Clinica Chimica Acta 498 (2019) 108–115

Contents lists available at ScienceDirect

Clinica Chimica Acta

journal homepage: www.elsevier.com/locate/cca

Evaluation of algorithm development approaches: Development of biomarker panels for early detection of colorectal lesions

Susan H. Gawel[a], Michael Lucht[a], Heather Gomer[b], Patrick Treado[b], Ib J. Christensen[c], Hans J. Nielsen[c], Gerard J. Davis[a,*], for the Danish Research Group on Early Detection of Colorectal Cancer

[a] Abbott Cancer Core R&D, Abbott Diagnostics Division, Abbott Park, IL, USA
[b] ChemImage Life Sciences, ChemImage Corporation, Pittsburgh, PA, USA
[c] Department of Surgical Gastroenterology, Hvidovre Hospital, Denmark

ARTICLE INFO

Keywords:
Colon cancer
Early detection
Biomarkers
Algorithm methodologies

ABSTRACT

Introduction: Colorectal cancer (CRC) is the third most common cancer in the U.S. Early detection of CRC can substantially increase survival rates. Test compliance may be improved by offering a blood-based test option.
Methods: Endoscopy II trial specimens were tested for AFP, CA19-9, CEA, hs-CRP, CyFra 21-1, Ferritin, Galectin-3, and TIMP-1 levels. These biomarkers, as well as patient demographic information (e.g., age, gender), were included in algorithm development. Six statistical methods were utilized to develop algorithms that would discriminate cancer vs. noncancers. Statistical methods included logistic regression, adaptive index modeling, partial least-squares discriminant analysis, feature vector (weighted and unweighted), and random forest. The performance of these algorithms was compared against benchmark criteria established for stool-based tests.
Results: Using several statistical methods, the presence of CRC and high-risk adenomas was detected with an AUCs of at least 0.65–0.76, with a few of models approaching the stool-based tests benchmark performance. Further, common markers were utilized across the different statistical techniques, with model complexities ranging from 3 to 9 markers.
Conclusions: Predictive models identified subjects with CRC and high-risk adenomas with the similar levels of statistical accuracy. Clinical performance differences were minimal across the statistical techniques, although the intuitive interpretations, model complexity, clinical adoption and implementation varied.

1. Introduction

Colorectal cancer (CRC) is the third most common type of cancer and the second leading cause of cancer death in the United States, with 2014 CRC-related deaths projected to be nearly 50,000 in the U.S. [1]. CRC screening effectiveness has led to early stage detection [2] and improvements in population-based CRC screening programs are estimated to save 18,800 lives per year in the U.S. [3]. Detection of CRC in earlier stages substantially increases the 5-year survival rate (Fig. 1). This increase in the 5-year survival rate has been used as the basis for implementing numerous CRC screening programs in many countries, including in the U.S. In 2013, the U.S. Preventative Services Task Force (USPSTF) gave a Grade A recommendation for CRC screening using fecal occult blood testing, sigmoidoscopy, or colonoscopy in adults, beginning at age 50 years and continuing until age 75 y [4].

Compliance to CRC screening recommendations remain challenging for numerous reasons, resulting in approximately 1/3 of U.S. screening-eligible patients not being screened [1]. Compliance may be improved by offering a blood-based test option to those refusing colonoscopy [5] or in place of a stool-based screening test. Further, there has been recent focus on developing blood-based algorithms to aid in the early detection of CRC [6]. A biomarker panel/algorithm approach may increase clinical sensitivity over single marker approaches since cancers, including CRC, are known to be of diverse etiologies [7]. Biomarker algorithm approaches are becoming more mainstream, including in CRC screening, with the recent U.S. approval of Cologuard (Exact Sciences), which utilizes three biomarkers in its stool-based test [8].

In the present study, an exploratory strategy was used to develop biomarker panels/algorithms, utilizing blood-based biomarkers and available patient information, in order to discriminate patients with CRC and high-risk adenomas from those without (Fig. 2). Eight biomarkers were assessed in conjunction with age, gender, and medical

* Corresponding author at: Abbott Laboratories, Diagnostics Division R&D, D09GP, Bdg. AP20/1, 100 Abbott Park Road, Abbott Park, IL 60064, USA.
E-mail address: gerard.davis@abbott.com (G.J. Davis).

https://doi.org/10.1016/j.cca.2019.08.007
Received 27 April 2019; Received in revised form 1 August 2019; Accepted 13 August 2019
Available online 13 August 2019
0009-8981/

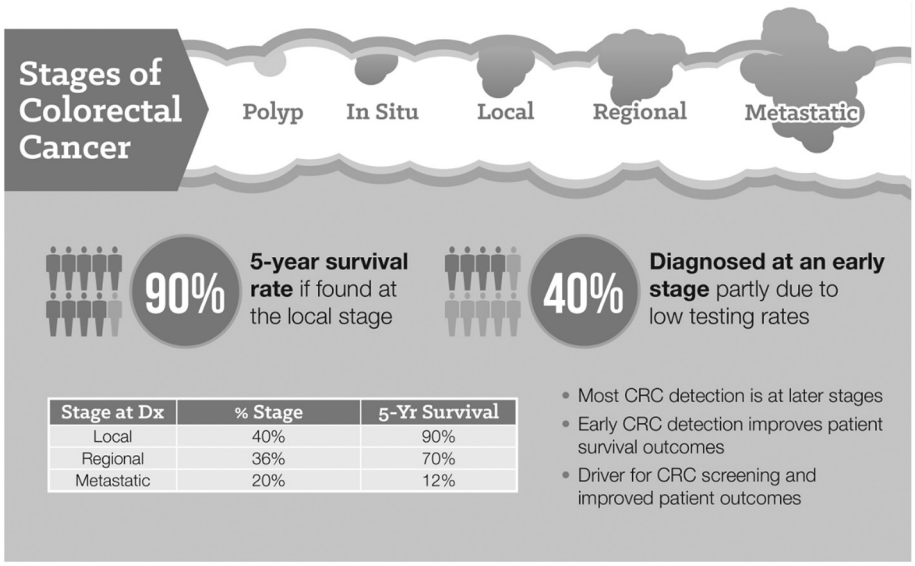

Fig. 1. Patient cancer stage and 5-year survival prognosis.

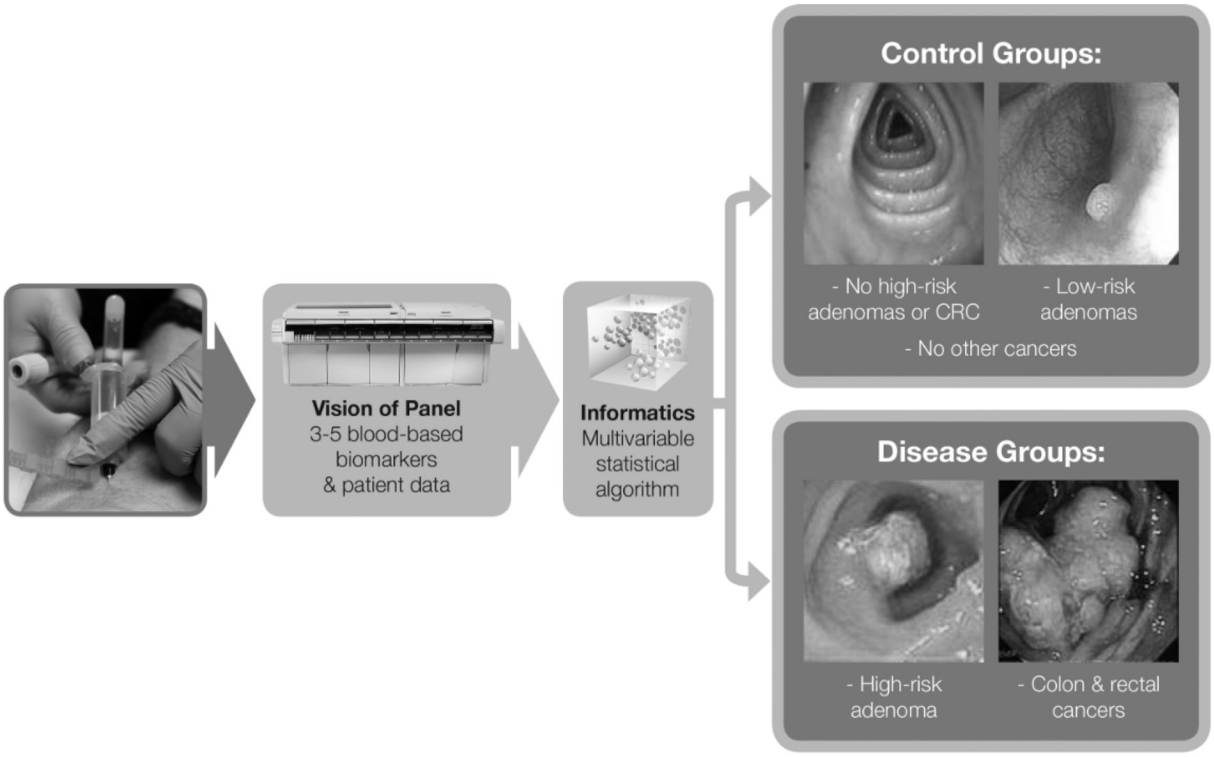

Fig. 2. Primary algorithm indication.

histories to develop combination algorithms that optimally discriminate between groups [6].

The objective of the current study is to compare the clinical performance of different statistical approaches by utilizing a common trial dataset. Two algorithms were developed for each statistical method: a "primary algorithm" and secondary "reflex algorithm" (Fig. 3). The "primary algorithm" was developed to discriminate the presence of CRC and high-risk adenomas from those without (i.e. provide a "positive" or "negative" result). A clinical challenge for screening tests is how to manage false positive results on a primary test, e.g. a patient is positive for the primary test but negative for colonoscopy. As the biomarkers included in these algorithms can be elevated in some other non-CRC cancers as well, an approach was taken to address those false positives in the "primary" algorithm by also generating a "reflex" algorithm for the likelihood of other non-CRC cancers. The "reflex algorithm" was developed to help address a challenge in the biomarker early detection setting, where biomarkers are generally not organ specific. Clinical use CRC screening tests, such as EpiPro Colon (Epigenomics) and Cologuard (Exact Sciences), report that their CRC screening tests can be positive for some lung cancers as well as for other cancers [8,9]. To help address this organ specificity challenge, a "reflex algorithm" approach was taken to identify patients that may have a non-CRC cancer, were initially positive for the "primary algorithm" but had no subsequent colorectal lesion identified upon colonoscopy. "Reflex" algorithms could be developed since 177 other non-CRC cancers were identified during subsequent examination of patients in this trial [6,10].

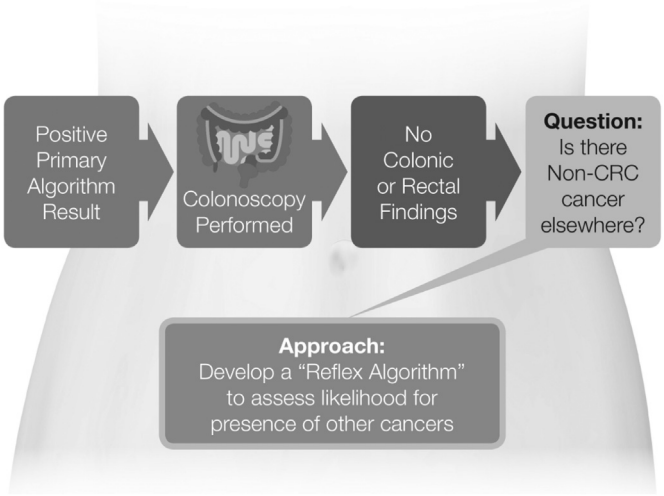

Fig. 3. Reflex algorithm indication.

2. Methods

2.1. Endoscopy II Trial

The Endoscopy II Trial, conducted in 4698 symptomatic (i.e., pain, changed bowel movement habits, visible blood on the stool, weight loss, anemia, and bleeding) subjects who underwent their first colonoscopy and were definitively assessed for rectal and colonic lesions. As previously published, findings included 1978 subjects with no colorectal findings, 1342 subjects with nonmalignant findings (diverticulitis, ulcerative colitis etc.), 174 subjects with rectal adenomas (123 high-risk), 515 subjects with colonic adenomas (276 high-risk), 177 subjects with non-CRC cancers, 193 subjects with rectal cancers, and 319 subjects with colon cancers [6]. In this trial, colonoscopy identified high risk adenomas (HRA), defined as > 1 cm, > 3 adenomas, presence of villous elements, or high-grade neoplasia. Adenomas identified by colonoscopy that did not meet these criteria were classified as low-risk adenomas (LRA). Staging was defined by the criteria of the American Joint Commission on Cancer.

2.2. Marker measurements and outcomes

The Architect® *i2000SR* immunoassay and *c8000* clinical chemistry systems (Abbott Labs.) were utilized to measure AFP, CA19-9, CEA, hs-CRP, CyFra 21-1, Ferritin, Galectin-3, and TIMP-1 levels from Endoscopy II trial EDTA plasma specimens [6,11]. These 8 biomarkers, as well as reliable and audited patient demographic information, such as age, gender, and common co-morbid diseases, were considered as potential markers in the algorithm development. Two outcomes were investigated: 1) the presence of CRC and high risk adenomas versus all others (excluding extra-colonic cancers) and 2) the presence of a non-CRC, or extra-colonic cancer, versus all others (excluding CRCs and high risk adenomas).

2.3. Training and validation data

In order to develop the statistical algorithms and test the performance of these algorithms on "future" patients, the Endoscopy II trial data was split into a training and a validation dataset. The data was split by patient ID number, which allowed for a non-random split of the data in order to avoid simply replicating results in a smaller validation dataset. Eighty percent of the whole cohort was used as the unblinded training dataset, and 20% was left out and blinded to those developing the statistical models. The same training and validation datasets were used to develop each of the models developed from the six different statistical techniques. Both the training and validation datasets had approximately 20% prevalence of the outcome under study, which is a sufficient sample size to develop models for predication.

2.4. Statistical methods

Six statistical methods evaluated the Endoscopy II data; these methods include logistic regression (LR), adaptive index modeling (AIM), partial least-squares discriminant analysis (PLS-DA), feature vector (weighted (IWBF) and unweighted (UVF)), and random forest (RF). These models were built by use of SAS (ver. 9.3), R (ver. 3.03), and MatLab (ver. R2016a) software. Each statistical technique utilizes mechanisms during the training process to optimize, but not over fit, the model. These mechanisms can include internal k-fold cross-validation or resampling techniques. Additionally, some methodologies may re-balance the ratio of samples (those with and without the outcome) to develop a prediction model. The methods are described below.

2.4.1. Logistic regression

Logistic regression is a robust statistical methodology that can be used in a variety of clinical domains and is easily understood [12]. It can be used to model data in which the outcome of interest is binary (i.e., has cancer, does not have cancer). The probability of that outcome can be estimated by the set of predictor variables included in the model. The assumption of this method assumes that the predictor variables are related in a linear fashion to the log odds of the outcome and interactions amongst predictors are difficult to detect. Variable and model selection was based on maximizing the likelihood score [13]. Validation of the selected models were performed using goodness of fit, residuals and 10-fold cross validation of the training dataset. A probability threshold for each algorithm was chosen to classify patients as having the outcome.

2.4.2. Adaptive index model (AIM)

Adaptive index model is a supervised learning method that can be used for linear, logistic or Cox regression analysis [14]. This method adaptively selects a subset of binary rules constructed from predictor variables at an optimal cutoff, selecting variables based on maximizing score test statistics. A 10-fold cross-validation is performed to estimate the best model size. Using the best model, a total AIM score for each patient is calculated, which is the number of binary rules that a patient meets. A threshold value can be set, i.e., the total AIM score that needs to be exceeded in order to classify a patient as having the outcome. The threshold can be selected to maximize a performance characteristic. In this study, the threshold was selected to meet a minimum sensitivity and specificity of 70%, based on inputs provided by key opinion leaders discussions at the U.S. National Cancer Institute's Early Detection Research Network workshops.

2.4.3. Partial least-squares discriminant analysis (PLS-DA)

This method is a multivariable least-squares supervised classification method used for dimensionality reduction and classification. PLS finds components maximize the covariance between the input variables and the dependent variable. The PLS components are linearly combined to predict the dependent variable. PLS-DA can be especially beneficial when there exists a large number of predictors relative to the number of patients or when the variables are noisy and correlated [15].

The training set was used to fit several models, each varying the number of PLS components. To accommodate uneven class sizes while optimizing the classification threshold for a specific model, performance of both classes is taken into account. The threshold is selected to be the one that corresponds to the optimal operating point on the ROC curve, or the point closest to an ideal sensor with sensitivity and specificity equal to 100%. The model with the highest performance was selected as the best model and used to predict the performance in the validation dataset [16], where patients above the optimized threshold

Table 1
(a) Criteria for clinical performance discrimination. (b) Published clinical performance measures for stool-based tests.

(a)

Performance	Sensitivity (%)	Specificity (%)	Sensitivity + Specificity (%)	Performance Basis
High Performance	≥90	≥95	≥185	Similar to Cologuard test (Imperiale et al., [8])
Intermediate Performance	≥70	≥70	≥140	Meets EDRN discussion's "minimum threshold"
Low Performance	< 70	< 70	< 140	Less than EDRN discussion's "minimum threshold"

(b)

Test	Comparison	Sensitivity (%)	Specificity (%)	Sens + Spec (%)	Reference
Cologuard	CRC all stages	92	87	179	Imperiale et al. [8]
	CRC & high risk dysplasia	84	87	171	Imperiale et al. [8]
	Advanced precancerous lesions	42	87	129	Imperiale et al. [8]
FIT	CRC all stages	74	95	169	Imperiale et al. [8]
	CRC all stages	70	95	165	Lee et al. [19]
	CRC all stages*	42	95	137	Adler et al. [5]
	CRC & high risk dysplasia	64	95	159	Imperiale et al. [8]
	CRC & high risk dysplasia*	38	95	133	Adler et al. [5]
	Advanced precancerous lesions	24	95	119	Imperiale et al. [8]
	Advanced precancerous lesions*	14	95	109	Imperiale et al. [8], Adler et al. (2016)

were classified as having the outcome.

2.4.4. Feature vector

In feature vector fusion, predictors are mathematically combined using a set of weights that modulate the predictors. When unweighted (UVF), the predictors are combined with equal weights, giving the same importance to each predictor for discrimination. Alternatively, Image Weighted Bayesian Fusion (IWBF) uses Monte Carlo methods to find a set of weights, such that when combined, weighted predictors optimize outcome discrimination based on sensitivity and specificity [17]. IWBF inherently performs feature selection, so training was done on all available markers. Patients above the optimized threshold were classified as having the outcome.

2.4.5. Random forest

Random forest is an ensemble learning method that can process a large number of predictors simultaneously [18]. An ensemble of trees is developed from both a random sample of patients as well as a random set of predictors in each split. This allows for a diverse creation of trees that can easily allow for nonlinear associations amongst predictors. The ensemble of classification trees is then aggregated for prediction, with the prediction for a patient being the majority classification of the ensemble. The backward elimination method based on feature importance is performed to select the best markers. Only the model with the highest out-of-bag AUC is presented. Additionally, since the share of CRC-positive samples in the dataset is lower than the share of negative samples, an additional analysis of the random forest algorithm is estimated. Instead of a bootstrap sample, each tree is generated on a random sample with an equal number of positive and negative subjects.

2.4.6. Performance evaluation

The diagnostic performance of each statistical method was characterized by several measures of diagnostic accuracy: sensitivity, specificity, sensitivity + specificity, overall diagnostic accuracy, diagnostic odds ratio (DOR), area under receiver operator curve (AUC), and positive predictive value (PPV) and negative predictive value (NPV). Although AUC has been popularized in published literature as an objective single measure used for model performance, a single measure is not sufficient to adequately assess model performance. These aforementioned measures each have specific features related to either discrimination (classification of patients between those who are and those who are not diseased) or predication (estimation of the post-test probability of disease) and together provide a wholistic assessment of model performance.

Sensitivity and specificity can be weighted, prioritizing one over the other based on clinical considerations and needs. Generally, higher sensitivity comes at a specificity cost, and a higher specificity comes at a sensitivity cost. One way to compare sensitivity and specificity performances across algorithms is to add sensitivity and specificity estimates for each algorithm and compare this sum across algorithms. In addition, overall accuracy provides an estimate of total misclassification. The DOR is a measure of the effectiveness and is defined as the ratio of the odds of the algorithm being positive if the patient has the outcome relative to the odds of the algorithm being positive if the patient does not have the outcome. Lastly, the area under the curve provides an overall measure of the diagnostic differentiation of a test.

Predicative measures are most useful for predicting the probability of disease in an individual once the test result is known. PPV and NPV are the proportions of positive and negative results that are true positive and true negative results, respectively.

Using the same training data, optimal or best models were developed independently for each of the 6 different statistical methodologies. The blinded validation data was then used to assess the predictive performance of each of the statistical algorithms developed as an assessment of "future" patients. Further, the predictive performance of the methodologies was benchmarked against criteria established for stool-based tests (Table 1) including Fecal Immunochemical Test or FIT [19], and the published performance of various stool-based testing methodologies that are clinically utilized (Table 1).

3. Results

Results are presented for the optimized best model for each of the 6 statistical techniques using the independent validation dataset. Benchmarking against the criteria established for stool-based tests, the six different statistical methodologies were assessed for their diagnostic performance, as displayed in Table 2a and b. In the primary algorithm models (Table 2a), few of the models meet the criteria established in Table 1a, although the discrimination performance is not directly comparable to the Imperiale et al. report [8] since the disease group endpoints are different than utilized for our studies [5,8]. They do not combine CRC and HRA into one class for a sensitivity calculation, and they do not include greater than or equal to 3 adenomas in the high-risk adenoma class. Generally, colonic adenomas are more difficult lesions to detect as compared to CRC. Imperiale et al. [8] do not include high-risk adenomas defined as greater than or equal to 3 adenomas; rather

Table 2
(a) "Primary Algorithm": CRC & HRA vs. All Others (excluding other non-CRC cancers) (b) "Reflex Algorithm": Non-CRC cancers vs. All Others (excluding CRCs).

							Stratified			
							Female	Male	Female	Male
Performance Measure	**LR**	**PLS-DA**	**RF**	**B - RF**	**UFV**	**IWBF**	**IWBF**	**IWBF**	**AIM**	**AIM**
Sensitivity	77	64	21	50	67	63	64	68	51	80
Specificity	60	55	98	83	57	67	68	57	80	51
Sens + Spec	137	120	119	132	125	130	132	125	131	131
Accuracy	63	62	82	76	59	66	67	60	75	52
DOR	5.0	2.2	12.4	4.7	2.8	3.5	3.7	2.8	4.2	4.2
PPV	33	28	72	42	29	33	27	35	33	35
NPV	91	85	83	87	87	88	91	84	89	89
ROC	0.76	0.65	0.72	0.74	0.67	0.70	0.72	0.70	0.66	0.72
N	896	896	896	896	896	896	490	406	490	406
Number Positive	181	181	181	181	181	181	81	100	81	100
Prevalence 1%										
PPV	2	1	10	3	2	2	2	2	3	2
NPV	100	99	99	99	99	99	99	99	99	100
Prevalence 5%										
PPV	9	7	36	13	8	9	10	8	12	8
NPV	98	97	96	97	97	97	97	97	97	98
Prevalence 10%										
PPV	34	19	49	33	22	26	28	23	30	30
NPV	91	90	93	91	90	91	91	90	91	91
Prevalence 15%										
PPV	45	27	61	43	31	36	38	32	40	41
NPV	86	85	89	86	86	86	86	86	86	86

							Stratified			
							Female	Male	Female	Male
Performance Measure	**LR**	**PLS-DA**	**RF**	**B - RF**	**UFV**	**IWBF**	**IWBF**	**IWBF**	**AIM**	**AIM**
Sensitivity	70	78	35	56	76	73	35	67	71	59
Specificity	60	64	99	91	79	84	92	76	80	78
Sens + Spec	130	142	134	147	155	157	165	143	152	137
Accuracy	60	78	96	89	79	83	91	76	80	77
DOR	3.5	6.3	95.2	13.1	11.8	13.9	30.5	6.4	9.8	5.1
PPV	9	12	79	28	14	17	26	13	16	16
NPV	97	98	96	97	99	99	99	98	98	96
ROC	0.75	0.75	0.83	0.82	0.84	0.86	0.92	0.82	0.79	0.72
N	758	758	758	758	758	758	430	328	430	328
Number Positive	43	43	43	43	43	43	21	22	21	22
Prevalence 1%										
PPV	2	2	26	6	4	4	4	3	3	3
NPV	99	100	99	100	100	100	99	100	100	99
Prevalence 5%										
PPV	8	10	65	25	16	19	19	13	16	12
NPV	97	98	97	98	98	98	96	98	98	97
Prevalence 10%										
PPV	27	39	68	52	53	56	37	39	48	34
NPV	91	91	96	92	91	92	91	91	91	91
Prevalence 15%										
PPV	37	50	77	64	64	67	48	50	60	45
NPV	86	86	94	88	87	87	87	86	87	86

they report advanced precancerous lesion with sensitivity of 46% at specificity of 87% which is substantially lower than CRC only endpoint performance.

Additionally, there is no benchmark for the reflex algorithm.

Regardless of statistical technique, every best model utilized CEA, Cyfa21–2 and hs-CRP, and all except one used Ferritin, for the primary algorithm model. Similarly, for the reflex algorithm, every best model utilized Cyfa21–2 and hs-CRP, and all except one built the best model using CEA. Model complexity ranged from 3 to 9 markers, including clinical parameters of gender and age (Table 3).

Table 3
Model Comparisons for (a) "Primary Algorithm": CRC & HRA vs. All Others (excluding other non-CRC cancers) (b) "Reflex Algorithm": Non-CRC cancers vs. All Others (excluding CRCs).

(a)

Model	Stratification	Training		Validation		Gender	Age	AFP	CA19–9	CEA	Cyfra 21-1	Ferritin	hs-CRP	Galectin-3	TIMP-1	# Markers	Output	Optimization	Positive Result
		N	No. Positive	N	No. Positive														
Logistic Regression	None	3623	730	896	181	x	x			x	x	x	x			6	Probability	60% specificity	> 15.5% probability considered positive
PLS-DA	None	3623	730	896	181			x	x	x	x	x	x	x	x	8	Probability	Maximized SE and SP	> optimized threshold considered positive
Random Forest	None	3623	730	896	181	x	x			x	x	x	x	x		7	Each tree votes	Maximized SE and SP	Majority classification
	Balanced	1460	730	896	181	x	x			x	x	x	x	x		7	Each tree votes	Maximized SE and SP	Majority classification
UVF	None	3623	730	896	181			x	x	x	x	x	x	x	x	8	Fusion score	Maximized SE and SP	> optimized threshold considered positive
IWBF	Females	1887	308	490	81		x	x	x	x	x	x	x	x	x	9	Fusion score	Maximized SE and SP	> optimized threshold considered positive
	Males	1736	422	406	100		x	x	x	x	x	x	x	x	x	9	Fusion score	Maximized SE and SP	> optimized threshold considered positive
AIM	Females	1887	308	490	81					x	x		x			3	AIM Score	Minimum 70% SE and SP	≥1 considered positive
	Males	1736	422	406	100		x		x	x	x	x	x			6	AIM Score	Minimum 70% SE and SP	≥2 considered positive

(b)

Model	Stratification	Training		Validation		Gender	Age	AFP	CA19–9	CEA	Cyfra 21-1	Ferritin	hs-CRP	Galectin-3	TIMP-1	# Markers	Output	Optimization	Positive Result
		N	No. Positive	N	No. Positive														
Logistic Regression	None	3027	134	758	43	x	x			x	x	x	x			6	Probability	60% specificity	> 2.2% probability considered positive
PLS-DA	None	3027	134	758	43			x	x	x	x	x	x	x	x	8	Probability	Maximized SE and SP	> optimized threshold considered positive
Random Forest	None	3027	134	758	43	x	x		x	x	x	x	x	x		8	Each tree votes	Maximized SE and SP	Majority classification
	Balanced	268	134	758	43	x	x		x	x	x	x	x	x		8	Each tree votes	Maximized SE and SP	Majority classification
UVF	None	3027	134	758	43			x	x	x	x	x	x	x	x	8	Fusion score	Maximized SE and SP	> optimized threshold

(*continued on next page*)

Table 3 (*continued*)

(b)

Model	Stratification	Training		Validation		Gender	Age	AFP	CA19–9	CEA	Cyfra 21-1	Ferritin	hs-CRP	Galectin-3	TIMP-1	# Markers	Output	Optimization	Positive Result
		N	No. Positive	N	No. Positive														
IWBF	Females	1636	57	430	21		x	x	x	x	x	x	x	x	x	9	Fusion score	Maximized SE and SP	considered positive > optimized threshold considered positive
	Males	1391	77	328	22		x	x	x	x	x	x	x	x	x	9	Fusion score	Maximized SE and SP	considered positive > optimized threshold considered positive
AIM	Females	1636	57	430	21					x	x		x			3	AIM Score	Minimum 70% SE and SP	≥ 1 considered positive
	Males	1391	77	328	21						x		x		x	3	AIM Score	Minimum 70% SE and SP	≥ 1 considered positive

4. Discussion

The objective of this study was to evaluate several common and emerging statistical methodologies applied to a common set of clinical data to develop algorithms for clinical discrimination of high risk adenoma and CRC malignancy patients from all other subjects (not having these colonic lesions), to help assess which statistical methodology, given their strengths and weaknesses, may perform better than others. This head-to-head approach provided comparisons for six methodologies to help understand how these methodologies performed in the aforementioned clinical discrimination indications using a validation dataset to simulate future patients. Two clinical discriminations were evaluated (Fig. 3), including a "primary algorithm" as a front-line discrimination of HRA and CRC from all others, and a "reflex algorithm" to address those that had been positive by the primary algorithm but negative upon colonoscopy to assess the likelihood of the presence of other non-CRC malignancies. These pilot algorithms are focused on addressing early CRC detection with a blood-based test as opposed to stool-based methodologies currently used in clinical settings.

These algorithms are discussed further below in terms of their clinical accuracy, complexity and interpretability, and ease of clinical adoption and implementation.

4.1. Clinical accuracy

In a clinical setting, the key concern is whether the diagnostic procedure predicts disease status [20]. Although popular, sensitivity and specificity describe how the disease predicts diagnostic test results. It is the NPV and PPV values that inform the clinician and the patient the probability of disease given the test result. For the primary algorithm, NPV represents the proportion of patients correctly predicted to be without neoplasia, while PPV represents the proportion of patients correctly predicted to have CRC. A high PPV indicates that the model accurately classifies a CRC patient, and thus is a desirable clinical characteristic for screening. In this population where prevalence is approximately 20%, evaluating the six statistical methodologies with respect to NPV and PPV indicates that all methods perform comparably regarding NPV on the primary algorithm, but only the RF model performs well on PPV. Further, for the reflex algorithm, a high NPV would be optimal to spare patients from undergoing a subsequent and unnecessary colonoscopy. The results of the models in settings with different prevalence rates are also shown for comparison in Table 2a and b.

4.2. Complexity and interpretability

Partial Least-Squares Discriminant Analysis (PLS-DA), Feature Vectors and Random Forest methods are considered "black-box" statistical techniques, since the logic to determine the best model is not transparent to the clinician. Although these methods can perform better than traditional statistical techniques, this performance is not uniform. If the same predictive accuracy can be attained by a method such as logistic regression, which is generally understood by clinicians and statisticians alike, it is difficult to justify the use of such methods in a clinical setting. Model complexity can also be judged based on variable stratification, which may aid in clinical interpretation, and the number of markers included in the model. Model complexity ranged from 3 to 9 markers, including clinical parameters of gender and age.

4.3. Clinical adoption and implementation

To gain clinical adoption and implementation, the proposed model needs to be clinically meaningful. Clinical adoption and implementation in a clinical setting would be dependent on the familiarity with the statistical method, the kind and number of markers in the model, the understanding of the interpretation of results and the manner in which

the clinician could receive the final interpretation. For data mining methods, statistical software would be required to interact with the collection of data in order to calculate and generate a result for the clinician. Implementation of these "black-box" solutions in the laboratory thus may require hardware, software and IT personnel involvement. Currently, other methods developed here, such as logistic regression and AIM, are much easier to implement, interpret and adopt in a clinical setting. Other strategies, such as stratifying by a key variable, can also aid in implementing and interpreting the model in a clinical setting. However, with development of newer software and digital methodologies, implementing machine learning models may become more prevalent in clinical settings.

5. Conclusion

Like other cancers, CRC is considered a heterogenous cancer, and it is generally assumed that a single biomarker will not be able to detect all cases of CRC. Thus, combining multiple biomarkers and clinical information has a greater chance of success in detecting CRC. The current study has yielded "proof of concept" algorithms for the early detection of HRA and CRC utilizing patient demographic and biomarker information. The clinical performance of these algorithms need further enhancement, but this evaluation positively indicates that combining clinical and biomarker data combined in a statistical algorithm can be used to address challenging clinical problems. Future investigations will be to expand the base of biomarkers from which to develop algorithms to further increase the diagnostic performance. Biomarkers that are complementary markers in terms of biological or pathological mechanisms to existing ones would be the focus of this expansion.

Acknowledgement

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The clinical study received financial support from The Augustinus Foundation, The Becket Fund, The Inger Bonnén Fund, The Hans & Nora Buchard Fund, CEO Jens Bærentsen (private donation), The Walter Christensen Family Fund, The P.M. Christiansen Family Fund, The Aase & Ejnar Danielsen Fund, The Erichsen Family Fund, The Knud & Edith Eriksen Fund, The Svend Espersen Fund, The Elna and Jørgen Fagerholt Fund, The Sofus Carl Emil Friis Fund, The Torben & Alice Frimodt Fund, The Eva & Henry Frænkel Fund, The Gangsted Fund, The Thora & Viggo Grove Fund, The Humanitarian Foundation, The Erna Hamilton Fund, The Sven & Ina Hansen Fund, The Søren & Helene Hempel Fund, The Henrik Henriksen Fund, The Jørgen Holm Family Fund, Foundation Jochum, The KID Fund, The Kornerup Fund, The Linex Fund, The Dagmar Marshall's Fund, The "Midtjyske Bladfund," The Axel Muusfeldt Fund, The Michael Hermann Nielsen Fund, The Arvid Nilsson Fund, The Obel Family Fund, The Krista & Viggo Petersen Fund, The Willy & Ingeborg Reinhard Fund, The Kathrine & Vigo Skovgaard Fund, The Toyota Fund, The Vissing Fund, The Wedell-Wedellsborg Fund and Hvidovre University Hospital (The Capital Region of Denmark). Abbott Laboratories Inc., Abbott Park, IL, USA, sponsored the protein analyses at their Center of Excellence, VUMC, Amsterdam, The Netherlands. The study was partly sponsored by an unrestricted grant from Abbott Laboratories Inc. to Hvidovre Hospital (H.J.N.). The ChemImage research was supported, in part, by Abbott and, in part, by ChemImage.

References

[1] American Cancer Society, Colorectal Cancer Facts & Figures 2014–2016, American Cancer Society, Atlanta, 2014https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/colorectal-cancer-facts-and-Fig.s/colorectal-cancer-facts-and-Fig.s-2014-2016.pdf.

[2] C.H. Kubisch, A. Crispin, U. Mansmann, B. Göke, F.T. Kolligs, Screening for colorectal cancer is associated with lower disease stage: a population-based study, Clin. Gastroenterol. Hepatol. 14 (2016) 1612–1618.

[3] M.V. Maciosek, L.I. Solberg, A.B. Coffield, N.M. Edwards, M.J. Goodman, Colorectal cancer screening: health impact and cost effectiveness, Am. J. Prev. Med. 31 (2006) 80–89.

[4] U.S. Preventive Services, Task force recommendations on Colon Cancer, Screening 315 (23) (2013) 2564–2575 https://www.uspreventiveservicestaskforce.org/Page/Document/RecommendationStatementFinal/colorectal-cancer-screening.

[5] A. Adler, S. Geiger, A. Keil, H. Bias, P. Schatz, T. deVos, J. Dhein, M. Zimmermann, R. Tauber, B. Wiedenmann, Improving compliance to colorectal cancer screening using blood and stool bases tests in patients refusing screening colonoscopy in Germany, BMC Gastroenterol. 14 (183) (2014), http://www.biomedcentral.com/1471-230X/14/183.

[6] M. Wilhelmsen, I.J. Christensen, L. Rasmussen, L.N. Jørgensen, M.R. Madsen, J. Vilandt, T. Hillig, M. Klærke, K.T. Nielsen, S. Laurberg, N. Brünner, S. Gawel, X. Yang, G. Davis, A. Heijboer, F. Martens, H.J. Nielsen, Detection of colorectal neoplasia: combination of eight blood-based, cancer-associated protein biomarkers, Int. J. Cancer 140 (6) (2017) 1436–1446.

[7] D. Hanahan, R.A. Weinberg, Hallmarks of cancer: the next generation, Cell 144 (2011) 646–674.

[8] T.F. Imperiale, D.F. Ransohoff, S.H. Itzkowitz, T.R. Levin, P. Lavin, G.P. Lidgard, D.A. Ahlquist, B.M. Berger, Multitarget stool DNA testing for colorectal-cancer screening, N. Engl. J. Med. 370 (14) (2014) 1287–1297.

[9] R. Grutzmann, B. Molnar, C. Pilarsky, J.K. Habermann, P.M. Schlag, H.D. Saeger, S. Miehlke, et al., C. Lofton-Day, Sensitive detection of colorectal cancer in peripheral blood by Septin 9 DNA methylation assay, PLoS One 3 (11) (2008) e3759.

[10] M. Wilhelmsen, I.J. Christensen, L.N. Jørgensen, M.R. Madsen, J. Vilandt, T. Hillig, et al., Clean colorectum at diagnostic colonoscopy: subsequent detection of extracolonic malignancies by plasma protein biomarkers? Biomarkers Cancer 10 (2018) 1–9.

[11] F.A. Quinn, Architect® i2000® and i2000$_{SR}$® analyzers, in: D. Wild (Ed.), The Immunoassay Handbook, 3rd ed., Elsevier Ltd, Amsterdam, Netherlands, 2005, pp. 406–411.

[12] V. Bewick, L. Cheek, J. Ball, Statistics review 14: logistic regression, Clin. Care 9 (1) (2005) 112–118.

[13] G.M. Furnival, R.W. Wilson, Regression by leaps and bounds, Technometrics 16 (1974) 499–511.

[14] L. Tian, R. Tibshirani, Adaptive index models for marker-based risk stratification, Biostatistics 12 (1) (2011) 68–86.

[15] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2) (2001) 109–130 ISSN 0169-7439 https://doi.org/10.1016/S0169-7439(01)00155-1.

[16] T. Hastie, R. Tibshirani, J.H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, 2013.

[17] F. Cremer, K. Schutte, J.G.M. Schavemaker, E. den Breejen, A comparison of decision-level sensor-fusion methods for anti-personnel landmine detection, Information Fusion 2 (3) (2001) 187–208 ISSN 1566-2535 https://doi.org/10.1016/S1566-2535(01)00034-3.

[18] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[19] J.K. Lee, E.G. Liles, S. Bent, T.R. Levin, D.A. Corley, Accuracy of fecal immunochemical tests for colorectal cancer, Ann. Intern. Med. 160 (2014) 171-181.

[20] U.M. Okeh, C.N. Okoro, Evaluating measures of indicators of diagnostic test performance: fundamental meanings and formulars, J. Biomet. Biostat. 3 (2012) 132.