SUPPLEMENTAL METHODS

Natural language processing (NLP) was used in three instances of our study: cohort identification, outcome ascertainment, and eradication ascertainment. NLP is a tool to leverage unstructured, free text data, such as pathology reports or clinic notes, in an efficient manner, increasingly used for research purposes.[1,2]

For cohort identification, we used NLP to identify pathology reports positive for *H. pylori* (HP). This included identifying all pathology reports that indicated HP testing was performed and sentiment analysis to ensure that testing was positive (by identification of organisms or specific stains).[3] Steps included spell correction, data reduction, data transformation including disambiguation of abbreviations, regular expression matching, negative expression matching, and finally, validation. Validation was performed by taking a sample of random pathology reports, generated randomly across institutions. Over 300 pathology reports were manually reviewed to identify presence or absence of HP. This was compared to results of natural language processing. We found 100% positive predictive value for HP if classified as positive by our natural language processing algorithm. To note, we made a conscious decision to focus not on sensitivity or negative predictive value for this, as it was not our intent to determine prevalence of HP diagnosis in the Veterans Health Administration, but to focus on truly capturing those with HP infection.

For outcome ascertainment, we relied both on the Veterans Affairs Central Cancer Registry and/or ICD 9/10 codes for non-proximal gastric adenocarcinoma (ICD-9: 151.1-151.9; ICD-10: C16.1-C16.9). We filtered our searches in the Veterans Health Administration (VHA) Corporate Data Warehouse (CDW), which includes data from the unified electronic medical record of all VHA facilities (i.e., hospitals and outpatient) since 10/01/1999. The searches were filtered to include intestinal type non-cardia cancers, to avoid capturing non-adenocarcinomas and

cardiac/gastroesophageal junction tumors, which are less clearly associated with HP.[4-6] We initially identified 4,709 patients by ICD code, and had pathology reports for 2,071 (44%). Sample validation of over 100 charts (as described above) was performed after additional natural language processing, and showed the following test characteristics:
>90% positive predictive value of distal adenocarcinomas, 95% negative predictive value, 95% sensitivity, and >90% specificity.

We sought to evaluate eradication status after treatment. Eradication was based on having either a negative stool antigen, urea breath test, and/or pathology (gastric biopsy on endoscopy) upon repeat testing. Failed eradication was defined as a positive stool antigen, urea breath test, and/or pathology, or a positive HP test after a prior negative test given that true re-infection is exceedingly rare. Patients without any eradication testing were considered as 'unknown' eradication status. HP status on pathology was determined by repeat natural language processing, with resultant sample validation of 100 chart showing the following test characteristics >90% sensitivity and specificity and >90% negative and positive predictive values.

Supplemental Methods Table 1: Medication regimens used to identify prescription of *H. pylori* eradication regimens

Medication Regimen (concomitant use of the medications in the regimen)	a) amoxicillin (1000 mg twice a day, or metronidazole 500 mg four times a day for penicillin allergic patients), clarithromycin (500 mg twice a day) and a proton pump inhibitor (PPI) b) *or* quadruple therapy with bismuth subsalicylate (525 mg four times a day), metronidazole or clarithromycin, tetracycline (500 mg four times a day) and a PPI	Previously validated in: Thirumurthi, S., et al., *Identification of Helicobacter pylori infected patients, using administrative data.* Aliment Pharmacol Ther, 2008. **28**(11-12): p. 1309-16.
Alternative Medication Regimens (concomitant use of the medications in the regimen)	a) Clarithromycin, amoxicillin or nitroimidazole / metronidazole, with PPI b) PPI, bismuth, tetracycline, and a nitroimidazole / metronidazole c) PPI, clarithromycin, amoxicillin and a nitroimidazole / metronidazole d) PPI and amoxicillin for 5–7 days followed by a PPI, clarithromycin, and a nitroimidazole / metronidazole for 5–7 days e) PPI and amoxicillin for 7 days followed by a PPI, amoxicillin, clarithromycin and a nitroimidazole / metronidazole for 7 days f) PPI, levofloxacin, and amoxicillin g) PPI and amoxicillin for	Appropriate eradication therapy, as defined by American College of Gastroenterology

	5–7 days followed by a PPI, fluoroquinolone, and nitroimidazole / metronidazole for 5–7 days h) PPI, amoxicillin, and rifabutin i) PPI and amoxicillin for 14 days	

REFERENCES

1. Kreimeyer K, Foster M, Pandey A, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *J Biomed Inform.* 2017;73:14-29.
2. Yim WW, Yetisgen M, Harris WP, Kwan SW. Natural Language Processing in Oncology: A Review. *JAMA Oncol.* 2016;2(6):797-804.
3. Hirschberg J, Manning CD. Advances in natural language processing. *Science.* 2015;349(6245):261-266.
4. Helicobacter, Cancer Collaborative G. Gastric cancer and Helicobacter pylori: a combined analysis of 12 case control studies nested within prospective cohorts. *Gut.* 2001;49(3):347-353.
5. Kim JH, Cheung DY. Must-Have Knowledge about the Helicobacter pylori-Negative Gastric Cancer. *Gut Liver.* 2016;10(2):157-159.
6. Kumar S, Long JM, Ginsberg GG, Katona BW. The role of endoscopy in the management of hereditary diffuse gastric cancer syndrome. *World J Gastroenterol.* 2019;25(23):2878-2886.

Supplemental Table 1. Risk factors for development of gastric cancer after positive diagnostic test, considering treatment status, using multivariable competing risk time to event model

	SHR	*P*-value
Age[a]	1.20 (1.14-1.26)	*P*<0.001
Race		
White	REFERENCE	
Black or African American	1.61 (1.19-2.20)	
American Indian or Alaskan Native	0.75 (0.10 – 5.45)	
Asian	1.37 (0.19 – 9.78)	0.02
Native Hawaiian or other Pacific Islander	0.74 (0.10 – 5.34)	
Unknown	0.78 (0.47 – 1.29)	
Smoking history	1.42 (1.05-1.92)	0.02
Treated after diagnosis	1.16 (0.74-1.83)	0.51

a. Per 5-year increase in age at *H pylori* diagnosis

Other covariates tested but not included in the final multivariable models as they were not significant ($p \geq 0.1$) were: gender, ethnicity, method of *H pylori* diagnosis, zip code poverty level where patient resided at *H pylori* diagnosis

Supplemental Table 2. Risk factors for development of gastric cancer after positive diagnostic test, considering eradication status, using multivariable competing risk time to event model

	SHR	*P*-value
Age[a]	1.21 (1.15-1.28)	*P*<0.001
Race		
White	REFERENCE	0.009
Black or African American	1.62 (1.19-2.21)	
American Indian or Alaskan Native	0.75 (0.10-5.39)	
Asian	1.37 (0.19-9.90)	
Native Hawaiian or other Pacific Islander	0.70 (0.19-5.05)	
Unknown	0.78 (0.52-1.17)	
Smoking history	1.39 (1.03-1.88)	0.03
Eradication status		
Confirmed eradication	0.24 (0.15-0.41)	*P*<0.001
Unknown eradication status	0.16 (0.10 – 0.25)	

a. Per 5-year increase in age at *H pylori* diagnosis

Other covariates tested but not included in the final multivariable models as they were not significant ($p \geq 0.1$) were: gender, ethnicity, method of *H pylori* diagnosis, zip code poverty level where patient resided at *H pylori* diagnosis, and whether the patient received prescription of eradication regimen

Supplemental Table 3. Risk factors for development of gastric cancer using multivariable competing risk time to event model, excluding those incident cancers within 6 months of *H pylori* diagnosis

	SHR (95% CI)	*P*-value
Age [a]	1.12 (1.10-1.14)	*P*<0.001
Method of *H pylori* diagnosis		
Positive diagnostic test	REFERENCE	0.02
RX, no serum Ab	0.95 (0.79-1.14)	
ICD, no serum Ab	1.04 (0.73-1.50)	
RX, with serum Ab	0.80 (0.53-1.21)	
ICD, with serum Ab	0.74 (0.59-0.93)	
Ethnicity		
Not Hispanic or Latino	REFERENCE	*P*<0.001
Hispanic or Latino	1.61 (1.33-1.94)	
Unknown	1.03 (0.82-1.30)	
Race		
White	REFERENCE	*P*<0.001
Black or African American	2.02 (1.78-2.29)	
American Indian or Alaskan Native	1.52 (0.90-2.59)	
Asian	3.21 (2.04-5.08)	
Native Hawaiian or other Pacific Islander	0.78 (0.41-1.51)	
Unknown	1.10 (0.89-1.36)	
History of smoking	1.34 (1.20-1.51)	*P*<0.001
Female gender	0.59 (0.44-0.78)	*P*<0.001
Poverty level of zip code where patient resided at *H pylori* diagnosis		
< 10% residing below poverty level	REFERENCE	0.40
10 – 24.9% residing below poverty level	0.91 (0.79-1.05)	
25 – 49.9% residing below poverty level	0.97 (0.83-1.14)	
≥50% residing below poverty level	1.06 (0.77-1.45)	
Unknown	0.81 (0.60-1.09)	

a. Age is per 5-year incremental increase in year

RX = prescription therapy; ICD = International Classification of Diseases (administrative codes); Ab = antibody

Supplemental Table 4. Risk factors for development of gastric cancer after positive diagnostic test, considering treatment status, using multivariable competing risk time to event model, excluding those incident cancers within 6 months of *H pylori* diagnosis

	SHR	*P*-value
Age[a]	1.17 (1.09-1.24)	*P*<0.001
Race		
White	REFERENCE	*P*<0.001
Black or African American	1.47 (1.00-2.14)	
American Indian or Alaskan Native	8.79e-08 (6.52e-08-1.19e-09)	
Asian	2.02 (0.28-14.5)	
Native Hawaiian or other Pacific Islander	1.06 (0.15-7.70)	
Unknown	0.76 (0.41-1.41)	
Treated after diagnosis	2.00 (1.09-3.69)	0.02

a. Per 5-year increase in age at *H pylori* diagnosis

Other covariates tested but not included in the final multivariable models as they were not significant ($p \geq 0.1$) were: gender, ethnicity, method of *H pylori* diagnosis, smoking status, zip code poverty level where patient resided at *H pylori* diagnosis

Supplemental Table 5. Risk factors for development of gastric cancer after positive diagnostic test, considering eradication status, using multivariable competing risk time to event model, excluding those incident cancers within 6 months of *H pylori* diagnosis

	SHR	*P*-value
Age[a]	1.19 (1.10-1.27)	*P*<0.001
Race		
White	REFERENCE	*P*<0.001
Black or African American	1.47 (1.01-2.15)	
American Indian or Alaskan Native	1.54e-09 (1.12e-09-2.10e-09)	
Asian	2.25 (0.31-14.15)	
Native Hawaiian or other Pacific Islander	1.02 (0.14-7.43)	
Unknown	0.81 (0.44-1.50)	
Treated after diagnosis of *H pylori*	1.80 (0.97-3.33)	0.06
Eradication status		
Confirmed eradication	0.25 (0.14-0.47)	*P*<0.001
Unknown eradication status	0.14 (0.08-0.25)	

a. Per 5-year increase in age at *H pylori* diagnosis

Other covariates tested but not included in the final multivariable models as they were not significant ($p \geq 0.1$) were: gender, ethnicity, method of *H pylori* diagnosis, smoking status, zip code poverty level where patient resided at *H pylori* diagnosis, and whether the patient received prescription of eradication regimen

Supplemental Table 6. Risk factors for development of gastric cancer using multivariable competing risk time to event model, excluding those incident cancers within 12 months of *H pylori* diagnosis

	SHR (95% CI)	*P*-value
Age [a]	1.12 (1.10-1.15)	*P*<0.001
Method of *H pylori* diagnosis		
Positive diagnostic test	REFERENCE	0.06
RX, no serum Ab	0.91 (0.75-1.12)	
ICD, no serum Ab	0.87 (0.57-1.33)	
RX, with serum Ab	0.77 (0.49-1.21)	
ICD, with serum Ab	0.72 (0.57-0.93)	
Ethnicity		
Not Hispanic or Latino	REFERENCE	
Hispanic or Latino	1.64 (1.33-2.00)	*P*<0.001
Unknown	0.91 (0.70-1.19)	
Race		
White	REFERENCE	
Black or African American	1.97 (1.71-2.25)	*P*<0.001
American Indian or Alaskan Native	1.62 (0.93-2.80)	
Asian	3.35 (2.07-5.43)	
Native Hawaiian or other Pacific Islander	0.80 (0.40-1.61)	
Unknown	1.08 (0.85-1.37)	
History of smoking	1.37 (1.21-1.56)	*P*<0.001
Female gender	0.60 (0.44-0.82)	0.001
Poverty level of zip code where patient resided at *H pylori* diagnosis		
< 10% residing below poverty level	REFERENCE	0.71
10 – 24.9% residing below poverty level	0.96 (0.82-1.12)	
25 – 49.9% residing below poverty level	1.01 (0.85-1.20)	
≥50% residing below poverty level	1.09 (0.77-1.53)	
Unknown	0.83 (0.60-1.15)	

a. Age is per 5-year incremental increase in year

RX = prescription therapy; ICD = International Classification of Diseases (administrative codes); Ab = antibody

Supplemental Table 7. Risk factors for development of gastric cancer after positive diagnostic test, considering treatment status, using multivariable competing risk time to event model, excluding those incident cancers within 12 months of *H pylori* diagnosis

	SHR	*P*-value
Age[a]	1.16 (1.08-1.25)	*P*<0.001
Race		
White	REFERENCE	*P*<0.001
Black or African American	1.55 (1.05-2.35)	
American Indian or Alaskan Native	5.05 e-08 (3.65e-08-6.99 e-09)	
Asian	2.75 (0.38-19.8)	
Native Hawaiian or other Pacific Islander	1.38 (0.19-10.1)	
Unknown	0.94 (0.49-1.79)	
Female gender	0.16 (0.02-1.13)	0.01
Treated after diagnosis	2.36 (1.19-4.68)	0.02

a. Per 5-year increase in age at *H pylori* diagnosis

Other covariates tested but not included in the final multivariable models as they were not significant (p ≥ 0.1) were: gender, ethnicity, method of *H pylori* diagnosis, smoking status, zip code poverty level where patient resided at *H pylori* diagnosis

Supplemental Table 8. Risk factors for development of gastric cancer after positive diagnostic test, considering eradication status, using multivariable competing risk time to event model, excluding those incident cancers within 12 months of *H pylori* diagnosis

	SHR	*P*-value
Age[a]	1.17 (1.10-1.26)	*P*<0.001
Race		
White	REFERENCE	*P*<0.001
Black or African American	1.57 (1.05-2.37)	
American Indian or Alaskan Native	5.17e-08 (3.72e-08-7.20e-08)	
Asian	2.89 (0.40-20.72)	
Native Hawaiian or other Pacific Islander	1.36 (0.19-9.96)	
Unknown	0.98 (0.51-1.87)	
Female gender	0.15 (0.02-1.09)	0.06
Treated after diagnosis of *H pylori*	2.16 (1.09-4.23)	0.03
Eradication status		
Confirmed eradication	0.53 (0.24-1.19)	0.001
Unknown eradication status	0.26 (0.12-0.57)	

a. Per 5-year increase in age at *H pylori* diagnosis

Other covariates tested but not included in the final multivariable models as they were not significant ($p \geq 0.1$) were: gender, ethnicity, method of *H pylori* diagnosis, smoking status, zip code poverty level where patient resided at *H pylori* diagnosis, and whether the patient received prescription of eradication regimen

Supplemental Table 9. Risk factors for development of gastric cancer using multivariable competing risk time to event model

	SHR (95% CI)	*P*-value
Age [a]	1.11 (1.08-1.14)	*P*<0.001
Method of *H pylori* diagnosis		
Positive diagnostic test	REFERENCE	0.04
RX, no serum Ab	1.09 (0.83-1.41)	
ICD, no serum Ab	1.02 (0.59-1.77)	
RX, with serum Ab	1.09 (0.63-1.90)	
ICD, with serum Ab	0.74 (0.54-1.04)	
Ethnicity		
Not Hispanic or Latino	REFERENCE	
Hispanic or Latino	1.86 (1.41-2.46)	*P*<0.001
Unknown	1.59 (1.23-2.06)	
Race		
White	REFERENCE	
Black or African American	2.04 (1.70-2.47)	*P*<0.001
American Indian or Alaskan Native	0.73 (0.23-2.28)	
Asian	1.52 (0.3-3.68)	
Native Hawaiian or other Pacific Islander	0.96 (0.40-2.34)	
Unknown	1.11 (0.86-1.43)	
History of smoking	1.54 (1.30 – 1.84)	*P*<0.001
Female gender	0.49 (0.31-0.78)	0.002
Poverty level of zip code where patient resided at *H pylori* diagnosis		
< 10% residing below poverty level	REFERENCE	
10 – 24.9% residing below poverty level	0.87 (0.71-1.06)	0.41
25 – 49.9% residing below poverty level	0.91 (0.73-1.40)	
≥50% residing below poverty level	1.00 (0.63-1.59)	
Unknown	0.67 (0.43-1.06)	
BMI [b]		
Underweight	REFERENCE	
Normal	0.57 (0.41-0.79)	*P*<0.001
Overweight	0.38 (0.27-0.53)	
Obese	0.30 (0.21-0.43)	

a. Age is per 5-year incremental increase in year
b. The cohort is missing BMI data for 171,212 patients (46.0%). The missingness is not markedly different between those veterans who go onto develop cancer and those who do not, yet given the degree of missingness, it was not evaluated in the primary model.

RX = prescription therapy; ICD = International Classification of Diseases (administrative codes); Ab = antibody