ORIGINAL ARTICLE

A systematic evaluation of big data-driven colorectal cancer studies

Eslam Bani Mohammad[1], Muayyad Ahmad[2]

[1]Department of Applied Science/Nursing, Al-Balqa Applied University, [2]Clinical Nursing Department; School of Nursing; University of Jordan

ABSTRACT

Aim To assess machine-learning models, their methodological quality, compare their performance, and highlight their limitations.

Methods The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) recommendations were applied. Electronic databases Science Direct, MEDLINE through (PubMed, Google Scholar), EBSCO, ERIC, and CINAHL were searched for the period of January 2016 to September 2023. Using a pre-designed data extraction sheet, the review data were extracted. Big data, risk assessment, colorectal cancer, and artificial intelligence were the main terms.

Results Fifteen studies were included. A total of 3,057,329 colorectal cancer (CRC) health records, including those of adult patients older than 18, were used to generate the results. The curve's area under the curve ranged from 0.704 to 0.976. Logistic regression, random forests, and colon flag were often employed techniques. Overall, these trials provide a considerable and accurate CRC risk prediction.

Conclusion An up-to-date summary of recent research on the use of big data in CRC prediction was given. Future research can be facilitated by the review's identification of gaps in the literature. Missing data, a lack of external validation, and the diversity of machine learning algorithms are the current obstacles. Despite having a sound mathematical definition, area under the curve application depends on the modelling context.

Key words: artificial intelligence, adults, oncology, review, systematic

Corresponding author:
Muayyad Ahmad
Clinical Nursing Department,
School of Nursing,
University of Jordan
Queen Rania St., Amman 11942
Jordan
E-mail: mma4@ju.edu.jo; mma4jo@yahoo.com
Eslam Bani Mohammad ORCID ID:
https://orcid.org/0000-0003-3569-7875

Original submission:
12 October 2023;
Revised submission:
26 November 2023;
Accepted:
10 December 2023
doi: 10.17392/1684-23

Med Glas (Zenica) 2024; 21(1):63-77

INTRODUCTION

Colorectal cancer (CRC) is a diverse illness caused by molecular changes that result in tumour initiation, development, and invasiveness in the colorectal portion of the gastrointestinal tract (1). Colorectal cancer is the third deadliest cancer in the world after lung cancer and breast cancer (2). In the years 2015 to 2020, around 5.25 million people globally were living with CRC, and approximately 0.94 million deaths attributable by CRC will occur worldwide in 2020 (3). According to the American Cancer Society (ACS) estimates, approximately 153,020 individuals will be diagnosed with CRC and 52, 550 will die from the disease in Untied State (4). In addition, the World Health Organization (WHO) urged the prioritization of cancer care, particularly CRC because it is present in all nations regardless of poverty (5). Globally, it is essential to employ future solutions for the CRC burden control (6).

Artificial intelligence has improved data utilization for researches that is needed to reduce mortality and morbidity of CRC (7). Big data in healthcare pertains to extensive volumes of health-related information gathered from diverse origins (8). Contrary to conventional healthcare data, big data includes non-clinical information like wearables, patient-generated data, social media posts, environmental data, and non-clinical data from electronic health records (EHRs), medical imaging, and genetic sequencing. Using big data in healthcare has the potential to significantly enhance service provision, patient outcomes, and the overall efficacy of the medical industry (9). Data privacy, security, the ethical use of information, and the necessity for effective data governance structures are a few of the obstacles it poses (10). Big data, that can be analysed by artificial intellegence (AI), can generate new knowledge, support clinical decision-making, and develop treatment recommendations (11). Big data facilitates proactive care planning by developing risk prediction models, which predict performance status, treatments, and severe symptoms for cancer patients (12,13). Furthermore, big data can improve nursing information systems using clinical decision system supportive tools, saving clinical documentation, and maps nursing documentation (9).

Using big data for building predictive model was effective screening tool for early intervention, prevention, and early prediction of CRC risks (14,15). Big data has the promise to improve cancer patient care by having more diverse information about health-related issue, social, and economic issues (16). Furthermore, big data analysis can extract knowledge and value from complex dynamic environment with automatic and accurate detection of variables comparing to traditional data techniques which is less efficient (17). However, predictive modelling has been used widely with large-scale data to forecast a hidden pattern of information which helps in predicting future outcomes (18).

The ability of machine learning (ML) algorithms, a subset of AI, to manage the volume and heterogeneity of big data to develop predictive models that provide insight about the progression, risk factors, and management of diseases, has been particularly helpful (19). Big data based on ML for analysis, which gives meaningful outcomes for patients, identifies patients' needs and improves their quality of life (20).

Although many studies have evaluated the effectiveness of machine learning algorithms in predicting the risk of CRC, the methodological quality of existing studies remains unclear. The purposes of this review were to identify machine-learning models, evaluate their methodological quality, compare their performance, and identify their limitations. This review thus helps in identifying research gaps for improving future studies and clinical practice. Furthermore, the findings from this systematic review help to understand the utility of machine learning models for the prediction of CRC.

MATERIALS AND METHODS

Materials and study design

An electronic search was performed using the following online databases: Science Direct, Medline through (PubMed), EBSCO, ERIC (Education Resources Information Centre) and Cumulative Index to Nursing & Allied Health Literature (CINAHL) for the period January 2016-September 2023. The key words were: machine learning, big data, risk prediction, colorectal cancer, area under the curve (AUC), model and artificial intelligence.

The systematic review was conducted and reported according to the Preferred Reporting Items

for Systematic Reviews and Meta-Analysis (PRISMA) guidelines (21).

The study was approved the Research and Ethics Committee at the School of Nursing, University of Jordan.

Two independent authors were responsible for removing the duplicates, screening titles and abstracts, and analysing the full content of the studies in accordance with inclusion criteria. The inclusion criteria were studies published in the English language, articles related to colorectal or colon cancer, use of ML within the prediction, full text articles, have outcome CRC risk and survival predication, and published between 2016 and 2023. Exclusion criteria were papers of protocols, conference papers, abstracts, posters, and letters to editors and editorials. This electronic search was conducted by two reviewers (MM-A and EHB-M), who screened the titles and abstracts independently. Disagreements and clarifications were discussed between the reviewers.

To determine the eligibility criteria, the Population, Intervention, Control, Outcome, Study (PICOS) criteria (22) were used to determine studies that would be included in the review (Table 1).

Table 1. Population, Intervention, Control, Outcome, Study (PICOS) criteria for inclusion

P	Population	Patients with CRC
I	Intervention	Build predictive model based on data
C	Comparison	Comparison between patients with CRC and patients with non-CRC
O	Outcomes	AUC, odds ratio, confidence interval
S	Study	cohort design, case-controlled retrospective, cross-sectional retrospective

CRC, colorectal cancer; AUC, area under the curve

Methods

Data from the included studies were extracted by using the standardized extraction form developed by Joanna Briggs Institute (23) including: first author, year of publication, country, characteristics of the participants, area under the curve (AUC), AI model type, study design, the period of study, number of features, and the aim of study. The included articles were discussed within the research team to divide them based on the type of the predictive models. We used the extracted data to determine key components, which include model performance, CRC predictors, and the number of participants in the included studies, as well as the key components of the included interventions. To reduce bias, data extraction was performed by two researchers independently and the type of extracted data was specified in study protocol.

Statistical analysis

The performance metric used to assess the CRC predictive model's discriminatory power was the area under the curve (AUC) of the receiver-operator characteristic (ROC) plot, known as the concordance index.

The risk of bias assessment used was Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies (PROBAST) (22). Two independent researchers (MM-A and EHB-M) assessed the risk of bias. PROBAST is a structured tool comprising 20 questions for assessing the risk of bias and applicability across the four domains of outcome, participants, analysis, and predictors (22).

RESULTS

After an initial database search, a total of 10,800 studies were identified. After title, abstract, duplication removal, and full content screening, 15 studies were finally included in the review (15, 24-37) (Figure 1).

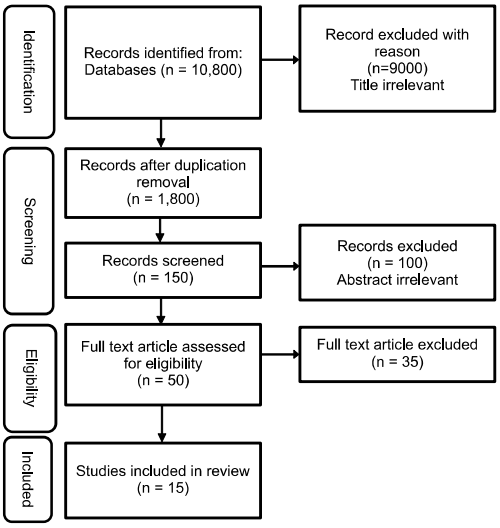

Figure 1. PRISMA Flow diagram of literature search

Of all 15 included studies, three used cohort design, one used case-controlled retrospective design, four were cross-sectional retrospective and seven used retrospective design. Overall, these 15 included studies reported the results from

3,057,329 CRC health records including adult CRC patients above18 years of age. In terms of country of the study, United States (n=4), United Kingdom (n=2), and China (n=3) were represented by more than one study, while South Korea, Japan, Netherland, Brazil, Austral, Canada, and Taiwan had one study each.

In terms of AI models, 25 models were used, which include Random Forest (RF), Naive Bayes (NB), XGBoost, Colon Flag, One-Class Embedding Classifier (OCEC), Neural Network (NN), Logistic Regression (LR), Light Gradient Boosting Machine (LGBM), Machine Learning (ML), nomogram, artificial intelligence (AI), artificial neural network (ANN), Support-Vector Machine (SVM), Linear Discriminant Analysis (LDA), Decision Tree (DT), K-neighbours classifier (KN), Gaussian Naive Bayes (GNB), Multinomial Naive Bayes (MNB), C-support vector classifier (SVC), Stochastic Gradient Descent Classifier (SGD), Gradient Boosting Classifier (GB), Extreme Gradient Boost Classifier (XGB), Decision Tree - Discrete and Sparse Multimodal Instance Learning (DT-DSMIL), Multi-Omics Multi-Cohort Assessment (MOMA), joint models (Cox sub-mode). Seven studies compared between more than one model (Table 2).

Model performance. The model performance was assessed by a range of measures and was not consistently reported across the included studies. The most frequently reported measure was AUC accuracy, odd ratio, and sensitivity. The confidence intervals for model performance measures were not always reported. Access to supplementary materials was not always obtained.

The outcome was about AUC, odds ratio, and accuracy. The area under the curve varied between 0.704 and 0.976. Each model had a varying degree of accuracy, sensitivity, and specificity. There were statistical differences in the primary metric of AUC between the models. Although AUC is well defined mathematically, the practical use of this metric depends on the modelling context. Furthermore, each model has its benefits and drawbacks that are determined by its utilization based on the nature of data used to analyse the hidden pattern of stored data.

Risk of bias. Risk-of-bias assessment with PROBAST signalling questions in four domains and 20 items: 1.1. Were appropriate data sources used? 1.2. Were all inclusions and exclusions of participants appropriate? 2.1. Were predictors defined and assessed in a similar way for all participants? 2.2. Were predictor assessments made without knowledge of outcome data? 2.3. Are all predictors available at the time the model is intended to be used? 3.1. Was the outcome determined appropriately? 3.2. Was a prespecified or standard outcome definition used? 3.3. Were predictors excluded from the outcome definition? 3.4. Was the outcome defined and determined in a similar way for all participants? 3.5. Was the outcome determined without knowledge of predictor information? 3.6. Was the time interval between predictor assessment and outcome determination appropriate? 4.1. Was there a reasonable number of participants with the outcome? 4.2. Were continuous and categorical predictors handled appropriately? 4.3. Were all enrolled participants included in the analysis? 4.4. Were participants with missing data handled appropriately? 4.5. Was the selection of predictors based on univariable analysis avoided? 4.6. Were complexities in the data accounted for appropriately? 4.7. Were relevant model performance measures evaluated appropriately? 4.8. Were model overfitting and optimism in model performance accounted for? 4.9. Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?

The first domain for the PROBAST risk of bias assessment tool was used to assess the patient domain; different studies had small sample size (26, 33, 35). A study with 471 patients was judged as having a high risk of bias due to small sample size and lack of external validation (33). The same was with another study with only 94 patients (26). In predictors domain, the predictors were defined and assessed in a similar way for all participants in all studies. Similarly, the outcomes were determined appropriately and defined in a similar way for all participants in the selected studies. The analysis of predictor selection also influenced this domain. Most predictors were categorical, either nominal or ordinal, which may limit discriminative ability and set limitations in the interpretability and accuracy of data (31). All validation studies were ranked as having a high risk of bias in this domain because of the selection bias, and the lack of external validation (33).

Table 2. Application of Artificial Intelligence (AI) in colorectal cancer (CRC) prediction, diagnosis and treatment

Country (reference number)	Study Design (Period)	No of patients (age range)	Features (number of features)	Model type	Outcome	Recommendation	Limitations
UK (25)	Cross-sectional retrospective (Jan 2000 – April 2015)	2,550,119 (25,430 CRC) (>40 years)	Age, gender, full blood count (23)	RF	AUC=0.776 95% CI: 0.771, 0.781 PPV at 99.5% specificity 8.8% with NPV 99.6%	Further research of the risk algorithm in the routine care setting	Family history and MSI status are not included Limited sample size
Brazil (26)	Retrospective (2000 – 2021)	31,916 (unknown)	Age, gender, residence, category of care, previous diagnosis and treatment, stage, code of combination of treatments, difference in days between the consultation and diagnosis dates, difference in days between consultation and treatment dates, difference in days between treatment and diagnosis dates, year of diagnosis, education level (25)	RF, NB, XGBoost models	RF: Accuracy: 77.2%, AUC= 0.85. NB: Accuracy 50.0%, AUC= 0.78. XGBoost models: Accuracy 77.7%, AUC = 0.86	More studies are needed to compare the performance of AI models with the most common statistical models for prediction	Used models, which do not allow the inclusion of patients lost to follow-up
China (27)	Cohort (April 2020 - January 2021)	252 (59 HC, 99 AA, 94 CRC) (21-85)	Age, gender, smoking, stage, BMI, Differentiation, tumour location, cancer family history, Serum Igg (11)	AA CRC	AUC = 0.847 AUC = 0.844	Need a larger and more comprehensive study cohort for validation	Limited sample size
Canada (28)	Cross-sectional retrospective (Jan 2013 – June 2015)	17,676 (PCL 1014, CRC 60) (50-75)	N= 23 Age, gender, full blood count (23)	Colon Flag	OR: CRC 5.1, PCL 2.0	Increasing the number of risk factors	Detailed information on a patient's personal or family history of CRC or polyps was not available/not complete characterization of all sessile serrated polyps
USA (29)	Case-control retrospective (Jan 2000 – Dec 2013)	17,095 (900 CRC) (40–89 years)	Age, gender, full blood count (23)	RF	AUC= 0.80 OR=34.7 (28.9– 40.4)	Identifying characteristics predictive of undiagnosed cancer risks such as age, gender, last BMI, and length of time since last physician visit.	Unknown
Netherlands (30)	Cross-sectional retrospective (Jul 2006 – Dec 2011)	90,000 (588 CRC) (at least 30 years)	Consultation notes, demographics, medical history, medications, blood parameters, referrals (50)	LR	AUC= 0.896	Study other diseases, datasets from other GP information systems	The language used: Dutch
South Korea (31)	Cross-sectional retrospective (Unknown)	1628,522 (2845 CRC)	Demographics, Medical History, Family history, Questionnaire results, medical records (39)	OCEC NN LR RF LGBM	AUC= 0.796 AUC= 0.791 AUC= 0.769 AUC= 0.704 AUC= 0.794	Comparative effectiveness research is needed	Depending on the country, the performance of the developed algorithms can differ
UK (31)	Retrospective (2013)	528,060 (59-79 years)	Age Surgery (30)	LR RF XGBoost	AUC= 0.730 AUC =0.757 AUC =0.748 AUC ML range =0.748 to 0.757	Future research needed if interventions derived from ML prediction lead to significant savings.	Interpretability and accuracy of data - most predictors are categorical (nominal or ordinal), limiting discriminative ability Lacking clinical details
China (32)	Retrospective (2010 - 2016)	57,835 CRC (Above 18 years)	N=22 age at diagnosis, gender, marital status, race/ethnicity, insurance status, histology type, primary tumour site, grade, tumour size, TNM stage, CEA level, surgery primary site, surgery metastasis site, survival time, survival status, gene, and causes of death.	OS nomogram CSS nomogram	C-index = 0.67 (95% CI 0.662–0.678) for training group 0.658 (95% CI 0.646–0.670) for validation groups C-index = 0.692 (95% CI 0.682–0.702) for training group 0.646 (95% CI 0.622–0.670) for validation groups	Use prospective design	Further verification through prospective studies SEER database does not contain some important information
Japan (33)	Retrospective (2004 - 2015)	471 58–76	N=27 Age, gender, stage, chemotherapy, recurrence number, tumour size, lymph nodes, pathology, histology, CT, tumour location, tumour markers: CEA and CA19-9.	AI model	AUC=0.7245 [95% confidence interval (CI) 0.6707–0.7783	Perform an external validation	Selection bias Old pathological slides Small sample size

Table 2. Application of Artificial Intelligence (AI) in colorectal cancer (CRC) prediction, diagnosis and treatment (continued)

USA (15)	Retrospective NHIS (1997 to 2017) PLCO (1993 and July 2001)	1,077,653 (18-75 years)	N=22 Age, first-degree relatives, BMI, screening, NSAID use, diet, inflammatory bowel disease, alcohol/tobacco use, physical activity, family history, obesity, screening, diet (multivitamin, alcohol, vegetables, and red meat consumption), height, physical activity, pharmaceuticals	ANN LR NB RF SVM LDA	ROC curves Family history available ANN=0.75 (0.03) LR=0.63 (0.11) NB=0.69 (0.14) RF=0.59 (0.04) SVM=0.60 (0.04) DT=0.58 (0.03) LDA=0.50 (0.03) Family history unavailable ANN=0.70 (0.02) LR=0.60 (0.05) NB=0.71 (0.14) RF=0.57 (0.08) SVM=0.61 (0.04) DT=0.55 (0.06) LDA=0.50 (0.01)	Recording these strong predictors more regularly	Missing data
Australia (34)	Retrospective (1994 – 2010)	1236 mean age 67.1- 71.2	N=118 Age, gender, stage, metastasis, recurrence within 5 years, tumour length, width, depth, lymph nodes, perineural indicator, lymphovascular invasion, chemotherapy	LR, DT, RF, KN, GNB, MNB, SVC, SGD, GB, LGBM, and XGB	The LR algorithm was the top model in AUC= 0.850 (0.014 SD, 95% CI 0.840-0.860) for the 1-year AUC= 0.872 (0.014 SD, 95% CI 0.861-0.882) for the 5-year survival prediction. Using only the 5 most important predictor variables, the corresponding values are 0.793 (0.020 SD, 95% CI 0.778-0.807) and 0.794 (0.011 SD, 95% CI 0.785-0.802).	Potential incorporation of the developed model in to clinical practice needs to be further investigated	Some of the variables such as the chemotherapy cycles administered and the type of radiation therapy, were not used. A small number of patients For the short-term survival predictions, the dataset was unbalanced
China (35)	Retrospective (January 2019 and January 2021)	357	Lymph nodes	DT-DSMIL	Single lymph node classification, AUC= 0.976 (95% CI: 0.9607–0.9891) Lymph nodes with micro-metastasis AUC= 0.9816 (95% CI: 0.9659–0.9935). Macro-metastasis, AUC= 0.9902 (95% CI: 0.9787–0.9983). Accuracy= 95.3%	Follow up researches, developing the diagnostic system, and classifying single lymph nodes	Model memory inefficiency
USA Taiwan (36)	Cohort study (2022)	1888 (Unknown)	N=2048 demographic compositions, pathology images Stage, gene	MOMA	C-index = 0.74	NA	Unclear inclusion criteria
UK (37)	Cohort study (January 2000 - January 2014)	1,327,996 (at least 40 years)	N=5 (Age, gender Hb, MCV, Platelets)	Joint models (Cox sub-model) Colon Flag	AUC (males) Joint model: 0.751 Colon Flag: 0.762 AUC (females) Joint model: 0.763 Colon Flag: 0.761 PPV=0.61-1.62 NPV=99.68–99.86%	-larger staging subgroups -further risk factors -include other common types of blood tests	Only blood testing. False positives results Small Sample size for some subgroups. Tumour staging was missing.

N, number; CRC, colorectal cancer; RF, Random Forest; AUC, area under curve; CI, Confidence Interval; PPV, Positive Predictive Value; NB, naive Bayes; XGB, Extreme Gradient Boost Classifier; HC, Healthy Control; AA, Advanced Adenomas; BMI, Body Mass Index; PCL, Pre-Cancerous Lesion; OR, Odd Ratio; LR ,Logistic Regression; GB, Gradient Boosting Classifier; OCEC , One-Class Embedding Classifier; NN, neural network; LGBM, Light Gradient Boosting Machine; ML, machine learning; N, Number of Features; OS, Overall Survival; CSS ,Cancer-Specific Survival; C-index , Consistency Index; CT, Computed Tomography; CEA, Carcinoembryonic Antigen; CA19-9, Carbohydrate Antigen; NHIS , National Health Interview; PLCO, Pancreatic, Lung; Colorectal; NSAID, nonsteroidal anti-inflammatory drug; ANN, artificial neural network; SVM , Support-Vector Machine; LDA , Linear Discriminant Analysis; ROC, receiver operating characteristics; SVM, Support-Vector Machine; DT, Decision Tree; LDA, Linear Discriminant Analysis; KN, K-neighbours classifier; GNB, Gaussian Naive Bayes; MNB, multinomial naive Bayes; SVC, C-support vector classifier; DT-DSMIL, Decision Tree - Discrete and Sparse Multimodal Instance Learning; MOMA, Multi-Omics Multi-Cohort Assessment; Hb, Haemoglobin; MCV, Mean Corpuscular Volume; NPV, Negative Predictive Value

Furthermore, the relevant model performance measures were evaluated appropriately (Table 3).

Model comparison of ML algorithm's role in predicting CRC. Most commonly used ML methods included random forests (n=7), logistic regression (n=5), and Colon Flag (n=2). This study used data mining approaches to predict CRC.

Several metrics features identified the most accurate models to predict CRC, namely, sensitivity, specificity, accuracy, AUC, and receiver operating characteristics (ROC) (24, 28, 37). When model accuracy (the proportion of totally used datasets that are correctly predicted out of the total instances) increased, the quality of interpretation increased too (38, 39). The AUC measures the overall performance of the model; it captures the discriminatory ability of a model, estimates the probability and the performance over a series of thresholds (40). The AUC value was within range (0.5-1.0), the maximum value represented perfect model and the minimum value represented the performance of random model (41).

Colon Flag is a machine learning algorithm that uses basic patient information and CBC to identify the elevated risk of CRC (42). A five years' prospective cohort study used AI to develop predictive model for CRC based on full blood count (FBC), these models were joint models and Colon Flag mode.

The total sample in Wang et al. study (42) was 1,327,996 patients: The developed models could predict the risk of CRC based on blood test that included full blood count (FBC) results, lowering haemoglobin concentration, lowering mean corpuscular volume concentration, and rising in platelet measurement increase risk of CRC incidence. The AUC of joint model was 0.751 for males and 0.763 for females, AUC of Colon Flag was 0.762 for males and 0.761 for females, meaning the models could discriminate high-risk from low-risk patients using only earlier data prior to two years before diagnosis (37). Furthermore, according to Hornbrook, Goshen, Choman, O'Keeffe-Rosetti, Kinar, Liles and Rust (28) Colon Flag model (AUC=0.80) identified individuals with a higher risk of undiagnosed colorectal cancer at curable stages (0/I/II), predicted colorectal tumours 180–360 days prior to usual clinical diagnosis, and it is more accurate at identifying right-sided CRC. Moreover, the predicted relative risks of the model according to the CRC stage were: 12.1 for carcinoma *in situ* and 16.7 for Stage I, at 99% specificity, and 54.1 for Stage II, 12.1 for Stage III, and 40.4 for Stage IV (28). Furthermore, Colon Flag was an effective model that uses routine blood test results to determine individuals at elevated risk for high risk precancerous polyps and colorectal cancer; in 17,676 individuals who had a screening colonoscopy there were 1,014 (5.7%) with a high risk precancerous lesion (PCL) and 60 (0.3%) had colorectal cancer; the odds ratio for CRC was 5.1 and for PCL it was 2.0 (27).

The Decision. Tree model constructed a flowchart of factors to predict CRC. It easily understood and is desirable in a clinical setting; it constructs the base of the tree, then used less informative variables at higher branches; this mode is known for its interpretability and robustness

Table 3. Prediction model risk of bias assessment tool (PROBAST) assessment of predictive modelling studies

Reference No	Participants		Predictors			Outcomes						Analysis									Overall
	1.1	1.2	2.1	2.2	2.3	3.1	3.2	3.3	3.4	3.5	3.6	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	
(24)																					
(25)																					
(26)																					
(27)																					
(28)																					
(29)																					
(30)																					
(31)																					
(32)																					
(33)																					
(15)																					
(34)																					
(35)																					
(36)																					
(37)																					

Light grey, low; black, high; dark grey, unclear risk of bias

(43). A retrospective study found that the ROC of Decision Tree (DT) was 0.58 (0.03) and it was prone to overfitting; in this study missing data affect prediction model (15).

Tan, Li, Yu, Zhou, Wang, Niu, Li and Li (35) predict CRC metastasis from slide images using Decision Tree - Discrete and Sparse Multimodal Instance Learning (DT-DSMIL) for 357 patients from 2019 to 2021; this model achieves accuracy of 95.3% and AUC= 0.9762 (95% CI: 0.9607–0.9891), meaning that the model could identify the most likely metastases (35). Although this model has the highest AUC, it has memory inefficiency (35).

Random forests (RF) model is a collection of random trees, each tree in a forest down to a terminal node which assigns it a class (44). Moreover, it determines outcome predictions using binary splits on predictor variables by splitting "high" versus "low" values of a predictor related to outcome (45). A retrospective study demonstrated that ROC curves of RF were only 0.59 (0.04), which RF model is prone to lack the transparency and information that DTs have in making their classifications (15). Cross-sectional retrospective study about predicting CRC using FBC data found that the AUC of RF model was 0.776, and most of the predictive power was due to age >40 years (24).

Logistic regression (LR) is a classification model based on the probability of a discrete outcome given an input variable (46, 47). Susič, Syed-Abdul, Dovgan, Jonnagaddala and Gradišek (34) found that LR algorithm has the highest AUC of 0.850 (0.014 SD, 0.840-0.860 95 % CI) for the 1-year, and 0.872 (0.014 SD, 0.861-0.882 95% CI) for the 5-year survival prediction. Using only 5 most important predictor variables, the corresponding values were 0.793 (0.020 SD, 0.778-0.807 95% CI) and 0.794 (0.011 SD, 0.785-0.802 95% CI). A cross-sectional retrospective study predicts CRC from the consultation notes, demographics, medical history, medications, blood parameters, and referrals; the results suggested that AUC=0.896 for LR and the important remark is that the combination of age and gender is highly predictive for CRC (29). Furthermore, a retrospective study that analysed 528,060 CRC patients demonstrated that big data of surgical colon cancer patients can be utilized to build ML models; AUC and CI for the developed model were LR (AUC 0.730, 95% confidence interval - CI: 0.725-0.735), ML algorithms (0.748 and 0.757), the RF model (AUC 0.757, 95% CI: 0.752-0.762), outperforming XGBoost (AUC 0.756, 95% CI: 0.751-0.761), and XGBoost using Synthetic Minority Over-sampling Technique (SMOTE) data (AUC 0.748, 95% CI: 0.743-0.753). However, these methods have the potential to enhance surgical patients with CRC risk prediction (31). Whereas, another study that used ML to predicate CRC survival rate demonstrated that the algorithm with the best survival prediction was XGBoost with more than 77% of accuracy, followed by RF and NB, which had the worst performance among those used; the results of the predictors were around 77% of accuracy, with AUC close to 0.86, and the most important column was the clinical staging in all of them (25).

One model used artificial neural network (ANN), a simplified model that is based on a huge number of interconnected unites. It can evaluate the digital pathology images and demographic data for the diagnosis of CRC cancer (48). According to Nartowt, Hart, Muhammad, Liang, Stark and Deng (15), after comparing ANN, LR, NB, DT, RF, SVM, and LDA models in predicting CRC based on personal health data, the results suggested that the ANN was the best model with expectation-maximization imputation, having a concordance or AUC of 0.70±0.02 sensitivity of 0.63±0.06, and specificity of 0.82±0.04; however, drops in an individual's risk score in response to better personal health habits were used as a non-invasive and cost-effective tool to screen the CRC risk in large populations effectively (15).

Another model used to predict CRC was the Nomogram model that is widely used in tumour-related research, and it predicts the probability of patient's clinical events based on multivariate regression analysis, which can quickly and intuitively predict the prognosis of patients (32). A retrospective study in China, with a sample of 57,835 CRCs, used Nomogram model to identify factors that lead to poorer prognosis, which include newly diagnosed CRC patients with distant metastasis and deliver appropriate treatment; the result was older age, unmarried status, poorly differentiated or undifferentiated grade, larger tumour size, N2 stage, right colon site, more metastatic sites, and elevated carcinoembryonic

antigen (CEA), might increase the CRC risk. In this study, two models were used, overall survival (OS) and cancer-specific survival (CSS), and the reliability and accuracy of the prediction models were assessed using a C index. The C-index of the OS nomogram prediction model in the training and validation groups was 0.67 (95% CI 0.662–0.678) and 0.658 (95% CI 0.646–0.670), the C-index of the CSS nomogram prediction model in the training and validation groups was 0.692 (95% CI 0.682–0.702) and 0.646 (95% CI 0.622–0.670), respectively. The limitations of this study were low level of research evidence and missing data (32).

Multi-omics Multi-cohort Assessment (MOMA) machine learning model aims to predict cancer genomics, proteomics, and important clinical outcomes (36). A cohort study based on histopathology images to predict CRC survival demonstrated that MOMA model successfully predicts patients' progression-free survival outcomes with concordance index C=0.74 (36). It predicted survival of early-stage (stage I and stage II) and stage III colorectal cancer patients; stage IV was excluded due to multiple treatments (36).

Another retrospective study used digital pathological images to build AI model that predicts patients with a high risk of recurrence of stage I–III CRC. The AI model is an inexpensive, non-invasive method that can be implemented using only clinically existing materials. However, the sample size of 471 patients was small and AUC of AI model of 0.7245 with 95% CI (0.6707–0.7783) was found (33).

In a cohort study that used serum IgGN to predict CRC, the total sample was 252, which were classified into three groups: 59 healthy control (HC), 99 advanced adenomas (AA), and 94 CRCs; the results demonstrated that combined index GlycoF (was developed to provide a potential early diagnostic biomarker in discriminating simultaneously) AA (AUC = 0.847) and CRC (AUC = 0.844) from HC; serum IgG N-glycans analysis provided powerful early screening biomarkers that can efficiently differentiate CRC and AA from HC (26).

Our results found six studies that utilized multimodel to predict CRC. For instance, a cross-sectional retrospective study utilized five models that showed acceptable performance level as the following: LR (0.769), RF (0.704), LGBM (0.794), OCEC (0.796), and NN (0.791) (30). Another study including 528,060 patients and ML models, including RF and XGBoost, were built and compared with conventional LR; this study found that building ML models using big data to improve outcome prediction can enhance CRC risk prediction; AUC for LR (AUC 0.730, 95% CI: 0.725-0.735), AUC for ML algorithms was between 0.748 and 0.757, the RF model (AUC 0.757, 95% CI: 0.752-0.762), outperforming XGBoost (AUC 0.756, 95% CI: 0.751-0.761) and XGBoost using Synthetic Minority Over-sampling Technique (SMOTE) data (AUC 0.748, 95% CI: 0.743-0.753). In sum, RF had the highest AUC (31).

DISCUSSION

There are several studies that applied ML algorithms to predict and determine the recurrence of the CRC disease (49, 50). Our study identifies trends in the use of predictive analytics and big data for CRC.

Machine learning algorithms and big data analysis techniques

Machine learning algorithms have been applied in many medical fields, which help health care providers to make optimal decision, for example, it can be used for early disease prediction such as the prediction of chronic diseases (51), heart disease (49), and CRC (31). There are many applications of AI in fighting against CRC, which are based on developing predictive models for diagnosing disease by extracting big data from endoscopes, genetics, CT, MRI, and pathological test. However, how treatment approaches to CRC can be enhanced when applying AI is still under discussion (52). Using a large registry of surgical colon cancer patients to build ML models, can improve outcome prediction which enhance risk prediction, leading to improved strategies to mitigate those risks (31). But, it is essential to consider the greatest weakness: the data accuracy of big data and the discriminative ability that is related to the fact that the majority of predictors are categorical, either nominal or ordinal, and the lack of clinical details with predictive power, such as estimated blood loss or the presence of a disease (31).

When comparing big data to traditional models, ML models require a large amount of data and

large number of observations, the performance of predictive model is optimized at the expense of interpretability (31). According to Nwosu, Collins and Mason (53), big data analysis aims to support the process of improving the quality of service, reducing medical errors, promoting consultation, and providing answers for health care inquires.

Depending on quality of data, the performance of the developed algorithms can differ (30). For the prediction of the recurrence of breast cancer using ML algorithms in 1475 patient records, the extracted features included tumour grade, molecular subtype, cancer focality, menopause, age, and greatest dimension of primary tumour as predictors of breast cancer recurrence using different machine learning algorithms ((49, 54).

Machine learning algorithms and big data analysis benefits and drawbacks

Big data utilization in health care had many advantages such as improving safety and quality, designing new standardized protocols and evidence based practice in healthcare (55). Big data analysis was effective in utilizing a large amount of data that were derived from public health surveillance to identify drugs, diagnostic and prognostic features (56). Moreover, it emphasises the needs for developing adequate healthcare services throughout estimating the prevalence of life-limiting diseases (57, 58).

On the one hand, big data facilitates proactive care planning by developing risk prediction models with their prediction model of poor performance status and severe symptoms (34). The availability of big data provides many opportunities to find out valuable knowledge and applications for policy and practice such as identify those at risk of adverse outcomes and inappropriate treatment (59). Currently, the use of big data analysis has been increased in health care, which is gathered from the healthcare system for decision making and improving quality of health care (60). Also, the limitations that slowed the adoption of big data analytics in healthcare include privacy concerns, limited resources, security risks, and the difficulty of big data analysis (60). The limitations of a study that used big data analysis are incomplete data entry, the sample was not enough, and data have missing values at ministry of health database (9); in addition, limited sample size and missing data (15, 24-26, 33, 34, 37), as well as unclear inclusion criteria (36).

Some studies encountered challenges related to potential bias due to the absence of comprehensive information and missing data regarding colorectal cancer (27, 31, 32). Some of the variables such as the chemotherapy cycles administered and the type of radiation therapy, were not used in Susic et al. study (34). Only blood testing, false positives results, tumour staging were missing in Virdee et al. study (37). Another problem with using big data to improve patient care is that not enough information is gathered. For example, data on family support, patient experience, and death rates are not collected well enough (61). Furthermore, political and economic issues, such as the refusal of organizations for the need of technical and adaptive changes and inadequate engagement among political leaders, healthcare providers, and the technological industry were barriers for adopting advanced electronic health system, which is necessary to create big data sources (62).

Also, the performance of the developed algorithms can be different depending on the country (30). Another issue is the predictors, categorical either nominal or ordinal, which may limit discriminative ability of the predictive model (31).

The challenges faced by individuals and organizations in the process of utilizing big data in healthcare are data privacy, security, ownership, expertise requirements, clinical data linkage, data storage and processing issues (53, 63). Furthermore, there are several challenges in predicting the CRC using AI that include medical and technical challenges. The technical challenges include patients' privacy and reliability of the models; medical challenges such as lack of awareness and knowledge about AI (64).

Big data is important for healthcare providers. According to a systematic review that investigated eight papers between 2015 and 2018, big data is essential to prepare nurses and improve patient outcomes by improving safety, quality and outcomes; it provides a holistic view of a patient's health status. Big data should be adopted by nurses as it is essential for their development in researches, practice, and education (65).

Ethical implications and machine learning models

Using patient data in healthcare to train machine learning models has a number of ethical ramifications that must be thoroughly evaluated. It is vital to ensure that patient information is de-identified and anonymized to maintain privacy. By doing so, every personal identity is eliminated from the dataset.

Although ethical concerns may be less apparent in systematic review studies in comparison to research involving human subjects in primary settings, they continue to be substantial. The researchers in this systematic review have incorporated data from previous studies while taking into account various ethical considerations.

We tried to keep the data from the original studies accurate. This means giving accurate results and not picking and choosing which results to show to support a certain conclusion. In the presentation process each step used to collect and combine data is kept clear and can be repeated so that other people can check the results or do the review again. Furthermore, we tried our best to make sure that the reviewed studies had no ethical problems. The privacy and anonymity of the people who took part in the original research are maintained in this review. Ethical considerations have been meticulously managed throughout this systematic review; the review process is transparent, unbiased, and in accordance with the ethical norms of the included research. In addition, a potential consequence of an excessive dependence on machine learning techniques in healthcare is the erosion of clinical abilities and a diminished value placed on human judgment.

Colorectal cancer risk prediction

Colorectal cancer risk prediction was linked to certain demographic and characteristics such as age, gender, education, employment, date of first diagnosis, location, stage, treatment modality, route of diagnosis comorbidities, the primary site, tumour size, histological type, and number of lymph nodes affected (15, 29, 30, 32-34). The most important variables for five-year prediction were the number of residual, distant metastasis, stage, probable recurrence, and tumour length, whereas biomarkers do not appear among the top 20 most important ones (34).

Understanding the contribution of modifiable and non-modifiable risk factors to CRC burden and their trends over time is important to provide better care for those patients. The non-modifiable risk factors are age, gender, and hereditary factors (32, 33, 66). These factors were utilized in big data analysis to predict CRC (25). Other modifiable risk factors for developing CRC are alcohol intake, dietary patterns (diet with processed foods, diets low in fruit and vegetables) (67). In this review one study analysed diet risk factors such as multivitamin, alcohol, vegetables, and red meat consumption (15). The modifiable factors such as environmental and lifestyle factors, obesity, physical activity, smoking, high salt and red meat increased the risk of colorectal cancer (15, 66).

Colorectal cancer survival rates are associated with age at diagnosis, and patient characteristics, such as race, ethnicity and socioeconomic status according the American Cancer Society (ACS) (68). The strongest predictor of cancer mortality was advanced age (69). The age is an important predictor for CRC occurrence (34). Low survival rates are associated with older age (50 years or older), poor differentiation and right-sided cancers (70). Colorectal cancer survival rates among Jordanian patients for five and ten years were 58.2% and 51.8%, respectively (70). Furthermore, CRC survival is better when colorectal cancer is diagnosed while being still at localized stage; the 5-year survival rate for localized CRC was 72.1%, regional 53.8%, and for distant stage 22.6% (70). Low survival rates are associated with older age (50 years or older), poor differentiation, right-sided cancers, and advanced cancer stage (70).

In order to achieve meaningful utilization of big data to improve cancer patient care, it is essential to collect adequate information such as mortality data, risk predictor, diagnosis, and treatment modality, which facilitate early diagnosis to reduce mortality rates and to determine the effective therapeutic interventions (51).

Staging of colorectal cancer (CRC)

Our results showed that the stage of CRC enabled early detection and early relapse prediction of CRC (25, 26, 32-34). Although tumour staging is important, it was missing which cause the limitation in a cohort study aimed at the development and validation of a dynamic prediction model

(37). The most important predictor of the CRC survival is stage at diagnosis (68): the 5-year survival rate is 71% and 14% for those diagnosed with regional and distant stages, respectively, but it increased to 90% for 39% of patients diagnosed with localized-stage disease (68). Furthermore, CRC survival is better when colorectal cancer is diagnosed while being still at a localized stage; the 5-year survival rate for localized CRC was 72.1%, 53.8% for regional, and 22.6% for the distant metastases stage (70).

Types of colorectal cancer (CRC)

The distribution of CRC by topography includes cecum, appendix, ascending colon, hepatic flexure of colon, transverse colon, splenic flexure of colon, sigmoid colon, descending colon, rectosigmoid junction, rectum, anus and anal canal (71, 72). Tumour site was used to predict distant metastasis pattern, prognostic prediction model of CRC patients, prediction of recurrence based on big data analysis (32, 33). In Jordan, rectal cancer was the most common site (22.6%), other types were sigmoid colon (21.7%), unknown sites (12.3%), recto-sigmoid (9.9 %), and cecum (7.9%), respectively (70). According to the American Cancer Society (ACS (68), males have rectal cancer (60%) more often than colon cancer (20%), and females are more likely to develop adenomas in the proximal colon than men.

Histology type was used to predict distant metastasis pattern and prognostic prediction model of CRC patients using big data mining (32). Similarly, histology was utilized in artificial intelligence-based prediction of CRC recurrence after curative resection (33). Histological types were divided into primary adenocarcinoma, which constituted the majority of cases, and others, which included lymphomas, carcinoid tumours, and gastrointestinal stromal tumours (73). Adenocarcinoma was the commonest morphology (85%) followed by mucinous (colloid) adenocarcinoma (8.4%), other carcinomas (4.6%), signet ring adenocarcinoma (0.9%), carcinoids (0.7%), adenocarcinoma in adenomatous polyps (0.2%), and adenocarcinoma in villous adenoma (0.2%) (70). Many AI studies used laboratory tests to develop and evaluate prediction model for colorectal cancer such as CBC (24, 26-29, 32).

Colorectal cancer screening

Factors that decrease the incidence of cancer are early detection and screening, which help in diagnosing CRC at an early stage and chances of survival become better (5). Screening strategies are needed for early detection of colon adenomas and colorectal cancer (15, 70). The American Cancer Society (ACS) recommends the following screening tests that can find CRC: annual faecal immunochemical test (FIT), annual faecal occult blood test (FOBT) or Stool DNA test every 3 years (74). The three categories commonly used as prediction and screening methods for colorectal cancer in clinical practice include: stool test, imaging examination, gut microbiome and colonoscopy (75).

Imaging examination such as spiral CT and MRI is fast, but it is difficult to detect early intestinal lesions (75). However, CT scan results can be used for predicting survival of colorectal cancer patients using RF, NB, XGBoost models (33). While colonoscopy can directly observe colorectal lesions, the disadvantages of colonoscopy examination include the need to empty the bowel, thus it may cause electrolyte imbalance (75). Furthermore, colonoscopy is an invasive examination, which may cause intestinal perforation, intestinal infection, and acute peritonitis. That means that the existing screening methods suffer from shortages in accuracy, feasibility sensitivity, and patient experience (75). For that, AI model prediction could be safer than routine screening methods.

A cohort study demonstrated that growing tumours often cause changes in blood test results that may help in earlier detection; blood test includes complete blood count (CBC) results of lowering haemoglobin concentration, lowering mean corpuscular volume concentration, and a rise in platelet measurement concentration (37).

Colorectal cancer predictive models can be integrated into routine practice effectively as they can be used as a cost-effective and non-invasive method to screen the CRC risk in large populations, which are based on personal health data only. For instance, the ANN model, which classifies participants into low-, medium-, and high CRC-risk groups, can be used as an effective screening tool for early intervention and prevention of CRC in routine practice (15).

It is recommended to compare the effectiveness of the predictive model between CRC and other types of cancer. Avoiding data missing and incomplete data entry is essential to test the predictive model, ML models require a large amount of data in comparison to traditional models. With a large number of observations, the predictive performance is optimized. Further verification through prospective studies is recommended. Furthermore, performing external validation for models is essential.

In conclusion, colorectal cancer is the leading cause of death worldwide. For that, it is essential for early detection and medical intervention to reduce the mortality rate. Better utilization of big data stored in medical health records is important for health planning to prevent CRC by identifying attributing factors to colorectal cancer using predictive model which may decrease the proportion of late diagnoses. Machine learning algorithms can predict specific attributes for deferent types of cancer including CRC. There are different models used to predict CRC and it is important to assess accuracy, sensitivity and specificity of each model. Furthermore, big data analysis is important to improve patient outcome, quality of care, and safety, to facilitate decision making regarding treatment options, decrease mortality rates, improve quality of care, and reduce the financial burden on health care institutions.

FUNDING

No specific funding was received for this study.

TRANSPARENCY DECLARATION

Competing interests: None to declare.

REFERENCES

1. Picard E, Verschoor CP, Ma GW, Pawelec G. Relationships between immune landscapes, genetic subtypes and responses to immunotherapy in colorectal cancer. Front Immunol 2020; 11:369.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021; 71:209-49.
3. Xi Y, Xu P. Global colorectal cancer burden in 2020 and projections to 2040. Transl Oncol 2021; 14:101174.
4. Siegel RL, Wagle NS, Cercek A, Smith RA, Jemal A. Colorectal cancer statistics 2023. CA Cancer J Clin 2023; 73:233-54.
5. WHO. WHO Report on Cancer: Setting Priorities, Investing Wisely and Providing Care for All. 2020; https://www.who.int/publications/i/item/9789240001299 (15 October 2023)
6. Sawicki T, Ruszkowska M, Danielewicz A, Niedźwiedzka E, Arłukowicz T, Przybyłowicz KE. A review of colorectal cancer in terms of epidemiology, risk factors, development, symptoms and diagnosis. Cancers 2021; 13:2025.
7. Kanth P, Inadomi JM. Screening and prevention of colorectal cancer. BMJ 2021; 374.
8. Ahmad M, Alhalaiqa F, Subih M. Constructing and testing the psychometrics of an instrument to measure the attitudes, benefits, and threats associated with the use of Artificial Intelligence tools in higher education. JALT 2023; 6:114-20.
9. Hani SB, Ahmad M. Effective prediction of mortality by heart disease among women in jordan using the Chi-Squared Automatic Interaction Detection Model: retrospective validation study. JMIR Cardio 2023; 7:e48795.
10. Ahmad M, Hani SHB, Sabra MA, Almahmoud O. Big data can help prepare nurses and improve patient outcomes by improving quality, safety, and outcomes. Front Nurs 2023; 10:241-8.
11. Agrawal R, Prabakaran S. Big data in digital healthcare: lessons learnt and recommendations for general practice. Heredity 2020; 124:525-34.
12. Seow H, Tanuseputro P, Barbera L, Earle CC, Guthrie DM, Isenberg SR, Juergens RA, Myers J, Brouwers M, Tibebu S, Sutradhar R . Development and validation of a prediction model of poor performance status and severe symptoms over time in cancer patients (PROVIEW+). Palliat Med 2021; 35:1713-23.
13. Hani HSB, Ahmad MM. Large-scale data in health care: a concept analysis. Georgian Med News 2022; 325:33-6.
14. Zhang L, Zheng C, Li T, Xing L, Zeng H, Li T, Yang H, Cao J, Chen B, Zhou Z. Building up a robust risk mathematical platform to predict colorectal cancer 2017. Complexity; 2017. https://doi.org/10.1155/2017/8917258 (20 October 2023)
15. Nartowt BJ, Hart GR, Muhammad W, Liang Y, Stark GF, Deng J. Robust machine learning for colorectal cancer risk prediction and stratification. Front. Big Data 2020; 3:6.
16. Morin L, Onwuteaka-Philipsen BD. The promise of big data for palliative and end-of-life care research. In. Vol 35: SAGE Publications Sage UK: London, England 2021; 1638-40.
17. Oussous A, Benjelloun F-Z, Lahcen AA, Belfkih S. Big Data technologies: A survey. J King Saud Univ - Comput Inf Sci 2018; 30:431-48.
18. Park H, Kang Y. AI-Big Data-Mobile System development of measuring nursing workloads using wearable device and real time location information. 2023. https://doi.org/10.21203/rs.3.rs-2802548/v1

19. Knevel R, Liao KP. From real-world electronic health record data to real-world results using artificial intelligence. Ann Rheum Dis 2023; 82:306-11.
20. Palanisamy V, Thirunavukarasu R. Implications of big data analytics in developing healthcare frameworks–A review. J King Saud Univ - Comput Inf Sci 2019; 31:415-25.
21. Moher D, Liberati A, Tetzlaff J, Altman D. Preferred Reporting items for Systematic and Meta-Analysis (PRISMA) Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg 2010; 8:336-41.
22. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S; PROBAST Group. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019; 170:51-8.
23. Aromataris E, Fernandez R, Godfrey CM, Holly C, Khalil H, Tungpunkom P. Summarizing systematic reviews: methodological development, conduct and reporting of an umbrella review approach. Int J Evid Based Healthc 2015; 13:132-40.
24. Birks J, Bankhead C, Holt TA, Fuller A, Patnick J. Evaluation of a prediction model for colorectal cancer: retrospective analysis of 2.5 million patient records. Cancer Med 2017; 6:2453-60.
25. Buk Cardoso L, Cunha Parro V, Verzinhasse Peres S, Curado MP, Fernandes GA, Wünsch Filho V, Natasha Toporcov T. Machine learning for predicting survival of colorectal cancer patients. Sci Rep 2023; 13:8874.
26. Gu Y, Duan B, Sha J, Zhang R, Fan J, Xu X, Zhao H, Niu X, Geng Z, Gu J, Huang B. Serum IgG N-glycans enable early detection and early relapse prediction of colorectal cancer. Int J Cancer 2023; 152:536-47.
27. Hilsden RJ, Heitman SJ, Mizrahi B, Narod SA, Goshen R. Prediction of findings at screening colonoscopy using a machine learning algorithm based on complete blood counts (ColonFlag). PLoS One 2018; 13:e0207848.
28. Hornbrook MC, Goshen R, Choman E, O'Keeffe-Rosetti M, Kinar Y, Liles EG, Rust KC. Early colorectal cancer detected by machine learning model using gender, age, and complete blood count data. Dig Dis Sci 2017; 62:2719-27.
29. Hoogendoorn M, Szolovits P, Moons LM, Numans ME. Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer. AIM 2016; 69:53-61.
30. Lee E, Jung SY, Hwang HJ, Jung J. Patient-level cancer prediction models from a nationwide patient cohort: model development and validation. JMIR Med Inform 2021; 9:e29807.
31. Leonard G, South C, Balentine C, Porembka M, Mansour J, Wang S, Yopp A, Polanco P, Zeh H, Augustine M. Machine learning improves prediction over logistic regression on resected colon cancer patients. J Surg Res 2022; 275:181-93.
32. Liu C, Wang T, Yang J, Zhang J, Wei S, Guo Y, Yu R, Tan Z, Wang S, Dong W. Distant metastasis pattern and prognostic prediction model of colorectal cancer patients based on big data mining. Front Oncol 2022; 12:878805.
33. Nakanishi R, Morooka KI, Omori K, Toyota S, Tanaka Y, Hasuda H, Koga N, Nonaka K, Hu Q, Nakaji Y, Nakanoko T. Artificial intelligence-based prediction of recurrence after curative resection for colorectal cancer from digital pathological images. Ann Surg Oncol 2023; 30:3506-14.
34. Susič D, Syed-Abdul S, Dovgan E, Jonnagaddala J, Gradišek A. Artificial intelligence based personalized predictive survival among colorectal cancer patients. Comput Methods Programs Biomed 2023; 231:107435.
35. Tan L, Li H, Yu J, Zhou H, Wang Z, Niu Z, Li J, Li Z. Colorectal cancer lymph node metastasis prediction with weakly supervised transformer-based multi-instance learning. Med Biol Eng Comput 2023; 61:1565-80.
36. Tsai PC, Lee TH, Kuo KC, Su FY, Lee TL, Marostica E, Ugai T, Zhao M, Lau MC, Väyrynen JP, Giannakis M. Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients. Nat Commun 2023; 14:2102.
37. Virdee PS, Patnick J, Watkinson P, Holt T, Birks J. Full blood count trends for colorectal cancer detection in primary care: development and validation of a dynamic prediction model. Cancers 2022; 14:4779.
38. Joloudari JH, Saadatfar H, Dehzangi A, Shamshirband S. Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection. IMU 2019; 17:100255.
39. Liu B, Udell M. Impact of accuracy on model interpretations. arXiv CS - Machine Learning 2020; 201109903.
40. Muschelli J. ROC and AUC with a binary predictor: a potentially misleading metric. J Classif 2020; 37:696-708.
41. Melo F. Area under the ROC Curve. Encyclopedia of systems biology 2013; 38-9.
42. Wang Y, He X, Nie H, Zhou J, Cao P, Ou C. Application of artificial intelligence to the diagnosis and therapy of colorectal cancer. Am. J Cancer Res 2020; 10:3575.
43. Costa VG, Pedreira CE. Recent advances in decision trees: An updated survey. Artif Intell Rev 2023; 56:4765-800.
44. Biau G. Analysis of a random forests model. JMLR 2012; 13:1063-1095.
45. Speiser JL, Miller ME, Tooze J, Ip E. A comparison of random forest variable selection methods for classification prediction modeling. Expert Syst Appl 2019;134:93-101.
46. Connelly L. Logistic regression. Medsurg Nurs 2020; 29:353-4.
47. Nusinovici S, Tham YC, Yan MY, Ting DS, Li J, Sabanayagam C, Wong TY, Cheng CY. Logistic regression was as good as machine learning for predicting major chronic diseases. JCE 2020; 122:56-69.
48. Mangal S, Chaurasia A, Khajanchi A. Convolution neural networks for diagnosing colon and lung cancer histopathological images. ArXiv CS - Machine Learning2020; 200903878.
49. Bani Hani SH, Ahmad MM. Machine-learning Algorithms for Ischemic Heart Disease Prediction: A Systematic Review. Curr Cardiol Rev 2023; 19:87-99.
50. Jones OT, Matin RN, van der Schaar M, Bhayankaram KP, Ranmuthu CK, Islam MS, Behiyat D, Boscott R, Calanzani N, Emery J, Williams HC. Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review. Lancet Digit Health 2022; 4:e466-e476.

51. Dlamini Z, Francies FZ, Hull R, Marima R. Artificial intelligence (AI) and big data in cancer and precision oncology. CSBJ 2020; 18:2300-2311.
52. Yu C, Helwig EJ. The role of AI technology in prediction, diagnosis and treatment of colorectal cancer. Artif Intell Rev 2022:1-21.
53. Nwosu AC, Collins B, Mason S. Big data analysis to improve care for people living with serious illness: the potential to use new emerging technology in palliative care. Palliat Med 2018; 32:164-6.
54. Sammour F, Alkailani H, Sweis GJ, Sweis RJ, Maaitah W, Alashkar A. Forecasting demand in the residential construction industry using machine learning algorithms in Jordan. Constr Innov 2023.
55. El Khatib M, Hamidi S, Al Ameeri I, Al Zaabi H, Al Marqab R. Digital disruption and big data in healthcare-opportunities and challenges. CEOR 2022; 563-74.
56. Bragazzi NL, Dai H, Damiani G, Behzadifar M, Martini M, Wu J. How big data and artificial intelligence can help better manage the COVID-19 pandemic. IJERPH 2020; 17:3176.
57. Stefanicka-Wojtas D, Kurpas D. eHealth and mHealth in Chronic Diseases—identification of barriers, existing solutions, and promoters based on a survey of EU stakeholders involved in Regions4PerMed (H2020). J Pers Med 2022; 12:467.
58. Ruiters S, Mombaerts I. Applications of three-dimensional printing in orbital diseases and disorders. Curr Opin Ophthalmol 2019; 30:372-9.
59. Storick V, O'Herlihy A, Abdelhafeez S, Ahmed R, May P. Improving palliative care with machine learning and routine data: a rapid review. HRB Open Research 2019; 2.
60. Srivastava D, Pandey H, Agarwal AK. Complex predictive analysis for health care: a comprehensive review. BEEI 2023; 12:521-531.
61. Sangaiah AK, Rezaei S, Javadpour A, Zhang W. Explainable AI in big data intelligence of community detection for digitalization e-healthcare services. Appl Soft Comput 2023; 136:110119.
62. Pastorino R, De Vito C, Migliara G, Glocker K, Binenbaum I, Ricciardi W, Boccia S. Benefits and challenges of Big Data in healthcare: an overview of the European initiatives. Eur J Public Health 2019; 29(Suppl 3):23-7.
63. Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med 2019; 25:37-43.
64. Alboaneen D, Alqarni R, Alqahtani S, Alrashidi M, Alhuda R, Alyahyan E, Alshammari T. Predicting colorectal cancer using machine and deep learning algorithms: challenges and opportunities. BDCC 2023; 7:74.
65. Ahmad M, Hani SHB, Sabra MA, Almahmoud O. Big data can help prepare nurses and improve patient outcomes by improving quality, safety, and outcomes. Front Nurs 2023; 10:241-8.
66. Lewandowska A, Rudzki G, Lewandowski T, Stryjkowska-Góra A, Rudzki S. Risk factors for the diagnosis of colorectal cancer. Cancer Control 2022; 29:10732748211056692.
67. Cervantes A, Adam R, Roselló S, Arnold D, Normanno N, Taïeb J, Seligmann J, De Baere T, Osterlund P, Yoshino T, Martinelli E. Metastatic colorectal cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Ann Oncol 2023; 34:10-32.
68. American Cancer Society. Colorectal cancer facts & figures 2020–2022. Atlanta Am Cancer Soc 2020; 66:1-41.
69. Amarin JZ, Mansour R, Nimri OF, Al-Hussaini M. Incidence of cancer in adolescents and young adults in Jordan, 2000-2017. JCO Glob Oncol 2021; 7:934-46.
70. Sharkas GF, Arqoub KH, Khader YS, Tarawneh MR, Omar F Nimri Nimri OF, Al-Zaghal MJ, Subih HS . Colorectal cancer in Jordan: survival rate and its related factors. J Oncol 2017; 2017.
71. Bazira PJ. Anatomy of the caecum, appendix, and colon. Surgery (Oxford) 2022; 41:1-6
72. Billmann F, Keck T, editors. Essentials of Visceral Surgery: For Residents and Fellows. Berlin: Springer Nature; 2023
73. Awad H, Abu-Shanab A, Hammad N, Atallah A, Abdulattif M. Demographic features of patients with colorectal carcinoma based on 14 years of experience at Jordan University Hospital. Ann Saudi Med 2018; 38:427-432.
74. American Cancer Society. Colorectal cancer, early detection, diagnosis, and staging. 2023 https://www.cancer.org/cancer/types/colon-rectal-cancer/detection-diagnosis-staging.html. 2023 (29 June 2023).
75. Sun Y, Fan X, Zhao J. Development of colorectal cancer detection and prediction based on gut microbiome big-data. Med Microecol 2022; 12:100053.