (17.3)</td>
</tr>
<tr>
<td>Female, n (%)</td>
<td>51.9 (16.9)</td>
<td>51.4 (16.5)</td>
<td>52.9 (17.2)</td>
</tr>
</tbody>
</table>

**Table 2. Bootstraped Estimates of Median and 95% CI Model Performance**

<table>
<thead>
<tr>
<th>Diagnostic Metric</th>
<th>Model Performance</th>
<th>Area under ROC curve (AUC)</th>
<th>L1 Regularization</th>
<th>Random forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>UC</td>
<td>0.641 (95% CI: 0.632-0.650)</td>
<td>0.668 (95% CI: 0.656-0.680)</td>
<td>0.675 (95% CI: 0.660-0.690)</td>
</tr>
<tr>
<td>Specificity</td>
<td>UC</td>
<td>0.225 (95% CI: 0.216-0.234)</td>
<td>0.237 (95% CI: 0.228-0.246)</td>
<td>0.221 (95% CI: 0.212-0.231)</td>
</tr>
<tr>
<td>Positive predictor value</td>
<td>UC</td>
<td>0.557 (95% CI: 0.546-0.567)</td>
<td>0.585 (95% CI: 0.574-0.595)</td>
<td>0.562 (95% CI: 0.551-0.573)</td>
</tr>
<tr>
<td>Negative predictor value</td>
<td>UC</td>
<td>0.443 (95% CI: 0.433-0.454)</td>
<td>0.415 (95% CI: 0.404-0.425)</td>
<td>0.438 (95% CI: 0.427-0.448)</td>
</tr>
</tbody>
</table>

**Figure 1. SHAP summary plot of 10 most important predictors of IBD flare.**

**Figure 2. SHAP Value (Impact on Model Output)"