

**Title:** Accurate, Robust, and Scalable Machine Abstraction of Mayo Endoscopic Subscores from Colonoscopy Reports

**Authors:** Anna L. Silverman, MD<sup>1,2,9</sup>, Balu Bhasuran, PhD<sup>3,9</sup>, Arman Mosenia, MSE<sup>4</sup>, Fatema Yasini, BS<sup>5</sup>, Saransh Gupta, BS<sup>5</sup>, Taline Mardrossian, BS<sup>5</sup>, Rohan Narain, BS<sup>5</sup>, Justin Sewell, MD, PhD<sup>6, 8</sup>, Atul J. Butte, MD, PhD<sup>3,7</sup>, Vivek A. Rudrapatna, MD, PhD<sup>3,8</sup>

<sup>1</sup>Division of Gastroenterology and Hepatology, Department of Medicine, Mayo Clinic, Phoenix, AZ, 85054

<sup>2</sup>Department of Medicine, University of California, San Diego, La Jolla, CA, 92093

<sup>3</sup>Bakar Computational Health Sciences Institute, San Francisco, CA, 94143

<sup>4</sup>School of Medicine, University of California San Francisco, San Francisco, CA 94143

<sup>5</sup>Department of Computer Science, University of California Berkeley, Berkeley, CA, 94720

<sup>6</sup>Division of Gastroenterology, Department of Medicine, Zuckerberg San Francisco General Hospital, San Francisco, CA, 94110

<sup>7</sup>Center for Data-Driven Insights and Innovation, University of California Health, Oakland, CA, 94607

<sup>8</sup>Division of Gastroenterology and Hepatology, Department of Medicine, University of California, San Francisco, San Francisco, CA, 94143

<sup>9</sup>These authors contributed equally to this work

**Corresponding Author**

Vivek A. Rudrapatna, MD PhD

Assistant Professor of Medicine

University of California, San Francisco Bakar Institute, Box 2993

490 Illinois Street, Floor 2

San Francisco, CA 94143

Email: [vivek.rudrapatna@ucsf.edu](mailto:vivek.rudrapatna@ucsf.edu)

## Key Points

**Question:** Is natural language processing (NLP) a viable alternative to manually abstracting disease activity from procedure notes?

**Findings:** We compared different methods for abstracting the ulcerative colitis Mayo endoscopic subscore from colonoscopy reports. Classifiers trained using automated machine learning (autoML) achieved the greatest accuracy (97%), recognized when to abstain, generalized well to other health systems, required limited effort for annotation and programming, demonstrated fairness, and had a small carbon footprint.

**Meaning:** NLP methods like autoML appear to be sufficiently mature technologies for clinical text classification, and thus are poised to enable many downstream endeavors using electronic health records data.

## Abstract

**Importance:** Electronic health records (EHR) data are growing in importance as a source of evidence on real-world treatment effects. However, many clinically important measures are not directly captured as structured data by these systems, limiting their utility for research and quality improvement. Although this information can usually be manually abstracted from clinical notes, this process is expensive and subject to variability. Natural language processing (NLP) is a scalable alternative but has historically been subject to multiple limitations including insufficient accuracy, data hunger, technical complexity, poor generalizability, algorithmic unfairness, and an outsized carbon footprint.

**Objective:** Compare different algorithmic approaches for classifying colonoscopy reports according to their ulcerative colitis Mayo endoscopic subscores

**Design:** Other observational study – NLP algorithm development and validation

**Setting:** Academic medical center (UCSF) and safety-net hospital (ZSFG) in California

**Participants:** Patients with ulcerative colitis

**Exposures:** Colonoscopy

**Main Outcomes and Measures:** The primary outcome was accuracy in identifying reports suitable for Mayo subscore (binary yes/no) and then separately assigning a Mayo subscore where relevant (ordinal). Secondary outcomes included learning efficiency from training data, generalizability, computational costs, fairness, and sustainability.

**Results:** Using automated machine learning (autoML) we trained a pair of classifiers that were 98% [91-99%] accurate at determining which reports to score and 97% [88-99%] accurate at assigning the correct Mayo endoscopic subscore. The binary classifiers trained on UCSF data achieved 96% accuracy on hold-out test data from ZSFG. Training these classifiers required 4 hours of computation on a standard laptop. Classification errors were not associated with either gender or area deprivation index. The carbon footprint of this approach was 24x less than current deep learning algorithms for clinical text classification.

**Conclusions and Relevance:** We identified autoML as an efficient and robust method for training clinical text classifiers. AutoML-trained classifiers demonstrated many favorable properties including generalizability, limited effort needed for data annotation and algorithm training, fairness, and sustainability. More generally, these results support the feasibility of using unstructured EHR data to generate real-world evidence and drive continuous

improvements in learning health systems.

## Introduction

Real-world evidence (RWE) refers to the use of observational data for the study of diseases and treatments.<sup>1,2</sup> Interest in using electronic health records (EHRs) for RWE continues to grow, as these systems capture detailed data on patient diagnoses, treatments and outcomes over time. As such, they have the potential to be useful for many purposes, from regulatory evaluations of drugs and devices to scalable assessments of healthcare quality and value.<sup>3,4</sup>

A challenge with using these data is that many critical elements are captured in a free-text form rather than analysis-ready data. Thus, many EHR-based analyses utilize the structured data alone and ignore the content in notes. Although this approach is attractive because of its feasibility, it increases the risks of confounding and other biases in observational research.<sup>5,6</sup> This is particularly the case for chronic conditions where disease assessments are complex and captured as free-text, and where these assessments inform decision making. Manual abstraction of these elements from notes has been a common strategy to minimize this bias. However, this approach is fundamentally unscalable. It also requires expertise and is subject to interrater variability and reviewer fatigue.

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 for several decades, with dozens of software packages now freely available.<sup>7-10</sup> While these technologies have opened opportunities for scalable and robust analyses of real-world data, they too are not

without their challenges. These methods have generally not been sufficiently developed or validated for most abstraction tasks due to their sheer number. Thus, dedicated efforts by individual groups are typically needed to evaluate and enhance these methods for specific uses. Doing this can require significant technical capabilities, including programming experience and specialized hardware. Many modern NLP methods can be very data inefficient, placing large burdens on annotators and decreasing the overall scalability of the algorithm-development process.

Lastly, these methods can be associated with many negative societal consequences. Many black box methods can overfit on clinically immaterial features like socioeconomic status. As a result, they can propagate healthcare disparities and threaten algorithmic robustness unless dedicated precautions are taken.<sup>11-14</sup> Algorithmic training can also be unsustainable for the environment, with single algorithms being associated with a carbon footprint equivalent to the lifetime emission of five cars.<sup>15</sup>

We sought to determine if current NLP methods can be considered a viable alternative to manual abstraction from the EHR. 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 RWE studies. We selected a use case of ulcerative colitis (UC), a chronic inflammatory disease of the large intestine. Using the colonoscopy reports from two medical centers, we compared different classifiers by their ability to process colonoscopy reports and abstract the Mayo endoscopic subscore, a disease activity measure commonly used in UC registrational trials.<sup>16</sup> Our primary endpoint was accuracy on the sequential tasks of identifying which reports could be scored using the Mayo endoscopic subscore and assigning a subscore if

appropriate. Secondary endpoints included learning efficiency, generalizability, programming effort, algorithmic fairness, and carbon emissions.

## **Methods**

### *Procedure Reports*

To identify colonoscopy reports for classifier training and evaluation, we accessed the EHRs at two 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 the back end EHR databases to identify all patients who had ever been assigned an ICD-10 code for inflammatory bowel disease (K50\*, K51\*), and extracted all corresponding colonoscopy reports from the 2017-2020 period.

### *Annotation Procedure*

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

In the second stage, suitable reports were assigned a Mayo endoscopic subscore, an 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 (Supplemental Methods).<sup>16</sup>

The interrater agreement of this process was assessed on a separate set of 50 uniformly sampled reports.

### *Algorithm Development and Validation*

We developed and evaluated four standard methods for abstracting information from notes: cTAKES<sup>7</sup>-based concept recognition, bag-of-words models using sklearn<sup>8</sup> and automated machine learning (autoML<sup>9</sup>), as well as three models related to BERT<sup>10</sup> (Supplemental Methods). These methods vary in their underlying technique, requirements for training data, tendencies towards robust and generalizable learning, and ease of use.

We used these methods to separately train a binary classifier (to predict which procedure reports were Mayo scorable) and an ordinal classifier (to predict the correct Mayo endoscopic subscore 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 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.<sup>17</sup> 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.

### *Carbon Emissions*

We used a public machine learning (ML) emission calculator to estimate the environmental impact of our algorithms (Supplemental Methods).<sup>18</sup>

### *Statistics*

We computed exact binomial confidence intervals 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 a Wald test. We performed all computing using *R* 4.1.3 and *Python* 3.

### *Ethics*

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

## **Results**

### *Procedure Reports*

The source corpus consisted of 3,769 notes from UCSF and 835 from ZSFG, all authored between 2017 and 2020 (Figure 1). The manually annotated corpus consisted of 499 notes of which 282 were from UCSF and 217 from ZSFG. 302 notes were eligible for Mayo endoscopic scoring, with 151 notes from each site. Inter-annotator agreement was 96% [86-100%] for the binary task (N=50) and 88% [69-97%] for the ordinal task (N=25).

### *Algorithmic Accuracy*



The autoML-trained classifiers achieved the overall highest accuracy. They were 98% [91-99%] accurate at identifying Mayo scorable reports and 97% [88-99%] accurate at assigning a Mayo subscore (Table 2). 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. To our surprise our 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%).

#### *Learning efficiency, technical feasibility*

Although high accuracy is typically the primary objective of most endeavors involving NLP, practicality is an almost equally important consideration. Many supervised learners like BERT can require hundreds to thousands of expert annotations, an unrealistic requirement in many contexts. 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% [82-96%] accurate despite training on only 240 notes (60% of the training data) (Table 3). On the ordinal task, the accuracy dropped from 97% [88-99%] to 70% [66-74%] when reducing the training data from 242 notes to 194 notes (80% of the dataset). This was consistent with our expectation that more data would be required for the ordinal task, given the multiplicity of

target classes and degree of class imbalance. These results supported the utility of autoML as a method for efficiently developing text classifiers in practical contexts.

AutoML appeared favorable from a technical perspective as well. We produced these classifiers with 4 hours of computation on a standard laptop. By contrast, the BERT-related models required substantial technical troubleshooting on a hardware-intensive environment.

### *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 re-training them on just the UCSF data and evaluating it 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 above.

### *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. In addition, the training of some models such as BERT can be surprisingly unsustainable for the environment, with carbon footprints equivalent to the lifetime exhaust produced from five cars.<sup>15</sup> We assessed our autoML classifiers along these two dimensions: fairness and sustainability.

We used linked EHR data to map procedure reports to patient gender, and mapped area deprivation index (ADI) via residential zip code. We found no evidence of bias by either of these factors, with p-values of 0.65 and 0.80 respectively. We could not assess other variables of *a priori* importance like race and ethnicity due to severe class imbalance.

Finally, we used a previously published method for quantifying the carbon emissions of machine learning algorithms.<sup>18</sup> AutoML training required 0.0216 kilogram equivalents of CO<sub>2</sub>, comparable to charging a smartphone 3 times.<sup>19</sup> This was 24 times less costly than the process of fine tuning ClinicalBERT on the annotated dataset, and roughly 5,000 times less costly than training a dedicated, BERT-base algorithm.

## Discussion

We compared several computational methods for transforming routinely documented clinical text into quantitative, analysis-ready data. As our use case we selected two sequential tasks related to the Mayo endoscopic subscore, a key measure of ulcerative colitis disease activity. The method yielding the best results across a wide range of metrics was automated machine learning (autoML), a computationally lean and powerful framework for training supervised learning models. These classifiers were highly accurate at assigning Mayo subscores to colonoscopy reports and recognizing when to abstain. They appeared to learn robust and generalizable predictive features while requiring only a limited amount of effort for annotations and programming. 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. Lastly, they demonstrated evidence of algorithmic fairness and environmental sustainability.

Recent years have seen a growing interest in the use of electronic health records data to close evidence gaps and improve quality and value in healthcare. However, a persistent bottleneck in the optimal use of these data has been the large burden of analytically inaccessible data captured as free text. Over the past several decades, clinical NLP has made substantial strides towards the goal of accurate and practical computational alternatives to manual data abstraction. It has explored a multitude of technical approaches, from rule-based expert systems incorporating clinical knowledge to data-driven, supervised machine learning. Progress in this field has generally been uneven, with different solutions typically reflecting tradeoffs between accuracy, flexibility, generalizability, learning efficiency, and technical expertise. Consequently, the use of NLP in clinical research and operations has largely remained in the domain of specialized technical laboratories but otherwise failed to achieve widespread adoption to date.

Our results provide early evidence that the field of “applied clinical NLP” has finally arrived. It appears that we finally have enough tools at our disposal to be able to solve many clinical text-related tasks in a scalable way, and perhaps NLP can realistically be considered as a first-line option for many clinical teams with limited resources and technical expertise. The advent of autoML technologies exemplifies this perspective particularly well, given their unique combination of performance and user friendliness.

However, we note that no algorithm will solve all problems equally well, and different solutions will be required for different types of problems.<sup>20</sup> We outline our approach in Table 4. Simple tasks that may not require textual context may benefit from keyword- and rule-based methods, whereas complicated classification tasks may require supervised approaches like

autoML and BERT. It is worth noting that it is not always easy to characterize the difficulty of tasks *a priori*. We expected the rule-based model to perform well because of the clear mapping between natural language descriptors and Mayo subscores. The reasons for its mediocre performance remain unclear. We elected not to investigate this further given the identification of far more performant models and the overall objectives of this study.

Another 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 Mayo 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.<sup>11</sup> 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.

Finally, we note that our autoML classifiers were associated with a small carbon footprint. The emissions associated with these models was 24 times less than a fine-tuned ClinicalBERT, and roughly 5,000 times less than a dedicated BERT model. We do acknowledge that the relative environmental stakes between the choice of an autoML model versus a pre-trained BERT model are modest, considering the innumerable and substantial sources of anthropogenic carbon emissions. However, we foresee the continued growth of machine learning and data science across a range of endeavors including healthcare. It is incumbent upon all of us using these technologies to do our part, raise awareness within our communities of practices, and curb these multiplicative effects.<sup>21</sup> We must also modernize our reporting standards, and treat the reporting of environmental self-audits with the same level of rigor as we have come to expect with other ethics-related disclosures in research.

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 to feasibility and other barriers to widespread adoption, including data hunger and technical requirements. We assessed the social impacts of our models, including algorithmic fairness and carbon footprint. We are releasing the analytic code to reproduce and extend these results for other uses.

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 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% confidence intervals are consistent with statistical equivalence, and the performance of these models on the generalizability assessment 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.

In conclusion, we conducted a multicenter assessment of computational methods for performing text classification. We adopted a use case of abstracting Mayo endoscopic scores from colonoscopy procedure reports. We found that classifiers trained using autoML performed well across a range of metrics, including accuracy, generalizability, learning efficiency, programming effort, technical requirements, fairness, and carbon footprint. We propose that this method be a key element in the toolkit of clinical and data science teams working with free-text, and that this overall technology be considered a viable option for situations where efficient, scalable, and consistent annotations of text are needed.

#### **Access to Data:**

The analytic code has been made publicly available at <https://github.com/rwelab/MayoClassifier>. The data used for this study contains 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.

## Disclosures:

VAR receives research support from the following for-profit entities: Janssen Research and Development, Alnylam, and Genentech. Atul Butte is a co-founder and consultant to Personalis and NuMedii; consultant to Mango Tree Corporation, and in the recent past, 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, and Merck, and Roche; is a shareholder in Personalis and NuMedii; 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 several other non-health related companies and mutual funds; and has received honoraria and travel reimbursement for invited talks from Johnson and 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. Atul Butte receives royalty payments through Stanford University, for several patents and other disclosures licensed to NuMedii and Personalis. Atul Butte's research has been funded by NIH, Peraton (as the prime on an NIH contract), Genentech, Johnson and Johnson, FDA, Robert Wood Johnson Foundation, Leon Lowenstein Foundation, Intervalien Foundation, Priscilla Chan and Mark Zuckerberg, the Barbara and Gerson Bakar Foundation, and in the recent past, the March of Dimes, Juvenile Diabetes Research Foundation, California Governor's Office of Planning and Research, California Institute for Regenerative Medicine, L'Oreal, and Progenity. These institutions had no influence



on the study or the manuscript. The remaining authors have nothing to disclose nor any conflicts of interest.

### **Funding Source:**

This study was supported by funding from the UCSF Bakar Computational Health Science Institute and the National Center for Advancing Translational Sciences of the NIH, grant number UL1TR001872. VAR was supported by funding from the NIH/National Center for Advancing Translational Sciences, grant number TL1TR001871.

### **Previous Presentation of Information:**

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

### **Works Cited**

1. Rudrapatna VA, Butte AJ. Opportunities and challenges in using real-world data for health care. *J Clin Invest*. 2020;130(2):565-574.
2. Rudrapatna VA, Glicksberg BS, Butte AJ. Utility of routinely collected electronic health records data to support effectiveness evaluations in inflammatory bowel disease: a pilot study of tofacitinib. *BMJ Health Care Inform*. 2021;28(1).
3. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products. 2021; <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory>. Accessed December 14, 2021.
4. Hou JK, Imler TD, Imperiale TF. Current and future applications of natural language processing in the field of digestive diseases. *Clin Gastroenterol Hepatol*. 2014;12(8):1257-1261.
5. Hou JK, Tan M, Stidham RW, et al. Accuracy of diagnostic codes for identifying patients with ulcerative colitis and Crohn's disease in the Veterans Affairs Health Care System. *Dig Dis Sci*. 2014;59(10):2406-2410.
6. Thirumurthi S, Chowdhury R, Richardson P, Abraham NS. Validation of ICD-9-CM diagnostic codes for inflammatory bowel disease among veterans. *Dig Dis Sci*. 2010;55(9):2592-2598.

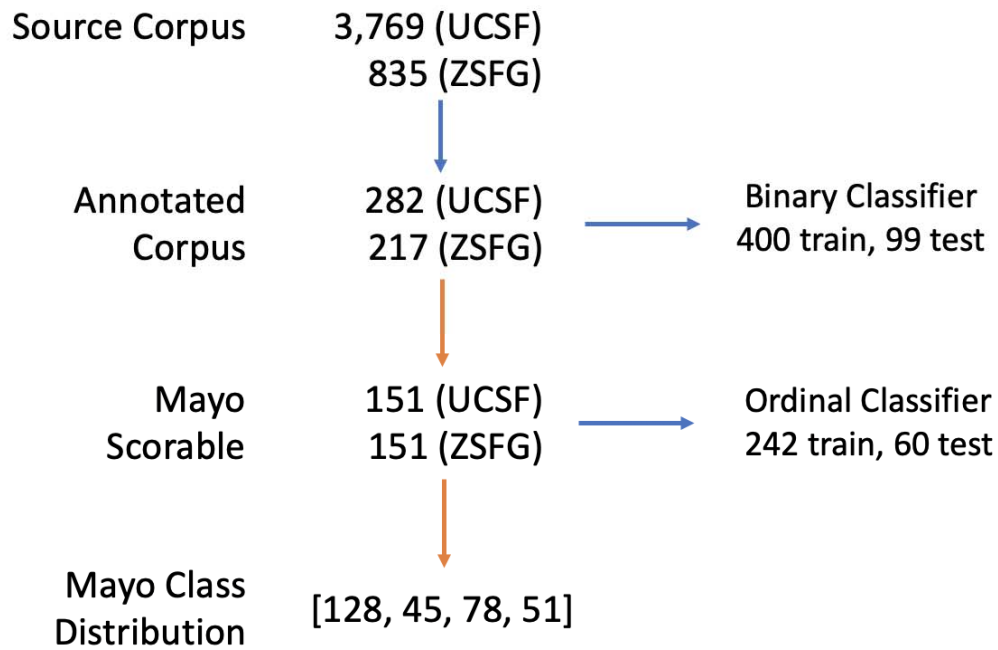
7. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*. 2010;17(5):507-513.
8. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*. 2011;12:2825-2830.
9. Erickson N, Mueller J, Shirkov A, et al. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:200306505*. 2020.
10. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:181004805*. 2018.
11. Jain A, Ravula M, Ghosh J. Biased models have biased explanations. *arXiv preprint arXiv:201210986*. 2020.
12. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*. 2018;51(5):1-42.
13. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. 2019;1(5):206-215.
14. Schramowski P, Turan C, Andersen N, Rothkopf CA, Kersting K. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*. 2022;4(3):258-268.
15. Strubell EG, Ananya; McCallum, Andrew. Energy and Policy Considerations for Deep Learning in NLP. *arXiv preprint arXiv:190602243*. 2019.
16. Ulcerative Colitis: Clinical Trial Endpoints Guidance for Industry. 2016; <https://www.fda.gov/files/drugs/published/Ulcerative-Colitis--Clinical-Trial-Endpoints-Guidance-for-Industry.pdf>. Accessed November 30, 2021.
17. 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*. 2016;4(3).
18. Lacoste A, Luccioni A, Schmidt V, Dandres T. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:191009700*. 2019.
19. Greenhouse Gas Equivalencies Calculator. 2022; <https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator#results>.
20. Wolpert DH, Macready WG. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*. 1997;1(1):67-82.
21. Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. 2022; <https://www.un.org/en/climatechange/ipcc-wgii-report>.

## Tables and Figures

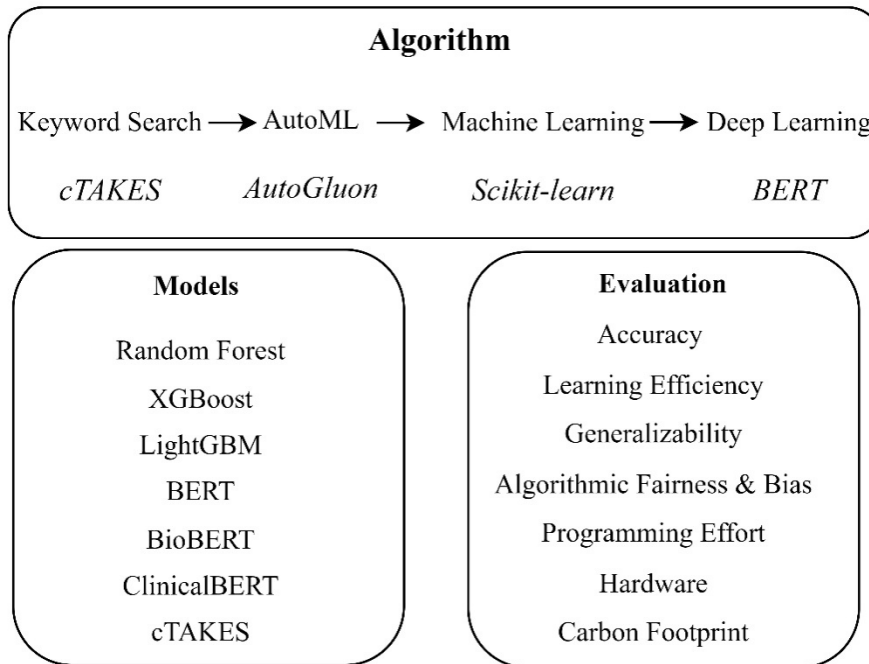
**Table 1: Mayo Endoscopic Score Keywords.** Keywords used to delineate each Mayo score for the keyword search-based algorithm. All procedure reports were classified according to the highest Mayo score for which a corresponding, non-negated concept was identified.

<b>Mayo Score</b>	<b>Keywords</b>
<b>0</b>	'normal', 'quiescent', 'scar'
<b>1</b>	'erythema', 'decreased vascular pattern', 'granularity', 'aphthous ulcer', 'aphthae', 'mild'
<b>2</b>	'friability', marked or extensive 'erythema', 'loss of vascularity', 'absent vascularity', 'erosions', 'moderate'
<b>3</b>	spontaneous 'bleeding', 'ulcer', 'ulcerated', 'severe'

**Figure 1: Corpus Flow Diagram and Algorithm Architectures**



Panel A: Flow diagram of the procedure reports selected for annotation. Orange arrows depict NLP prediction tasks. Numbers associated with the Mayo Class Distribution correspond to those scored as a 0, 1, 2 or 3 respectively.



Panel B: Software and algorithm architectures utilized. Arrows in the algorithm box depict increasing complexity, requirements for training data, and programming effort.

**Table 2: Algorithmic performance for the Mayo scorability (Binary) and the Mayo endoscopic subscore (Ordinal) prediction tasks.**

Algorithm	Framework	Type	Accuracy (95% CI)	F-Score	Programming Effort	Lines of Code	Hardware
RandomForest	autoML	Binary	<b>98% (91-99)</b>	<b>97.47</b>	1 week	12	CPU
LightGBM	Scikit-learn	Binary	<b>98% (89-98)</b>	97.00	4 days	124	CPU
ClinicalBERT	Transformer	Binary	93% (84-96)	93.00	3 weeks	146	GPU
Null Model	Scikit-learn	Binary	60% (50-70)	-	-	-	-
BERT	Transformer	Binary	47% (37-57)	44.68	3 weeks	146	GPU
BioBERT	Transformer	Binary	38% (28-48)	39.00	3 weeks	146	GPU
XGBoost	autoML	Ordinal	<b>97% (88-99)</b>	<b>97.00</b>	1 week	12	CPU
LightGBM	Scikit-learn	Ordinal	77% (63-87)	77.10	4 days	124	CPU
ClinicalBERT	Transformer	Ordinal	61% (48-74)	61.66	3 weeks	146	GPU
Null Model	Scikit-learn	Ordinal	41% (29-55)	-	-	-	-
BioBERT	Transformer	Ordinal	50% (37-63)	50.00	3 weeks	146	GPU
BERT	Transformer	Ordinal	38% (26-52)	38.33	3 weeks	146	GPU
Rule based using clinical concepts	cTAKES, RegEx	Ordinal	22% (12-37)	22.10	1 day	19	CPU

**Table 3: Learning efficiency of AutoML Algorithms**

**A)**

<b>Binary Classifier</b>	<b>Training Reports</b>	<b>Accuracy (95% CI)</b>
AutoML Light GBM	400 (100%)	98% (89-98)
AutoMLXGBoost	320 (80%)	96% (88-97)
AutoMLLightGBM	240 (60%)	95% (82-96)

A. Binary classifier performance with successively reduced training data for learning

**B)**

<b>Ordinal Classifier</b>	<b>Training Reports</b>	<b>Accuracy(95% CI)</b>
AutoMLLight GBM	242 (100%)	97% (88-99)
AutoMLXGBoost	194 (80%)	70% (66-74)
AutoMLLightGBM	145 (60%)	57% (51-62)

B. Ordinal classifier performance with successively reduced training data for learning

**Table 4: Selecting the best algorithmic approach for clinical information extraction and/or text classification**

<i>Clinical Task Complexity</i>	<i>Recommended procedure for classifier development</i>
<i>Simple: amenable to rule- and keyword-based classification</i>	<ul style="list-style-type: none"><li>• If clinical named entity recognition (NER) software are installed and readily available, then recommend using this to develop a concept-based classifier as a first step.</li><li>• If NER software is unavailable, recommend developing a baseline classifier using regular expressions. Regular expressions may also be useful to improve a sub-performant NER-based algorithm, particularly if the latter systematically misses certain kinds of concepts such as abbreviations</li><li>• Recognize that iterative improvement may be needed to achieve target performance. Maintain separate validation and test sets stratified by the outcome and important effect modifiers. Use the test set once, only if the algorithmic performance on the validation set surpasses the target performance with some margin.</li><li>• Monitor the time and effort spent on iterative enhancements as well as incremental gains and periodically reassess the need to switch to a supervised learning approach (see below). If an error-prone domain of the dataset can be clearly identified, consider the use of a hybrid approach (supervised learning just on the error-prone subset).</li></ul>
<i>Intermediate: More complex than keyword-based, some</i>	<ul style="list-style-type: none"><li>• Develop a training/validation dataset as separate from the test set. We recommend a few hundred notes to start.</li><li>• During the annotation process monitor the dataset for early evidence of significant class imbalance. If identified, consider</li></ul>



*context needed*

using keyword-based heuristics to preferentially sample and annotate notes from the minority class. Consider maintaining two test sets, one using uniform sampling and the other using preferential sampling as above.

- First, try an autoML package such as autoGluon<sup>9</sup> or mlJAR<sup>22</sup>
- We generally do not recommend trying an sklearn-based method unless there is a specific desire to modify or directly tune the modeling approach and the team has the necessary expertise.

*Complex:  
Context-  
dependent,  
potentially  
knowledge-  
dependent*

- Recommend using AutoML as the first approach given its relative advantages over other models. Consider using longer n-grams as textual features.
- If inadequate, consider using a clinical BERT model. However, you should anticipate the need for more annotated data (closer to a thousand examples) with class balance as well as greater technical expertise.

