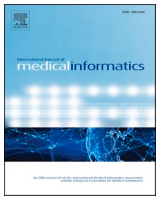
International Journal of Medical Informatics 178 (2023) 105194

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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



Development and validation of colorectal cancer risk prediction tools: A comparison of models

Duco T. Mülder [a,*], Rosita van den Puttelaar [a], Reinier G.S. Meester [a,b], James F. O'Mahony [a,c], Iris Lansdorp-Vogelaar [a]

[a] Department of Public Health, Erasmus Medical Center, Rotterdam, Netherlands
[b] Health Economics & Outcomes Research, Freenome Holdings Inc., San Francisco, CA, USA
[c] Centre for Health Policy & Management, Trinity College Dublin, Dublin, Ireland

ARTICLE INFO

Keywords:
Colorectal Cancer screening
Prognostic models
Risk stratification

ABSTRACT

Background: Identification of individuals at elevated risk can improve cancer screening programmes by permitting risk-adjusted screening intensities. Previous work introduced a prognostic model using sex, age and two preceding faecal haemoglobin concentrations to predict the risk of colorectal cancer (CRC) in the next screening round. Using data of 3 screening rounds, this model attained an area under the receiver-operating-characteristic curve (AUC) of 0.78 for predicting advanced neoplasia (AN). We validated this existing logistic regression (LR) model and attempted to improve it by applying a more flexible machine-learning approach.

Methods: We trained an existing LR and a newly developed random forest (RF) model using updated data from 219,257 third-round participants of the Dutch CRC screening programme until 2018. For both models, we performed two separate out-of-sample validations using 1,137,599 third-round participants after 2018 and 192,793 fourth-round participants from 2020 onwards. We evaluated the AUC and relative risks of the predicted high-risk groups for the outcomes AN and CRC.

Results: For third-round participants after 2018, the AUC for predicting AN was 0.77 (95% CI: 0.76–0.77) using LR and 0.77 (95% CI: 0.77–0.77) using RF. For fourth-round participants, the AUCs were 0.73 (95% CI: 0.72–0.74) and 0.73 (95% CI: 0.72–0.74) for the LR and RF models, respectively. For both models, the 5% with the highest predicted risk had a 7-fold risk of AN compared to average, whereas the lowest 80% had a risk below the population average for third-round participants.

Conclusion: The LR is a valid risk prediction method in stool-based screening programmes. Although predictive performance declined marginally, the LR model still effectively predicted risk in subsequent screening rounds. An RF did not improve CRC risk prediction compared to an LR, probably due to the limited number of available explanatory variables. The LR remains the preferred prediction tool because of its interpretability.

1. Introduction

The faecal immunochemical test (FIT) measures the faecal haemoglobin (f-Hb) concentration in stool samples and is used in screening for colorectal cancer (CRC). Screening programmes generally invite people for FIT-screening every 1 or 2 years and a follow-up colonoscopy is performed if the f-Hb concentration exceeds a certain threshold [1]. Although screening participants are repeatedly tested, only the current f-Hb concentration is typically used as an indicator of (advanced) colorectal neoplasia.

Previous work introduced a logistic regression (LR) model using sex, age and two preceding f-Hb concentrations to predict the risk of CRC in the next screening round [2]. Using this LR model, 5% of the participants with the highest predicted risk had a 6-fold risk of CRC compared to the population average [2]. Although the model was able to identify high-risk individuals, its validity for out-of-sample performance and subsequent screening rounds were unknown.

Valid risk prediction could permit effective risk-stratified screening, through personalizing the screening interval or the test positivity cut-off based on risk. For example, people with repeated negative FITs with f-Hb concentrations close to the positivity cut-off could be reinvited to FIT sooner or be referred to colonoscopy despite not meeting the

* Corresponding author.
E-mail address: d.t.mulder@erasmusmc.nl (D.T. Mülder).

https://doi.org/10.1016/j.ijmedinf.2023.105194
Received 20 December 2022; Received in revised form 5 July 2023; Accepted 8 August 2023
Available online 16 August 2023

conventional positivity criteria. Conversely, people without any traces of f-Hb in their stool could be reinvited after an extended interval. Several studies have indicated that risk-stratification can enhance the benefits of screening, but only when the risk prediction tool demonstrates adequate quality [3–5]. Thus, it is important that any proposed basis for risk stratification is carefully validated.".

Machine-learning offers alternative methods for accurate risk prediction. A random forest (RF) model – a machine-learning technique known for its parsimonious configuration and predictive ability - often outperforms regression models in terms of accuracy and precision [6]. Because an RF model combines multiple decision trees, this model can fit to input data closer than generalized linear models such as LR [7]. Previous research using clinical data showed that RFs can be used as prognostic systems for risk stratification in oropharyngeal and breast cancer [8–9]. RFs have so far not been used in the prediction of CRC risk based on prior f-Hb concentrations.

In this retrospective study, we aimed to validate the existing LR model and to investigate whether the model could be improved on with machine-learning. Most FIT-based CRC screening programmes have only started within the past decade and the number of screening rounds is still accumulating [10], thus validation of the model for subsequent screening rounds could particularly be relevant for risk-stratification purposes. This is the first validation study of the LR model using data from a nationwide screening programme and the first study that uses an RF to predict CRC risk based on previous f-Hb results.

2. Materials and methods

2.1. The Dutch screening programme

Data of the first four rounds of the Dutch CRC screening programme were used. The programme started in 2014 with a gradual implementation of FIT-based screening by birth cohort for all Dutch citizens aged between 55 and 75. A participant is invited for a follow-up colonoscopy if the f-Hb concentration exceeds a positivity threshold. This threshold was initially set at 15 micrograms of haemoglobin per gram (μg Hb/g) faeces, but was increased to 47 μg Hb/g faeces six months after the introduction of the screening programme due to colonoscopy capacity constraints [11]. During the follow-up colonoscopy, advanced neoplasia (AN) is considered as relevant finding. AN consists of the presence of either advanced adenomas or CRC. More details on the programme are described elsewhere [11].

2.2. Data

We split the data into a training set and two validation sets. For both the LR and the newly developed RF model, the training data contained all third-round participants up to 2018. Validation set 1 contained third-round participants between 2019 and 2021. Validation set 2 contained all fourth-round participants in 2020 and 2021. More details on the storage and format of the data can be found in the supplementary material (Appendix 1).

2.3. Prediction models

An individual's risk of AN or CRC was predicted based on age, sex, and the two most recent f-Hb concentrations before the last screening round: round 1–2f-Hb concentrations for prediction of round 3findings, and round 2–3f-Hb concentrations for the prediction of round 4 findings. The models were trained for two separate binary outcome variables: AN and CRC detection. The multivariate LR model used a discretization of the continuous f-Hb concentrations, similar to Meester et al: 0 μg/g, 0.1–9.9 μg/g, 10.0–19.9 μg/g, 20.0–29.9 μg/g, 30.0–39.9 μg/g and 40.0–46.9 μg/g [2]. The multivariate odds ratios (ORs) and 95% confidence intervals (CIs) were obtained for all covariates.

In addition to the LR, we applied an RF model. RFs are based on decision tree learning, by generating multiple decision trees using different subsets (with repeated samples) of the data [12]. The prediction is the average of the optimal decision trees [13]. The number of decision trees was set at 100, as a result of parameter tuning, considering performance and training speed, with 5 variables per decision tree as the default [14]. The RF model included the same inputs as the LR model, with the f-Hb inputs using original continuous values and an indicator variable of 1 if the two previous f-Hb concentrations were zero. Because RFs do not provide model parameters like LR, variable importance was evaluated using an earlier developed ranking method [15].

2.4. Model evaluation

The models' discriminate ability was assessed using receiver-operator-characteristic (ROC) curves, evaluating the true-positive and false-positive rates at different thresholds, and the corresponding area under the curve (AUC). The AUC indicates the probability that an individual with the outcome variable obtains a higher predicted risk than an individual without the outcome variable [16], with an AUC of 0.5 indicating random predictions and an AUC of 1 indicating perfect predictions. The variety in AUC labelling systems is substantial, but a value of 0.8 is generally considered to indicate a good performance [17]. CIs around the AUC were computed using the bootstrap method [18] with 100 iterations.

Observed relative risk plots were further used to visualize the models' discrimination between high-risk and low-risk groups. Participants were ranked by predicted risk and divided into equal-sized subgroups, where the proportion of participants with AN or CRC was computed and divided by the overall population risk. Models with good discriminative power yield high levels of observed relative risk for predicted high-risk subgroups and low relative risk for other groups. A relative risk of 1 corresponds to the population average. Calibration was assessed by plotting predicted relative risk against observed relative risk for all subgroups, aiming for close alignment. Using 20 subgroups, we were able to visually inspect both discrimination and calibration with our relative risk plots.

2.5. Sensitivity analysis

A sensitivity analysis was conducted by training the models only on participants with a FIT positivity cut-off of 47 μg Hb/g, hence excluding those with a cut-off of 15 μg Hb/g. This was compared to the base-case analysis in which those with a cut-off of 15 μg Hb/g were also included.

2.6. Software

All analyses were performed using R statistical software V4.0.4. The randomForest V4.6 package was used for the RF model.

3. Results

3.1. Study population

A total of 1,535,860 participants were identified in the third screening round between 2014 and 2020. We excluded participants with missing FITs in round 1 and 2 (n = 166,898), positive FITs in round 1 and 2 (n = 2,356), missing findings in the participant records on sex and/or age (n = 96) and participants with a positive FIT but no follow-up colonoscopy in round 3 (n = 9,650). The remaining 1,356,860 participants were split based on the year of invitation. Data from 219,258 third-round participants in 2018 and data from their two prior rounds in 2014–2016 were used to train our models. Data of the 1,137,602 third-round participants between 2019 and 2021 were used as validation data (Fig. 1).

There were 209,916 participants in the fourth screening round. Similar to round three, we excluded participants with missing FITs in

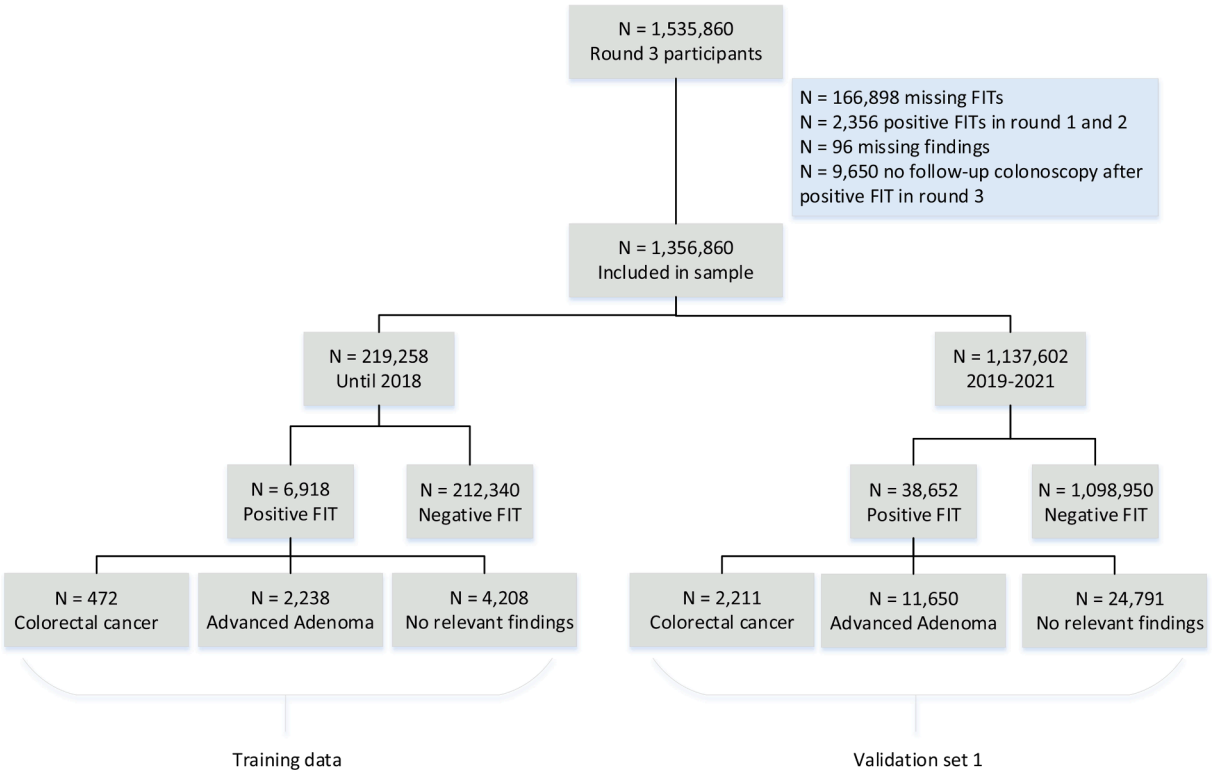


Fig. 1. Study flow diagram and outcomes of round 3 participants. FIT, faecal immunochemical test.

round 2 and 3 (n = 15,182), positive FITs in round 1, 2 or 3 (n = 378), missing findings (n = 13) and participants with a positive FIT, but no follow-up colonoscopy in round 4 (n = 1,550). The remaining 192,793 participants formed the sample of the second validation set (Fig. 2).

Descriptive statistics on the data can be found in Table 1.

3.2. Model specifications

The multivariate LR model showed that male sex and all categories > 0 µg Hb/g faeces of two previous f-Hb level were statistically significant predictors for both AN and CRC (Table 2). For AN, the multivariate

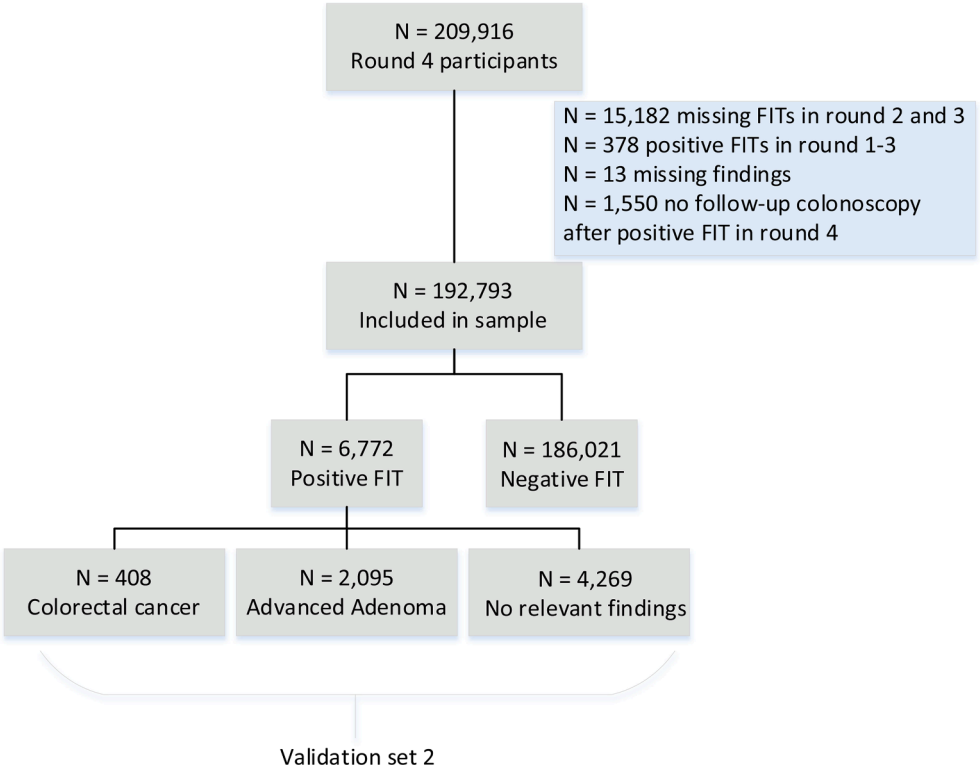


Fig. 2. Study flow diagram and outcomes of round 4 participants. FIT, faecal immunochemical test.

Table 1
Descriptive statistics.

		Training data		Validation set 1		Validation set 2	
		n	**%**	**n**	**%**	**n**	**%**
Total population	N	219,257	100%	1,137,599	100%	192,793	100%
Lesion	AA	2,237	1.0%	11,648	1.0%	2,095	1.1%
	CRC	472	0.2%	2,209	0.2%	408	0.2%
Sex	male	105,079	47.9%	533,189	46.9%	91,963	47.7%
	female	114,178	52.1%	604,410	53.1%	100,830	52.3%
	total	219,257	100%	1,137,599	100%	192,793	100%
Age	mean	69		68		71	
(at last available screening round)	SD	2		3		2	
	62–64	1,787	0.8%	165,177	14.5%	0	0.0%
	65–67	46,491	21.2%	462,425	40.6%	7,933	4.1%
	68–70	92,775	42.3%	192,631	16.9%	51,057	26.5%
	71–73	78,204	35.7%	247,276	21.7%	133,586	69.3%
	74–75	0	0.0%	70,090	6.2%	217	0.1%
f-Hb two rounds ago, µg/g	0	171,052	78.0%	961,891	84.6%	177,691	92.2%
	0.1–2.5	22,779	10.4%	58,536	5.1%	3,519	1.8%
	2.6–9.9	16,361	7.5%	68,125	6.0%	5,521	2.9%
	10–19.9	5,525	2.5%	27,965	2.5%	2,972	1.5%
	20–29.9	1,827	0.8%	10,389	0.9%	1,418	0.7%
	30–39.9	1,075	0.5%	6,756	0.6%	1,042	0.5%
	40–46.9	638	0.3%	3,937	0.3%	630	0.3%
f-Hb previous round, µg Hb/g	0	199,619	91.0%	1,054,056	92.7%	182,131	94.5%
	0–2.5	4,286	2.0%	18,228	1.6%	1,656	0.9%
	2.6–9.9	7,019	3.2%	27,334	2.4%	3,214	1.7%
	10–19.9	4,000	1.8%	16,316	1.4%	2,372	1.2%
	20–29.9	1,981	0.9%	9,527	0.8%	1,433	0.7%
	30–39.9	1,442	0.7%	7,441	0.7%	1,205	0.6%
	40–46.9	910	0.4%	4,697	0.4%	782	0.4%

Abbreviations: AA, advanced adenomas; CRC, colorectal cancer; f-Hb, faecal haemoglobin; SD, standard deviation; µg Hb/g, micrograms per gram.

Table 2
Specification multivariate regression model.

	Advanced Neoplasia			Colorectal cancer		
	OR	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
Age	1.0	1.0 to 1.0	0.25	1.0	1.0 to 1.1	0.1
Male sex	1.3	1.2 to 1.4	<0.001	1.2	1.0 to 1.5	0.0
f-Hb concentration, two rounds ago, µg/g						
0	Ref.			Ref.		
0.1–9.9	2.5	2.3 to 2.7	<0.001	1.9	1.6 to 2.4	<0.001
10.0–19.9	4.6	4.0 to 5.3	<0.001	3.6	2.5 to 5.0	<0.001
20.0–29.9	6.0	5.3 to 6.8	<0.001	5.6	3.7 to 8.6	<0.001
30.0–39.9	7.5	6.0 to 9.4	<0.001	3.8	2.0 to 7.1	<0.001
40.0–46.9	10.9	8.5 to 14.1	<0.001	7.6	4.2 to 13.7	<0.001
f-Hb concentration, previous round, µg/g						
0	Ref.			Ref.		
0.1–9.9	4.3	3.9 to 4.8	<0.001	3.6	2.8 to 4.7	<0.001
10.0–19.9	6.3	5.5 to 7.3	<0.001	5.6	4.1 to 7.7	<0.001
20.0–29.9	8.5	7.2 to 10.0	<0.001	4.1	2.5 to 6.7	<0.001
30.0–39.9	8.8	7.3 to 10.6	<0.001	5.5	3.4 to 9.0	<0.001
40.0–46.9	9.4	7.5 to 11.8	<0.001	5.6	3.1 to 10.2	<0.001

Abbreviations: OR, odds ratio; CI, confidence interval; f-Hb, faecal haemoglobin; µg/g, micrograms per gram; Ref., reference category.

OR (95% CI) varied between 2.5 (2.3–2.7) for the lowest level > 0 and 10.9 (8.5–14.1) for the highest level of measured f-Hb concentrations two prior rounds, compared to 0f-Hb. ORs (95% CI) for f-Hb concentrations of the previous round varied between 4.3 (CI: 3.9–4.8) and 9.4 (CI: 7.5–11.8). For CRC, the OR (95% CI) varied between 1.9 (1.6–2.4) and 7.6 (4.2–13.7) for f-Hb levels two prior rounds; and between 3.6 (2.8–4.7) and 5.6 (CI: 3.1–10.2) for the previous round. The RF model showed that the f-Hb values contributed most to the predictions for both AN and CRC. The feature importance of age and sex were limited.

3.3. Performance on validation set 1

In validation set 1, the LR model attained an AUC (95% CI) of 0.77 (0.76–0.77) for AN and 0.73 (0.72–0.75) for CRC. These values were similar to the values of the training set of 0.78 (0.77–0.79) and 0.73 (0.71–0.75) (Table 3). For both AN and CRC, we observed similar shapes of the ROC-curve for the LR and the RF model (Fig. 3). The RF model attained an AUC (95% CI) of 0.77 (0.77–0.77) for AN and 0.73 (0.72–0.74) for CRC (Table 3).

The discriminative performance of the LR model was similar to the RF model for AN and for CRC. Both models yielded relative risks less than 0.6 of AN and 0.9 of CRC for the 80% of the participants with lowest predicted risk compared to the population average. The 5% participants with the highest predicted risk had a 7.0-fold (6.9-fold) risk of AN and 5.8-fold (5.3-fold) risk of CRC compared to the population average according to the LR (RF). The predicted relative risks were close to the observed relative risks for both models investigated (Fig. 4).

3.4. Performance on validation set 2

In validation set 2 with fourth-round data, the LR model attained lower AUC values (95% CI) compared to the observed values for third-round data: 0.73 (0.72–0.74) for AN and 0.68 (0.66–0.71) for CRC (Table 3). For both AN and CRC, shapes of the ROC-curve for the LR and

Table 3
Values of the area under the curve (AUC) for the training data (third-round participants up 2014–2018), validation set 1 (third-round participants 2019–2021) and validation set 2 (fourth-round participants 2020–2021).

	Outcome	AUC Logistic Regression (95% CI)	AUC Random Forest (95% CI)
Training data	AN	0.78 (0.77–0.79)	0.79 (0.78–0.80)
	CRC	0.73 (0.71–0.75)	0.76 (0.74–0.78)
Validation set 1	AN	0.77 (0.76–0.77)	0.77 (0.77–0.77)
	CRC	0.73 (0.72–0.75)	0.73 (0.72–0.74)
Validation set 2	AN	0.73 (0.72–0.74)	0.73 (0.72–0.74)
	CRC	0.68 (0.66–0.71)	0.68 (0.65–0.72)

Abbreviations: AUC, area under the curve; CI, confidence interval; AN, advanced neoplasia; CRC, colorectal cancer.

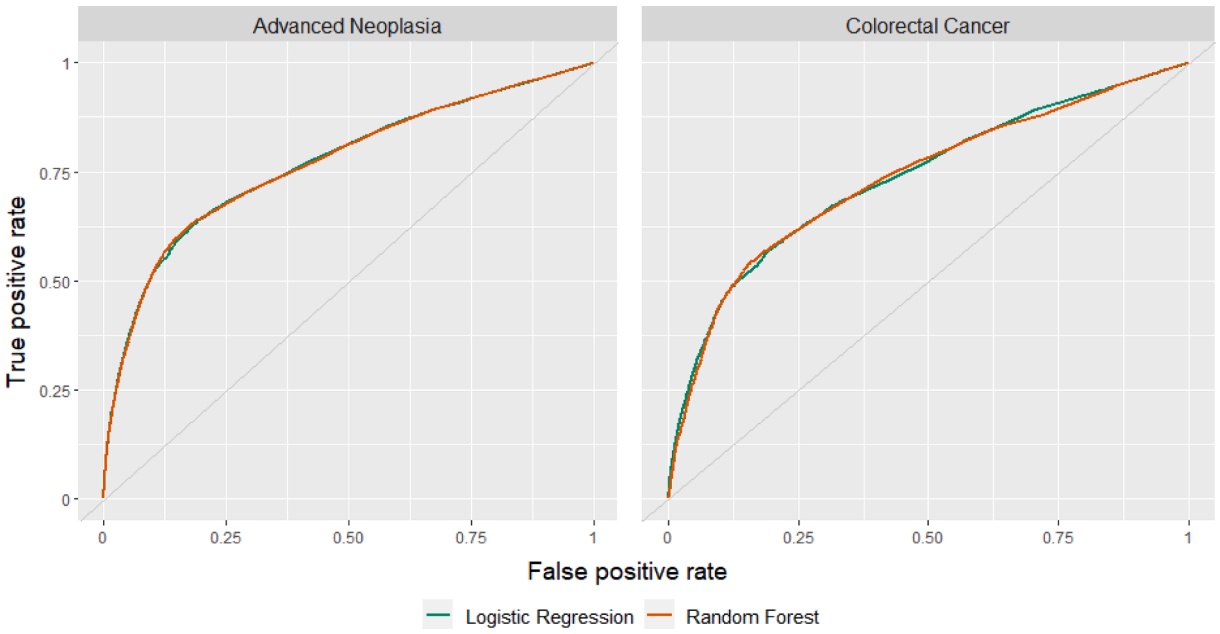


Fig. 3. Receiver-operating characteristic curves for predicted screening outcomes for third round participants after 2018. For predicting advanced neoplasia, the values of the area under the curve (AUC) were 0.77 (95% CI: 0.76–0.77) for the LR model and 0.77 (95% CI: 0.77–0.77) for the RF model. For predicting colorectal cancer, the values of the AUC were 0.73 (95% CI: 0.72–0.75) and 0.73 (95% CI: 0.72–0.74) for the LR and the RF respectively.

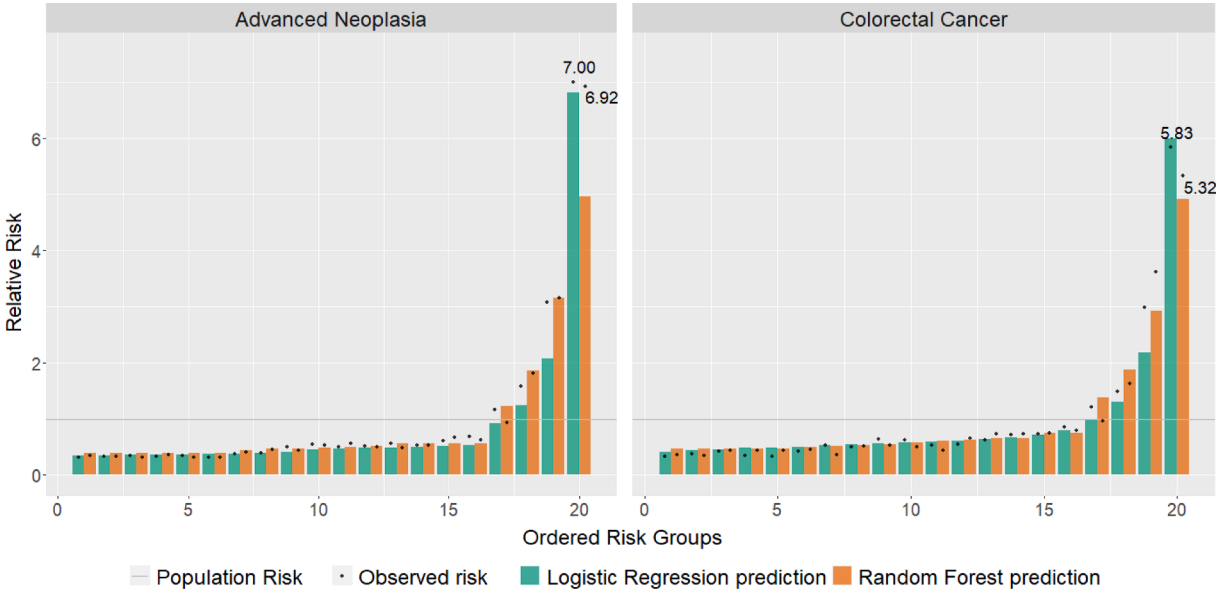


Fig. 4. Observed and predicted relative risks by risk groups ordered by predicted risk for third round participants after 2018. The bars represent the risk groups' predicted risk relative to the population risk. The black dots represent a risk group's observed risk to the predicted population average. The observed risk is determined by the composition of the risk group as predicted by the models. The observed relative risk of the highest risk group is reported numerically for both models. The grey horizontal line corresponds to a relative risk of one i.e. the mean risk of the total population. Each subgroup contains 5% of the population, corresponding to 56,880 individuals.

the RF model were similar (Fig. 5). The RF model attained an AUC (95% CI) of 0.73 (0.72–0.74) for AN and 0.68 (0.65–0.72) for CRC, also lower than the performance on third-round data (Table 3).

Given the similar AUCs, both models yielded relative risks of AN below the population average for the 80% of the participants with lowest predicted risk. The 5% participants with the highest predicted risk had a 6.1-fold (5.9-fold) risk of AN and 5.6-fold (5.1-fold) risk of CRC compared to the population average according to the LR (RF). The observed risks for subgroups with a lower predicted risk did not follow a smooth increasing pattern, indicating that the models were less effective in discriminating between subgroups with a lower predicted risk. The predicted relative risks slightly underestimated the observed relative risks (Fig. 6).

3.5. Sensitivity analysis

The results of the sensitivity analysis indicated that there were no significant differences in the model's performance when trained using all participants until 2018 compared to excluding those with a cut-off of 15 μg Hb/g (Supplementary material, Appendix 2).

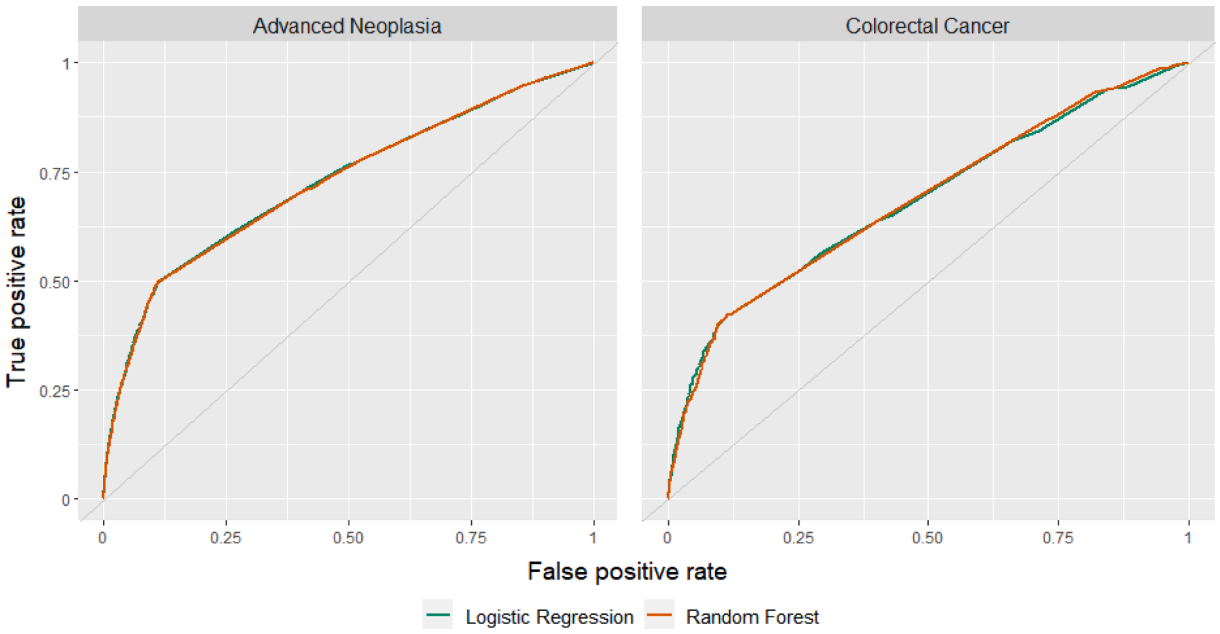


Fig. 5. Receiver-operating characteristic curves for predicted screening outcomes for the fourth screening round. For predicting advanced neoplasia, the values of the area under the curve (AUC) were 0.73 (95% CI: 0.72–0.74) for the LR model and 0.73 (95% CI: 0.72–0.74) for the RF model. For predicting colorectal cancer, the values of the AUC were 0.68 (95% CI: 0.66–0.71) and 0.68 (95% CI: 0.65–0.72) for the LR and the RF respectively.

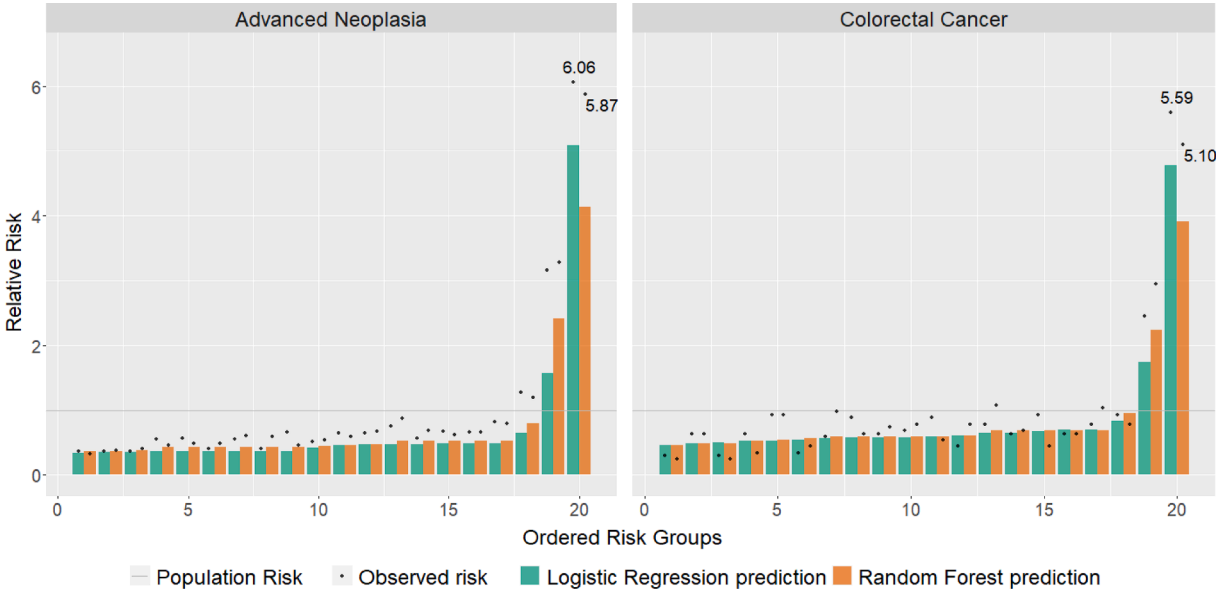


Fig. 6. Observed and predicted relative risks by risk groups ordered by predicted risk for fourth round participants. The bars represent the risk groups' predicted risk relative to the population risk. The black dots represent a risk group's observed risk relative to the predicted population average. The observed risk is determined by the composition of the risk group as predicted by the models. The observed relative risk of the highest risk group is reported numerically for both models. The dashed line corresponds to a relative risk of one i.e. the mean risk of the total population. Each subgroup contains 5% of the population, corresponding to 9,641 individuals.

4. Discussion

This study validated the performance of an existing CRC risk prediction tool based on prior f-Hb values using two separate validation sets. The shape and AUC of the ROC-curve for third-round participants after 2018 were similar to participants before 2018. The prognostic model also produced accurate risk predictions for outcomes in the fourth round. The RF model did not improve the predictions of CRC risk compared to the LR model for both validation sets.

To demonstrate generalizability across settings, the model developers validated the initial LR model against a trial from 2006 [2,19]. By contrast, we used nearly complete data from an ongoing national CRC screening programme for our validation. As the AUC values for third-round participants observed in our study are considered to indicate acceptable to good performance [17], our study offers direct evidence that the prediction model remains valid. The validation on fourth-round participants suggests that prognostic models may have some utility in predicting the risk of AN and CRC in later screening rounds. The performance however decreased compared to third-round predictions.

A possible cause for the decrease in performance in the fourth round is that participants were older (71 years ± 2 years) than third-round

participants up to 2018 (69 years ± 2 years). Previous studies associated age with a decrease in FIT sensitivity [20]. Because participants with false-negative FITs do not receive a follow-up colonoscopy, they were considered negative for AN and CRC in our analyses. An increase in the number of false-negatives might then lead to an apparent reduction in model performance. Another possible cause is that the f-Hb concentration measured in prevalence rounds should be interpreted differently from measurements in incidence rounds. The higher prevalence of detectable f-Hb in stool during the initial round observed in this study supports this explanation.

Although RF models have demonstrated success in risk prediction, even with small datasets [8–9,21], the limited number of covariates in our application resulted in insufficient complexity to fully leverage the added flexibility offered by machine learning. Many studies in other medical applications also found that machine-learning prediction tools did not outperform LR models under all circumstances [22–24]. Suggested reasons for this finding are the limited complexity of covariates in medical data, insufficient sample sizes and the lack of unstructured data. Moreover, due to its non-parametric character, RF models are less interpretable than regression models [25]. We therefore conclude that the LR model remains the preferred prediction tool.

Strengths of our study include the extensive validation of a prognostic model, the large size of the study population and the novel application of a machine-learning model for CRC risk prediction. The use of data of the nationwide Dutch screening programme with nearly complete capture of participants represents additional study strengths. However, a limitation of our study is that validation was conducted within the same screening programme as in which the LR model was built. More research is needed on the performance of our models in other settings. A second limitation is that at the time of this study, data on interval cancers were unavailable. Incorporating data of patients with interval cancers as participants with relevant outcomes would strengthen the validity of our study. Third, our validation is limited to the third and fourth round of the screening programme. Once available, a repetition of this study on data of future rounds could confirm validity and demonstrate the incremental value of added information over time. Finally, our proposed models do not explicitly consider the temporal nature and class imbalance of the data. Further investigation into the application of mixed-effect models and synthetic oversampling in this context could enhance the predictive performance.

This is the first study that temporally validated the LR as a CRC risk prediction tool. Although several studies have related prior negative FITs to the risk of advanced colorectal lesions [26–28], we only found one other prognostic model that was evaluated in terms of calibration and discrimination [29]. This model used explanatory variables, such as body mass index, alcohol consumption and family history of CRC in addition to f-Hb concentrations. In the absence of population wide surveys on risk factors, this model therefore cannot be used to stratify risk groups in screening programmes, nor can it be validated using the currently available screening data.

Our validation offers policy-makers additional evidence that prognostic models can be used for risk-stratification. The potential gains are under investigation in clinical trials in which the screening interval or test positivity threshold is personalized based on previous f-Hb concentrations [30–31]. Nevertheless, the current approaches in these trials rely on basic means of stratification that lack the level of precision provided by our risk prediction models. Adopting well-developed risk prediction models could therefore substantially enhance the potential benefits of risk stratifications compared to these relatively crude measures. The appropriate risk thresholds for inviting individuals for intensified or less intensified screening, and the corresponding benefits, are context-dependent and could be further investigated in simulation studies [3,32].

Our validation on fourth-round participants suggests that a model trained on the first screening rounds demonstrates potential applicability for subsequent screening rounds. This is particularly useful to new screening programs, for which the number of screening rounds is still expanding. That many FIT-based screening programmes have not reached their steady state yet [10], increases the relevance of valid risk predictions in subsequent rounds. Because we observed a decline in performance for the fourth round, further research is needed to investigate whether the LR model can be improved by training it on data from new screening rounds.

5. Conclusion

In conclusion, a prognostic model using sex, age and previously measured f-Hb concentrations is valid to predict the risk of AN and CRC. With this validation, the risk prediction tool is ready to be considered for implementation in screening programmes, for example to personalize the screening interval and/or test positivity cut-off in CRC screening programmes. An RF model does not improve CRC risk prediction compared to an LR model, probably due to the limited number of available explanatory variables. Therefore, the LR remains the preferred prediction tool because of its interpretability.

6. Authors' contributions

The authors confirm contribution to the paper as follows: DTM contributed to the design, data analysis, data interpretation, drafting and revision of the manuscript for important intellectual content. RP contributed to the data analysis and revision of the manuscript. RGSM contributed to the design, data interpretation and revision of the manuscript. JFOM contributed to the data interpretation and revision of the manuscript. ILV contributed to the conception of the work, data interpretation and revision of the manuscript. All authors accept full responsibility for the work and controlled the decision to publish. The Dutch Colorectal cancer screening working group *Landelijk Evaluatie team voor COlorectaal kanker bevolkingsonderzoek (LECO)* provided helpful comments on the paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The Dutch Colorectal cancer screening working group consists of Lucie de Jonge, Emilie Breekveldt, Hilliene Vandermeer, Esther Toes-Zoutendijk, Hanneke van Vuuren, Manon Spaander, Evelien Dekker, Christian Ramakers, Folkert van Kemenade, Iris Nagtegaal and Monique van Leerdam.

Statement on conflicts of interest

RGSM recently started a new position at Freenome. The other authors have nothing to declare.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2023.105194.

References

[1] F. Bénard, A.N. Barkun, M. Martel, D. von Renteln, Systematic review of colorectal cancer screening guidelines for average-risk adults: Summarizing the current global recommendations, World J. Gastroenterol. 24 (1) (2018) 124.

[2] R.G.S. Meester, H.J. van de Schootbrugge-Vandermeer, E.C.H. Breekveldt, L. de Jonge, E. Toes-Zoutendijk, A. Kooyker, et al., Faecal occult blood loss accurately predicts future detection of colorectal cancer. A prognostic model, Gut (2022).

[3] R. van den Puttelaar, R.G.S. Meester, E.F.P. Peterse, A.G. Zauber, J. Zheng, R. B. Hayes, et al., Risk-stratified screening for colorectal cancer using genetic and environmental risk factors: a cost-effectiveness analysis based on real-world data, Clin. Gastroenterol. Hepatol. (2023).
[4] D.R. Cenin, S.K. Naber, A.C. de Weerdt, M.A. Jenkins, D.B. Preen, H.C. Ee, et al., Cost-Effectiveness of personalized screening for colorectal cancer based on polygenic risk and family history personalized screening for colorectal cancer: cost-effectiveness analysis, Cancer Epidemiol. Biomark. Prev. 29 (1) (2020) 10–21.
[5] M.K. Thomsen, L. Pedersen, R. Erichsen, T.L. Lash, H.T. Sørensen, E.M. Mikkelsen, Risk-stratified selection to colonoscopy in FIT colorectal cancer screening: development and temporal validation of a prediction model, Br. J. Cancer 126 (8) (2022) 1229–1235.
[6] R. Couronné, P. Probst, A.-L. Boulesteix, Random forest versus logistic regression: a large-scale benchmark experiment, BMC Bioinf. 19 (1) (2018) 1–14.
[7] K. Kirasich, T. Smith, B. Sadler, Random forest vs logistic regression: binary classification for heterogeneous datasets, SMU Data Sci. Rev. 1 (3) (2018) 9.
[8] R.O. Alabi, A. Almangush, M. Elmusrati, I. Leivo, A.A. Mäkitie, An interpretable machine learning prognostic system for risk stratification in oropharyngeal cancer, Int. J. Med. Inf. 104896 (2022).
[9] B.O. Macaulay, B.S. Aribisala, S.A. Akande, B.A. Akinnuwesi, O.A. Olabanjo, Breast cancer risk prediction in African women using Random Forest Classifier, Cancer Treatment Res. Commun. 28 (2021), 100396.
[10] E.H. Schreuders, A. Ruco, L. Rabeneck, R.E. Schoen, J.J.Y. Sung, G.P. Young, et al., Colorectal cancer screening: a global overview of existing programmes, Gut 64 (10) (2015) 1637–1649.
[11] E. Toes-Zoutendijk, M.E. van Leerdam, E. Dekker, F. Van Hees, C. Penning, I. Nagtegaal, et al., Real-time monitoring of results during first year of Dutch colorectal cancer screening program and optimization by altering fecal immunochemical test cut-off levels, Gastroenterology 152 (4) (2017) 767–775.
[12] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.
[13] A. Prinzie, D. Van den Poel, Random forests for multiclass classification: Random multinomial logit, Expert Syst. Appl. 34 (3) (2008) 1721–1732.
[14] P. Probst, M.N. Wright, A.L. Boulesteix, Hyperparameters and tuning strategies for random forest, Wiley Interdisc. Rev.: Data Mining Knowledge Discov. 9 (3) (2019) e1301.
[15] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[16] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (8) (2006) 861–874.
[17] A.A.H. de Hond, E.W. Steyerberg, B. van Calster, Interpreting area under the receiver operating characteristic curve, Lancet Digital Health 4 (12) (2022) e853–e855.
[18] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, CRC Press, Boca Raton. FL, 1993.
[19] M. Van Der Vlugt, E.J. Grobbee, P.M.M. Bossuyt, E. Bongers, W. Spijker, E. J. Kuipers, et al., Adherence to colorectal cancer screening: four rounds of faecal immunochemical test-based screening, Br. J. Cancer 116 (1) (2017) 44–49.
[20] K. Selby, E.H. Levine, C. Doan, A. Gies, H. Brenner, C. Quesenberry, et al., Effect of sex, age, and positivity threshold on fecal immunochemical test accuracy: a systematic review and meta-analysis, Gastroenterology 157 (6) (2019) 1494–1505.
[21] S. Han, B.D. Williamson, Y. Fong, Improving random forest predictions in small datasets from two-phase sampling designs, BMC Med. Inf. Decis. Making 21 (1) (2021) 1–9.
[22] A. De Hond, W. Raven, L. Schinkelshoek, M. Gaakeer, E. Ter Avest, O. Sir, et al., Machine learning for developing a prediction model of hospital admission of emergency department patients: Hype or hope? Int. J. Med. Inf. 152 (2021), 104496.
[23] X. Song, X. Liu, F. Liu, C. Wang, Comparison of machine learning and logistic regression models in predicting acute kidney injury: a systematic review and meta-analysis, Int. J. Med. Inf. 151 (2021), 104484.
[24] R.J. Huang, N.S.-E. Kwon, Y. Tomizawa, A.Y. Choi, T. Hernandez-Boussard, J. H. Hwang, A comparison of logistic regression against machine learning algorithms for gastric cancer risk prediction within real-world clinical data streams, JCO Clin. Cancer Informatics 6 (2022), e2200039.
[25] S.I. Birbil, M. Edali, B. Yuceoglu, Rule covering for interpretation and boosting, 2020. arXiv preprint arXiv:200706379.
[26] C. Balamou, A. Koïvogui, C.M. Rodrigue, A. Clerc, C. Piccotti, A. Deloraine, et al., Prediction of the severity of colorectal lesion by fecal hemoglobin concentration observed during previous test in the French screening program, World J. Gastroenterol. 27 (31) (2021) 5272.
[27] A. Buron, M. Román, J.M. Augé, F. Macià, J. Grau, M. Sala, et al., Changes in FIT values below the threshold of positivity and short-term risk of advanced colorectal neoplasia: results from a population-based cancer screening program, Eur. J. Cancer 107 (2019) 53–59.
[28] S.-Y.-H. Chiu, S.-L. Chuang, S.-L.-S. Chen, A.-M.-F. Yen, J.-C.-Y. Fann, D.-C. Chang, et al., Faecal haemoglobin concentration influences risk prediction of interval cancers resulting from inadequate colonoscopy quality: analysis of the Taiwanese Nationwide Colorectal Cancer Screening Program, Gut 66 (2) (2017) 293–300.
[29] A.M.F. Yen, S.L.S. Chen, S.Y.H. Chiu, J.C.Y. Fann, P.E. Wang, S.C. Lin, et al., A new insight into fecal hemoglobin concentration-dependent predictor for colorectal neoplasia, Int. J. Cancer 135 (5) (2014) 1203–1212.
[30] E.C.H. Breekveldt, E. Toes-Zoutendijk, L. de Jonge, M.C.W. Spaander, E. Dekker, F. J. van Kemenade, et al., Personalized colorectal cancer screening: study protocol of a mixed-methods study on the effectiveness of tailored intervals based on prior f-Hb concentration in a fit-based colorectal cancer screening program (PERFECT-FIT), BMC Gastroenterol. 23 (1) (2023) 1–10.
[31] C. Senore (Ed.) Beyond Positive vs. Negative: Cumulative fecal Hb level for risk prediction. World Endoscopy Organization Colorectal Cancer Screening Committee, 2023, Chicago, USA.
[32] L.A. van Duuren, J. Ozik, R. Spliet, N.T. Collier, I. Lansdorp-Vogelaar, R.G. S. Meester, An evolutionary algorithm to personalize stool-based colorectal cancer screening, Front. Physiol. 2515 (2022).