Original Research Article

Health Informatics Journal

Health Informatics Journal
Volume 28: 1–11

DOI: 10.1177/14604582221134406
journals.sagepub.com/home/jhi
SAGE

Identification of colorectal cancer using structured and free text clinical data

Douglas F Redd, Yijun Shao and Qing Zeng-Treitler
Washington DC VA Medical Center, Washington, DC, USA
Biomedical Informatics Center, The George Washington University School of Medicine and Health Sciences, Washington, DC, USA

Laura J Myers
Richard L Roudebush VA Medical Center, Indianapolis, IN, USA
Indiana University School of Medicine, Indianapolis, IN, USA
Regenstrief Institute Inc, Indianapolis, IN, USA

Barry C Barker
Richard L Roudebush VA Medical Center, Indianapolis, IN, USA

Stuart J Nelson
Biomedical Informatics Center, The George Washington University School of Medicine and Health Sciences, Washington, DC, USA

Thomas F Imperiale
Richard L Roudebush VA Medical Center, Indianapolis, IN, USA
Indiana University School of Medicine, Indianapolis, IN, USA
Regenstrief Institute Inc, Indianapolis, IN, USA

Abstract

Colorectal cancer incidence has continually fallen among those 50 years old and over. However, the incidence has increased in those under 50. Even with the recent screening guidelines recommending that screening begins at age 45, nearly half of all early-onset colorectal cancer will be missed. Methods are needed to identify high-risk individuals in this age group for targeted screening.

Corresponding author:
Qing Zeng-Treitler, Department of Biomedical Informatics Center, Washington DC VA Medical Center, 2600 Virginia Ave, Suite 345, Washington, DC 20037, USA, Washington, DC, USA.
Email: zengq@gwu.edu

Colorectal cancer studies, as with other clinical studies, have required labor intensive chart review for the identification of those affected and risk factors. Natural language processing and machine learning can be used to automate the process and enable the screening of large numbers of patients. This study developed and compared four machine learning and statistical models: logistic regression, support vector machine, random forest, and deep neural network, in their performance in classifying colorectal cancer patients. Excellent classification performance is achieved with AUCs over 97%.

Keywords
Colon cancer, feature utilization, machine learning, model comparison, statistical models

Introduction

The incidence of colorectal cancer (CRC) increases beginning at the age of 50.[1] Screening for CRC is widely recommended for those of average risk in this age group.[2-4] As a result, incidence has continually fallen among those 50 and older.[5] CRC incidence increased by 50% in the age group 40–45 over the years 1987 to 2006.[6] Patients younger than 50 years tend to have a worse prognosis, presenting with more advanced disease.[7-10] For these reasons, several guideline organizations now recommend age 45 for screening.[11,12] Even if fully implemented and followed, half of all cases of CRC in persons younger than age 50 would be missed because of CRC instances in those below the age of 45. A method is needed to identify other high-risk individuals.

Traditionally, domain experts have identified risk factors by review of all available data on a subject. In medical settings, this is generally accomplished through a process of chart review;[13,14] chart review is labor-intensive, requiring large time commitments from skilled researchers and clinicians. Automating some of that effort can make the process more efficient, and more information can potentially be analyzed.

There are several prior studies related to the identification of CRC cases using EHR data.[15–18] The most pertinent to our study is a 2011 study by Xu et al.[19] It describes an algorithm combining machine learning and natural language processing to detect patients with colorectal cancer (CRC) from entire EHRs at Vanderbilt University Hospital. The algorithm achieved an excellent F-measure of 0.93. This study, however only focused on identifying CRC cases that had CRC ICD codes or CRC diagnosis in free text notes. It also did not take any risk factors into count. In contrast, the algorithm we report in this paper does not depend on the CRC diagnostic codes because our ultimate goal is to identify high-risk individuals for targeted screening who would not have CRC diagnosis codes present.

Known risk factors for early onset CRC include age, male gender, body mass index (especially in late adolescence and early adulthood), cigarette smoking, and alcohol consumption.[5–10] Other candidate risk factors include metabolic syndrome, low physical activity, processed meat consumption, and oral antibiotic use in childhood. Some of these variables (e.g., prior diagnoses) can be easily obtained from structured data tables in the electronic health records (EHR). Other variables including social history and lifestyle factors are often documented more in the free text notes. We thus developed and applied natural language processing and machine learning techniques, which have shown impressive results in many cases.[5–7]

To extract both known and unknown CRC risk factors, we used an unsupervised method of topic modeling, the Latent Dirichlet allocation (LDA) algorithm.[5] Topic modeling is typically used to

discovering common themes, called "topics," that are shared by documents in a large text corpus. These topics are technically represented as a series of words that are thought to be semantically related. LDA is one of the most widely used topic modeling approaches and makes assumptions that topics are probabilistic distributions over words and that documents are mixtures of topics. It has found applications in many areas, including biomedicine.[7,20–22]

There are many machine learning methods. We trained and evaluated the performance of logistic regression (LR), support vector machine (SVM), random forest (RF), and deep neural network (DNN) models in the identification of CRC in US military Veterans aged 35–49 from medical records.

Methods

Data set

Structured and unstructured data from the Veterans Administration's Corporate Data Warehouse (CDW) was used.[23] Statistical and ML models were trained and tested using a subset of CDW data that had undergone a chart review for a case-control study on risk factors for early-onset CRC in Veterans. In this dataset, study patients were initially classified into three groups: (1) Cases diagnosed with CRC during 2008–2015; (2) Colonoscopy controls, who underwent colonoscopy for diagnostic purposes (e.g., rectal bleeding) but were not diagnosed with CRC; and (3) Clinic controls, who did not undergo colonoscopy, were not diagnosed with CRC and were seen in a primary care clinic at least once per year for each of the 2 years immediately preceding the index date of a matched case. Inclusion and exclusion criteria are shown in Appendix 1. Colonoscopy controls and clinic controls were matched to cases on healthcare facility (based on the location of where the majority of primary care encounters occur) and year of index date in a 2:1 ratio (4 total controls per case), and then merged into a single control group. CRC diagnosis was determined by CRC oncology reports in CDW and through the VA's cancer registry. In total, there are 722 cases and 3617 controls.

Feature extraction

Features from structured data. We included 52 features representing medications, diagnoses, procedures, and relevant document types that were present in greater than 10% of the patients between 1 year before and 1 month after the index dates of the case and control patients. There were no procedures other than colonoscopy that occurred in 10% or more of the patients in our cohort; thus, procedure codes were not included. Additionally, we included the total number of documents, diagnoses, procedures, prescriptions, and visits for each patient during the period between 1 year before the index date and 1 month after the index date. A summary of the features is shown in Table 1.

Features from unstructured data. The Latent Dirichlet allocation (LDA) algorithm[22] was used to generate 1000 topics based on clinical notes (n = 33,135) between 1 year before and 1 month after the index dates of the case and control patients. To increase the size of the training set, an additional 2576 documents were included from 70 patients diagnosed with CRC and 10,634 documents from 1188 patients who underwent colonoscopies from the CDW. Documents were converted to lowercase, and words in a stopword list or words occurring 10 times or less were excluded. The stopword list used was a general-purpose list of 524 common English words, to which we added the

Table 1. Summary of variables.

	icd9_v65.9	Other and unspecified hyperlipidemia
	icd9_455.0	Obesity, unspecified
	icd9_569.3	Anxiety state, unspecified
	icd9_309.81	Tobacco use disorder, unspecified use
	icd9_v57.1	Posttraumatic stress disorder
	icd9_v81.2	Depressive disorder, not elsewhere classified
	icd9_v65.40	Presbyopia
	icd9_724.2	Unspecified essential hypertension
	icd9_278.00	Internal hemorrhoids without mention of complication
	icd9_719.46	Esophageal reflux
	icd9_250.00	Hemorrhage of rectum and anus
	icd9_v65.49	Blood in stool
	icd9_578.1	Pain in joint involving shoulder region
	icd9_v65.3	Pain in joint involving lower leg
	icd9_719.41	Lumbago
	icd9_305.1	Unspecified sleep apnea
	icd9_v76.51	Need for prophylactic vaccination and inoculation, influenza
	icd9_272.4	Care involving other physical therapy
	icd9_v04.81	Dietary surveillance and counseling
	icd9_311.	Other unspecified counseling
	icd9_530.81	Other specified counseling
	icd9_401.9	Unspecified reason for consultation
	icd9_367.4	Special screening for malignant neoplasms, colon
	icd9_300.00	Screening for other and unspecified cardiovascular conditions
Medications	Magnesium citrate	Binary value
	Gabapentin	
	Naproxen	
	Simvastatin	
	Bisacodyl	
	Omeprazole	
	Electrolytes/peg-3350	
	Acetaminophen/hydrocodone	
	Cyclobenzaprine	
	Trazodone	
	Ibuprofen	
	Supply	
	Docusate	
	Lisinopril	
	Tramadol	

(continued)

Table 1. (continued)

Document Type	Pathology		Numerical value (Integer)
	Hematology		
	Oncology		
	gi		
	Colonoscopy		
	All document count		Numerical value (Integer)
	All diagnosis count		
	All procedure count		
	All medication count		
	All visit count		
Topics	Topic	example topics	Numerical value (Percent)
	Topic_1	labs, elevated, panel, lipid, lab, cbc, tsh, ldl, liver, fasting, ordered, months, repeat, cholesterol, low, normal, start, daily, lfts, wnl	
	Topic_2	cancer, history, family, age, mother, father, years, colon, died, brother, social, children, sister, past, ago, yrs, lives, alive, colonoscopy, medical	
	……	colon, colonoscope, tissue, colonoscopy, consent, procedure, biopsy, disease, interventions, treatment/ procedure, wire, bleeding, polyp, removal, forceps, decompression, medication, performed, tube, physician	
	Topic_998	assessment, level, fall, normal, baseline, risk, consciousness, patients, score, status, pain, change, falls, oriented, scale, problems, admission, mental, acute, sounds	
	Topic_999	pain, colon, morphine, liver, mass, metastatic, previously, series, coll, prior, image, sigmoid, continue, lesion, pca, gist, port, regimen, cancer, daily	

25 words most common in our set of clinical documents including "patient," "date," and "time." Table 1 shows examples of some of the topics that were discovered. We did not filter or select the LDA topics based on their content or relation to the outcomes of interest. As such, the LDA served as an unsupervised feature extraction tool. The number of topics (1000) was chosen empirically after reviewing the topic content, after experimenting with smaller numbers of topics (250, 500, and 750). Because the documents were not only focused on colon cancer, many topics including diseases, social history, symptoms, medications, procedures, and tests are covered in the clinical notes, especially in the control group. Only when the number of topics reached 1,000, were we able to find topics specifically focused on colon and colonoscope.

The topic model was applied to all clinical notes on each patient recorded from 1 year before the index date to 1 month after. The proportions of each topic's representation in each patient note were averaged to give an overall proportion for each topic for each patient. These probabilities were used as the values for the topic features in the ML data set.

Machine Learning

Four supervised ML methods were applied and evaluated for their ability to classify patients as cases or controls. These included Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Deep Neural Network (DNN) mechanisms. For the first three methods (i.e., LR, SVM, RF), we used a Java ML library called Weka. We also used the default settings in Weka for the hyperparameters: LR:[24] ridge parameter=1.0E-8; SVM:[25] Kernel = Linear, complexity constant = 1.0; RF:[26] Size of each bag, as a percentage of the training set size = 100, Number of trees = 100, number of attributes to randomly investigate = 0, minimum number of instances = 1, minimum variance for split = 0.001, Seed for random number generator = 1, maximum depth of the tree = unlimited. Because of the large number of parameters, we cannot explain each one in detail but included the references for the algorithms.

For DNN, we implemented a Python deep learning library named Theano[27] together with a helper library called Lasagne.[28] The DNN was constructed of 5 hidden layers of sizes 200, 300, 200, 300, 200, and a single output using sigmoid activation, 300 epochs, batch size 100, and Nesterov momentum[29] with a constant learning rate of 0.001 and momentum of 0.9.

To avoid overfitting in the DNN training, the data was partitioned into 70% training, 20% validation, and 10% testing groups, with performance being measured on the testing group only. Training, validation, and testing groups were created using stratified sampling. After each pass over the whole training set, the DNN model was evaluated in the validation set. When performance on the validation dataset starts to degrade, we stopped further training the DNN model and applied it to the 20% testing set to assess final performance. This strategy is called "early stopping."

LR, SVM, and RF did not require a separate validation dataset. The evaluation was performed with 10-fold cross-validation in all cases, each fold consisting of 3905 training (including validation for DNN) patients (650 cases, 3255 controls) and 434 testing patients (72 cases, 362 controls). All performance measures including the confusion matrix results were computed as the micro average of the 10 evaluations.

Results

The overall performance of the models is summarized in Table 2. All models performed well, with SVM, RF, and DNN all performing better than LR in F-measure and AUC-ROC. RF and DNN showed the highest AUC-ROC at 0.975 and 0.965, respectively. DNN had the best accuracy of 0.985. The confusion matrixes of the models are shown in Table 3. Overall, DNN can be viewed as a better model because it performed well on all measures (recall, specificity, accuracy and AUC-ROC).

Table 2. Machine learning models and their performance in classifying colorectal cancer cases versus controls. Values are represented as mean.

	Recall	Specificity	Accuracy	AUC-ROC
Logistic regression	0.757	0.891	0.869	0.875
Support vector machine	0.884	0.995	0.976	0.939
Random forest	0.662	0.998	0.942	0.975
Deep neural network	0.911	0.997	0.982	0.965

Table 3. Confusion matrixes of LR, SVM, RF and DNN models.

	True				True		
LR	P	N		SVM		P	N
	Y	545	395		Y	636	18
	N	175	3225		N	84	3602
	True				True		
RF		P	N	DNN		P	N
	Y	477	7		Y	656	11
	N	243	3613		N	64	3609

Discussion

Main findings

This study demonstrated that multiple ML models can correctly classify patients ages 35–49 with colorectal cancer using a combination of structured and unstructured medical data. This process can greatly impact future colorectal cancer studies by reducing the effort required to perform a large-scale chart review.

Among the 3 ML methods (SVM, RF, and DNN), DNN can be considered to be the most powerful and complex while the least interpretable. The fact that all three achieved an excellent and comparable performance level (in terms of F measure and AUC) is not very surprising, as ML methods' performance tends to be task and dataset-specific.

Compared to the most pertinent prior study by Xu et al.,[19] our work focused on a separate age group (<50 years old). The low prevalence of CRC in the younger patient limited the number of cases we could use for learning. On the other hand, we utilized more features including procedures, medications, and comorbidities as well as a large number of topics as opposed to CRC ICD codes and keywords. As a result, were achieved a higher AUC (97.5% vs 93%). Even though we did not explore the feature importance or contribution in this paper, it is well known that models like DNN benefit from a large number of features.

Beyond CRC, there have been a number of prior studies that have utilized a combination of structured data and/or free text notes for case identification, including our own work[30,31] and work by other researchers.[32,33] Such efforts are often driven by the need for more accurate or consistent phenotype definitions beyond what diagnostic codes (often assigned for billing purposes) have to offer. This study, consistent with prior studies, showed that free text notes could be mined and combined with structured data to achieve high distinguishing power.

Limitations and future work

We selected the DNN parameters based on past experience and used the default settings in Weka for the hyperparameters for LR, SVM, RF. Since all ML methods reached a fairly high performance level, there was no need for extensive hyperparameter tuning. Without tuning the hyperparameters, we did not need to use a validation dataset for LR, SVM, and RF and went directly from training to testing. Before any ML models are used in a clinical context, additional validation would need to be performed. It is probable that these models will continue to perform well on a VA population meeting the same inclusion/exclusion criteria. Applying these models to other populations would

require adaption and additional validation, adding examples from the target population to the training set or using a hierarchical method to augment the original results.

In the real world, CRC under 50 years of age is extremely uncommon (33 per 100,000 in 45–49-year-olds vs. 59.5 per 100,000 in 50–54-year-olds)[34] but is associated with worse outcomes than in older patients. For that reason, we sought to identify high-risk patients for targeted screening. In terms of training a model, having extremely imbalanced classes is a known problem. We chose the 1:5 match to allow more effective learning from positive and negative cases. We also want to note that the AUC values we reported are not dependent on the prevalence of cases in a sample, while accuracy rates are. In future studies, we would like to apply this model to a large set of clinical data and evaluate the positive predictive values using different thresholds.

We utilized a large number of topics ($n = 1000$) and a smaller number of structured data elements ($n = 52$). We did not experiment with a model using only the topics or structured data, which can be explored in the future. Features from the topic modeling cover all findings in a clinical note, from personal history to medications and from mental health to functional status. Because we performed unsupervised topic modeling, the topics do not exactly correspond to a specific diagnosis or treatment. As such, we expect the topics to overlap but not completely overlap with the structured data variables.

From the perspective of interpreting a model, having fewer features is desirable. On the other hand, one of the key strengths of DNN is its ability to handle a large number of features and does not require feature engineering. In a related analysis not described in this paper, we explored the features contributing to the different models by examining the feature coefficients in LR, weight in SVM, importance measure[26] for random forest, and impact score[35] for DNN. Different topic features were among the highest contributing features. For example, the top feature in the DNN is a topic on colon and colonoscopy. What was intriguing is that there was almost no overlap in the top 15 features utilized by the 4 different models. Further studies are needed to interpret the models and understand why and how other models use different features.

This study focused on the classification of CRC cases; however, an informative follow-on study could investigate predictive modeling. This may be accomplished by restricting the data period to only include measurements from 3 months to 1 year prior to the diagnosis. With a predictive model, it may be possible to identify features corresponding to risk factors. This could allow validation by features mapping to known risk factors and also identifying possibly unknown risk factors.

Conclusion

Patients with colorectal cancer can be identified with a highly reliable performance from structured and unstructured medical records using ML. These ML models can be used to decrease the manual effort required for chart review to identify cases in clinical research.

Summary table

- The ML and statistical model identified colorectal cancer cases using EHR with the best AUCs over 97% and best accuracy over 98%.
- Overall, the DNN model performed the best.
- There was a very low correlation among the four models in terms of the top features used.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by a Veterans Administration Merit Review Grant (IIR 14-011).

Ethical approval

Due to the retrospective character of the study and because data were collected during routine healthcare procedures and patients were not subject to intervention, the Indiana University Institutional Review Board waived the need for informed consent (#1511734340).

Data availability

The datasets generated during and/or analyzed during the current study are not publicly available to protect the privacy of research participants. Still, aggregated datasets are available from the corresponding author on reasonable request.

ORCID iD

Qing Zeng-Treitler https://orcid.org/0000-0002-8353-7473

References

1. Edwards BK, Ward E, Kohler BA, et al. Annual report to the nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates. *Cancer* 2010; 116: 544–573. DOI: 10.1002/cncr.24760
2. Davidson KW, Barry MJ, Mangione CM, et al. Screening for colorectal cancer: U.S. preventive services task force recommendation statement. *Ann Intern Med* 2008; 149: 627–637.
3. Rex DK, Johnson DA, Anderson JC, et al. American College of Gastroenterology guidelines for colorectal cancer screening 2009 [corrected]. *Am J Gastroenterol* 2009; 104: 739–750. DOI: 10.1038/ajg.2009.104
4. Levin B, Lieberman DA, McFarland B, et al. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *Gastroenterology* 2008; 134: 1570–1595. DOI: 10.1053/j.gastro.2008.02.002
5. *Colorectal Cancer - Cancer Stat Facts*, NIH, (2020). https://seer.cancer.gov/statfacts/html/colorect.html
6. Wachter K. *Colorectal cancer rates up in people aged 40 to 44*, 4. GI & Hepatology News AGA Institute, 2010, pp. 1–4.
7. You YN, Xing Y, Feig BW, et al. Young-onset colorectal cancer: is it time to pay attention? *Arch Intern Med* 2012; 172: 287–289. DOI: 10.1001/archinternmed.2011.602
8. Fairley TL, Cardinez CJ, Martin J, et al. Colorectal cancer in US adults younger than 50 years of age 1998–2001. *Cancer* 2006; 107: 1153–1161.
9. Marble K, Banerjee S and Greenwald L. Colorectal carcinoma in young patients. *J Surg Oncol* 1992; 51: 179–182. DOI: 10.1002/jso.2930510311
10. O'Connell JB, Maggard MA, Liu JH, et al. Do young colon cancer patients have worse outcomes? *World Journal Surgery* 2004; 28: 558–562.

11. Wolf AMD, Fontham ETH, Church TR, et al. Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society. *CA Cancer J Clin* 2018; 68: 250–281. DOI: 10.3322/caac.21457
12. US Preventive Services Task Force. Screening for colorectal cancer: US preventive services task force recommendation statement. *JAMA* 2021; 325: 1965–1977. DOI: 10.1001/jama.2021.6238
13. Worster A and Haines T. Advanced statistics: understanding medical record review (MRR) studies. *Acad Emerg Med* 2004; 11: 187–192.
14. Findley TW and Daum MC. Research in physical medicine and rehabilitation. III. The chart review or how to use clinical data for exploratory retrospective studies. *Am Journal Physical Medicine Rehabilitation* 1989; 68: 150–157.
15. Denny JC, Choma NN, Peterson JF, et al. Natural language processing improves identification of colorectal cancer testing in the electronic medical record. *Med Decis Making* 2012; 32: 188–197. DOI: 10.1177/0272989X11400418
16. Parthasarathy G, Lopez R, McMichael J, et al. A natural language-based tool for diagnosis of serrated polyposis syndrome. *Gastrointest Endosc* 2020; 92: 886–890.
17. Vadyala SR and Sherer EA. Natural language processing accurately categorizes indications, findings and pathology reports from multicenter colonoscopy. arXiv preprint arXiv:210811034 2021.
18. Imler TD, Morea J, Kahi C, et al. Natural language processing accurately categorizes findings from colonoscopy and pathology reports. *Clin Gastroenterol Hepatol* 2013; 11: 689–694. DOI: 10.1016/j.cgh.2012.11.035
19. Xu H, Fu Z, Shah A, et al. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. *AMIA Annu Symp Proc* 2011; 2011: 1564–1572.
20. Menze BH, Kelm BM, Masuch R, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* 2009; 10: 213. DOI: 10.1186/1471-2105-10-213
21. Zhao W, Zou W and Chen JJ. Topic modeling for cluster analysis of large biological and medical datasets. *BMC Bioinformatics* 2014; 15: S11.
22. Blei DM, Ng AY and Jordan MI. Latent dirichlet allocation. *J Machine Learn Research* 2003; 3: 993–1022.
23. Health Services Research & Development. *Corporate Data Warehouse (CDW)*, https://www.hsrd.research.va.gov/for_researchers/cdw.cfm
24. Le Cessie S and Van Houwelingen JC. Ridge estimators in logistic regression. *J R Stat Soc Ser C (Applied Statistics)* 1992; 411: 191–201.
25. Platt J. Sequential minimal optimization: A fast algorithm for training support vector machines. In: Schoelkopf B, Burges C and Smola A (eds). *Kernel Methods - Support Vector Learning*, Microsoft, 1998.
26. Breiman L. Random forests. *Machine Learning* 2001; 45: 5–32.
27. Bergstra JBO, Bastien F, Lamblin P, et al. Theano: A CPU and GPU math expression compiler. In: *Proceedings of the Python for Scientific Computing Conference*. SciPy, 2010.
28. Dieleman SSJ, Raffel C, Olso E, et al. *Lasagne*. First release, 2015.
29. Sutskever I, Martens J, Dahl G, et al. On the importance of initialization and momentum in deep learning. *Proc 30th Int Conf Machine Learn PMLR* 2013; 28: 1139–1147.
30. Shao Y, Zeng QT, Chen KK, et al. Detection of probable dementia cases in undiagnosed patients using structured and unstructured electronic health records. *BMC Med Inform Decis Mak* 2019; 19: 128.
31. Walsh JA, Shao Y, Leng J, et al. Identifying axial spondyloarthritis in electronic medical records of US veterans. *Arthritis Care Res (Hoboken)* 2017; 69: 1414–1420. DOI: 10.1002/acr.23140

32. Hazlehurst B, Green CA, Perrin NA, et al. Using natural language processing of clinical text to enhance identification of opioid-related overdoses in electronic health records data. *Pharmacoepidemiol Drug Saf* 2019; 28: 1143–1151. DOI: 10.1002/pds.4810
33. Maarseveen TD, Meinderink T, Reinders MJT, et al. Machine learning electronic health record identification of patients with rheumatoid arthritis: algorithm pipeline development and validation study. *JMIR Med Inform* 2020; 8: e23930. DOI: 10.2196/23930
34. Siegel RL, Miller KD, Goding Sauer A, et al. Colorectal cancer statistics, 2020. *CA Cancer J Clin* 2020; 70: 145–164. DOI: 10.3322/caac.21601
35. Shao Y, Cheng Y, Shah RU, et al. Shedding light on the black box: explaining deep neural network prediction of clinical outcomes. *J Med Syst* 2021; 45: 5.

Appendix 1

Inclusion and exclusion criteria for cases, colonoscopy controls, and clinic controls. Note that for this study, the primary and secondary control groups (Clinic Controls and Colonoscopy Controls) are merged into a single control group.

Group	Inclusion	Exclusion
Cases	Colorectal cancer • CRC diagnosis during 2008–2015 • Ages 35-49 at CRC diagnosis	• Inflammatory bowel disease • Hereditary polyposis or non- polyposis syndrome • Surgical resection of any part of the colon prior to CRC diagnosis • High-risk family history of colorectal cancer
Colonoscopy controls	Colonoscopy without colorectal cancer • Ages 35–49 • Colonoscopy during 2008–2015	• Same exclusions as cases • Screening colonoscopy performed on patients at average risk for CRC • Colonoscopies that failed to reach cecum, terminal ileum, or appendiceal orifice or with bowel prep quality insufficient for adequate examination
Clinic controls	No colonoscopy and no colorectal cancer • Ages 35–49 • No history of colonoscopy • Received at least some of their health care through VA-based primary care clinics	• Same exclusions as cases • Prior colonoscopy