Natural Language Processing Accurately Identifies Colorectal Dysplasia in a National Cohort of Veterans with Inflammatory Bowel Disease

Jason Ken Hou (jkhou@bcm.edu)
Baylor College of Medicine https://orcid.org/0000-0003-0644-6778
Christopher C. Taylor
Michael E DeBakey VA Medical Center
Ergin Soysal
The University of Texas, School of Bioinformatics
Shubhada Sansgiry
Michael E DeBakey VA Medical Center
Peter Richardson
Michael E DeBakey VA Medical Center
Hua Xu
The University of Texas, School of Bioinformatics
Nader N. Massarweh
Michael E DeBakey VA Medical Center

Research article

Keywords: Inflammatory Bowel Disease, Ulcerative Colitis, Crohn's Disease, Natural Language Processing, Clinical Language Annotation, Modeling, and Processing (CLAMP), Low Grade Dysplasia, and High Grade Dysplasia,

Posted Date: October 24th, 2019

DOI: https://doi.org/10.21203/rs.2.16432/v1

Abstract

Background: Although practice guidelines recommend colorectal cancer surveillance for inflammatory bowel disease (IBD) patients, the natural history of patient with dysplasia is poorly described. Assembling large cohorts of IBD patients with dysplasia is difficult as administrative codes are lacking. The aim of this study was to use natural language processing (NLP) in a large electronic health records (EHR) to identify IBD patients with colonic dysplasia.

Methods: We conducted a retrospective cohort study using administrative data from the national Veterans Health Administration (VHA) Corporate Data Warehouse for patients with IBD. Full-text histopathology reports from patients who underwent colonoscopy in the VHA were obtained and a validation cohort was created using a random sample of 2000 reports. An NLP algorithm to identify the presence and grade of dysplasia was developed and performance tested in a validation cohort. The final NLP algorithm was applied to the entire IBD cohort to identify all cases of colonic dysplasia.

Results: We identified a total of 44,099 Veterans with IBD, with 22,431 colonoscopy related histopathology reports. NLP had an accuracy of 97.1% for detection of low grade dysplasia, with a precision of 87%, recall of 96.6%, and F- measure of 91.5%. When applied to the entire cohort, a total of 1,762 cases of colonic dysplasia were identified.

Conclusions: NLP accurately identifies colonic low-grade dysplasia in IBD patients from a national EHR. NLP can be used to identify large cohorts of IBD patients with dysplasia to further study the natural history and outcomes of colonic dysplasia in patients with IBD.

Background

Patients with inflammatory bowel disease (IBD) are at an increased risk of developing colorectal cancer, and currently practice guidelines recommend these patients undergo colonoscopies to detect dysplasia or cancer.[1, 2] If patients are found to have high-grade dysplasia or cancer, total proctocolectomy is recommended given the increased risk of synchronous dysplasia or cancer elsewhere in the colon.[1, 2] However, there is a paucity of data characterizing outcomes of patients with dysplasia detected on either surveillance or diagnostic colonoscopy. [3, 4]

One of the main challenges of studying IBD-related dysplasia is that large cohorts of IBD patients with dysplasia are difficult to identify in commonly used administrative data sources because there are no diagnostic or billing codes for non-cancer dysplasia.[5] Alternatively, applying Natural Language Processing (NLP) to an electronic health record (EHR) could be an efficient means of identifying large cohorts of patients across numerous hospitals within a health system. NLP utilizes computer science and linguistics to develop methods to automatically add structure to otherwise unstructured text.[6, 7] NLP has been previously applied to identify quality metrics of colonoscopy and has been applied to differentiate surveillance colonoscopy for IBD patients from a non-surveillance colonoscopy.[8, 9]

While primary data collection via chart review and prospective registries provide granular information, this approach is inefficient, time consuming and expensive. NLP provides a potential approach to efficiently create large cohorts of IBD patients that are likely better powered to conduct meaningful analyses of IBD-associated colorectal dysplasia outcomes. The aims of this study were to develop an NLP algorithm to identify cases of colonic dysplasia in patients with IBD and to then apply NLP to data from a national EHR in order to identify IBD patients with colonic dysplasia.

Methods

The Institutional Review Boards of Baylor College of Medicine and the R&D Committee of the Michael E. DeBakey VA Medical Center approved this study.

Data source:

Data were obtained from the Veterans Health Administration (VHA) Corporate Data Warehouse (CDW). CDW compiles all electronic health record (EHR) and administrative data from the VHA facilities nationwide, including diagnostic and procedural codes and full-text notes from the EHR. Full text-histopathology reports from colonoscopy of patients with IBD were obtained from 2003–2009. Cases of IBD were identified using a case finding algorithm using ICD–9 codes previously validated in the VHA.[10]

Study Design

A random sample of 2000 IBD colonoscopy pathology reports were manually adjudicated by 2 independent reviewers for the presence and level of dysplasia. Pathology reports describing any grade of colonic dysplasia or colonic adenomas were considered positive for dysplasia. Indefinite for dysplasia was not classified as dysplastic. Levels of dysplasia were classified as 1) low grade dysplasia, including adenomas without further grade of dysplasia defined, 2) high grade dysplasia, or 3) adenocarcinoma. Discrepancies between the reviewers were resolved by a third reviewer. The study was performed as a split-validation study. Overall, 557 of the 2000 reports were included in the NLP training cohort and the remaining 1443 reports were included in the NLP validation cohort.

NLP Algorithm Development and Validation

An NLP algorithm to identify the presence and grade of dysplasia was developed using an NLP training cohort of 557 reports. The NLP pipeline was created using the Clinical Language Annotation, Modeling, and Processing (CLAMP) Toolkit— the software identified diagnosis and related comments sections from reports which were processed for intended concepts and phrases by applying a generic NLP pipeline utilizing components from the CLAMP Toolkit.[11] Dysplasia status was assigned using rule-based methods as the final step. The performance of the NLP algorithm was tested on the remainder of the validation cohort and reported as accuracy, precision (estimate of specificity), recall (estimate of sensitivity), and F-measure (harmonic mean of precision and recall).

Application of IBD-Dysplasia NLP Algorithm

To estimate the total number of expected dysplasia cases, the IBD-Dysplasia NLP algorithm was applied to the entire cohort of IBD colon pathology reports. The total number of reports found with dysplasia and levels of dysplasia from the NLP algorithm were reported.

Results

In total, 22,431 patients had colonoscopy-related histopathology reports available for extraction. On manual adjudication of 2000 reports, 22 were indefinite for dysplasia and 325 were classified as having any level of dysplasia, including 288 cases of LGD, 13 cases of HGD, and 24 cases of adenocarcinoma. Among the 557 reports in the training cohort, there were 82 cases of LGD, 2 cases of HGD, and 7 cases of adenocarcinoma.

NLP Development and Validation

The training cohort was used to train the NLP algorithm in CLAMP. When applied to the validation cohort, the IBD-dysplasia algorithm was found to have an accuracy of 97.1% for detection of low grade dysplasia, with a precision of 87%, recall of 96.6%, and F- measure of 91.5% (Figure 1). For high grade dysplasia, the accuracy was 97.7%, precision 36.6%, recall 96.2%, and F-measure of 53.1%.

Estimation of cases of IBD-Dysplasia

The IBD-Dysplasia NLP algorithm was applied to the national VA IBD cohort of 74,258 unique IBD patients and 116,338 available pathology reports (Table 1). The algorithm identified colonic dysplasia among 3,969 unique IBD patients (5.3%), with 4,545 pathology reports with LGD and 513 reports with HGD.

Discussion

In this study, we developed and validated an NLP-based algorithm to identify presence and level of colorectal dysplasia in patients with IBD. Our work demonstrates that this algorithm can be applied to a large, national dataset to estimate the prevalence of colonic dysplasia among patients with IBD.

Patients with IBD are at an increased risk of developing CRC compared to persons without IBD.[1, 12] While the increased risk of CRC among patients with IBD has been consistently demonstrated, the reported magnitude of the 30-year cumulative risk varies greatly, from 2.1 to 33.2%.[13–16] IBD-associated CRC has distinct genetic and molecular pathways relative to sporadic CRC, and therefore the natural history of colonic dysplasia among patients with IBD likely differs compared to sporadic CRC and deserves independent study. However, studying the natural history of IBD-associated CRC and the manner in which IBD leads to the development of CRC remains relatively unexplored. Observational studies based on administrative claims data have been the standard for efficiently studying the natural history of diseases; however, progress in the field of IBD-associated colonic dysplasia has been limited in large part due to the absence of administrative codes for the presence or degree of IBD-associated colonic

dysplasia. The findings of this study can help to address this current knowledge gap by permitting the efficient identification and study of cases of dysplasia among large cohorts of patients with IBD.

Prior studies have demonstrated the utility of NLP for dysplasia detection and characterization in sporadic CRC with high levels of accuracy (92–100%).[9, 17–19] However, patterns of pathology reports for IBD and non-IBD patients differ greatly, with a higher number of pathology samples provided for IBD patients undergoing surveillance colonoscopy. Also, the identification of IBD-associated colonic dysplasia can be complicated by the description of background inflammation in histopathologic reports in NLP derivation for IBD-associated colonic dysplasia. Our study demonstrates that NLP accurately identifies both the presence and level of dysplasia in this population. There are numerous challenges in applying novel data analytic techniques, like NLP, to large administrative datasets, including intricacies of data architecture and transferability of NLP algorithms between datasets. The advantage of an NLP platform such as CLAMP is the portability to other datasets.

This study has several limitations. Cases of IBD were determined by administrative codes and hence this study is subject to limitations inherent to administrative database studies. We have previously validated the accuracy of administrative codes for IBD in this dataset. Classification of dysplasia is also inherently limited by the quality and variability of reporting in pathology reports. This dataset represents one of the largest datasets of IBD patients in the US, and through CDW we have the unique capability to extract and manually review full-text pathology reports linked to individual patients. Classification of dysplasia was determined by two reviewers and discrepancies were adjudicated by a third reviewer. Lastly, we were unable to differentiated if dye-spray chromoendoscopy was performed which may influence the likelihood of presence of level of dysplasia identified. However, due to the time period of the study period, dye-spray chromoendoscopy was not routinely performed in the VA.

Conclusion

In summary, NLP can accurately identify colonic dysplasia in IBD patients in a large EHR data repository and can be used to identify large cohorts of IBD patients with dysplasia. This work will be used as a first step toward filling an important gap in the current literature—specifically, the further study of the natural history and outcomes of colonic dysplasia in patients with IBD.

List Of Abbreviations

IBD - Inflammatory Bowel Disease

UC—Ulcerative Colitis

CD - Crohn’s Disease

NPL - Natural Language Processing

EHR - Electronic Health Records

VHA - Veterans Health Administration

CDW - Corporate Data Warehouse

CLAMP - Clinical Language Annotation, Modeling, and Processing

LGD—Low Grade Dysplasia

HGD—High Grade Dysplasia

CRC - Colorectal Cancer

Declarations

Ethics approval and consent to participate:

Not applicable

Consent for publication:

Not applicable

Availability of data and materials:

The datasets generated and/or analyzed during the current study are not publicly available as data was obtained by completing a Data User Agreement the Veterans Health Administration (VHA), Corporate Data Warehouse (CDW) which the data resides behind the VA firewall. CDW compiles all electronic health record (EHR) and administrative data from the VHA facilities nationwide, including diagnostic and procedural codes and full-text notes from the EHR. Cases of IBD were identified using a case finding algorithm using ICD–9 codes previously validated in the VHA. Data are available from the corresponding author on reasonable request however only available from and with permission of Veterans Health Administration.

Competing interests:

Authors' declaration of personal interests:

(i) JH has served as a speaker for Abbvie and Janssen, a consultant for Abbvie, and served on an advisory board for Pfizer, Abbvie, and Janssen. JH has received research funding from Abbvie, Janssen, Pfizer, Celgene, and Redhill Biopharma, and Eli-Lilly

(ii) CT, ES, SS, PR, HX, and NM have no financial disclosures.

Funding:

Financial Support

Declaration of funding interests:

(i) This material is based upon work supported by the Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, and the Center for Innovations in Quality, Effectiveness and Safety (CIN 13–413 [NM, JH]) and in part by the Agency for Healthcare Research and Quality (AHRQ K08 PA13–180 [JH]), and the Crohn's and Colitis Foundation Career Development Award (JH).

(ii) No writing assistance was used in the preparation of this manuscript.

Authors' contributions:

Jason Hou- contributed in study design, data analysis, and authorship of manuscript. He has approved of the final draft submitted.

Christopher Taylor- contributed to authorship of manuscript. He has approved of the final draft submitted.

Ergin Soysal- contributed in study design, programming, data abstraction, data analysis, and editorial input in the manuscript. He has approved of the final draft submitted.

Shubhada Sansgiry- contributed in study design, programming, data abstraction, data analysis, and editorial input in the manuscript. She has approved of the final draft submitted.

Peter Richardson- contributed in study design, programming, data abstraction, data analysis, and editorial input in the manuscript. He has approved of the final draft submitted.

Hua Xu- contributed in study design, data interpretation, and editorial input in the manuscript. He has approved of the final draft submitted.

Nader N Massarweh- contributed to the interpretation of the data and authorship and critical reviews of manuscript. He has approved of the final draft submitted.

The views expressed in this article are those of the author(s) and do not necessarily represent the views of the Department of Veterans Affairs.

Acknowledgements:

Not applicable

References

1.Kornbluth, A., D. B. Sachar, and P. P. C.o.t.A. C.o. Gastroenterology, *Ulcerative colitis practice guidelines in adults: American College Of Gastroenterology, Practice Parameters Committee.* The American Journal

of Gastroenterology, 2010. *105:* p. 501–523; quiz 524.

2.Farraye, F. A., et al., *AGA technical review on the diagnosis and management of colorectal neoplasia in inflammatory bowel disease.* Gastroenterology, 2010. *138:* p. 746–774, 774.e1–4; quiz e12–13.

3.Bae, S. I. and Y. S. Kim, *Colon cancer screening and surveillance in inflammatory bowel disease.* Clin Endosc, 2014. *47(*6): p. 509–15.

4.Fumery, M., et al., *Incidence, Risk Factors, and Outcomes of Colorectal Cancer in Patients With Ulcerative Colitis With Low-Grade Dysplasia: A Systematic Review and Meta-analysis.* Clin Gastroenterol Hepatol, 2017. *15(*5): p. 665–674 e5.

5.Ananthakrishnan, A. N., et al., *Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach.* Inflammatory Bowel Diseases, 2013. *19:* p. 1411–1420.

6.Savova, G. K., et al., *Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications.* Journal of the American Medical Informatics Association: JAMIA, 2010. *17:* p. 507–513.

7.D'Avolio, L. W., et al., *Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC).* Journal of the American Medical Informatics Association: JAMIA, 2010. *17:* p. 375–382.

8.Hou, J. K., et al., *Automated identification of surveillance colonoscopy in inflammatory bowel disease using natural language processing.* Digestive Diseases and Sciences, 2013. *58:* p. 936–941.

9.Raju, G. S., et al., *Natural language processing as an alternative to manual reporting of colonoscopy quality metrics.* Gastrointest Endosc, 2015. *82(*3): p. 512–9.

10.Hou, J. K., et al., *Accuracy of Diagnostic Codes for Identifying Patients with Ulcerative Colitis and Crohn's Disease in the Veterans Affairs Health Care System.* Digestive Diseases and Sciences, 2014.

11.Soysal, E., et al., *CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines.* J Am Med Inform Assoc, 2017.

12.Rutter, M., et al., *Severity of inflammation is a risk factor for colorectal neoplasia in ulcerative colitis.* Gastroenterology, 2004. *126:* p. 451–459.

13.Winther, K. V., et al., *Long-term risk of cancer in ulcerative colitis: a population-based cohort study from Copenhagen County.* Clinical Gastroenterology and Hepatology: The Official Clinical Practice Journal of the American Gastroenterological Association, 2004. *2:* p. 1088–1095.

14.Rutter, M. D., et al., *Thirty-year analysis of a colonoscopic surveillance program for neoplasia in ulcerative colitis.* Gastroenterology, 2006. *130:* p. 1030–1038.

15.Söderlund, S., et al., *Decreasing time-trends of colorectal cancer in a large cohort of patients with inflammatory bowel disease.* Gastroenterology, 2009. *136:* p. 1561–1567; quiz 1818–1819.

16.Kim, B. J., et al., *Trends of ulcerative colitis-associated colorectal cancer in Korea: A KASID study.* Journal of Gastroenterology and Hepatology, 2009. *24:* p. 667–671.

17.Imler, T. D., et al., *Multi-center colonoscopy quality measurement utilizing natural language processing.* Am J Gastroenterol, 2015. *110(*4): p. 543–52.

18.Hou, J. K., T. D. Imler, and T. F. Imperiale, *Current and Future Applications of Natural Language Processing in the Field of Digestive Diseases.* Clinical Gastroenterology and Hepatology: The Official Clinical Practice Journal of the American Gastroenterological Association, 2014. *12:* p. 1257–1261.

19.Nayor, J., et al., *Natural Language Processing Accurately Calculates Adenoma and Sessile Serrated Polyp Detection Rates.* Dig Dis Sci, 2018. *63(*7): p. 1794–1800.

20.Hou, J. K., et al., *Su1815 Natural Language Processing Accurately Identifies Colorectal Dysplasia in a National Cohort of Veterans with Infammatory Bowel Disease (Abstract).* Gastroenterology, 2016. *150(*4): p. 560–561.

Tables

Table 1: Demographics of IBD patients with colorectal dysplasia (n=74,258)

Age at dysplasia (n=3,960)	62.1 ±11.5 years
Age at IBD index (n=74,135)	59.1 ±15.6 years
Gender (% male)	92.9%
Race	
White	72.6%
Black	8.1%
Hispanic/Other	1.9%
Unknown/missing	17.4%
IBD type	
Ulcerative colitis	53.7%
Crohn's disease	34.9%
IBDU	12.4%
Mean colonoscopies (#)*	4.2 (±4.3)

*Among pts with ≥1 colonoscopy n=27,164 (mean±std)

Figures

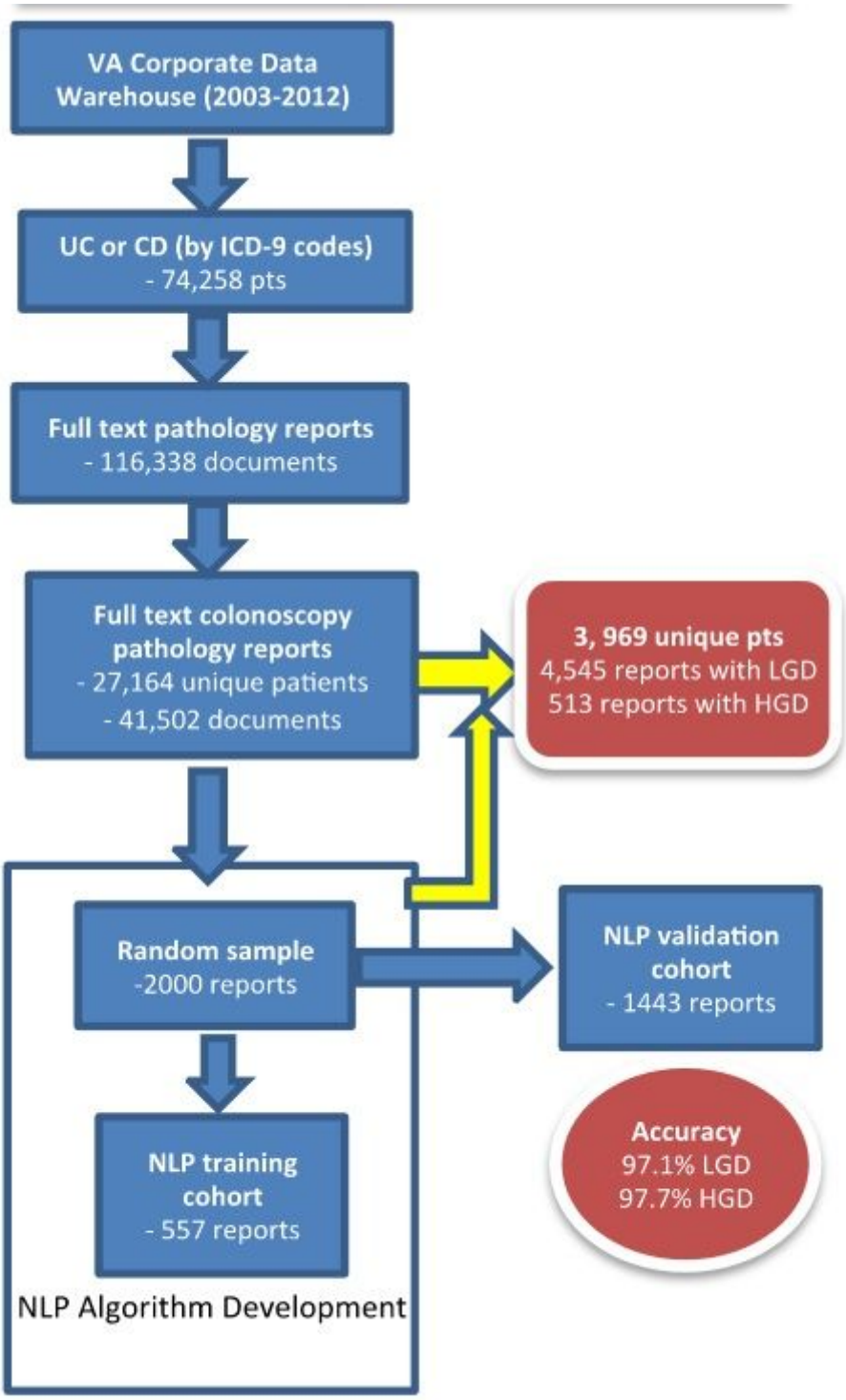
VA Corporate Data Warehouse (2003-2012)
UC or CD (by ICD-9 codes)
- 74,258 pts
Full text pathology reports
- 116,338 documents
Full text colonoscopy pathology reports
- 27,164 unique patients
- 41,502 documents
3, 969 unique pts
4,545 reports with LGD
513 reports with HGD
Random sample
-2000 reports
NLP validation cohort
- 1443 reports
NLP training cohort
- 557 reports
Accuracy
97.1% LGD
97.7% HGD
NLP Algorithm Development

Figure 1