ORIGINAL ARTICLE

Streptococcus Gallolyticus Infection and its Interrelation with Colorectal Cancer: Diagnostic Accuracy of Statistical and Machine Learning Models for Early Detection Algorithm

Edre MA[a], Hairul Aini H [b], Mohd Shaiful Ehsan S[c], Azmi MN[d], Che Muhammad Khairul Hisyam I[e]*

[a]Department of Community Medicine, Kulliyyah of Medicine, International Islamic University Malaysia
[b]Department of Basic Medical Sciences, Kulliyyah of Medicine, International Islamic University Malaysia
[c]Department of Family Medicine, Kulliyyah of Medicine, International Islamic University Malaysia
[d]Department of Surgery, Kulliyyah of Medicine, International Islamic University Malaysia
[e]Kulliyyah of Science, International Islamic University Malaysia

ABSTRACT

INTRODUCTION: Epidemiological studies have emphasized the role of Streptococcus *gallolyticus subspecies gallolyticus (Sgg)* infection in the development of colorectal cancer (CRC), yet it remains underappreciated. While statistical and machine learning (ML) models can enhance CRC prediction, direct comparisons between them are rare. This study aims to assess the diagnostic accuracy of stool polymerase chain reaction (PCR) for *Sgg* and immunochemical fecal occult blood test (iFOBT) for CRC detection and to compare multivariable statistical and ML models in predicting CRC. **MATERIALS AND METHODS:** A hospital-based case-control study with a reversed flow design was conducted, involving 33 CRC cases and 80 controls. The analysis incorporated Asia Pacific Colorectal Screening (APCS) risk factors into three predictive models: logistic regression (LR), decision tree (DT), and ensemble Bayesian boosted decision tree (BDT). **RESULTS:** Combined testing achieved a net sensitivity of 54%, outperforming individual tests (iFOBT=12.1%, Stool PCR=48.5%). Among the models, the ensemble BDT approach demonstrated the highest classification accuracy for CRC (BDT= 78.1%; DT=72.4%; LR=69.9%). The DT model identified iFOBT as the sole predictor, while the BDT ensemble model prioritized positive stool PCR for *Sgg* as the primary predictor, followed by normal to overweight body mass index and individuals aged over 53 years. **CONCLUSION:** The ensemble ML model incorporating *Sgg* infection demonstrated superior predictive performance. Screening for *Sgg* in stool samples has the potential as an early CRC detection strategy, particularly for individuals with a normal to overweight BMI and those above 53 years old.

Keywords
Streptococcus gallolyticus, colorectal cancer, diagnostic accuracy, machine learning, bayesian

Corresponding Author
Assoc. Prof. Dr. Edre Mohammad Aidid
Department of Community Medicine,
Kulliyyah of Medicine, International Islamic
University Malaysia, Jalan Sultan Ahmad
Shah, Bandar Indera Mahkota,
25200 Kuantan, Pahang, Malaysia
Email: edreaidid@iium.edu.my

Received: 10th April 2025; Accepted: 26th August 2025

Doi: https://doi.org/10.31436/imjm.v24i04.2895

INTRODUCTION

Colorectal cancer (CRC) is one of the most common cancers in developing countries.[1] Various studies look for lifestyle factors but do not highlight infection as the potential driver for CRC development. Many epidemiological studies highlighted the importance of the *Streptococcus galloyticus subspecies gallolyticus (Sgg)* infection in carcinogenesis of CRC.[2–5] It is regarded as a highly important and treatable disease. However, it is not often examined.

An increasing incidence of CRC coupled with late detection of the disease due to poor screening practice leads to a high burden of disease.[6] This is made worse by variable results of immunochemical faecal occult blood (iFOBT) as the first line screening which leads to false positive and false negative results. *Sgg* infection detected has been shown to increase the risk for CRC in many epidemiological studies.[4,5,7] However, there is a paucity of studies on screening diagnostic accuracy related to *Sgg.*

More and more patients are referred for unnecessary invasive colonoscopy that resulting from false positive iFOBT. Furthermore, current guidelines follow stratification based on family history. There has been a lack of risk stratification based on modifiable factors. Thus, in order to have more people correctly classified as having CRC, this infection should be highlighted and included as one of the risk factors apart from presence of blood in the stool (source).[2] In view of the multifactorial nature of CRC, a combination of risk factors is a more plausible explanation for its occurrence, which can be explained not just by statistical but machine learning models (ML).

ML has been used to predict cancers but not specifically looking at multivariable infection models. Various techniques such as decision tree, artificial neural network and naive bayes have been utilised, but suffers from either underfitting or overfitting. One potential technique to improve model accuracy is by ensemble methods such as bagging, boosting and stacking.[8] However, small sample sizes often deter the ability to make good predictions. Hence, the bayesian method of looking at prior knowledge helps to overcome this problem.

The research objectives in this study are twofold; First objective is to determine diagnostic accuracy of simultaneous testing of stool PCR for *Sgg* with iFOBT in detecting CRC and second objective is to predict the likelihood of CRC by comparing multivariable statistical and machine learning models. We hypothesised that the overall accuracy of bayesian ensemble machine learning model is better than the statistical model.

MATERIALS AND METHODS

This research was carried out over 3 year duration (from September 2019 until July 2022). It was a case-control study with reversed flow design with an allocation ratio of 1 case to 2 controls which followed Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement checklist. The study population were patients who came to the surgical clinic of Sultan Ahmad Shah Medical Centre IIUM (SASMEC @ IIUM) for colonoscopy. A case was defined as patients who attended the colonoscopy and diagnosed as histopathologically-confirmed CRC and controls were patients who attended the colonoscopy but were diagnosed as not having CRC.

Sample size was calculated based on PASS software version 19.03. For objective 1, the calculation for estimation of the minimum sample size for this study was calculated with estimated prevalence of 20% for 1 university hospital in peninsular Malaysia with similar setting,[9] power of 88.5% and sensitivity of 90%, the minimum sample size required is 95.[10] The power was chosen as it is the highest achievable power to detect to reject the null hypothesis of no association, hence able to reach up to 90% sensitivity based on the simulation.[10] For objective 2, the sample size was calculated based on 6 predictors using logistic regression formula,[11] where the number of predictors were multiplied by 15 to get 90. Since objective 1 has a higher estimated sample size (95), this estimated sample size is chosen as it can answer objective 1 and 2, thus final sample size after adjusting for 20% attrition rate, total sample size was 120. Thus, approximately 40 cases and 80 controls were required.

Quality control

For iFOBT, ABON FOB test was used, which is a simple, rapid qualitative immunochemical assay that detects 50 ng/ml of human blood haemoglobin done in controlled laboratory environment. If both test and control lines were visible, it indicated that the test was positive. If only the control line was visible, then the test was deemed negative. If only the test line is visible, the test was taken as invalid.

Concurrently, stool PCR were tested for evidence of *Sgg* infection. Two sets of primers were tested in the PCR assay. Following PCR optimization of both primers and gel electrophoresis, amplification using primer set 2 resulted in non-specific and multiple bands, while primer set 1 exhibited a specific and single band of DNA without primer dimer and the needs of pure culture bacteria. The PCR with primer set 1 showed an optimum outcome with the gradient annealing temperatures of 55^{o}C to 62.9^{o}C. DNA sequence alignment of randomly selected samples showed 100% similarity without any nucleotide mutation. BLASTN analysis of the targeted 167 bp amplified gene segments demonstrated that DNA sequences were 100% identical to the *tanB* gene of *Sgg* found elsewhere.[12]

The background characteristics of patients, which are age, gender, family history, BMI and smoking status were taken based on the validated modified Asia Pacific Colorectal Screening (APCS) score.[13]

Univariate data analysis

Chi square test and fisher's exact test was employed for univariate analysis in RStudio version 2022.07.2. Accuracy parameters were calculated for iFOBT and stool PCR in terms of sensitivity, specificity, positive predictive value, negative predictive value and diagnostic accuracy. The calculations were split into single testing for each test and simultaneous testing of both tests.

Multivariable data analysis

Data was analysed using RapidMiner Studio version 9.10.011. Continuous data was reported as mean (SD) and categorical as frequency (%). Classical logistic regression (LR) was used for the statistical model. For ML prediction, bayesian boosting operator methods were utilised using ensemble machine learning, where we compared bayesian-decision tree (BDT) ensemble versus decision tree (DT) model alone, to improve the accuracy and generate screening algorithm. The Bayesian boosting operator builds an ensemble of classifiers to predict boolean target attributes. During each iteration, the training data was reweighted to reduce the influence of previously learned patterns and incorporate prior knowledge. A base classifier, usually a rule-based or decision tree algorithm, was applied repeatedly in sequence, and the resulting models were merged into one comprehensive model. The total number of models generated was determined by the specified number of iterations. The steps involved were feature engineering, threshold setting, cross validation and model performance evaluation. After the data collection, the data was pre-processed and missingness was handled via multiple imputation techniques. The feature selection involved the variables identified by the APCS risk factors, as well as the main laboratory markers namely the iFOBT (in line with CPG) and Stool PCR for *Sgg* as the feature of interest. The threshold for prediction was fine-tuned at 0.6. If the prediction probability is greater than this, the class label to be predicted is CRC. A 10-fold cross validation method was done with shuffled sampling, coupled with random seed number to randomly choose the subsets for the training and testing model. Accuracy parameters were the same as the univariate data analysis, adding in another parameter which is area under the curve (AUC).

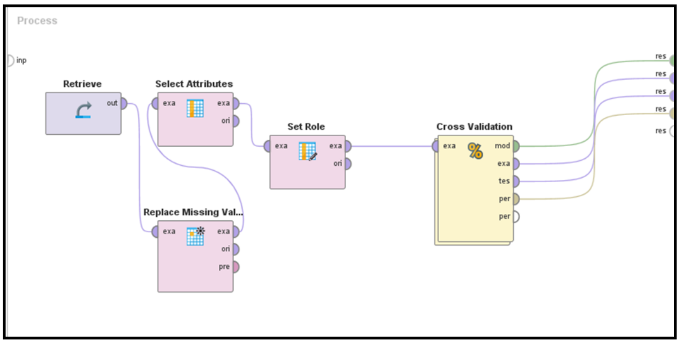

Figure 1: Process canvas for multivariable analysis using boxes called operators. The grey "retrieve" operator signifies the original dataset, the red "select attribute" operator signifies the variables selected for analysis, the red "replace missing values" operator signifies the imputation of missing values, the red "set role" operator is setting the CRC as the case and non-CRC as the control (Boolean) and the yellow "cross validation" operation is the method for data validation.

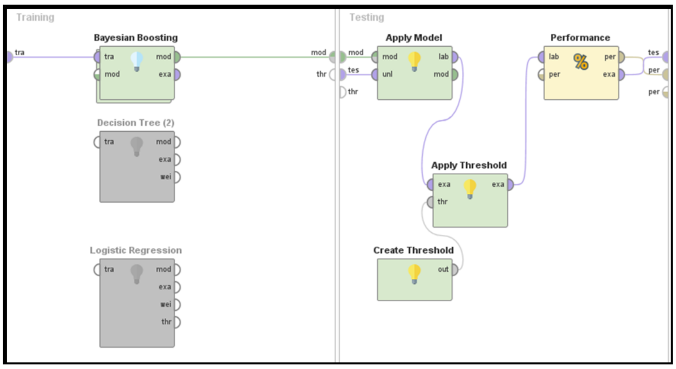

Figure 2: Cross validation for the three models (LR, DT and BDT)), tested at different stages. The training section on the left signifies the model used (the green is the active model), while the testing section on the right signifies the execution of the model based on 30% testing data (green operator) and performance metric generation (yellow operator such as sensitivity)

RESULTS

Objective 1

Thirty-three (33) CRC cases were obtained instead of 40 due to patient loss to follow-up. The proportion of cases with a positive iFOBT and stool PCR was higher than a negative test (P<0.05). Post hoc power analysis using

48.5% sensitivity of stool PCR for *Sgg* obtained shows the actual power was 92% even though there was loss to follow up. The simultaneous testing produced net sensitivity that was higher compared to a single test.

Table 1: Distribution of case and control according to iFOBT and stool PCR for *Sgg*

Test	Classification		P value
	Case, n(%)	Control, n(%)	
iFOBT			
Positive	4(80.0)	1(20.0)	0.03*
Negative	29(26.9)	79(73.1)	
Stool PCR for *Sgg*			
Positive	16(50.0)	16(50.0)	0.005*
Negative	17 (21.0)	64(79.0)	

*Significant at P<0.05

Table 2: Accuracy parameters for single and simultaneous testing

Accuracy parameter	iFOBT	Stool PCR for *Sgg*	Simultaneous testing
Sensitivity[a] (%)	12.1	48.5	54.7
Specificity[b] (%)	98.8	80.0	79.0
Positive predictive value (%)	80.0	50.0	-
Negative predictive value (%)	73.1	79.9	-
Diagnostic accuracy (%)	73.5	70.8	-

[a]For simultaneous testing, net sensitivity was calculated
[b]For simultaneous testing, net specificity was calculated

Objective 2

The performance of the BDT ensemble approach was superior (BDT accuracy=78.1%; DT accuracy=72.4%; LR accuracy=69.9%). The DT algorithm generated only iFOBT as the predictor without the inclusion of *Sgg*, whereas the ensemble approach produced positive stool PCR for *Sgg* as the main branch followed by normal to overweight body mass index and adults above 53 years of age.

Table 3: Performance metrics of the models compared

Parameter	Logistic Regression (LR)	Decision Tree (DT)	Bayesian-Decision Tree (BDT)
Accuracy (%)	69.0	72.4	78.1
Sensitivity (%)	36.0	15.8	56.6
Specificity (%)	82.2	96.0	85.8
Positive predictive value (%)	39.3	62.5	62.5
Negative predictive value (%)	76.5	73.2	84.3
Area under the curve (AUC)	0.74	0.53	0.74
Significant predictors (categorical unless stated otherwise)	Gender Stool PCR iFOBT	iFOBT	Stool PCR BMI (kg/m^2) Age (years)

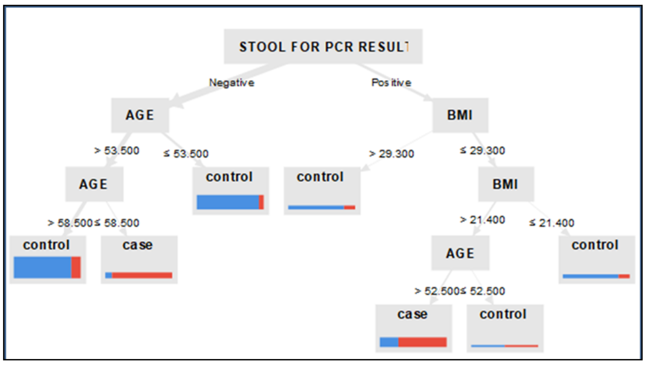

Figure 3: Best model algorithm generated from the BDT ensemble method.

DISCUSSION

Gist of the research

The sustainable development goals (SDG) number 3 focuses on ensuring wellbeing in terms of reducing burden due to NCDs. Goal 3.4.1 states the importance of reducing premature mortality from cancer by one third by early detection and prompt treatment.[14] This includes detecting the agents triggering cancer formation such as the *Sgg* infection. If the infection is detected early, it can be treated with antibiotics and prevents the occurrence of disease. Furthermore, the current evidence to risk-stratify the patients for screening does not consider infective agents such as *Sgg*.[15] Risk stratification allows for risk prediction to tell patients that early screening is important as it increases the risk for cancer. This in turn will assist the patients in making good decisions towards their health and wellbeing. Hence, it is of utmost importance for this research to detect new risk predictions for people to know and to act early. For clinicians, clinical decision support model and system can help in stratifying those who require invasive diagnostics such as colonoscopy.[16,17]

Outcome on single and simultaneous testing

Currently the clinical practice guideline focuses on iFOBT as the asymptomatic screening for CRC. Here, the proportion of CRC among positive iFOBT was higher than negative iFOBT by almost 60%, indicating the relevance of screening asymptomatic population in the guidelines. For stool PCR, it is not routinely done worldwide as a screening tool. However, our study proved its relevance as a potential screening marker as it was detected more in CRC cases, similar with previous studies.[2,4] The sensitivity of iFOBT was low at 12.1% as compared to a systematic review by the Malaysian Health

Technology Assessment Section team.[18] The iFOBT sensitivity depends on yield quality, timing of sample collection, even though the respondents have been briefed regarding the sample collection procedure. The briefing followed standard informed consent procedure and the enumerators were trained. On the other hand, the sensitivity of stool PCR was higher than iFOBT, indicating a negative test will rule out CRC. In addition, the negative predictive value was higher than iFOBT by almost 7%, signifying the probability of not having CRC following a negative test was higher. This gave a promising clue for stool PCR as a screening tool rather than diagnostic tool. However, the screening is predicted to be better in certain population such as those with high red meat consumption, where the bacteria reside and gets into the human gut. The challenges of using stool PCR are the testing requires PCR machines that are not readily available in all healthcare setting and stool sampling is yield-dependent as it may not contain the causative agent, if any.

When both tests are done simultaneously, gains in sensitivity were observed indicating that iFOBT and stool PCR, if done concurrently, can improve early detection rates for CRC. A systematic review showed the usefulness of simultaneous testing in detecting CRC.[19]

Statistical model building

Many research has focused on statistical models.[20,21] One of the most used is logistic regression. The model's ability to predict binary outcomes such as cancer and non-cancer has made it a popular choice for classification problems. Here, our model showed that male patients and positive iFOBT are independent risk factors of CRC. These risk factors are in favour of the current epidemiology of CRC in Malaysia.[6] However, the stool PCR here is of interest as a positive test will increase the odds to get CRC by 4 times as compared to a negative test controlling for APCS risk score and iFOBT. The method of detection here is improved by detecting the subspecies (*S. gallolyticus subsp. gallolyticus*) which are more accurate for Sgg-CRC interaction. Thus, molecular detection here plays an important role in establishing that interaction.[2,22] However, other microbiota such as *Fusobacterium nucleatum*, also linked with CRC are not tested here which might serve as a marker for carcinogenesis as well, hence serving as a selection bias.[1] This becomes the limitation in the study. Nevertheless, these findings were made after data cleaning and multiple imputation, which improved the model fit and accuracy in one study.[23]

Machine learning model building

Feature engineering has become one of the most important steps in building machine learning models. In our research, the missingness, if left untreated, could lead to biased estimates. Multiple imputation method was chosen as the multivariable nature of the data would benefit more from this method as compared to single imputation.[23]

Threshold for predicting the case class was set higher at 0.6 than the default 0.5 because we wanted to correctly detect true CRC cases and minimise the false positive rate, similar with a previous study.[24] This would in turn increase the precision of the model.

In terms of the performance metric, the most important parameter would be the AUC. The use of AUC is known to be best used when there is an imbalance in outcome because it measures the quality of the model's predictions irrespective of what classification threshold is chosen.[25] Thus, it is a robust performance measure.

When compared to the LR statistical model, ML outperforms it in few key areas such as accuracy, specificity and positive predictive value. This shows the value of ML models in multivariable analysis, which is more practical in real-world applications such as clinical decision support system (CDSS) for diagnosis of CRC. A study revealed the use of ML to aid diagnosing cancer and estimating prognosis to aid decision-making.[26]

We used the decision tree model as the base ML model because of the classification nature of the outcome. It has shown to be robust but tends to overfit. To overcome this, ensemble method which combined bayesian boosting and cross validation was utilised to combat overfitting.[8]

Performance evaluation of the three models

The main interest here is to determine which model is the best in predicting CRC. Here we had a variety of performance metrics but the gain and losses in these parameters were evaluated more in terms of its value in clinical decision making. For example, to diagnose high-burden diseases such as cancer requires more specificity than sensitivity. For screening, sensitivity is preferred. It is a trade-off. The statistical LR model showed good predictive power, whereby being male, positive stool PCR and positive iFOBT had significantly higher odds to get CRC. However, its other metrics are poorer than ML models. Statistical models are sometimes not as good as machine learning models in terms of optimization,[26] and quite strict in their test assumptions. ML models, on the other hand, are more robust but require model optimization to increase prediction accuracy.[27]

For the DT model, the best performance was specificity (96%) and iFOBT as the significant predictor. When BDT ensemble method was used, there was a positive trade-off where improvement in almost all parameters except specificity, as well as having stool PCR, age and BMI as the significant predictors. Thus, it can be inferred here that DT model accurately detects CRC by only iFOBT, but BDT was able to accurately predict CRC much better with multivariable inputs. Bayesian boosted decision trees integrate the advantages of Bayesian optimization and boosting with decision trees to build highly effective and resilient machine learning models. Bayesian optimization aids in selecting the best hyperparameters for the boosted trees, enhancing both accuracy and the ability to generalize. Boosting works by sequentially combining multiple weak decision trees to form a strong overall model capable of capturing intricate patterns in the data. Together, these techniques produce models that are typically more accurate, less prone to overfitting, and well-suited for handling complex datasets. The other traditional machine learning methods such as random forest and gradient-boosted machines lack this. The superiority of ensemble learning here was similar with another study on e-learning performance prediction[28] albeit in a different context. The BDT algorithm provided clues that a positive stool PCR patient should be further risk-stratified into normal and overweight BMI and age more than 53 years old to be advised for early colonoscopy. We also tested for other CRC risk factors based on APCS criteria but was not significant in the BDT algorithm.[13]

Novel theories/knowledge

The tool to detect *Sgg* infection as a marker for CRC can be classified to be either for diagnostic purpose or as a screening tool based on the diagnostic accuracy parameters. As described earlier, the results show promise more towards screening tool rather than diagnostic tool, as the *Sgg* infection is mainly asymptomatic. The availability of the primer for testing and the prevalence of SGG that is fairly high serves the potential for this organism to be developed as a rapid test kit for early clue and detection of CRC, especially when simultaneous testing with iFOBT is done as per result findings.[9]

The Bayesian prediction model that incorporates *Sgg* infection allows for a more accurate prediction for CRC. The advantage lies in the ability of Bayesian prediction to cater for small sample size using prior probability, posterior probability and likelihood.[29,30] Hence, a more realistic prediction and better risk stratification for early colonoscopy for patients can be shown.

Potential real-world application

CDSS can be implemented at clinic and hospital level to risk-stratify those going for colonoscopy. Antibiotic treatment can be given to patients with *Sgg* infection. Future development of vaccines towards *Sgg* may be able to prevent the infection to either induce CRC formation or delay the progress of the disease. A cost-effective screening that reduces cost of treatment will benefit the people in low socioeconomic status by correctly screening those at risk, where early diagnosis such as at stage I of the disease will advert the need for high-costing treatment such as chemotherapy.[6] Potential vaccine against the *Sgg* infection can be developed which will prevent CRC development. This promotes a healthier and more productive nation.

LIMITATIONS

The research has several limitations. First, causality cannot be ascertained for *Sgg* and CRC interaction. The single centre experience in obtaining the case and control might be limiting the strength of the relationship. The results for simultaneous testing, although better than single testing, were not proven with cost effectiveness calculations. The prediction using machine learning is data-dependent and operator-dependent, thus is prone to biases just like the statistical model. The missing data, although imputed, might be missing not at random but are difficult to prove. Hence, better primary data collection method and guidelines for testing should be outlined and developed in hospital setting for quality *Sgg* and CRC data to be obtained for more precise results. Having no external validation using external data serves as our limitation. Since this study in done mainly for model development, the error risk was reduced by including training (70%) and testing (30%) data using the same dataset. Together with cross validation, this serves to reduce overfitting and provide a better valid model.

CONCLUSION

Ensemble ML model incorporating *Sgg* infection was superior to the standard ML and statistical model in predicting CRC. *Sgg* screening for early CRC detection in those with normal to overweight BMI patients and aged above 53 years old has potential, provided further research with more robust techniques to minimize error is done in the future. Future studies are recommended to use the ensemble ML model to explore the dietary and environmental source of this infection towards CRC which may pose an undetected one health problem, as well as external validation using other CRC cohorts to enhance the model validity.

FUNDING

The research was funded by fundamental research grant scheme (FRGS) year 2019, as mentioned in the acknowledgements.

CONFLICT OF INTEREST

There were no conflicts of interests between the contributors of this research.

INSTITUTIONAL REVIEW BOARD (ETHICS COMMITTEE)

The study protocol was approved by the Medical Research Ethics Committee (MREC), Ministry of Health Malaysia (NMRR-19-3062-51372) and IIUM research ethics committee (IREC) (IREC2020-109) before the study.

ACKNOWLEDGEMENTS

We would like to thank the Ministry of Higher Education for the research grant provided (FRGS/1/2019/SKK05/UIAM/03/1). We also would like to thank the research management centre, IIUM for facilitating the grant management (FRGS19-177-0786).

REFERENCES

1. Young C, Wood HM, Seshadri RA, et al. The colorectal cancer-associated faecal microbiome of developing countries resembles that of developed countries. *Genome Med.* 2021;13(1). doi:10.1186/s13073-021-00844-8
2. Périchon B, Lichtl-Häfele J, Bergsten E, et al. Detection of Streptococcus gallolyticus and Four Other CRC-Associated Bacteria in Patient Stools Reveals a Potential “Driver” Role for Enterotoxigenic Bacteroides fragilis. *Front Cell Infect Microbiol.* 2022;12. doi:10.3389/fcimb.2022.794391
3. Kwong TNY, Wang X, Nakatsu G, et al. Association Between Bacteremia From Specific Microbes and Subsequent Diagnosis of Colorectal Cancer. *Gastroenterology.* 2018;155(2):383-390.e8. doi:10.1053/j.gastro.2018.04.028
4. Rezasoltani S, Asadzadeh Aghdaei H, Dabiri H, et al.. The association between fecal microbiota and different types of colorectal polyp as precursors of colorectal cancer. *Microb Pathog.* 2018;124:244-249. doi:10.1016/j.micpath.2018.08.035
5. Corredoira-Sánchez J, García-Garrote F, Rabunal R, et al. Association between bacteremia due to Streptococcus gallolyticus subsp. gallolyticus (Streptococcus bovis I) and colorectal neoplasia: A case-control study. *Clinical Infectious Diseases.* 2012;55(4):491-496. doi:10.1093/cid/cis434
6. Veettil SK, Lim KG, Chaiyakunapruk N, et al. Colorectal cancer in Malaysia: Its burden and

implications for a multiethnic country. *Asian J Surg.* 2017;40(6):481-489. doi:10.1016/j.asjsur.2016.07.005

7. Butt J, Romero-Hernández B, Pérez-Gómez B, et al. Association of Streptococcus gallolyticus subspecies gallolyticus with colorectal cancer: Serological evidence. *Int J Cancer.* 2016;138(7):1670-1679. doi:10.1002/ijc.29914
8. Ali MS, Hossain MM, Kona MA, Nowrin KR, Islam MK. An ensemble classification approach for cervical cancer prediction using behavioral risk factors. Healthcare Analytics. 2024 Jun 1;5:100324.
9. Al-Jashamy K, Murad A, Zeehaida M, Rohaini M, Hasnan J. Prevalence of colorectal cancer associated with Streptococcus bovis among inflammatory bowel and chronic gastrointestinal tract disease patients. Asian Pac J Cancer Prev. 2010;11(6):1765-8. PMID: 21338230.
10. Bujang MA, Adnan TH. Requirements for minimum sample size for sensitivity and specificity analysis. *Journal of Clinical and Diagnostic Research.* 2016;10 (10):YE01-YE06. doi:10.7860/JCDR/2016/18129.8744
11. Ann L, Llewellyn A, Foreword JC, et al. *DESIGNING AND CONDUCTING HEALTH SURVEYS A Comprehensive Guide THIRD EDITION.*
12. Pompilio A, di Bonaventura G, Gherardi G. An overview on streptococcus bovis/streptococcus equinus complex isolates: Identification to the species/subspecies level and antibiotic resistance. *Int J Mol Sci.* 2019;20(3). doi:10.3390/ijms20030480
13. Kong Y, Zhuo L, Dong D, et al. Validation of the Asia-Pacific colorectal screening score and its modified versions in predicting colorectal advanced neoplasia in Chinese population. *BMC Cancer.* 2022;22(1). doi:10.1186/s12885-022-10047-y
14. NCD Countdown 2030: pathways to achieving Sustainable Development Goal target 3.4. *The Lancet.* 2020;396(10255):918-934. doi:10.1016/S0140-6736 (20)31761-X
15. *Management of Colorectal Carcinoma.* http://www.moh.gov.myhttp://www.acadmed.org.myhttp://www.colorectalmy.orghttp://www.msgh.org.myhttp://www.malaysiaoncology.org
16. Keikes L, Kos M, Verbeek XAAM, et al. Conversion of a colorectal cancer guideline into clinical decision trees with assessment of validity. *International Journal for Quality in Health Care.* 2021;33(2). doi:10.1093/intqhc/mzab051
17. Militello LG, Saleem JJ, Borders MR, et al. Designing Colorectal Cancer Screening Decision Support: A Cognitive Engineering Enterprise. *J Cogn Eng Decis Mak.* 2016;10(1):74-90. doi:10.1177/1555343416630875
18. Fuzi, S. A. M., Hassan, M. R. A., Sabirin, J., & Bakri, R. (2015). Immunochemical faecal occult blood test for colorectal cancer screening: a systematic review. *Med J Malaysia*, *70*(1), 25.
19. Niedermaier T, Weigl K, Hoffmeister M, et al. Fecal Immunochemical Tests Combined with Other Stool Tests for Colorectal Cancer and Advanced Adenoma Detection: A Systematic Review. *Clin Transl Gastroenterol.* 2016;7(6). doi:10.1038/ctg.2016.29
20. Kaindi DWM, Kogi-Makau W, Lule GN, et al. Investigating the association between African spontaneously fermented dairy products, faecal carriage of Streptococcus infantarius subsp. infantarius and colorectal adenocarcinoma in Kenya. *Acta Trop.* 2018;178:10-18. doi:10.1016/j.actatropica.2017.10.018
21. Butt J, Werner S, Willhauck-Fleckenstein M, et al. Serology of Streptococcus gallolyticus subspecies gallolyticus and its association with colorectal cancer and precursors. *Int J Cancer.* 2017;141(5):897-904. doi:10.1002/ijc.30765
22. Boleij A, Roelofs R, Danne C, et al. Selective antibody response to Streptococcus gallolyticus pilus proteins in colorectal cancer patients. *Cancer Prevention Research.* 2012;5(2):260-265. doi:10.1158/1940-6207.CAPR-11-0321
23. Murcia O, Juárez M, Rodríguez-Soler M, et al. Colorectal cancer molecular classification using BRAF, KRAS, microsatellite instability and CIMP status: Prognostic implications and response to chemotherapy. *PLoS One.* 2018;13(9). doi:10.1371/journal.pone.0203051
24. Briggs E, de Kamps M, Hamilton W, et al. Machine Learning for Risk Prediction of Oesophago-Gastric Cancer in Primary Care: Comparison with Existing

Risk-Assessment Tools. *Cancers (Basel)*. 2022;14(20). doi:10.3390/cancers14205023

25. Silva A, Oliveira T, Neves J, et al. *ADCAIJ Author Submission Guidelines Treating Colon Cancer Survivability Prediction as a Classification Problem*. http://www.lifemath.net/cancer/coloncancer/outcome/index.php.
26. Cruz JA, Wishart DS. *Applications of Machine Learning in Cancer Prediction and Prognosis*. Vol 2.; 2006.
27. Adadi A. A survey on data-efficient algorithms in big data era. *J Big Data*. 2021;8(1). doi:10.1186/s40537-021-00419-9
28. Saleem F, Ullah Z, Fakieh Bet al. Intelligent decision support system for predicting student's e-learning performance using ensemble machine learning. *Mathematics*. 2021;9(17). doi:10.3390/math9172078
29. Moradzadeh R, Mansournia MA, Baghfalaki T, et al. Misclassification adjustment of family history of breast cancer in a case-control study: A Bayesian approach. *Asian Pacific Journal of Cancer Prevention*. 2016;16(18):8221-8226. doi:10.7314/APJCP.2015.16.18.8221
30. Martinez EZ, Louzada-Neto F, Achcar JA, et al. Bayesian estimation of performance measures of screening tests in the presence of covariates and absence of a gold standard. *Braz J Probab Stat*. 2009;23(1):68-81. doi:10.1214/08-BJPS006