Title

Accurate, Robust, and Scalable Machine Abstraction of Mayo Endoscopic Subscores From Colonoscopy Reports.

Journal

Inflammatory Bowel Diseases, 31(3)

Authors

Silverman, Anna
Bhasuran, Balu
Mosenia, Arman
et al.

Publication Date

2025-03-03

DOI

10.1093/ibd/izae068

Peer reviewed

Inflammatory Bowel Diseases, 2025, **31**, 665–670
https://doi.org/10.1093/ibd/izae068
Advance access publication 26 March 2024
Original Research Article - Clinical

Accurate, Robust, and Scalable Machine Abstraction of Mayo Endoscopic Subscores From Colonoscopy Reports

Anna L. Silverman, MD,[*,†,a] Balu Bhasuran, PhD,[‡,a] Arman Mosenia, MSE,[§] Fatema Yasini, BS,[¶] Gokul Ramasamy, MS,[‖,] Imon Banerjee, PhD,[‖,**] Saransh Gupta, BS,[¶] Taline Mardirossian, BS,[¶] Rohan Narain, BS,[¶] Justin Sewell, MD, PhD,[††,‡‡] Atul J. Butte, MD, PhD,[‡,§§,] and Vivek A. Rudrapatna, MD, PhD[‡,‡‡,]**

[*]Division of Gastroenterology and Hepatology, Department of Medicine, Mayo Clinic, Phoenix, AZ, USA
[†]Department of Medicine, University of California, San Diego, La Jolla, CA, USA
[‡]Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA, USA
[§]UCSF School of Medicine, University of California, San Francisco, San Francisco, CA, USA
[¶]Department of Computer Science, University of California, Berkeley, Berkeley, CA, USA
[‖]Department of Radiology, Mayo Clinic, Phoenix, AZ, USA
[**]School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, USA
[††]Division of Gastroenterology, Department of Medicine, Zuckerberg San Francisco General Hospital, San Francisco, CA, USA
[‡‡]Division of Gastroenterology and Hepatology, Department of Medicine, University of California, San Francisco, San Francisco, CA, USA
[§§]Center for Data-Driven Insights and Innovation, University of California Health, Oakland, CA, USA

[a]These authors contributed equally to this work.
Address correspondence to: Vivek A. Rudrapatna, MD, PhD, University of California, San Francisco Bakar Institute, Box 2993, 490 Illinois Street, Floor 2, San Francisco, CA 94143, USA (vivek.rudrapatna@ucsf.edu).

Background: The Mayo endoscopic subscore (MES) is an important quantitative measure of disease activity in ulcerative colitis. Colonoscopy reports in routine clinical care usually characterize ulcerative colitis disease activity using free text description, limiting their utility for clinical research and quality improvement. We sought to develop algorithms to classify colonoscopy reports according to their MES.

Methods: We annotated 500 colonoscopy reports from 2 health systems. We trained and evaluated 4 classes of algorithms. Our primary outcome was accuracy in identifying scorable reports (binary) and assigning an MES (ordinal). Secondary outcomes included learning efficiency, generalizability, and fairness.

Results: Automated machine learning models achieved 98% and 97% accuracy on the binary and ordinal prediction tasks, outperforming other models. Binary models trained on the University of California, San Francisco data alone maintained accuracy (96%) on validation data from Zuckerberg San Francisco General. When using 80% of the training data, models remained accurate for the binary task (97% [n = 320]) but lost accuracy on the ordinal task (67% [n = 194]). We found no evidence of bias by gender ($P = .65$) or area deprivation index ($P = .80$).

Conclusions: We derived a highly accurate pair of models capable of classifying reports by their MES and recognizing when to abstain from prediction. Our models were generalizable on outside institution validation. There was no evidence of algorithmic bias. Our methods have the potential to enable retrospective studies of treatment effectiveness, prospective identification of patients meeting study criteria, and quality improvement efforts in inflammatory bowel diseases.

Lay Summary

Our accurate pair of models automatically classify colonoscopy reports by Mayo endoscopic subscore and abstain from prediction appropriately. Our methods can enable large-scale electronic health record studies of treatment effectiveness, prospective identification of patients for clinical trials, and quality improvement efforts in ulcerative colitis.

Key Words: ulcerative colitis, endoscopic disease activity scores, natural language processing, healthcare applied AI

Introduction

Endoscopic outcomes are an important therapeutic target in the care of patients with ulcerative colitis (UC), often being critical components in the decision to continue, dose escalate, or change therapy.[1,2] To standardize the severity of endoscopic findings, registrational trials in UC often use the Mayo endoscopic subscore (MES) to measure disease activity and objective response to therapy. Colonoscopy reports in the electronic health record (EHR) document disease activity; however, these notes typically do not explicitly capture validated disease activity scores, like the MES, limiting their utility for prospective or retrospective research. Thus, a high-accuracy computational method for automatically scoring procedure reports according to their MES would have great potential value for IBD clinical research.

Received for publication: October 19, 2023. Editorial Decision: March 3, 2024

Key Messages

- What is already known?
  Endoscopic disease activity scores, like the Mayo endoscopic subscore (MES), are important in the objective quantification of disease activity in ulcerative colitis but are typically absent in colonoscopy reports in usual clinical care.
- What is new here?
  We have developed a sequential pair of models that automatically determine if a colonoscopy report is suitable for MES and then assigns MES if appropriate. Previously, the MES required manual abstraction. Our methods are highly accurate, generalize on outside-center data, and do not show evidence of algorithmic bias.
- How can this study help patient care?
  Our methods have the potential to enable large-scale retrospective studies of treatment effectiveness, prospective identification of patients meeting study criteria, and quality improvement efforts in inflammatory bowel diseases.

Natural language processing (NLP) refers to computational methods for analyzing language-related data. The use of NLP on clinical text has been an active field of research for several decades, with dozens of software packages now freely available.[3-6] To date, there have been few applications of NLP to IBD research.[7-13] Currently, no NLP models exist to transform routine colonoscopy report text into the MES.

We sought to determine if current NLP methods could accurately abstract MES and abstain from assigning an MES when appropriate. Here, we report the results of a comprehensive and comparative assessment of several methods for training text classifiers designed to assess their current utility for MES abstraction. Our primary endpoint was accuracy on the sequential tasks of identifying which reports could be scored using the MES and assigning a score if appropriate. Secondary endpoints included learning efficiency, generalizability, and algorithmic fairness.

Methods

Procedure Reports

To identify colonoscopy reports for classifier training and evaluation, we accessed the EHRs at 2 health systems in California: an academic medical center (University of California, San Francisco [UCSF]) and a safety-net hospital (Zuckerberg San Francisco General [ZSFG]). These institutions have different physician groups and use different endoscopy reporting software. We queried EHR databases to identify all patients who had ever been assigned an International Classification of Diseases–Tenth Revision code for inflammatory bowel disease (K50*, K51*) and extracted all corresponding colonoscopy reports from the 2017 to 2020 period.

Annotation Procedure

In the first stage of this procedure, 2 physicians (A.L.S. and V.A.R.) uniformly sampled and annotated reports as being suitable for MES or not. The main criteria for defining suitability included a clear diagnosis of UC, surgically unaltered anatomy, and completed procedure (Supplementary Methods). This was recorded as a binary variable. This procedure continued until at least 75 MES eligible reports per site and per annotator were annotated.

In the second stage, suitable reports were assigned an MES, the ordinal measure of UC disease activity that ranges from 0 through 3. Scores were assigned based on the most severely affected segment, and with any friability scored as a 2 or higher in line with the guidance put forth by the U.S. Food and Drug Administration (Supplementary Methods).[2] If a report explicitly contained an MES (25% of reports had an MES noted), the annotators did not utilize it in their annotation and instead used the description of findings in the report text to determine the annotated score.

The interannotator agreement of this process was assessed on a set of 50 uniformly sampled reports. These 50 notes were independently annotated by each annotator and subsequently compared with quantify the objectivity of this annotation procedure.

Algorithm Development and Validation

We developed and evaluated 4 standard methods for abstracting information from notes: cTAKES-based[3] concept recognition, bag-of-words models using sklearn,[4] and automated machine learning (autoML),[5] as well as 3 models related to BERT (Bidirectional Encoder Representations from Transformers) (Figure 1; Supplementary Methods).[6] These methods vary in their underlying technique, requirements for training data, tendencies toward robust and generalizable learning, and ease of use.

We used these methods to separately train a binary classifier (to predict which procedure reports were MES scorable) and an ordinal classifier (to predict the correct MES for scorable reports). As a control, we developed null classifiers that predict the dominant class for each task. All classifiers were evaluated on a 20% held-out test set (100 notes for the binary model and 60 notes for the ordinal model) stratified by score, annotator, and site.

The classifier achieving the highest accuracy was subjected to additional evaluations of generalizability and learning efficiency. To assess generalizability, we retrained the binary classifier on the data from UCSF alone and evaluated it on data from ZSFG. There were insufficient reports to adequately assess generalizability for the ordinal prediction task due to the multiplicity of classes and class imbalance.

Algorithmic Fairness

We evaluated the fairness of our algorithms by estimating their misclassification rate along lines of gender and social deprivation. We accessed patient-level structured data at UCSF to perform these analyses. We used the area deprivation index (ADI) mapped to residential zip codes as a proxy for social deprivation.[14] Sex assigned at birth was unavailable in our database, and we were unable to perform analyses by race or ethnicity due to insufficient procedure notes from each race/ethnicity.

Statistics

We computed exact binomial confidence intervals (CIs) for all results reported as a sample proportion. For analyses of algorithmic fairness, we used ordinal logistic regression to separately model the misclassification error as a function of either gender or ADI. We report the corresponding p-values from

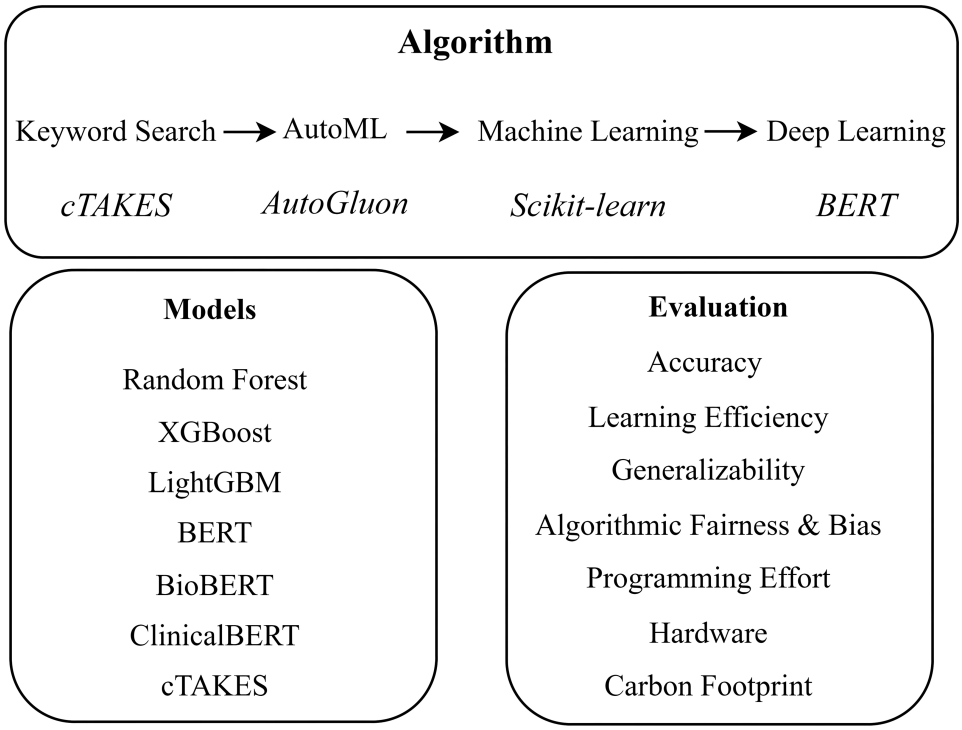

Figure 1. Software and algorithm architectures utilized. Arrows in the algorithm box depict increasing complexity, requirements for training data, and programming effort. autoML, automated machine learning; BERT, Bidirectional Encoder Representations from Transformers.

a Wald test. We performed all computing using R 4.1.3 (R Foundation for Statistical Computing) and Python 3 (Python Software Foundation).

Ethics

The study was approved by the UCSF Institutional Review Board (#18-24588).

Patient and Public Involvement

Patients were not involved in the study design outside of manuscript authors, who happen to be patients with IBD.

Results

Procedure Reports

The source corpus consisted of 3769 colonoscopy reports from UCSF and 835 from ZSFG, all authored between 2017 and 2020 (Figure 2). The manually annotated corpus consisted of 500 notes, of which 282 were from UCSF and 218 were from ZSFG. A total of 302 notes were eligible for the MES, with 151 notes from each site. Interannotator agreement was 96% (95% CI, 86%-100%) for the binary task (n = 50) and 88% (95% CI, 69%-97%) for the ordinal task (n = 25).

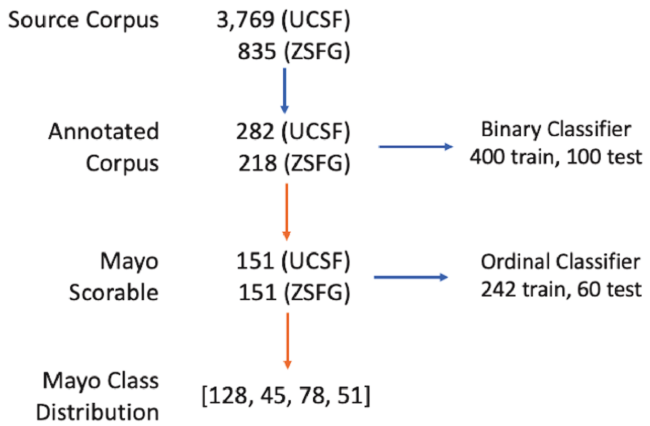

Figure 2. Flow diagram of the procedure reports selected for annotation. Orange arrows depict natural language processing prediction tasks. Numbers associated with the Mayo endoscopic subscore. Class distribution correspond to those scored as a 0, 1, 2, or 3 respectively. UCSF, University of California, San Francisco; ZSFG, Zuckerberg San Francisco General.

Algorithmic Accuracy

The autoML-trained classifiers achieved the overall highest accuracy. They were 98% (95% CI, 91%-99%) accurate at identifying MES scorable reports and 97% (95% CI, 88%-99%) accurate at assigning an MES (Table 1). The relative ordering of algorithmic performance was preserved across both tasks, with the sklearn classifiers consistently outperforming the BERT-based classifiers. ClinicalBERT was substantially more performant than BioBERT and BERT-base, presumably reflecting the sensitivity of these algorithms to pretraining data: ClinicalBERT was trained on clinical notes, whereas BioBERT and BERT-base were trained on biomedical journal articles and general Wikipedia articles, respectively. The manually designed, rule-based approach utilizing cTAKES-recognized clinical concepts was the least accurate (22%). It performed worse than a null model that predicts the most common subscore for all reports (42%). For the binary classifier, the model thought that 1 report was scorable when it was not, while it thought that 2 reports where not scorable

Table 1. Algorithmic performance for the MES scorability (binary) and prediction tasks (ordinal).

Algorithm	Framework	Type	Accuracy (95% CI) (%)	F score
Random forest	autoML	Binary	98 (91-99)[a]	97.47[a]
LightGBM	Scikit-learn	Binary	98 (89-98)[a]	97.00
ClinicalBERT	Transformer	Binary	93 (84-96)	93.00
Null model	Scikit-learn	Binary	60 (50-70)	—
BERT	Transformer	Binary	47 (37-57)	44.68
BioBERT	Transformer	Binary	38 (28-48)	39.00
XGBoost	autoML	Ordinal	97 (88-99)[a]	97.00[a]
LightGBM	Scikit-learn	Ordinal	77 (63-87)	77.10
ClinicalBERT	Transformer	Ordinal	61 (48-74)	61.66
Null Model	Scikit-learn	Ordinal	41 (29-55)	—
BioBERT	Transformer	Ordinal	50 (37-63)	50.00
BERT	Transformer	Ordinal	38 (26-52)	38.33
Rule based using clinical concepts	cTAKES, RegEx	Ordinal	22 (12-37)	22.10

Abbreviations: autoML, automated machine learning; BERT, Bidirectional Encoder Representations from Transformers; CI, confidence interval; MES, Mayo endoscopic subscore.
[a]Denotes highest performing models.

when they were. For the ordinal classifier, the model correctly classified all Mayo 0 and Mayo 1 reports. One Mayo 2 report and 1 Mayo 3 report were incorrectly classified as lower.

Learning Efficiency

We measured the learning efficiency of the autoML classifiers by measuring their performance using decreasing subsets of the training data. The binary classifier remained 95% (95% CI, 82%-96%) accurate despite training on only 240 notes (60% of the training data) (Table 2). On the ordinal task, the accuracy dropped from 97% (95% CI, 88%-99%) to 70% (95% CI, 66%-74%) when reducing the training data from 242 notes to 194 notes (80% of the dataset).

Generalizability

Many machine learning models are prone to overfitting on irrelevant predictive features and thus fail to generalize to data from other health systems. We assessed the robustness of the finalized autoML classifiers by retraining them on just the UCSF data and evaluating them on data from ZSFG. On the binary prediction task, classifiers trained on the 282 UCSF reports remained 96% accurate when evaluated on the 217 ZSFG notes. We did not assess the ordinal classifier due to insufficient data (only 151 available reports), considering the learning efficiency results as reported previously.

Social Impacts

There has been a growing awareness of the impacts that artificial intelligence has on society in recent years. For example, the rise of automation and our trust in it has the potential to propagate existing social disparities, a phenomenon known as algorithmic unfairness. We assessed our autoML classifiers fairness.

We used linked EHR data to map procedure reports to patient gender and mapped the ADI via residential zip code. We found no evidence of bias by either of these factors, with *P* values of .65 and .80, respectively. We could not assess other variables of a priori importance like race and ethnicity due to severe class imbalance.

Table 2. Learning efficiency of autoML algorithms with successively reduced training data.

Type	Classifier	Training reports	Accuracy (95% CI) (%)
Binary	AutoML LightGBM	400 (100)	98 (89-98)
Binary	AutoML XGBoost	320 (80)	96 (88-97)
Binary	AutoML LightGBM	240 (60)	95 (82-96)
Ordinal	AutoML LightGBM	242 (100)	97 (88-99)
Ordinal	AutoML XGBoost	194 (80)	70 (66-74)
Ordinal	AutoML LightGBM	145 (60)	57 (51-62)

Values are n (%), unless otherwise indicated.
Abbreviations: autoML, automated machine learning; CI, confidence interval.

Discussion

We developed a highly accurate pair of NLP models capable of classifying colonoscopy reports by their MES and recognizing when to abstain from making a prediction. The method yielding the best results across a wide range of metrics was autoML, a computationally lean and powerful framework for training supervised learning models. Our autoML models appeared to learn robust and generalizable predictive features while requiring only a limited amount of effort for annotations. They outperformed BERT-based classifiers, which hold state-of-the-art status on many NLP tasks, as well as cTAKES, a well-established software suite for clinical NLP. Last, they demonstrated evidence of algorithmic fairness.

There are no published, validated models for extracting MES from colonoscopy report free-text description of inflammation. There are published models for automated disease activity scoring of colonoscopy photos[15] and colonoscopy videos[16]; however, high-quality colonoscopy photos and video are variably available in routine clinical practice and are subject to high cost of computation compared with our text-based approach. Our future plans for this work include updating the annotation and training procedure to adapt the

models to the modified MES to allow for the output to include the subscore for each colonic segment and the colonic segment exhibiting the most severe disease.

We acknowledge several limitations. Some of our analyses were underpowered or lacked sufficient data to be analyzed, such as the assessments of algorithmic fairness. Our test set was relatively small. Although we followed standard model development procedures to limit overfitting and employed stratified sampling of the test set, we cannot rule out the possibility of some overfitting and residual bias from the algorithm procedure itself. Nonetheless, we believe that our primary results pertaining to the rank ordering of algorithms are likely to remain robust to this. We noted that the accuracy of the ordinal classifier (97%) was greater than the point estimate of interrater reliability on this task (88%). However, the 95% CIs are consistent with statistical equivalence, and the performance of these models on the generalizability assessment on data from an outside center not seen in training suggest that any overfitting is likely small. Finally, we cannot comment on the degree to which these findings will apply to a broader range of real-world tasks. More work is needed to investigate the generalizability of our findings.

An issue that deserves mention is class imbalance, a common feature in real-world data. In our study, there was a significant class imbalance for the MES 3 reports. We suspect that the suboptimal performance of the BERT models for the ordinal classification task was precisely for this reason. By contrast, ClinicalBERT performed extremely well on the class-balanced, binary classification task. Future solutions to this problem could include preferential sampling with keyword searches, to selectively annotate notes from the minority class.

Algorithms, especially those that rely on gold-standard annotations such as ours, are susceptible to the influence of bias similar to humans.[17] We tested and found no association between our classifier's misclassification rates and either gender or ADI, a measure of socioeconomic status. However, we acknowledge that our study was insufficiently powered to exclude important degrees of bias. Moreover, we note that our assessment only evaluated algorithmic fairness relative to the gold-standard produced by the annotators. This study design cannot exclude possible bias by the original clinician who documented the procedure report. Nonetheless we propose that all algorithms, especially those applied to healthcare, undergo formal evaluations for possible bias. Future work is needed to establish standards of fairness and ensure that unchecked algorithmic biases do not propagate at scale.

Our study has several strengths. We utilized a multicenter corpus of procedure notes encompassing differences in physicians and their documentation styles, patients, and procedure reporting software. We ensured acceptable agreement between expert annotators. In addition to typical metrics like accuracy, we paid attention barriers to widespread adoption, including data hunger. We assessed the social impact of our models, including algorithmic fairness. We are releasing the analytic code to allow other centers to reproduce and extend these results for other uses.

Conclusions

We conducted a multicenter assessment of computational methods for performing text classification of colonoscopy reports to automate the MES. We found that classifiers trained using autoML performed well across a range of metrics, including accuracy, generalizability, learning efficiency, and fairness. Our models open the door to a future of automated endoscopic disease activity scoring to facilitate a number of clinical investigations ranging from large-scale evaluation of treatment outcomes across multiple EHR systems without the limitations of using billing claims codes and hospitalizations as surrogates of endoscopic disease activity to automated clinical trial inclusion criteria evaluation. In addition, these models could allow IBD practices to quickly identify all patients who are well or unwell and design targeted outreach strategies to improve the health of the IBD population. We envision that tools like ours will allow large-scale understanding of UC real-world outcomes and help drive continuous improvements in healthcare.

Supplementary data

Supplementary data is available at *Inflammatory Bowel Diseases* online.

Acknowledgments

Prior versions of this work have been presented at the American College of Gastroenterology Meeting 2020 and Digestive Diseases Week 2022.

Author Contributions

A.L.S.: study concept and design; lead acquisition of data; analysis and interpretation of data; lead drafting of the manuscript; critical revision of the manuscript for important intellectual content. B.B.: study concept and design; acquisition of data; analysis and interpretation of data; drafting of the manuscript; critical revision of the manuscript for important intellectual content; model development. A.M.: acquisition of data; analysis and interpretation of data; technical support; critical revision of the manuscript for important intellectual content. F.Y.: acquisition of data; analysis and interpretation of data; technical support; critical revision of the manuscript for important intellectual content. G.R.: critical revision of the manuscript for important intellectual content; model development. I.B.: critical revision of the manuscript for important intellectual content; model development. S.G.: acquisition of data; analysis and interpretation of data; technical support; critical revision of the manuscript for important intellectual content

T.M.: acquisition of data; analysis and interpretation of data; technical support; critical revision of the manuscript for important intellectual content. R.N.: acquisition of data; analysis and interpretation of data; technical support; critical revision of the manuscript for important intellectual content. J.S.: acquisition of data; critical revision of the manuscript for important intellectual content. A.J.B.: acquisition of data; critical revision of the manuscript for important intellectual content. V.A.R.: study concept and design; acquisition of data; analysis and interpretation of data; technical support; critical revision of the manuscript for important intellectual content; study supervision.

Funding

This study was supported by funding from the UCSF Bakar Computational Health Science Institute and the National

Center for Advancing Translational Sciences of the National Institutes of Health (NIH) (grant number UL1TR001872). V.A.R. was supported by funding from the NIH/National Center for Advancing Translational Sciences (grant number TL1TR001871) and National Library of Medicine of the NIH (award number K99LM014099).

Conflicts of Interest

V.A.R. has received research support from the following for-profit entities: Janssen Research and Development, Alnylam, Merck, Blueprint Medicines, Stryker, Mitsubishi Tanabe, Takeda, and Genentech. V.A.R. is also a shareholder of ZebraMD. A.J.B. is a co-founder and shareholder of and has served as a consultant for Personalis and NuMedii; has served as a consultant for Mango Tree Corporation, Samsung, 10x Genomics, Helix, Pathway Genomics, and Verinata (Illumina); has served on paid advisory panels or boards for Geisinger Health, Regenstrief Institute, Gerson Lehman Group, AlphaSights, Covance, Novartis, Genentech, Merck, and Roche; is a minor shareholder in Apple, Meta (Facebook), Alphabet (Google), Microsoft, Amazon, Snap, 10x Genomics, Illumina, Regeneron, Sanofi, Pfizer, Royalty Pharma, Moderna, Sutro, Doximity, BioNtech, Invitae, Pacific Biosciences, Editas Medicine, Nuna Health, Assay Depot, and Vet24seven; and has received honoraria and travel reimbursement for invited talks from Johnson & Johnson, Roche, Genentech, Pfizer, Merck, Lilly, Takeda, Varian, Mars, Siemens, Optum, Abbott, Celgene, AstraZeneca, AbbVie, Westat, and many academic institutions, medical or disease specific foundations and associations, and health systems. A.J.B. has received royalty payments through Stanford University for several patents and other disclosures licensed to NuMedii and Personalis; and research support from the National Institutes of Health, Peraton (as the prime on a National Institutes of Health contract), Genentech, Johnson & Johnson, the Food and Drug Administration, the Robert Wood Johnson Foundation, the Leon Lowenstein Foundation, the Intervalien Foundation, Priscilla Chan and Mark Zuckerberg, the Barbara and Gerson Bakar Foundation, the March of Dimes, the Juvenile Diabetes Research Foundation, the California Governor's Office of Planning and Research, the California Institute for Regenerative Medicine, L'Oreal, and Progenity, and these institutions had no influence on the study or the manuscript. The remaining authors have nothing to disclose.

Data Availability

The analytic code has been made publicly available at https://github.com/rwelab/MayoClassifier. The data used for this study contain protected health information and thus have not been made available for reuse. However, a machine-redacted version of the data can be made available to requesting researchers by mutual agreement and following the execution of a data use agreement.

References

1. Turner D, Ricciuto A, Lewis A, et al.; International Organization for the Study of IBD. STRIDE-II: an update on the Selecting Therapeutic Targets in Inflammatory Bowel Disease (STRIDE) initiative of the International Organization for the Study of IBD (IOIBD): determining therapeutic goals for treat-to-target strategies in IBD. *Gastroenterology*. 2021;160(5):1570-1583.
2. Food and Drug Administration. Ulcerative Colitis: Clinical Trial Endpoints Guidance for Industry. 2016. Accessed November 30, 2021. https://www.fda.gov/files/drugs/published/Ulcerative-Colitis--Clinical-Trial-Endpoints-Guidance-for-Industry.pdf
3. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*. 2010;17(5):507-513.
4. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12(85):2825-2830.
5. Erickson N, Mueller J, Shirkov A, et al. Autogluon-tabular: robust and accurate automl for structured data. arXiv 2003.06505. doi:10.48550/arXiv.2003.06505, 13 Mar 2020, preprint: not peer reviewed.
6. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv 1810.04805. doi:10.48550/arXiv.1810.04805, 11 Oct 2018, preprint: not peer reviewed.
7. Stidham RW, Yu D, Zhao X, et al. Identifying the presence, activity, and status of extraintestinal manifestations of inflammatory bowel disease using natural language processing of clinical notes. *Inflamm Bowel Dis*. 2023;29(4):503-510.
8. Ananthakrishnan AN, Cai T, Savova G, et al. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. *Inflamm Bowel Dis*. 2013;19(7):1411-1420.
9. Cai T, Lin T-C, Bond A, et al. The association between arthralgia and vedolizumab using natural language processing. *Inflamm Bowel Dis*. 2018;24(10):2242-2246.
10. Gomollón F, Gisbert JP, Guerra I, et al.; Premonition-CD Study Group. Clinical characteristics and prognostic factors for Crohn's disease relapses using natural language processing and machine learning: a pilot study. *Eur J Gastroenterol Hepatol*. 2022;34(4):389-397.
11. Hou JK, Chang M, Nguyen T, et al. Automated identification of surveillance colonoscopy in inflammatory bowel disease using natural language processing. *Dig Dis Sci*. 2013;58(4):936-941.
12. Gundlapalli AV, South BR, Phansalkar S, et al. Application of natural language processing to VA electronic health records to identify phenotypic characteristics for clinical and research purposes. *Summit Transl Bioinform*. 2008;2008:36-40.
13. Montoto C, Gisbert JP, Guerra I, et al.; PREMONITION-CD Study Group. Evaluation of natural language processing for the identification of Crohn disease–related variables in Spanish electronic health records: a validation study for the PREMONITION-CD Project. *JMIR Med Inform*. 2022;10(2):e30345.
14. Knighton AJ, Savitz L, Belnap T, Stephenson B, VanDerslice J. Introduction of an area deprivation index measuring patient socioeconomic status in an integrated health system: implications for population health. *EGEMS (Wash DC)*. 2016;4(3):1238.
15. Stidham RW, Liu W, Bishu S, et al. Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis. *JAMA Netw Open*. 2019;2(5):e193963.
16. Yao H, Najarian K, Gryak J, et al. Fully automated endoscopic disease activity assessment in ulcerative colitis. *Gastrointest Endosc*. 2021;93(3):728-736.e1.
17. Jain A, Ravula M, Ghosh J. Biased models have biased explanations. arXiv 2012.10986. doi:10.48550/arXiv.2020.10986, 20 Dec 2020, preprint: not peer reviewed.