

## ORIGINAL RESEARCH—CLINICAL

# Leveraging Natural Language Processing to Accurately Identify Histologically Confirmed Esophageal and Gastric Premalignant and Malignant Conditions Using a Large-Scale Health-Record Linked Genomic Biobank



Aaron Giera,<sup>1</sup> Kyriacos Markianos,<sup>1</sup> Andrew J. Gawron,<sup>2,3</sup> VA Million Veteran Program, Samir Gupta,<sup>4,5</sup> Saiju Pyarajan,<sup>1,6</sup> Purushotham Karnam,<sup>1</sup> Lin Liu,<sup>4,7</sup> and Shailja C. Shah<sup>4,5</sup>

<sup>1</sup>Center for Data and Computational Sciences (C-DACS), VA Boston Healthcare System, Boston, Massachusetts;

<sup>2</sup>Gastroenterology Section, VA Salt Lake City Health Care System, Salt Lake City, Utah; <sup>3</sup>Division of Gastroenterology, Department of Internal Medicine, University of Utah, Salt Lake City, Utah; <sup>4</sup>Jennifer Moreno Department of Veterans Affairs Medical Center, San Diego, California; <sup>5</sup>Division of Gastroenterology and Hepatology, University of California, San Diego, California; <sup>6</sup>Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; and <sup>7</sup>Division of Biostatistics and Bioinformatics, Herbert Wertheim School of Public Health and Human Longevity Science, University of California, San Diego, California

**BACKGROUND AND AIMS:** Our understanding of the epidemiology and natural history of gastroesophageal precancerous conditions is limited by a lack of robust analyses using large-scale individual-level data. We aimed to develop and validate a natural language processing (NLP) algorithm to identify and phenotype esophageal and gastric precancerous conditions and cancer and apply this to the Million Veteran Program (MVP), a uniquely powerful nationwide genomic biobank linked to individual-level electronic health record data. **METHODS:** We identified 121,808 individuals in MVP who underwent upper endoscopy with biopsies. From these, 426 pathology notes from 426 individuals were used to develop and manually validate an NLP rule-based algorithm identifying intestinal metaplasia, dysplasia, and tumors of the stomach or esophagus. Anatomic subsite and “qualifier” terms were also evaluated (eg, dysplasia grade). Performance metrics were calculated. **RESULTS:** The algorithm identified all prespecified conditions with excellent accuracy, ranging from 97.6% to 100% (Bonferroni-corrected 95% lower bound 94.5%–98.5%). For gastric intestinal metaplasia, the algorithm achieved 91.7% precision, or positive predictive value and 86.8% recall, or sensitivity (F1 score 89.2%), with 99.0% specificity and 98.3% negative predictive value; while for Barrett's esophagus, it achieved 98.9% precision and recall, or sensitivity (F1 score 98.9%), with 99.6% specificity and 99.6% negative predictive value. When applied to the full MVP cohort (N = 121,808), 13.2% had gastric intestinal metaplasia (mean age 65 years) and 14.5% had Barrett's esophagus (mean age 64 years). **CONCLUSION:** This study confirms the ability to use NLP on large-scale unstructured data linked to robust genetic and clinical data for future gastroesophageal precancer analyses and to inform targeted prevention and/or early detection interventions (eg, endoscopic surveillance).

**Keywords:** Artificial intelligence; gastrointestinal neoplasm; helicobacter pylori; gastric intestinal metaplasia; Barrett's esophagus

## Introduction

Gastric cancer and esophageal cancer remain leading causes of cancer and cancer death globally.<sup>1</sup> Both esophageal and gastric adenocarcinoma are preceded by precancerous conditions—namely intestinal metaplasia and dysplasia—that are amenable to endoscopic surveillance for early cancer detection. Detection of early-stage disease is associated with substantially improved survival.<sup>2,3</sup> While the majority of people with gastric intestinal metaplasia (GIM) and esophageal intestinal metaplasia (Barrett's esophagus (BE)) will not progress to cancer, we are not able to accurately predict who will vs will not progress based on our current incomplete understanding of mechanisms driving progression and lacking robust clinical prediction models, particularly for GIM. Endoscopy is a limited, costly resource that is not without human risk. There is an unmet need to develop robust prediction models that ideally leverage genetic and nongenetic environmental and clinical determinants to precisely discriminate between individuals at high- vs low-risk for malignant progression.

**Abbreviations used in this paper:** BE, Barrett's esophagus; EGD, esophagogastroduodenoscopy; EHR, electronic health record; GIM, gastric intestinal metaplasia; IND, indefinite for dysplasia; LB, lower bound; LLMs, Large language models; LGD, low-grade dysplasia; MVP, Million Veteran Program; NLP, natural language processing; NPV, negative predictive value; PPV, positive predictive value; VA, Veterans Affairs; VHA, Veterans Health Administration.

Most current article

Published by Elsevier Inc. on behalf of the AGA Institute. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

2772-5723

<https://doi.org/10.1016/j.gastha.2025.100768>

There are limited large-scale studies analyzing histologically confirmed gastroesophageal precancerous conditions linked to longitudinal individual-level data. This literature gap relates to the fact that structured data cannot be used to accurately identify GIM<sup>4</sup>; in fact, before October 1, 2021, GIM did not have an associated International Classification of Diseases code. While structured data may achieve acceptable positive predictive value (precision) for a BE diagnosis,<sup>5</sup> the overall accuracy as well as negative predictive value (NPV), particularly among at-risk populations, is suboptimal. Moreover, structured data cannot achieve the comprehensive phenotyping for gastroesophageal precancerous conditions that is critical for risk stratification and, ultimately, clinical application.

Natural language processing (NLP) can be used to accurately identify conditions that rely on unstructured data, as well as refine phenotypes. NLP has been applied extensively to colonoscopy and associated pathology reports to comprehensively phenotype clinical conditions such as colorectal polyps (size, histology, location, resection), cancer, and inflammatory bowel disease.<sup>6–8</sup>

The Veterans Affairs (VA) Million Veteran Program (MVP) is an electronic health record (EHR)-linked genomic biobank that continues to prospectively enroll Veterans nationwide and link their genotype information in a de-identified manner to their comprehensive longitudinal health-care data, including GI procedures, histopathology reports, validated cancer registry and death data. MVP is a representative sample of the nationwide Veterans Health Administration (VHA) database, which contains the longitudinal clinical health-care data for over 12 million Veterans. Herein, we aimed to leverage MVP to (1) develop and validate a rule-based NLP approach to accurately identify and phenotype gastroesophageal precancerous conditions and cancer, including characterization of anatomic location (esophagus, gastroesophageal junction, stomach and its subsites) and histologic features and (2) apply this validated algorithm to nationwide MVP database to assemble a cohort of patients with histologically confirmed gastroesophageal precancer who can be evaluated in future studies, ideally to develop precise risk prediction models that incorporate genetic and environment factors.

## Methods

### Data Source/Corpus

We compiled clinical data from the VHA Corporate Data Warehouse, which includes the longitudinal patient data for all US Veterans receiving care through any VHA facility in the US and its territories (N ~ 12 million). Longitudinal data elements include demographics, comorbidities, labs, pharmacy prescription data, clinical encounter notes, procedural occurrences and notes, pathology reports, cancer registry data, death data, among other data elements, and all associated dates. From the Corporate Data Warehouse, we selected the cohort of Veterans also enrolled in the MVP, given one important objective of this work is to identify relevant gene

**Table 1.** Gastroesophageal Anatomic Locations and Conditions of Interest

| Anatomic location                                                                                                                                                                                                                                                                        | Conditions and qualifiers                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Stomach</b> <ul style="list-style-type: none"> <li>• Noncardia <ul style="list-style-type: none"> <li>◦ Antrum (includes pylorus)</li> <li>◦ Incisura</li> <li>◦ Body (includes fundus)</li> <li>◦ Other/not otherwise specified<sup>a</sup></li> </ul> </li> <li>• Cardia</li> </ul> | <b>Intestinal metaplasia</b> <ul style="list-style-type: none"> <li>• Qualifiers <ul style="list-style-type: none"> <li>◦ Focal vs extensive</li> <li>◦ Incomplete vs complete</li> </ul> </li> </ul> <b>Dysplasia</b> <ul style="list-style-type: none"> <li>• Qualifiers <ul style="list-style-type: none"> <li>◦ Indefinite (IND)</li> <li>◦ Low-grade (LGD)</li> <li>◦ High-grade (HGD)</li> </ul> </li> </ul> <b>Tumor (“cancer”)</b> <ul style="list-style-type: none"> <li>• Qualifiers <ul style="list-style-type: none"> <li>◦ Adenocarcinoma</li> <li>◦ Lymphoma</li> <li>◦ Neuroendocrine tumor/carcinoid</li> <li>◦ Gastrointestinal stromal tumor</li> <li>◦ Other</li> </ul> </li> </ul> |
| <b>Esophagus</b><br><b>Gastroesophageal junction<sup>b</sup></b>                                                                                                                                                                                                                         | <b>Intestinal metaplasia</b><br><b>Dysplasia</b> <ul style="list-style-type: none"> <li>• Qualifiers <ul style="list-style-type: none"> <li>◦ Indefinite (IND)</li> <li>◦ Low-grade (LGD)</li> <li>◦ High-grade (HGD)</li> </ul> </li> </ul> <b>Tumor (“cancer”)</b> <ul style="list-style-type: none"> <li>• Qualifiers <ul style="list-style-type: none"> <li>◦ Adenocarcinoma</li> <li>◦ Lymphoma</li> <li>◦ Neuroendocrine tumor/carcinoid</li> <li>◦ Gastrointestinal stromal tumor</li> <li>◦ Other</li> </ul> </li> </ul>                                                                                                                                                                       |

<sup>a</sup>Examples include “gastric lesser curvature” or “gastric greater curvature” without mention of antrum or body.

<sup>b</sup>For analysis purposes, the gastroesophageal junction was combined with the esophagus (e.g., gastroesophageal junction intestinal metaplasia was considered under Barrett’s esophagus). “Synonyms” of gastroesophageal junction included z-line and squamocolumnar junction.

and gene-environment interactions associated with gastroesophageal precancer and cancer, as well as progression. From the MVP cohort (MVP version 22.1 Release 4), we selected all individuals who had undergone at least one esophagogastroduodenoscopy (EGD) (*Common Procedural Terminology codes*, Table A1) and had a pathology report within 30 days of the procedure date that contained at least one mention of a “stomach” or “esophagus” term (Table 1; Table A2) anywhere in the report (N = 121,808). These pathology reports comprised the corpus for the development and validation phases and were confirmed to contain at least one biopsy specimen from the esophagus, gastroesophageal junction, or stomach location. Of these, 460 notes from 460 individuals were selected for algorithm development and validation and reviewed by a gastroenterologist with expertise in gastroesophageal precancer/cancer (S.C.S.); 34 (7.4%) were

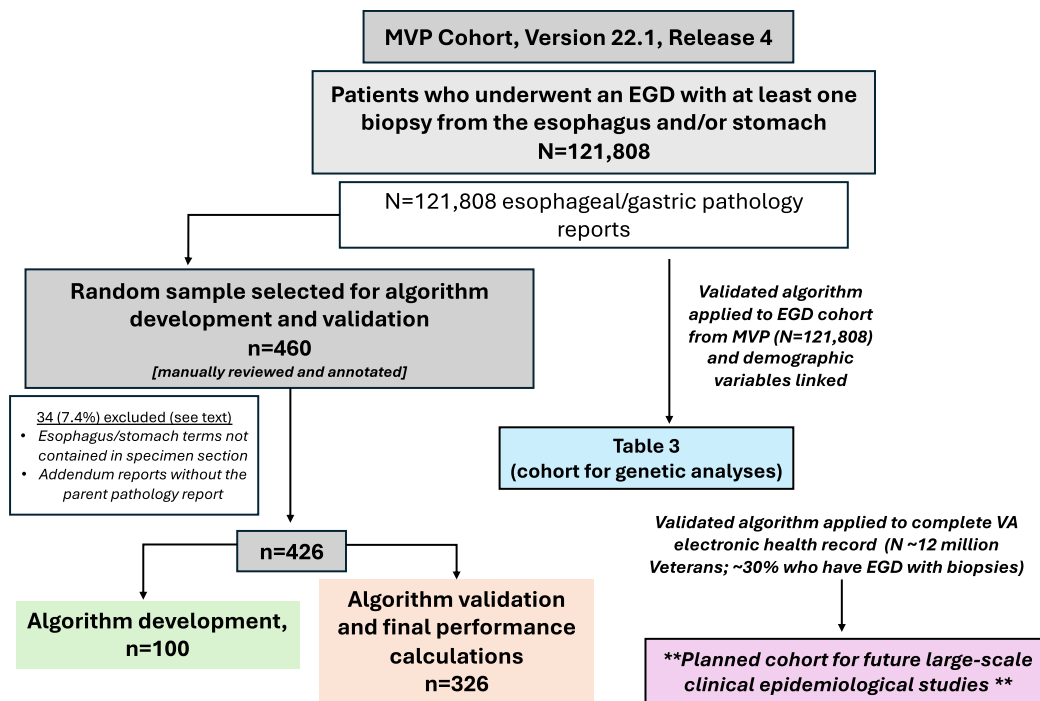


**Figure 1.** Flow diagram depicting the selection of development and validation test sets and the application of the algorithm to the genetic (MVP) and clinical databases contained within the VHA to create study cohorts.

subsequently removed due to VA Informatics and Computing Infrastructure provisioning errors, as these reports either did not contain histologic information from the stomach or esophagus (eg, a stomach or esophagus term occurred in the clinical history section, but not in the histologic section of interest) or they were addendum reports without the parent pathology report. From the remaining 426 independent notes, 100 were used in the algorithm development phase, and 326 were used in the validation phase to formally test the algorithm's performance (see below, and Figure 1).

### Annotation and Manual Classification

An expert gastroenterologist (S.C.S.) created a dictionary of gastric and esophageal precancerous and cancerous concepts to be extracted from the pathology reports by the NLP pipeline, along with an associated annotation guide to standardize concept extraction from the pathology notes (Table 1, Supplementary Material). This dictionary included the condition (eg, intestinal metaplasia) along with relevant anatomic locations, as well as “qualifiers”. Qualifiers were features considered to be relevant for clinical and risk stratification purposes. For example, GIM histology characterization as incomplete (colonic-type intestinal epithelium) or complete (small intestinal-type epithelium) subtype has implications for gastric cancer risk.<sup>9,10</sup> A complete list of conditions, anatomic location, and qualifiers is provided in Table 1, and further detailed below and in the Supplementary Material.

An expert gastroenterologist (S.C.S.) manually annotated each of the 426 reports that were appropriate for inclusion. An independent nonclinician reviewer (A.G.) used the annotation guide to annotate a random sample of 50 notes from this set to determine interobserver agreement (Cohen's kappa coefficient

98%–100% for all phenotypes, with 100% agreement after discussion of 1 discrepant report). Pathology reports where no gastric biopsies were obtained were considered “unknown” status for GIM, as GIM can only be reliably identified or ruled out based on gastric biopsies in routine US practice. A positive designation for any of the concepts necessitated histopathologic confirmation (eg, BE was only marked as positive if the presence of intestinal metaplasia in the esophagus was diagnosed).

The manually annotated set of 426 notes was randomly split into 100 for iterative development and training of the NLP pipeline, while the remaining 326 were reserved for the testing set (validation phase). The 100 training set notes were randomly split further into 2 subsets of 34 and 66 notes to allow for 2 iterative rounds of development. In the development phase, discordant findings were manually reviewed, and iterative modifications were made to the NLP pipeline as appropriate. Once saturation was reached and no further changes in the pipeline were deemed necessary, the NLP algorithm was applied to the 326 annotated notes in the test set to calculate the final performance metrics for the validation phase (Figure 1).

### NLP Algorithm Development Approach

We used the MedSpaCy framework to develop the NLP pipeline. This was the preferred framework because it is customizable without compromising speed or accuracy (see Supplementary Material). The vast majority, if not all, VA pathology notes associated with GI procedures are structured in the same general format (see Supplementary Material). From the outset, we preprocessed the corpus to pare pathology notes down to the most relevant sections to improve the

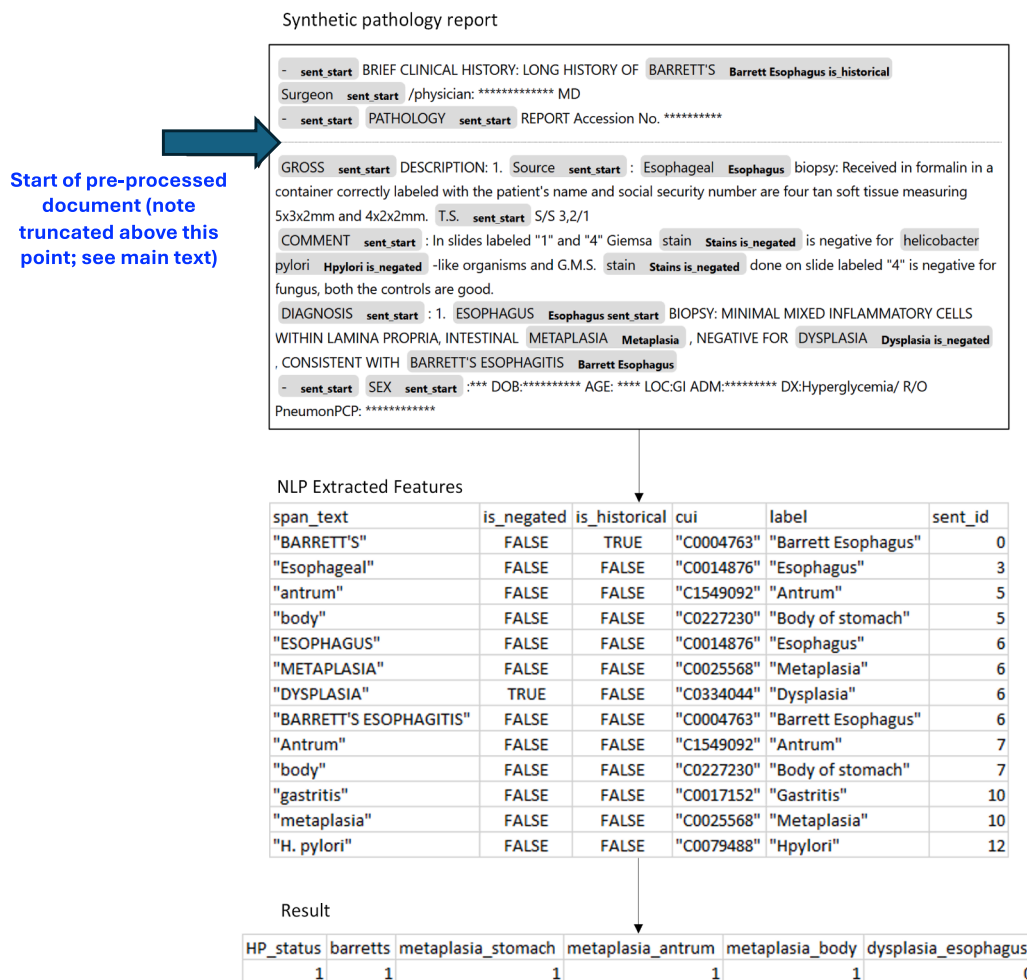


**Figure 2.** Example of a pathology note preprocessed and postprocessed through the NLP algorithm pipeline, along with NLP-extracted features.

algorithm performance and capture active diagnoses as opposed to historical diagnoses or suspected/hypothetical conditions. Figure 2 illustrates an example pathology note processed through the NLP pipeline.

A positive mention of a condition in the report is considered a positive occurrence even if there was another biopsy in the same report with a negative occurrence. For example, if the gastric antrum biopsy showed “intestinal metaplasia” but the gastric body biopsy showed “normal oxyntic mucosa”, the algorithm recorded this as “gastric metaplasia POSITIVE” (ie, any stomach location demonstrated GIM), “antrum” location for metaplasia POSITIVE, and “body” location for metaplasia NEGATIVE.

The main anatomic sites of interest were the esophagus and stomach (gastroesophageal was classified as esophagus location in postanalysis); stomach anatomic subsites were cardia, antrum, incisura, and body. In addition to the pipelines for the primary conditions of interest (intestinal metaplasia, dysplasia, tumor/cancer) along with their corresponding anatomic locations, we also captured ‘qualifiers’ of the conditions. Of note, the “tumor/cancer” condition captured both noninvasive and invasive (malignant) conditions, as some of the Unified Medical Language System (UMLS) terms for “cancer” include nondiagnostic descriptive terms such as “tumor”

and “mass”. This was mostly relevant for the stomach where mass-like lesions are not always invasive adenocarcinoma; accordingly, we used qualifiers (see below) to characterize the histology for any masses/tumors described. Qualifiers for dysplasia were low-grade (LGD), high-grade, and/or indefinite for dysplasia (IND). Qualifiers for intestinal metaplasia were focal, extensive, incomplete-type, and complete-type. Qualifiers for “cancer” included the specific histology (eg, adenocarcinoma, neuroendocrine/carcinoid, lymphoma, gastrointestinal stromal, etc.). Qualifiers were not mutually exclusive. Therefore, if a pathology report had esophageal biopsies showing both IND and LGD, both qualifiers were retained. These concept names were expanded to include ‘synonyms’ to improve name entity recognition (eg, “mild dysplasia” was considered under “LGD”, “indeterminate dysplasia” was considered under “indefinite” or “IND”) (see Table 1).

### Evaluation of NLP Pipeline Performance

We evaluated the final performance of the MedSpaCy NLP framework for correctly identifying the conditions of interest and their respective gastroesophageal anatomic location. Each of the condition\_location pairs (eg, metaplasia\_stomach) was categorized as algorithm “positive”, “negative”, or “unknown”

and then compared for concordance with the annotated notes, which were considered the reference standard. The standard NLP evaluation metrics are precision (positive predictive value, PPV), recall (sensitivity), F1 score (harmonic mean of precision and recall), and accuracy (overall correctness of predictions, mathematically calculated as  $[\text{true positive} + \text{true negative}] / [\text{true positive} + \text{true negative} + \text{false positive} + \text{false negative}]$ ).<sup>11</sup> The most “appropriate” evaluation metric depends on the goal of the NLP algorithm in the context of its anticipated application. We decided *a priori* that accuracy is the primary metric of interest given that correctly classifying individuals with vs without these conditions is core to the overarching research goal of this work. Performance metrics that achieve  $\geq 90\%$  are considered “excellent”, 80%–90% “very good”, and 70%–80% “good”. The *a priori* goal was to achieve a 1-sided 95% confidence lower bound (LB) for accuracy  $> 90\%$ ; we applied the Bonferroni correction given the multiple primary outcomes tested. In addition to the common metrics above, we also evaluated NPV.

We iteratively modified the NLP pipeline in the development phase until the *a priori* accuracy threshold was achieved for each parent condition\_location pair (metaplasia\_stomach; metaplasia\_esophagus; dysplasia\_stomach; dysplasia\_esophagus; cancer\_stomach; cancer\_esophagus). For analysis purposes, “gastroesophageal” was considered together with “esophageal” location (eg, gastroesophageal intestinal metaplasia was analyzed with esophageal intestinal metaplasia) given the clinical overlap. Once the accuracy threshold was achieved for all conditions, we applied the NLP pipeline to the testing set ( $n = 326$ ) to estimate final NLP performance metrics.

As an exploratory secondary outcome, we evaluated performance metrics for the gastric anatomic subsite of each condition and the qualifier terms described above.

### Cohort Assembly

We then applied the final validated MedSpaCy pipeline to the cohort of individuals within MVP who had undergone EGD with biopsies ( $N = 121,808$ ) to report on the prevalence of each of the precancerous conditions and basic demographic details of individuals identified as ‘cases’. For individuals classified as GIM cases and/or BE cases, we queried the linked longitudinal EHR to capture demographic details, date of index EGD, and number of subsequent EGDs following the index EGD diagnosing the precancerous condition of interest.

## Results

Six rounds of iterative modifications occurred during the development phase to achieve the *a priori* thresholds. Modifications included delineation of sentence structure, addition or modifications of UMLS terms (Table A3), the addition of an “uncertain” category for terms such as “pending” and “to be reported” which had no definitive outcome listed in the report, and preprocessing actions. The final performance of the NLP algorithm for the main condition\_location pairs is detailed in Table 2. The algorithm identified the predetermined precancerous conditions (metaplasia, dysplasia) and “cancer” of the esophagus

and stomach with excellent accuracy, ranging from 97.5% to 100%, with the Bonferroni corrected 95% LB of 95.2%–98.9% thus exceeding the *a priori*  $> 90\%$  threshold for the 95% LB for accuracy (Table 2). For GIM, the algorithm achieved 91.7% precision (PPV) and 86.8% recall (F1 score 89.2%), with 99.0% specificity and 98.3% NPV, reflecting 3 false positives and 5 false negatives among the 326 reports. Manual review of incorrectly classified pathology reports demonstrated that these were related to spelling errors (eg, “stoamch”); use of terms such as “suspicious for” yielding an uncertain diagnosis; truncation errors due to “gross description” appearing after the “microscopic description” (only occurred in 2 reports, thus no code modifications were made); and one false positive due to “pyloric metaplasia” in the duodenum being classified as gastric metaplasia. For BE, the algorithm achieved 98.9% for both precision and recall (F1 score 98.8%), with 99.6% for both specificity and NPV. The algorithm achieved 100% precision, recall, F1 score, specificity and NPV for gastric dysplasia and “cancer”. While the algorithm achieved excellent sensitivity (100%), specificity (98.3%–99.4%), and NPV (100%) for esophageal dysplasia and cancer, precision was suboptimal (71.4%–83.3%) driven by a few false positive reports, which consequently lowered the overall F1 score (83.3%–90.9%).

The results of the secondary analysis characterizing the anatomic subsite locations of gastric metaplasia are provided in Table A4. The accuracy for each anatomic subsite (antrum, body, incisura, and cardia) was 99.7%–100% with Bonferroni-corrected 95% LB 98.3%–98.9%. Precision, recall, F1 score, specificity and NPV were all high, although it should be noted that there were limited pathology reports for some locations (eg, GIM in the incisura).

Table A5 details the performance of specific “qualifier” terms, with a focus on precancerous conditions (secondary analysis). Overall, of the 38 reports with GIM, 19 commented on focal vs extensive GIM in the biopsy specimen, with most ( $n = 15$ ) reporting focal GIM; by contrast, few reported on incomplete vs complete histologic subtype ( $n = 4$ ). Regarding dysplasia diagnoses, all except one case were qualified by severity (IND, LGD, high-grade); most (80%) were IND or LGD.

### Prevalence and Core Demographics of Gastroesophageal Precancer and Cancer in MVP

We applied the final algorithm to the complete corpus of 197,589 EGD pathology reports from the 121,808 distinct individuals who underwent at least one EGD with at least one gastric and/or esophageal biopsy. The total processing time was 34 minutes. The frequency of the conditions and associated core demographic variables of this select EGD cohort are detailed in Table 3. In this cohort, the overall prevalence of GIM was 13.2% ( $n = 16,038$ ; mean age [standard deviation]: 65.1 (10.3) years old), the majority of which was of noncardia location (antrum, body, incisura). Of the individuals with GIM, 45%

**Table 2.** NLP Algorithm Performance Metrics for Stomach and Esophageal Precancer and Cancer, Based on the Validation Test Set of 326 Pathology Notes

| Phenotype                  | Frequency<br>(development + test set)    | Accuracy (%)<br>[95% LB <sup>a</sup> ] | Recall (sensitivity)<br>(%) [95% LB <sup>a</sup> ] | Specificity (%)<br>[95% LB <sup>a</sup> ] | Precision (PPV)<br>(%) [95% LB <sup>a</sup> ] | NPV (%)<br>[95% LB <sup>a</sup> ] | F1<br>score (%) |
|----------------------------|------------------------------------------|----------------------------------------|----------------------------------------------------|-------------------------------------------|-----------------------------------------------|-----------------------------------|-----------------|
| Metaplasia, esophagus (BE) | 26.7%                                    | 99.4 [96.6]                            | 98.9 [97.7]                                        | 99.6 [99.1]                               | 98.9 [97.7]                                   | 99.6 [99.2]                       | 98.9            |
| Metaplasia, stomach (GIM)  | 11.7%                                    | 97.5 [95.2]                            | 86.8 [81.3]                                        | 99.0 [98.4]                               | 91.7 [87.1]                                   | 98.3 [97.5]                       | 89.2            |
| Dysplasia, esophagus       | 7.7%                                     | 98.5 [96.5]                            | 100 [97.6]                                         | 98.3 [97.6]                               | 83.3 [76.5]                                   | 100 [100]                         | 90.9            |
| Dysplasia stomach          | 1.5%                                     | 100 [98.9]                             | 100 [100]                                          | 100 [100]                                 | 100 [100]                                     | 100 [100]                         | 100             |
| "Cancer", esophagus        | 1.5% <sup>b</sup> (0.9% <i>adenoCA</i> ) | 99.4 [97.8]                            | 100 [98.9]                                         | 99.4 [98.9]                               | 71.4 [54.4]                                   | 100 [100]                         | 83.3            |
| "Cancer", stomach          | 2.8% <sup>c</sup> (0.6% <i>adenoCA</i> ) | 100 [98.9]                             | 100 [100]                                          | 100 [100]                                 | 100 [100]                                     | 100 [100]                         | 100             |

<sup>a</sup>The Bonferroni correction for the 6 primary phenotypes was applied to the 1-sided 95% LB of all performance metrics. The exact method was used to calculate the LB.  
<sup>b</sup>Two adenocarcinoma, one squamous cell carcinoma, one carcinoma in situ.  
<sup>c</sup>Two adenocarcinoma, 2 gastrointestinal stromal tumor, 4 carcinoid tumor, one 'inflammatory' tumor. Note: the use of "qualifier" terms accurately discriminated adenocarcinoma vs other histology (see Table 1).

had at least 1 follow-up EGD with biopsies, with a median of 358 days between the 2 exams. Regarding BE, the overall prevalence was 14.5% (n = 17,700; mean (standard deviation): 63.9 (9.9) years old), with 55.5% having at least 1 follow-up EGD with biopsies, with a median of 636 days between the 2 exams.

Discussion

In this study, we successfully developed and validated a rule-based NLP algorithm that performed with excellent, >97.5%, accuracy for gastric and esophageal precancerous conditions and cancer. This high-performing algorithm also discriminated anatomic subsite locations and important risk stratification parameters of high clinical relevance, such as complete vs incomplete-type GIM, extensive vs limited GIM, and dysplasia grade, with similarly excellent accuracy. We further demonstrated that we can readily link this information to structured individual-level longitudinal data to create large nationwide cohorts of individuals with vs without gastroesophageal precancer to conduct subsequent robust genetic and clinical epidemiological analyses. Indeed, among the Veterans who had undergone EGD with esophageal and/or gastric biopsies and had genotype data available in MVP (version 22.1, approximately 650,000 of 850,000 as of this writing), 13.2% and 14.5% had histologically confirmed GIM and BE, respectively, underscoring the large number of individuals at high risk for gastroesophageal cancer; furthermore, 45%–55% of these individuals had at least one endoscopic exam with biopsies several months after their index endoscopic exam, which will allow for longitudinal studies. A prior VA study reported an overall BE prevalence of 8.6% (95% CI, 6.4%–11.3%) among 513 asymptomatic participants invited to undergo an EGD for research purposes, which is in contrast to our retrospective study reporting the prevalence of BE in individuals with a clinical indication for EGD.<sup>12</sup> The observed GIM prevalence is on par with estimates from cross-sectional US studies (13.8%, 95% CI, 10.4%–18.2%) that included a mix of asymptomatic and symptomatic individuals.<sup>13</sup> Indeed, in future studies, we will apply these algorithms to the complete VA nationwide database that currently contains over 12 million Veterans with comprehensive longitudinal health-care data to enable robust epidemiological analyses evaluating the burden and natural history of these conditions, so as to inform clinical practice and lay the groundwork for intervention trials.

GIM and BE are established precancerous conditions associated with an increased risk of intestinal-type gastric adenocarcinoma and esophageal adenocarcinoma, respectively. Appropriate risk stratification is of utmost importance<sup>14,15</sup>; for example, the risk of GIM progression to gastric cancer varies from negligible or minimal risk to as great as 20-fold higher risk depending primarily on the anatomic extent, histologic severity and GIM subtype.<sup>16,17</sup> Indeed, international guidelines are aligned in the

**Table 3.** Frequency of Gastroesophageal Precancer Among Individuals Enrolled in MVP Who Underwent at Least 1 EGD With Esophageal or Gastric Biopsies (N = 121,808)

|                            | Number of patients with histologically confirmed diagnosis (frequency) <sup>a</sup> | Mean age (standard deviation) | Percent of patients with follow-up EGD + biopsies after index EGD | Median days between index EGD and follow-up EGD with biopsies |
|----------------------------|-------------------------------------------------------------------------------------|-------------------------------|-------------------------------------------------------------------|---------------------------------------------------------------|
| Metaplasia, esophagus (BE) | 17,700 (14.5%)                                                                      | 63.9 (9.9)                    | 55.5%                                                             | 636                                                           |
| Metaplasia, stomach (GIM)  | 16,038 (13.2%)                                                                      | 65.1 (10.3)                   | 44.9%                                                             | 358                                                           |
| Dysplasia, esophagus       | 3329 (2.7%)                                                                         | 65.9 (9.1)                    | 73.4%                                                             | 191                                                           |
| Dysplasia stomach          | 1344 (1.1%)                                                                         | 67.4 (9.9)                    | 69.1%                                                             | 98                                                            |

<sup>a</sup>Denominator = 121,808 individual patients in MVP, who underwent EGD with gastric and/or esophageal biopsies (see text).

recommendation against *routine* endoscopic surveillance for low-risk GIM and, similarly, routine endoscopic surveillance of BE has also come into question given the high burden of endoscopy generated, yet lackluster impact on cancer mortality.<sup>14,15</sup> To date, studies evaluating machine learning algorithms for GIM have been limited to single-center studies that, while able to achieve high accuracy for GIM diagnosis, have not extracted anatomic and histologic features critical to determining risk of neoplastic progression.<sup>18</sup> The reliable determination and granular characterization of risk factors for gastroesophageal precancerous conditions and their progression has been limited by (1) the lack of large nationwide endoscopic surveillance databases, in part because histologically confirmed gastroesophageal precancerous conditions cannot be reliably determined using *structured* data alone (for example, International Classification of Diseases codes) and (2) the need for linkage to comprehensive, accurate individual-level endoscopic, histologic, and clinical data for robust epidemiological analyses. The approach and data sources leveraged in this study address both factors and open the door for important clinical and genetic epidemiological analyses, especially regarding GIM.

Prior studies have relied on manual chart review for individual-level data extraction. Manual chart review is labor- and resource-intensive, subject to human errors (eg, inadvertent errors of omission, mistranscription), and simply not feasible for very large cohorts. The nationwide VHA database contains longitudinal health-care records for over 12 million Veterans encompassing all medical encounters, laboratory data, procedure data, prescription data, deployment, and active-duty service data, among other relevant elements. VHA data can also be linked to cancer registry data, National Death Index data, Medicare, and other data-rich sources. We and others have demonstrated that phenotype definitions can readily be developed and validated using the VHA infrastructure and applied to the larger database thereby overcoming the need for manual data extraction to define outcomes, exposures, and covariates for analyses.<sup>16</sup> Here, we successfully applied the algorithm to 197,589 unstructured pathology reports from 121,808 distinct individuals and generated summary demographics for the cohort in about 30 minutes. Accordingly, the algorithm presented herein will

enable a pipeline for evaluating epidemiological risk factors for gastroesophageal precancer, cancer, and progression (or lack thereof), including the impact of military exposures, as well as evaluation of genetic and gene x environment determinants through leveraging MVP, the VHA's EHR linked genomic biobank.

This study has several strengths. The MedSpaCy framework provided a highly customizable pipeline while not sacrificing performance or accuracy, and also provided the highest likelihood of generalizability outside the VHA data infrastructure; to this end, this framework is built on top of the spaCy framework, the most widely used and supported Python NLP framework. We considered deep learning methods from the outset, but such approaches require massive training sets and would not be feasible for this task. Large language models (LLMs) with prompt engineering might perform well and are worth considering for future applications and in settings with sufficient graphics processing units. That said, rule-based NLP still has advantages over LLMs, particularly related to the task presented herein. NLP easily incorporates domain-specific rules that can be modified based on the topic and data source of interest. Unlike NLP, LLMs may not perform consistently well for repeated tasks that necessitate effectively analyzing, interpreting, and extracting specific information from unstructured text. This study also has limitations and special considerations. We provided the full spectrum of common metrics for evaluating the performance of NLP algorithms and nearly all performed at the "excellent" or "very good" threshold. Nevertheless, these metrics have intrinsic limitations. For example, the F1 score treats precision and recall equally, which has implications depending on the specific research question, and also does not consider true negatives, which can be misleading for imbalanced (rare) conditions, as demonstrated in our analysis for esophageal cancer. For this reason, we *a priori* favored accuracy (overall correctness) as the primary performance metric of interest, although accuracy may overestimate the performance for rare conditions since specificity will naturally be high. This study is also limited by intrinsic clinical factors including the possibility of sampling error when obtaining biopsies.

## Conclusion

We successfully developed and validated a high-performing NLP algorithm that identifies gastroesophageal precancer and cancer with excellent accuracy. We applied the algorithm to individuals in the MVP genomic biobank (demographically representative of the full VHA database) and identified a high prevalence of gastroesophageal precancerous conditions. Subsequent near-term plans include applying this algorithm to the complete VHA database (N~12 million Veterans with longitudinal care), which is anticipated to identify well over 1.5 million individuals with gastric and/or esophageal precancer. Accordingly, the results of this study open the door for a powerful pipeline fostering the creation and analysis of robust cohorts evaluating clinical, genetic, and gene x environment factors underlying gastroesophageal precancer and cancer development, ultimately informing primary and secondary prevention initiatives and interventions.

## Supplementary Materials

Material associated with this article can be found in the online version at <https://doi.org/10.1016/j.gastha.2025.100768>.

## References

- Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2024;74:229–263.
- National Cancer Institute (SEER). Cancer stat facts. Available at: <https://seer.cancer.gov/statfacts/>. Accessed November 13, 2024.
- Huang RJ, Epplin M, Hamashima C, et al. An approach to the primary and secondary prevention of gastric cancer in the United States. *Clin Gastroenterol Hepatol* 2022;20:2218–2228.
- American Gastroenterological Association (AGA). New gastric intestinal metaplasia diagnosis codes for ICD-10-CM, 2021. Available at: <https://gastro.org/news/new-gastric-intestinal-metaplasia-diagnosis-codes-for-icd-10-cm/#:~:text=K31A13,leading%20cause%20of%20cancer%20deaths>. Accessed November 13, 2024.
- El-Serag HB, Naik AD, Duan Z, et al. Surveillance endoscopy is associated with improved outcomes of esophageal adenocarcinoma detected in patients with Barrett's esophagus. *Gut* 2016;65:1252–1260.
- Benson R, Winterton C, Winn M, et al. Leveraging natural language processing to extract features of colorectal polyps from pathology reports for epidemiologic study. *JCO Clin Cancer Inform* 2023;7:e2200131.
- Gupta S, Earles A, Bustamante R, et al. Adenoma detection rate and clinical characteristics influence advanced neoplasia risk after colorectal polypectomy. *Clin Gastroenterol Hepatol* 2023;21:1924–1936.
- 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;58:936–941.
- Shah SC, Gawron AJ, Mustafa RA, et al. Histologic subtyping of gastric intestinal metaplasia: overview and considerations for clinical practice. *Gastroenterology* 2020;158:745–750.
- Wei N, Zhou M, Lei S, et al. A meta-analysis and systematic review on subtypes of gastric intestinal metaplasia and neoplasia risk. *Cancer Cell Int* 2021;21:173.
- Makhoul J, Kubala F, Schwartz R, et al. Performance measures for information extraction. In: *Proceedings of DARPA broadcast news workshop*. Herndon, Virginia: Morgan Kaufmann Publishers; 1999:249–252.
- Nguyen TH, Thrift AP, Rugge M, et al. Prevalence of Barrett's esophagus and performance of societal screening guidelines in an unreferred primary care population of U.S. veterans. *Gastrointest Endosc* 2021;93:409–419.
- Mülder DT, Hahn AI, Huang RJ, et al. Prevalence of gastric precursor lesions in countries with differential gastric cancer burden: a systematic review and meta-analysis. *Clin Gastroenterol Hepatol* 2024;22:1605–1617.
- Shaheen NJ, Falk GW, Iyer PG, et al. Diagnosis and management of Barrett's esophagus: an updated ACG guideline. *Am J Gastroenterol* 2022;117:559–587.
- Dinis-Ribeiro M, Shah S, El-Serag H, et al. The road to a world-unified approach to the management of patients with gastric intestinal metaplasia: a review of current guidelines. *Gut* 2024;73:1607–1617.
- Liu L, Bustamante R, Earles A, et al. A strategy for validation of variables derived from large-scale electronic health record data. *J Biomed Inform* 2021;121:103879.
- Lee JWW, Zhu F, Srivastava S, et al. Severity of gastric intestinal metaplasia predicts the risk of gastric cancer: a prospective multicentre cohort study (GCEP). *Gut* 2022;71:854–863.
- Parbhu SK, Shah SC, Sossenheimer MJ, et al. Index diagnoses of gastric intestinal metaplasia in the United States: patient characteristics, endoscopic findings, and clinical practice patterns at a large tertiary care center. *Therap Adv Gastroenterol* 2022;15:17562848221117640.

Received March 12, 2025. Accepted August 20, 2025.

### Correspondence:

Address correspondence to: Shailja C. Shah, MD, MPH, Gastroenterology Section, VA San Diego Healthcare System, Division of Gastroenterology, University of California, San Diego, 3350 La Jolla Village Drive, 3<sup>rd</sup> Floor South, GI Section (Mail Code 111D), San Diego, California 92161. e-mail: [s6shah@health.ucsd.edu](mailto:s6shah@health.ucsd.edu).

### Authors' Contributions:

Aaron Giera: Acquisition of data, analysis and interpretation of data, critical revision of the manuscript for important intellectual content, statistical analysis. Kyriacos Markianos: Analysis and interpretation of data, critical revision of the manuscript for important intellectual content. Andrew J. Gawron: Critical revision of the manuscript for important intellectual content. Samir Gupta: Analysis and interpretation of data, critical revision of the manuscript for important intellectual content. Saiju Pyarajan: Administrative, technical, or material support, obtained funding, study supervision. Purushotham Kamam: Administrative, technical, or material support. Lin Liu: Analysis and interpretation of data, critical revision of the manuscript for important intellectual content, statistical analysis. Shailja C. Shah: First draft of manuscript, study

concept, acquisition of data, analysis and interpretation of data, critical revision of the manuscript for important intellectual content, obtained funding, study supervision.

**Conflicts of Interest:**

Dr Shailja Shah is a consultant for RedHill Biopharma, Phathom Pharmaceuticals, and third-party consulting firms GLG and Guidepoint. Dr Shah is an unpaid advisory board member for Ilico Genetics, Inc. The other authors disclose no conflicts.

**Funding:**

Dr Shah receives funding from a Veterans Affairs Career Development Award (ICX002027A). This work was also supported by Veterans Affairs MVP000.

**Ethical Statement:**

This study was approved by the committee on research ethics at VA San Diego and through the centralized Institutional Review Board for studies involving the Million Veteran Program.

**Data Transparency Statement:**

Data access is available to approved VA investigators with approved data access privileges. Data will not be made publicly available.

**Reporting Guidelines:**

For this manuscript, we followed the “Standards for Reporting Diagnostic accuracy studies (STARD)” guidelines. The checklist is provided in the supplementary online materials. The guidelines are available at <https://www.equator-network.org>.