Original Article

Data mining-based model and risk prediction of colorectal cancer by using secondary health data: A systematic review

Hailun Liang[1*], Lei Yang[2*], Lei Tao[3], Leiyu Shi[4], Wuyang Yang[5], Jiawei Bai[6], Da Zheng[7], Ning Wang[2], Jiafu Ji[8]

[1]School of Public Administration and Policy, Renmin University of China, Beijing 100872, China; [2]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Beijing Office for Cancer Prevention and Control, Peking University Cancer Hospital & Institute, Beijing 100142, China; [3]Department of Public Policy, City University of Hong Kong, Hong Kong SAR, 999077, China; [4]Johns Hopkins Primary Care Policy Center, Baltimore, MD 21205, USA; [5]Department of Neurosurgery, Johns Hopkins Medicine, Baltimore, MD 21205, USA; [6]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA; [7]Department of Computer Science, Johns Hopkins Whiting School of Engineering, Baltimore, MD 21205, USA; [8]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Center of Gastrointestinal Surgery, Peking University Cancer Hospital & Institute, Beijing 100142, China

*These authors contributed equally to this work.

Correspondence to: Jiafu Ji, MD, PhD. Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Center of Gastrointestinal Surgery, Peking University Cancer Hospital & Institute, Beijing100142, China. Email: jijiafu@hsc.pku.edu.cn; Ning Wang. Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Beijing Office for Cancer Prevention and Control, Peking University Cancer Hospital & Institute, Beijing 100142, China. Email: bjwangning@126.com.

Abstract

Objective: Prevention and early detection of colorectal cancer (CRC) can increase the chances of successful treatment and reduce burden. Various data mining technologies have been utilized to strengthen the early detection of CRC in primary care. Evidence synthesis on the model's effectiveness is scant. This systematic review synthesizes studies that examine the effect of data mining on improving risk prediction of CRC.

Methods: The PRISMA framework guided the conduct of this study. We obtained papers via PubMed, Cochrane Library, EMBASE and Google Scholar. Quality appraisal was performed using Downs and Black's quality checklist. To evaluate the performance of included models, the values of specificity and sensitivity were comparted, the values of area under the curve (AUC) were plotted, and the median of overall AUC of included studies was computed.

Results: A total of 316 studies were reviewed for full text. Seven articles were included. Included studies implement techniques including artificial neural networks, Bayesian networks and decision trees. Six articles reported the overall model accuracy. Overall, the median AUC is 0.8243 [interquartile range (IQR): 0.8050−0.8886]. In the two articles that reported comparison results with traditional models, the data mining method performed better than the traditional models, with the best AUC improvement of 10.7%.

Conclusions: The adoption of data mining technologies for CRC detection is at an early stage. Limited numbers of included articles and heterogeneity of those studies implied that more rigorous research is expected to further investigate the techniques' effects.

Keywords: Systematic review; colorectal cancer; disease detection; data mining

Submitted Jan 19, 2020. Accepted for publication Apr 01, 2020.
doi: 10.21147/j.issn.1000-9604.2020.02.11
View this article at: https://doi.org/10.21147/j.issn.1000-9604.2020.02.11

Introduction

Globally, colorectal cancer (CRC) is the third most commonly diagnosed malignancy and the second leading cause of cancer death in the world, accounting for more than 1.8 million new cases and 900,000 deaths in 2018 (1,2). The global burden of CRC is expected to increase by 60% by 2030 (3). However, the burden of CRC can be reduced through prevention and early detection (4). Screening aims to identify individuals who are asymptomatic for the CRC and refer them promptly for diagnosis and treatment (5). Several screening tests have been developed and implemented to help doctors find CRC early, including both visual examinations (such as colonoscopy and sigmoidoscopy), and stool-based tests (like guaiac-based fecal occult blood test). The US Preventive Services Task Force's (USPSTF) updated recommendations in 2016 and recommended screening for CRC starting at age 50 years and continuing until age 75 years (6). Despite the availability of several effective options and proven benefit, a large fraction of eligible people does not participate in CRC screening programs (7). Possible reasons included an undifferentiated screening approach implemented to all average-risk persons and relatively greater time commitment over a short period required for examinations (6).

It is conceivable that the all average population could be stratified across a spectrum of predicted risk levels depending on the presence or absence of risk factors (8,9). Multiple demographic and clinical risk factors for CRC have been identified, such as age, gender, race, family history and behavioral risk factors (2). Moreover, several models that use these risk factors have been developed to predict the risk of CRC (8-11), which could be used to tailor the CRC screening schedules of people considered at average risk, and thus to optimize both outcomes and allocation of screening resources (12-14).

With advantages in data sizes and a large number of potential predictor variables, secondary data, such as electronic health records (EHR) and medical claims data, have been increasingly utilized to build and validate risk prediction models (15). The growing availability of health care datasets facilitates the development of analytic tools to stratify the population with average risk and prioritize patients at high risk of having CRC. Simple approaches based on single red flag symptoms, risk scores or indexes built on symptom complexes are likely to miss discrimination and calibration (16,17). Recent advances in population data and data science technologies have offered automated methods to extract information from large secondary datasets to identify individuals at increased risk, thereby potentially enhancing early detection of CRC at primary care settings.

Data mining analysis is the process to extract implicit, previously unknown and potentially useful patterns from data, which is also referred to as knowledge discovery in databases (18-20). Various data mining methods have been utilized for the detection of CRC (21,22). The mining makes powerful algorithms that can model previously unknown relationships in complex secondary datasets, and adapt to dynamic data environments (23). Undoubtedly, data mining approaches in CRC are of great concern when it comes to early detection, prevention, management and other related clinical administration aspects. Hence, in the framework of this study, efforts were made to review the current literature on data mining approaches in CRC risk prediction.

Materials and methods

Search strategy

A systematic search for journal article was conducted in March 2018 on four electronic database: PubMed, Cochrane Library, EMBASE and Google Scholar. The specific search strategy was as follows: ("data mining" [Title/Abstract] OR "machine learning"[Title/Abstract] OR "data driven"[Title/Abstract] OR "algorithm" [Title/Abstract] OR "text mining"[Title/Abstract] OR "natural language processing"[Title/Abstract] OR "support vector machines"[Title/Abstract] OR "decision trees" [Title/Abstract] OR "Bayesian networks"[Title/Abstract] OR "artificial neural networks"[Title/Abstract]) AND (colorectal cancer[Title/Abstract] OR oncology[Title/Abstract] OR neoplasms[Title/Abstract]).

Definition of data mining

Data mining is described as the process of selection, exploration, and modeling of a large database to discover unknown models or patterns (21). Typical data mining algorithms include decision trees, naive Bayes classifiers, support vector machine, and artificial neural network, etc. There is an apparent difference between a data mining model and a traditional prediction model. As depicted in *Figure 1*, in a traditional prediction program, the prediction model is given which will produce the desired output when

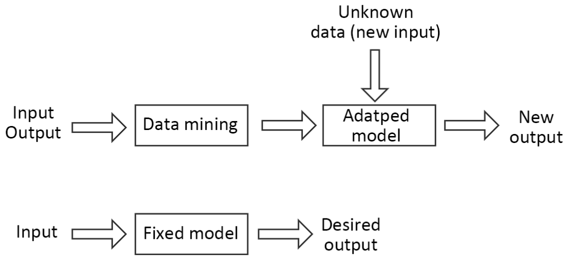

Figure 1 Difference between traditional model and data mining.

entering some particular input variables. While in a data mining process, computer algorithms will generate a series of rules when simultaneously specifying input and output variables (22).

Study selection

Figure 2 describes the selection process of the included articles. Initial records were transferred to the reference management software Endnote 8.0 (Thomson ResearchSoft, Stanford, USA), and duplicates were excluded automatically. Then two independent investigators (HL and LT) identified articles by reading the title and abstract and assessed them according to the following selection criteria: 1) articles related to CRC; 2) using data mining methods, such as decision trees, random forests and deep neural networks; and 3) theoretical articles are excluded. Any disagreements were resolved through discussion among investigators.

Articles that meet these selection criteria would be kept for the next full-text screening stage. Two investigators read the full texts of the remaining articles to assess them based on the following criteria: 1) research should be limited to cancer prevention or prediction, hence we excluded articles aiming at cancer treatment or diagnosis; 2) we included studies that use any single or combined aforementioned data mining models; and 3) a prediction model in detecting CRC must use at least one of the three types of features as input variables, including clinic features, imaging features, and molecular features (24). For the data availability in the primary care settings, we only select the models based on clinic and treatment features.

Data synthesis and analysis

Data extraction was conducted by two investigators (HL and LT). For each article, we extracted information shown in *Table 1*. Due to the high heterogeneity for quantitative analysis, descriptive statistics were presented for each included study. Moreover, we compared the values of specificity and sensitivity and computed the median of overall area under the curve (AUC) of included studies,

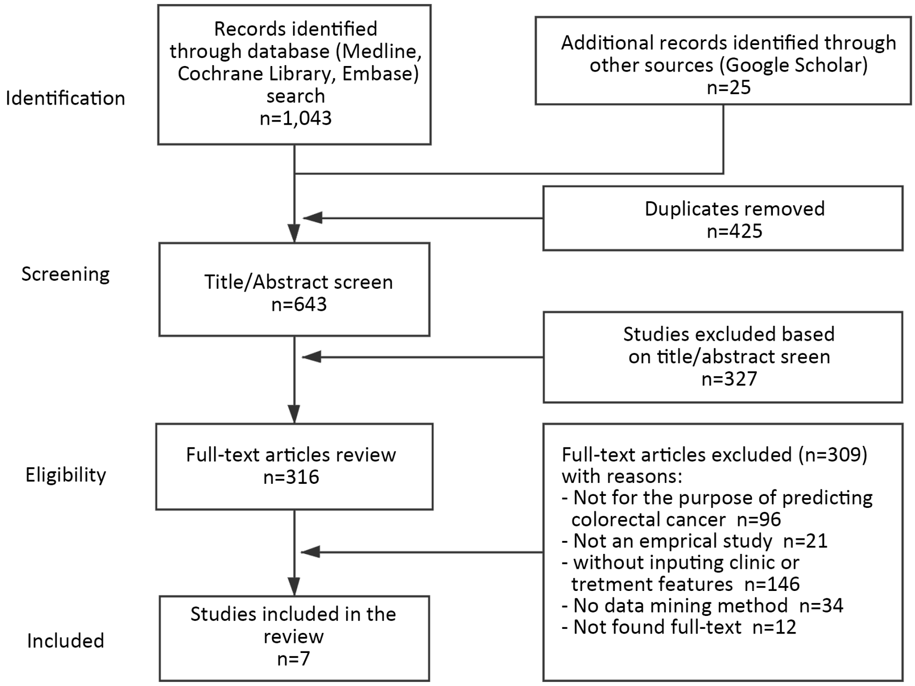
Identification
Records identified through database (Medline, Cochrane Library, Embase) search n=1,043
Additional records identified through other sources (Google Scholar) n=25
Duplicates removed n=425
Screening
Title/Abstract screen n=643
Studies excluded based on title/abstract sreen n=327
Eligibility
Full-text articles review n=316
Full-text articles excluded (n=309) with reasons:
- Not for the purpose of predicting colorectal cancer n=96
- Not an emprical study n=21
- without inputing clinic or tretment features n=146
- No data mining method n=34
- Not found full-text n=12
Included
Studies included in the review n=7

Figure 2 Screening process for eligible studies.

Table 1 Study characteristics

Author	Year	Country	Data source	Data mining methods	Sample	Input	Output	Validation method	Performance
Kinar *et al.* (25)	2016	Israel	MSH and the United Kingdom Health Improvement Network (THIN)	DT, GB, RF	451,535 samples aged 50−75 years old	Age, sex, CBC data	CRC risk scores	10-fold cross-validation	AUC: 0.81; OR for the association of high-risk scores with CRC at 90% specificity: 34; Specificity at 50% sensitivity was 88±2%
Hoogendoorn *et al.* (26)	2016	Netherlands	Primary care dataset in the region of Utrecht, the Netherlands	CART, LR, RF, NLP	90,000 samples aged over 30 years old	Age, gender, consults, lab tests, medication, and referrals, consultation notes	Occurrence of CRC	5-fold stratified cross-validation	AUC: 0.900 (95% CI: 0.886−0.914); Precision: 3.9%; Sensitivity at 90% specificity: 70%; F1-score: 0.039
Kop *et al.* (27)	2016	Netherlands	EMR dataset from three urban regions in the Netherlands	CART, LR, RF	263,879 samples aged over 30 years old	Age, gender, general practitioner (GP) consultations, drug prescriptions, specialist or additional diagnostic procedure referrals and lab test	Occurrence of CRC	5-fold stratified cross validation	AUC: 0.891 (95% CI: 0.879−0.903); Sensitivity at 90% specificity: 64.2%; F1-score at 10% false positive rate: 0.058

Kinar *et al.* (28)	2017	Israel	Maccabi Healthcare Services (MHS) electronic medical records (EMRs) and Israel Cancer Registry	DT, GB	112,584 samples aged 50−75 years old	Age, sex and CBC reports	CRC risk scores	10-fold cross-validation	Sensitivity at 1% percentile cutoff: 17.3% (23/135);
									Precision at 1% percentile cutoff: 2.1% (23/1,094);
									OR for the association of high-risk scores with CRC at 1% percentile cutoff: 21.8 (95% CI: 13.8−34.2)

Kop *et al.* (29)	2015	Netherlands	EMR dataset from the Utrecht region in the Netherlands	CART, LR, RF	219,447 samples aged over 30 years old	Age, gender, GP consults, drug prescriptions, specialist referrals, comorbidity, and lab test outcomes.	Occurrence of CRC	5-fold stratified cross-validation	AUC: 0.881 (95% CI: 0.864−0.898)

Table 1 (*continued*)

Table 1 (*continued*)

Author	Year	Country	Data source	Data mining methods	Sample	Input	Output	Validation method	Performance
Birks *et al.* (30)	2017	UK	Clinical Practice Research Datalink (CPRD) from the UK	RF	2,914,589 samples aged over 40 years old	Sex, year of birth, and CBC results	CRC risk scores	2-fold cross-validation	OR for a diagnosis of CRC at 99.84 points cutoff: 26.5 (95% CI: 23.3, 30.2); AUC: 0.776 (95% CI: 0.771, 0.781); Sensitivity at 95.5% specificity for the 18–24 month interval: 3.91%; Specificity at 50% Sensitivity for the 18–24 month interval: 82.73%
Hornbrook *et al.* (31)	2017	USA	Kaiser Permanente Northwest Region (KPNW) electronic medical record system and KPNW Tumor Registry	DT	17,095 samples aged 40–89 years old	Gender, year of birth, and CBC	CRC risk scores	10-fold cross-validation	AUC: 0.80 (95% CI: 0.79–0.82); Sensitivity for the 50–75 and 40–89 years old in the 0–180 days' window: 34.5%, 39.9% respectively; OR for the association of high-risk scores with CRC at 99% specificity: 34.7 (95% CI: 28.9–40.4)

DT, decision trees; GB, gradient boosting; CBC, complete blood count; CRC, colorectal cancer; OR, odds ratio; 95% CI, 95% confidence interval; AUC, area under the curve; RF, random forests; CART, classification and regression tree; NLP, natural language processing.

which was a common index to evaluate the performance and accuracy of the data mining model (22). We select the AUC value to evaluate the existing data mining model for CRC prediction based on the following considerations: First, although precision and recall are considered as two single indicators in practice, simply employing the above two indicators will cause great robustness problems. That is, when the data mining threshold setting was adjusted, the value of precision and recall indicators will be changed accordingly. However, AUC is not affected by its absolute value, which makes the evaluation of the classification ability of the model more robust. Secondly, the AUC value is obtained by calculating the sensitivity and specificity of each sample, which is independence from prevalence of abnormality. The prevalence of CRC is very small. In other words, the imbalance of positive and negative samples is significant in CRC patients. We believe that using AUC as a comprehensive index for evaluating the prediction performance of machine learning models for CRC is appropriate. In fact, AUC is also the focus of the existing literature.

Results

Results of search

We initially identified 1,043 citations based on our search criteria and obtained 25 citations from other sources, including theses, preprints and working papers. As described in *Figure 2* (PRISMA form), after removing duplicate articles, a total of 643 articles were left. First, we reviewed the title and abstract of each article, 327 were excluded. In the next stage, we reviewed the full text of the remaining 316 articles, and 309 studies were excluded from them, leaving a total of 7 articles. Articles that are unrelated to CRC prevention or prediction, or that use the image and genetic techniques for modeling, or that are unable to identify one of the data mining methods were excluded.

Study characteristics

We finally included a total of 7 articles on predicting CRC using data mining techniques. As shown in *Table 1*, these articles mainly originated from four countries including the United States, Israel, the United Kingdom and the Netherlands. The sample sized ranged from 90,000 to 451,535, representing more than 500,000 people. The time span was from 2015 to 2017.

With regard to the data mining methods, both supervised learning and non-supervised learning algorithm were employed among the included studies. The most commonly used data mining approaches are decision trees, random forests, and natural language processing. A large number of studies selected more than one algorithm for predictive modeling (25-28). In terms of types of data source, databases were mainly from highly reliable national databases or hospital clinical data systems, such as the Macabbi Health Services (MHS) EHRs, the Israel Cancer Registry, the Kaiser Permanente Northwest Region (KPNW) EHRs, and Clinical Practice Research Datalink (CPRD) dataset. Using the validation process is a common way to increase the robustness of the data mining model. Cross-validation was specified in all articles. The most frequently used validation procedure is the 5-fold and 10-fold validation method.

Overall prediction performance

A total of six articles reported the overall model accuracy (25-27,29-31). Several performance indicators included AUC, precision, recall and odds ratio measure to compare the CRC prevalence of individuals whose model score is above or below the threshold. In this review, we calculated the median AUC of these six studies. Overall, as shown in *Figure 3*, the median AUC is 0.8243 [interquartile range (IQR): 0.8050–0.8886; range: 0.776–0.900].

Four articles used age, gender and blood data from the complete blood count reports as features (25,28,30,31), and the output was the risk scores to assess the probability of having CRC among the suspected population. And three out of the five articles reported the overall AUC, ranged from 0.776 to 0.810. Three included articles used the same data mining algorithms developed by a Dutch team (26,27,29). Their input features included not only formatted data such as gender, age, but also unformatted

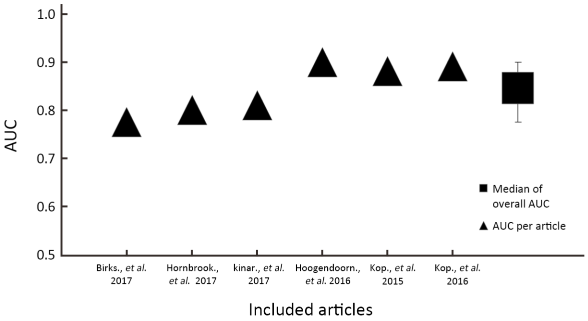

Figure 3 Summary plot and median of overall AUC. AUC, area under the curve.

forms of physician consultation and notes. The output is the probability for a patient to have CRC. The AUC ranged between 0.881 and 0.900. Due to the different sample sizes, the F1-score was particularly larger than the other two articles, with 0.058 in Kop *et al.* (27) and 0.074 in Hoogendoorn *et al.* (26).

Model performance by using subsets of samples

Three included articles reported model performance (25,30,32). For the gender subsets, three articles reported performance differences. In Hornbrook *et al.* (31), the odds ratio for association of a high-risk detection score with CRC was 39.6 [95% confidence interval (95% CI): 30.3−50.6] for women and 31.5 (95% CI: 24.3−39.5) for men at 99% of specificity, respectively. Similarly, Kinar *et al.* (28) found in the Israeli dataset, the odds ratio between men and women is 1:1.15.

In terms of follow-up time subsets, three articles reported performance at different follow-up times. In Hornbrook *et al.* (31), the odds ratio of detection of CRC in patients with a specificity of 99% was 34.7 in the window of 0−180 days, while was 20.4 in the window of 181−365 days. Birks *et al.* (30) also detected AUC in the UK dataset was 0.844 (95% CI: 0.839−0.849) for the 90−180 days' *vs.* 0.813 (95% CI: 0.809−0.818) for the 180−360 days' window at the 99.5% specificity level.

Comparison with other models

Two articles compared data mining models with different traditional models including logistic regression estimation, Bristol-Birmingham equation and passive learning (27,29). Kop *et al.* (27) found that the AUC of the data mining algorithm model was higher compared with the traditional model, with a value of 0.891 (95% CI: 0.879−0.903) and 0.864 (95% CI: 0.851−0.877), respectively. The similar results were also found in the Kop *et al.* (29) study. The AUC is 0.881 (95% CI: 0.864−0.898) and 0.796 (95% CI: 0.775−0.817) in random forest and regression model respectively, which shows the performance of random forest is 10.7% higher than regression. Overall, in the above two articles, the data mining method performed better than the traditional models, with the best AUC improvement of 10.7%.

Discussion

This systematic review summarizes the available CRC risk prediction models using data mining methods. One of the main strengths of this study is the broad search strategies and the systematic approaches used to identify studies and extract data. A total of six articles reported the overall accuracy of a model. The discriminatory power of the most currently available models is relatively high. The median AUC 0.8243 (IQR: 0.8050−0.8886) represents, therefore, a summary of overall performance. In the two articles that reported comparison results with traditional models, the data mining method performed better, with the best AUC improvement of 10.7%. In terms of quality of included studies, all included articles were in high quality based on the Downs and Black's quality checklist, and most models were cross-validated. This may conclude the tendency for current research into CRC risk prediction tools to focus not only on model development, but also validation and impact in clinical practice.

Data mining already has been well studied in many areas of clinical research. Several reviews have evaluated the application of data mining for prediction tasks (32-35). Cruz and Wishart (33) estimated that the data-driven approach substantially improves the accuracy of predicting cancer mortality and recurrence. Abbod *et al.* (34) reviewed and found that the application of artificial intelligence performed well for improving the diagnosis, staging and prognostic prediction of urological cancers. Kang *et al.* (35) presented a review of the use of machine learning and data mining to predict clinical outcomes in radiation oncology therapies but did not conclude an overall evaluation of models' performance. For CRC, evidence suggests that high-risk individuals are more likely to adhere to physician recommendations and receive CRC screening. Thus, identifying a high-risk group could optimize both outcomes and allocation of screening resources. The data mining-based risk models identified in this review have the potential to improve the early detection of CRC by helping health care providers to identify those patients presenting with risk of possible CRC, in whom further screening and examination are most appropriate. The potential advantages of risk prediction models in this context were that the included features were risk factors rather than symptoms, which were easily obtainable variables in many secondary datasets. These variables could be easily incorporated into practice. Moreover, these models overcome the limitations of symptom-based models, in which symptoms may present only at an advanced stage of CRC.

In contrast with certain other cancers with a single dominant risk factor, there is no single dominant predictor of CRC risk. In the absence of a single dominant risk factor, data mining is likely to remain the most powerful prediction method to identify the risk of developing the disease for population at average risk. In terms of the features selected through data mining models, gender and age are the most retained CRC risk factors, though no univariate associations have been reported. Three articles reported performance differences between the female and male subset of samples, but the results were mixed with respect to sensitivity and the positive predictive value of an abnormal score. Moreover, features related to blood count also seem to be another very crucial feature utilized in predicting CRC risk, since unexplained iron deficiency anemia due to bleeding is a "red flag" for CRC risk, especially among the elderly.

The overall results implied that a wide variety of data mining algorithms and techniques are used in the field of risk prediction of CRC. Included studies appeared to implement techniques almost all of the common known classification algorithms. However, the most commonly used ones are random forests and decision trees. A lot of articles used supervised learning approaches. More specifically, in all included research articles, the identified subsets of features were evaluated through appropriate procedures as splitting the dataset into train and test sets or through cross-validation. Moreover, after the feature selection, investigators have conducted comparative analysis on different data mining algorithms to compare the predictive performance and finally choose the most efficient one. The possible reason is likely that the accuracy of an algorithm depends heavily on the type of data dimensionality and origin (21).

Our study has limitations. First, non-English language studies were excluded, which could lead to language bias. Additionally, the methodological heterogeneity of the identified studies did not allow for meta-analysis. Furthermore, due to the limitation of the secondary dataset, models concerning the genetic background and environmental factors affecting the onset and progression of the disease were beyond the scope of the current study. Last, it should be cautious when applying our synthetic evidence to guide the real-world practice. AUC, the indicator that we selected to assess the overall performance of data mining model, may not be good enough to meet the needs of the clinical practice, as a high specificity is usually the priority to be secured for CRC screening.

It remains to be defined what role the currently emerging models can have, and what barriers existed to the incorporation of a data-driven CRC risk prediction model into practice. First, it is imperative for data mining studies that a dataset be sufficiently large for the algorithm to be trained appropriately. Furthermore, models are needed to be validated in diverse populations. Current studies rely largely on the health records or clinic data from the hospital, which makes the data mining model in CRC prediction only take effects among the hospital population rather than at the community-level. There is still a big gap between computer simulation based on hospital clinical data and the application of data mining methods in the real community setting. Additionally, ethical frameworks should be created to support the collection of training data and validation on heterogeneous settings before deployment. Although data mining models may become a valuable aid for clinical decision making, we do not foresee that data-driven algorithms replace human judgment. Rather, data mining has great potential as a complementary source of information to help guide the process of trusting decision making.

Conclusions

In this systematic review of data-driven risk prediction models for CRC in asymptomatic populations, a systematic effort was made to identify and review models that have been developed by using data mining to predict CRC risk. Many of these have shown better discrimination and accuracy, and most contain variables are easily obtainable. Moreover, many models have been validated in external populations, indicating that it is robust and applicable to populations from different countries. The current study showed machine learning-based algorithms were superior to the traditional ones, however, further research is still needed before these models can be incorporated into routine clinical practice. The advent of medical science and informatics is expected to give rise to further in-depth exploration toward early detection and prevention of CRC.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 71804183).

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

References

1. World Health Organization, Cancer Fact Sheet, 2018. Available online: https://www.who.int/news-room/fact-sheets/detail/cancer
2. American Cancer Society. Lifetime risk of developing or dying from cancer. Available online: https://www.cancer.org/cancer/cancer-basics.html
3. Ferlay J, Soerjomataram I, Ervik M, et al. GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012 v1.0. IARC CancerBase No. 11. Internatioanl Agency for Research on Cancer, 2013. Available online: http://globocan.iarc.fr/Pages/online.aspx
4. American Cancer Society. Cancer Prevention & Early Detection Facts & Figures 2017-2018. Atlanta: American Cancer Society, 2017.
5. American Cancer Society. Colorectal Cancer Facts & Figures 2017-2019. Atlanta: American Cancer Society, 2017.
6. US Preventive Services Task Force. Screening for colorectal cancer: US Preventive Services Task Force Recommendation Statement. JAMA 2016;315:2564-75.
7. Centers for Disease Control and Prevention (CDC). Vital signs: colorectal cancer screening test use--United States, 2012. MMWR Morb Mortal Wkly Rep 2013;62:881-8.
8. Driver JA, Gaziano JM, Gelber RP, et al. Development of a risk score for colorectal cancer in men. Am J Med 2007;120:257-63.
9. Freedman AN, Slattery ML, Ballard-Barbash R, et al. Colorectal cancer risk prediction tool for white men and women without known susceptibility. J Clin Oncol 2009;27:686-93.
10. Ma E, Sasazuki S, Iwasaki M, et al. 10-Year risk of colorectal cancer: development and validation of a prediction model in middle-aged Japanese men. Cancer Epidemiol 2010;34:534-41.
11. Cai QC, Yu ED, Xiao Y, et al. Derivation and validation of a prediction rule for estimating advanced colorectal neoplasm risk in average-risk Chinese. Am J Epidemiol 2012;175:584-93.
12. Selvachandran SN, Hodder RJ, Ballal MS, et al. Prediction of colorectal cancer by a patient consultation questionnaire and scoring system: a prospective study. Lancet 2002;360:278-83.
13. Adelstein BA, Macaskill P, Turner RM, et al. The value of age and medical history for predicting colorectal cancer and adenomas in people referred for colonoscopy. BMC gastroenterol 2011;11:97.
14. Vega P, Valentin F, Cubiella J. Colorectal cancer diagnosis: Pitfalls and opportunities. World J Gastrointest Oncol 2015;7:422-33.
15. Goldstein BA, Navar AM, Pencina MJ, et al. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017;24:198-208.
16. Jones R, Latinovic R, Charlton J, et al. Alarm symptoms in early diagnosis of cancer in primary care: cohort study using General Practice Research Database. BMJ 2007;334:1040.
17. Denny JC, Peterson JF, Choma NN, et al. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. J Am Med Inform Assoc 2010;17:383-8.
18. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. Ai Mag 1996;17:37-54.
19. Hastie T, Tibshirani R, Friedman J. Unsupervised learning. In: The Elements of Statistical Learning. New York: Springer, 2009:485-585.
20. Koh HC, Tan G. Data mining applications in healthcare. J Healthc Inf Manag 2005;19:64-72.
21. Kavakiotis I, Tsave O, Salifoglou A, et al. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 2017;15:104-16.
22. Senders JT, Staples PC, Karhade AV, et al. Machine learning and neurosurgical outcome prediction: A systematic review. World Neurosurg 2017;109:476-86.
23. Ehrenstein V, Nielsen H, Pedersen AB, et al. Clinical epidemiology in the era of big data: new opportunities, familiar challenges. Clin Epidemiol 2017;9:245-50.
24. Bibault JE, Giraud P, Burgun A. Big Data and machine learning in radiation oncology: State of the art and future prospects. Cancer Lett 2016;382:110-7.
25. Kinar Y, Kalkstein N, Akiva P, et al. Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective

study. J Am Med Inform Assoc 2016;23:879-90.

26. Hoogendoorn M, Szolovits P, Moons LMG, et al. Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer. Artif Intell Med 2016;69:53-61.
27. Kop R, Hoogendoorn M, Teije AT, et al. Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records. Comput Biol Med 2016;76:30-8.
28. Kinar Y, Akiva P, Choman E, et al. Performance analysis of a machine learning flagging system used to identify a group of individuals at a high risk for colorectal cancer. PLoS One 2017;12:e0171759.
29. Kop R, Hoogendoorn M, Moons L, et al. On the Advantage of Using Dedicated Data Mining Techniques to Predict Colorectal. Artificial Intelligence in Medicine 2015;9105:133-42.
30. Birks J, Bankhead C, Holt TA, et al. Evaluation of a prediction model for colorectal cancer: retrospective analysis of 2.5 million patient records. Cancer Med 2017;6:2453-60.
31. Hornbrook MC, Goshen R, Choman E, et al. Early colorectal cancer detected by machine learning model using gender, age, and complete blood count data. Dig Dis Sci 2017;62:2719-27.
32. Kourou K, Exarchos TP, Exarchos KP, et al. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2014;15:8-17.
33. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform 2007;2:59-77.
34. Abbod MF, Catto JW, Linkens DA, et al. Application of artificial intelligence to the management of urological cancer. J Urol 2007;178:1150-6.
35. Kang J, Schwartz R, Flickinger J, et al. Machine learning approaches for predicting radiation therapy outcomes: a clinician's perspective. Int J Radiat Oncol Biol Phys 2015;93:1127-35.

Cite this article as: Liang H, Yang L, Tao L, Shi L, Yang W, Bai J, Zheng D, Wang N, Ji J. Data mining-based model and risk prediction of colorectal cancer by using secondary health data: A systematic review. Chin J Cancer Res 2020;32(2):242-251. doi: 10.21147/j.issn.1000-9604.2020.02.11