Patricia Diana Soerensen*, Henry Christensen, Soeren Gray Worsoe Laursen, Christian Hardahl, Ivan Brandslund and Jonna Skov Madsen

Using artificial intelligence in a primary care setting to identify patients at risk for cancer: a risk prediction model based on routine laboratory tests

<https://doi.org/10.1515/cclm-2021-1015>

Received July 21, 2021; accepted October 1, 2021;

published online November 30, 2022

Abstract

Objectives: To evaluate the ability of an artificial intelligence (AI) model to predict the risk of cancer in patients referred from primary care based on routine blood tests. Results obtained with the AI model are compared to results based on logistic regression (LR).

Methods: An analytical profile consisting of 25 predefined routine laboratory blood tests was introduced to general practitioners (GPs) to be used for patients with non-specific symptoms, as an additional tool to identify individuals at increased risk of cancer. Consecutive analytical profiles ordered by GPs from November 29th 2011 until March 1st 2020 were included. AI and LR analysis were performed on data from 6,592 analytical profiles for their ability to detect cancer. Cohort I for model development included 5,224 analytical profiles ordered by GP's from November 29th 2011 until the December 31st 2018, while 1,368 analytical profiles included from January 1st 2019 until March 1st 2020 constituted the “out of time” validation test Cohort II. The

main outcome measure was a cancer diagnosis within 90 days.

Results: The AI model based on routine laboratory blood tests can provide an easy-to use risk score to predict cancer within 90 days. Results obtained with the AI model were comparable to results from the LR model. In the internal validation Cohort IB, the AI model provided slightly better results than the LR analysis both in terms of the area under the receiver operating characteristics curve (AUC) and PPV, sensitivity/specificity while in the “out of time” validation test Cohort II, the obtained results were comparable.

Conclusions: The AI risk score may be a valuable tool in the clinical decision-making. The score should be further validated to determine its applicability in other populations.

Keywords: artificial intelligence (AI) blood; cancer; predictive; score.

Introduction

Risk prediction models aim to assist healthcare providers in the process of clinical decision making by estimating the probability of specific outcomes in a population. Traditionally, parametric logistic regression analyses (LR) have dominated and improved risk prediction in healthcare for decades [1]. However, the increased opportunities of managing large and complex datasets have encouraged the application and the development of new models and tools based on artificial intelligence (AI) [2].

In a primary care setting one of the main challenges is to ensure an early diagnosis of cancer, as this entails better prognosis and lower mortality [3].

Many of the symptoms associated with malignant disease are non-specific, vague or imprecise and relative low risk. Even when it comes to classical “alarm” symptoms, the positive predictive value (PPV) for an underlying

*Corresponding author: Patricia Diana Soerensen, Department of Clinical Biochemistry and Immunology, Lillebaelt Hospital, University Hospital of Southern Denmark, Vejle, Denmark, E-mail: Patricia.Diana.Sorensen@rsyd.dk

Henry Christensen, Department of Clinical Biochemistry and Immunology, Lillebaelt Hospital, University Hospital of Southern Denmark, Vejle, Denmark

Soeren Gray Worsoe Laursen, The Danish Cancer Society, Copenhagen, Denmark

Christian Hardahl, SAS Institute A/S, Aarhus, Denmark

Ivan Brandslund, Department of Regional Health Research, University of Southern Denmark, Odense, Denmark

Jonna Skov Madsen, Department of Clinical Biochemistry and Immunology, Lillebaelt Hospital, University Hospital of Southern Denmark, Vejle, Denmark; and Department of Regional Health Research, University of Southern Denmark, Odense, Denmark. <https://orcid.org/0000-0001-6668-4714>

malignant disease is low [4]. While cancer biomarkers are routinely used in hospital settings, applied to the low risk population in a primary care setting, they have a low PPV towards detecting cancer, and at the same time high false positive rates.

Given the relatively low PPV of individual blood tests, two main approaches to assess cancer risk based on tests performed from blood samples have emerged. One approach is based on detecting circulating free DNA (cfDNA) in a blood sample, whereas the other is based on applying artificial intelligence to detect non-obvious and latent relationships in routine blood based laboratory test results.

The approach using cfDNA released to the blood in order to detect possible cancer is a field in rapid growth. Thus, a noninvasive blood test (CancerSEEK) was shown to perform with greater than 99% specificity and with sensitivities ranging from 69 to 98% for the detection of five cancer types—ovarian, liver, stomach, pancreas, and esophageal—for which there are no current screening tests available for average-risk individuals [5]. In addition, a noninvasive blood test based on circulating tumor DNA methylation (PanSeer) was reported to be able to detect cancer up to four years before standard diagnosis in a longitudinal study [6].

Schneider et al. validated a predictive model generated by a machine-learning algorithm that used complete blood cell count and demographic data from individuals at ages 50–75 years with the purpose of identifying individuals at increased risk for colorectal cancer. At a specificity of 97% corresponding to a high score from the developed algorithm they obtained a sensitivity of 35.4% for a colorectal cancer diagnosis within the next 6 months and had an area under the receiver operating characteristics curve (AUC) of 0.78 [7].

Thus, routine laboratory test results may contain far more information than recognized by even the most experienced clinician and detection of such non-obvious interrelationships are suitable to analysis by artificial intelligence in order to provide individual risk scores.

In January of 2008, Lillebaelt Hospital introduced a gender specific analytical profile based on routine laboratory tests to be used in the primary care setting by general practitioners (GP) as an additional tool for patients with non-specific symptoms to identify individuals at increased risk of cancer.

In the current study, we evaluate the ability of an AI model to provide an individual cancer risk score based on these routine laboratory tests. In addition, the risk scores obtained in the AI model are compared to results obtained by standard logistic regression (LR).

Materials and methods

Study population and laboratory tests

The uptake population area is located in the Region of Southern Denmark with around 350,000 inhabitants served by 106 GPs. In a joint collaboration between the local Clinical Biochemistry laboratory at the Lillebaelt Hospital and the GPs, a specific analytical profile containing routine blood tests was provided as an additional tool in the GPs diagnostic arsenal meant for patients consulting their physician with common or non-specific symptoms and where the GP suspected possible hidden cancer. As an initiative prompted by Denmark's third national cancer plan, the urgent referral for unspecific, serious symptoms was implemented nationally by the National Board of Health and Danish Regions in 2011. The pathway consists of a two-step approach with a filter function performed by the GP and, if still relevant, a referral to a diagnostic center. The filter function is a battery of diagnostic investigations consisting of anamnesis, blood and urine tests and diagnostic imaging. It is this predefined routine laboratory set of blood tests that is the subject matter of our study.

The GP could order an analytical profile labeled "Suspicion of Hidden Cancer/Woman" or "Suspicion of Hidden Cancer/Man". Thus, the set of blood tests was drawn from patients where no obvious tentative diagnoses for a specific cancer or other diseases identified by the GP.

The analytical profile was introduced January 2008 and consisted of the following components:

- In both men and women: B-hemoglobin, Mean corpuscular volume (MCV), Mean corpuscular hemoglobin (MCH), B-leukocytes with differential count, B-reticulocytes, B-platelets, P-C-reactive protein, P-sodium, P-potassium, P-calcium total, P-albumin, P-creatinine, P-carbamide, P-urate, P-glucose, P-bilirubin, P-alanine transaminase, P-basic phosphatase, P-amylase pancreatic specific, P-lactate dehydrogenase, P-immunoglobulins A, G and M (IgA, IgG and IgM), P-thyroidea stimulating hormone,
- In men: in addition P-prostate-specific antigen
- In women: in addition P-cancer antigen-125

We did the leukocyte subpopulation quantifications on the Sysmex hematology systems and they were quantified as total Leucocytes, Neutrophils, Eosinophils, Basophils, Lymphocytes and Monocytes.

During the whole study period, we have used equipment from Roche for routine biochemistry analysis and from Sysmex for the hematology instruments. However, for the Roche instruments, there has been both instrument and methodological upgrades in the period. According to our routine procedures for continuous quality insurance, the validation of each laboratory component was performed, including the investigation of a potential bias between the previous and the current modules. Thus, the upgrades in instrument and/or methodology did not have an impact on the results reported here.

Study cohort

Due to changes in the laboratory information system, only data after November 29th 2011 were available. The total eligible study cohort included 6,592 consecutive analytical profiles ordered from GPs: 5,224

were included from November 29th 2011 until the December 31st 2018 (Cohort I) and 1,368 from January 1st 2019 until March 1st 2020 (Cohort II).

The following exclusion criteria were applied: Individuals <18 years of age; individuals without information on gender; patients diagnosed with cancer within the last 5 years; for individuals with more than one analytical profile ordered prior to being diagnosed with cancer, only the first was included for data analysis. Finally, cases with too few laboratory tests results within the analytical profile were also excluded (initial analytical profiles with ≤ 10 laboratory test results, and again in the final model analytical profiles ≤ 30 laboratory tests results).

A presentation of the total study cohort is provided in Figure 1.

Outcome measures

Any cancer diagnosed within 730 days of follow up after study inclusion was registered. All types of cancer were included (both primary and metastatic).

The main outcome measure in this study was “cancer within 90 days” which we considered a sufficient follow period up to ensure valid data for Cohort II regarding registered cancer diagnosis as well.

Ethics

Project approval, data transfer and data safety: The project was carried out by permission from the Danish Board for Patient Safety (number 3-3013-2954/1) and Hospital Lillebaelt. All data were anonymized and transferred to the “in house” SAS Viya platform. No data left the hospital regional database.

Data management and statistics

AI model development was performed based on Cohort I and final validation of the model was performed in the “out of time” validation test Cohort II. Cohort I was divided into training and validation sections. This splitting was done outside of the modelling processes to ensure that all models used the same criteria for this procedure, thereby allowing for a comparison of performance results. The data was assigned to training and validation sections respectively by an 80/20 split by stratified random sampling. The 80/20 split was chosen to permit many training observations as machine learning techniques are “data hungry” and the Cohort I contained only 5,224 observations.

Cohort IA for training consisted of 4,182 analytical profiles while Cohort IB for validation consisted of 1,042 profiles.

Transformations: Data was transformed in the manner below to accommodate the types of models that are incapable of handling missing values and are sensitive to variables of different scales:

- Missing biomarkers were imputed/replaced by their mean value (from the training section).
- All input variables were z-normalized (subtraction of sample mean, division by sample standard deviation – both statistics estimated on the training section).

The implemented code allowed for imputation and transformations to be gender specific. At first, a gender specific model development was performed including PSA for males and Ca125 for females. A drawback

with this approach, however, was that our cohorts, both the training and validation, were rather small for the purpose. Thus, also a full model was developed, without regard to gender. The full model performed as well as the gender specific models and was therefore chosen as the final model.

AI model selection: AI model selection was performed on SAS® Viya® (V.03.04, Denmark).

All model parameters were estimated in the training section while performance measures were obtained from the validation section. The main criteria for model selection were the Area Under the curve and the receiver operating characteristics curve (AUC/ROC) measured on the validation section for both genders.

Determining the AI models: It was decided to evaluate two initial strategies for model development prior to determining which one to choose for further fine-tuning.

- A Random forest based method which has the advantage of frequently allowing good results with a minimum of imputations and transformations as this model handles missing values explicitly without being affected by variable scale or magnitude.
- A Neural network model, which requires imputation and a standardization method, but allows for more flexibility in the modelling process.

Both model types were tested in order to determine which method seemed most promising. As the first round of modelling indicated a greater fit for the Neural Network model, this was selected for further fine-tuning. Thus, multiple hyper parameters were tested in order to obtain the best possible AUC in the neural network model.

Applying the bootstrap procedure: The Bootstrap procedure was applied with the purpose of evaluation of the imprecision of the methodology through the following procedure:

- A number of groups were created based on the assigned score from the model. Each score belongs to one group.
- A Bootstrap procedure was used to estimate the event rate distribution within each data partition and group as described below. For each partition the following procedures were performed:
 - All blood samples in the partition were scored and assigned to the relevant group by comparing the individual score to the cutoffs of the groups.
 - From the partition one thousand samples (with a size equal to the partition size) were drawn with replacement (i.e. 1,000 Bootstrap replicates). The Bootstrap samples were generated as balanced Bootstrap samples, i.e. each original observation is represented the same number of times in the final samples.
 - For each replicate the empirical event rate in each group was estimated
 - For each group the distribution of the 1,000 samples of the event rate was used to compute the mean and relevant fractions of the distribution.

The end goal of the procedure was to obtain group estimates for the empirical event rate on the test and validation section.

Standard statistical analyses: Standard statistical analyses were performed on the SAS platform; logistic regression (LR) was performed for the standard ROC curve determination on the development Cohort I

Cohorts

Initial exploration - blood samples panels

Exclusion criteria

Too few biomarkers

Panels with ≤ 10 laboratory test results removed

Non adults

Patients < 18 years of age at the time of the blood sampling excluded

Previous cancer

Patients with a previous cancer diagnoses within 5 years (5×365 days) prior to the date for the current blood sample panels removed

Unknown gender

Cases without information of gender removed

Repeated patient

Panels from the patients that had previously received the blood sample panel within the last 2 years removed.

Final model

Panels with ≤ 30 laboratory test results removed

Final cohorts

Cohort IA for training
Cohort IB for validation
Cohort II for test

Figure 1: Total study cohort. The three cohorts of the study are shown.

(training Cohort IA and validation Cohort IB) and on the “out of time” validation test Cohort II using “cancer within 90 days” as target.

Test performance presentation: With the purpose of facilitating a better understanding of the results for the clinicians, data was also calculated to generate sensitivity, specificity and predictive values according to a model described in the work of Gerhardt et al. [8]. These calculations were done for both the AI model and the LR model.

The AI risk was generated by machine learning in a supervised process using the outcome data. Based on an arbitrary risk scale from 0 to 100, the relative risk was determined. Based on the observed absolute risk of being diagnosed with cancer within 90 days, this score was converted to absolute risk for cancer. The performance of the AI score to predict cancer within 90 days was calculated for the “out of time” validation test Cohort II at different thresholds. The formula $TP/(TP + FP)$ was used for calculation of the threshold for positive test, where TP is true positive and FP is false positive. The formula $FN/(FN + TN)$ was used for calculation of the threshold for negative test, where FN is false negative and TN is true negative.

Results

The total number of “cancer within 90 days” of study inclusion was 5.67% for training Cohort IA, 5.28% for validation Cohort IB and 6.14% for validation test Cohort II respectively.

Table 1: Cancer types diagnosed within 90 days of study inclusion.

Cancer type	Cases in total cohort, n (%)
Prostate cancer	72(1.00%)
Cancer of the upper lobe of the lung	22(0.306%)
Lung cancer	16(0.223%)
Multiple myeloma	14(0.195%)
Ovarian cancer	14(0.195%)
Prostate cancer with metastases	10(0.139%)
Diffuse large cell B-cell lymphoma	9(0.125%)
Cancer of the sigmoid colon	9(0.125%)
Rectal cancer	9(0.125%)
Cancer of the bronchi and lung spanning multiple localizations	8(0.111%)
Cancer of the ascending colon	8(0.111%)
Cancer of the lower lobe of the lung	8(0.111%)
Breast cancer	8(0.111%)
Chronic B-cell type (B-CLL) chronic lymphocytic leukemia	7(0.098%)
Hepatocellular carcinoma	7(0.098%)
Kidney cancer	7(0.098%)
Bladder cancer	6(0.084%)
Remote metastasis of bone or bone marrow	6(0.084%)
Uterine cancer	6(0.084%)
Pancreatic cancer	5(0.070%)
Cancer of the gastric cardia	5(0.070%)
Stomach cancer	4(0.056%)

The frequency of the 22 most common cancer types diagnosed in the total Cohort (IA, IB and II) within 90 days of study inclusion is presented in Table 1.

AI ROC curves for training and validation in Cohort I with the primary outcome “cancer within 90 days” are presented in Figure 2.

The AUC results obtained by the AI analysis in Cohort I (both the training Cohort IA and the validation Cohort IB) and in the “out of time” validation test Cohort II are presented in Table 2 (training Cohort IA, validation Cohort IB and validation test Cohort II, overall and by gender: female and male. Target: cancer within 90 days).

Bootstrap performed in both the validation Cohort IB and in the validation test Cohort II shows risk categories indicated as very low, low, medium, high and very high and the estimation of the uncertainty of the result (Table 3).

The distribution of patients across AI risk scores obtained in the validation test Cohort II is presented in Figure 3A together with the observed incidence of cancer within 90 days in Figure 3B.

Prediction scores for the AI model covering a range of scores from 0 to 100 were calculated and data obtained at different thresholds in the validation Cohort IB are presented in Table 4.

For each AI score, data are presented with the corresponding results regarding true positive number (TP(n)), false negative number (FN(N)), true negative number (TN(n)), false positive number (FP(n)), sensitivity%, specificity%, positive predictive value (PPV)% and negative predictive value (NPV)%.

Similar, the developed AI model was applied to the “out of time” validation test Cohort II. The absolute risk for having a cancer diagnose within 90 days at different thresholds according to the AI score is provided in Table 5.

Standard statistical analysis

The AUC obtained by the LR analysis was 0.80, 0.80 and 0.79 for the training Cohort IA, the validation Cohort IB and the “out of time” validation test Cohort II, respectively. Data is presented in Figure 4.

Data regarding sensitivity, specificity, PPV and NPV at different scores calculated from LR method is presented in Table 6 for the validation Cohort IB and in Table 7 for validation test Cohort II.

Discussion

This study demonstrates that the developed AI model based on routine laboratory tests from GP's in a primary

Figure 2: The AUC results obtained by the AI analysis in Cohort IA for training and validation Cohort IB. ROC curves obtained in the development cohort, with the training Cohort IA (left) and the validation Cohort IB (right). Green: Males, red: females, Blue: both genders.

Table 2: The AUC results obtained by the AI model.

AUC			
Gender	Cohort I for development		Corhort II for out of time validation test n=1,368
	Training n=4,182	Validation n=1,042	
Overall	0.91	0.86	0.79
Females	0.90	0.87	0.81
Males	0.92	0.84	0.75

Cohort I constituted a training Cohort IA and a validation Cohort IB. Cohort II served the “out of time” validation test cohort. Outcome measure was individuals being diagnosed with cancer within 90 days.

care setting can provide a specific risk score for the prediction of cancer comparable to, and in some respects slightly better than using standard statistical measures such as logistic regression (LR).

The AI model provided slightly better results than the LR analysis when looking at the ROC curves obtained in the validation Cohort IB. In this case, the AI model had an AUC of 0.86 compared to 0.80 obtained in the LR model. In the “out of time” validation test Cohort II however, the obtained AUC results were comparable, with an AUC for

the AI model of 0.79 compared to 0.79 for the LR model. A reason for the lower AUC obtained with the AI model in the validation test Cohort II than in the validation Cohort IB may be due to the model being “overfitted” with regards to the training and validation sections. Though the validation Cohort IB was not used to train the neural network, it was used in the selection of neural networks.

The corresponding results, when comparing the two models in performance on PPV, specificity, sensitivity and NPV (data presented in Tables 4–7) confirmed that results from the AI model turned out to be slightly better than the LR model in the validation Cohort IB. However, when applied to the “out of time” validation test Cohort II results obtained with the AI model were comparable to results based on the LR model. As mentioned above this is probably due to “overfitting” in the AI model regarding results in the validation Cohort IB.

This challenge with “overfitting” has been demonstrated in a multitude of previous studies, where the AI model provided good PPV within the dataset from which they were derived, but underperformed when applied to an external validation cohort [9].

This demonstrates and underscores the general importance of performing internal and external validation

A: Distribution of patients across AI risk scores
 B: Observed incidence of cancer across AI risk scores

A

B

Figure 3: AI risk score and cancer incidence in validation test Cohort II.

(A) The upper panel shows the distribution of patients across AI scores in validation test Cohort II. Thus, it appears that most patients had an AI score below 1, and only a total of 31 patients had an AI score above 50. (B) This panel provides the incidence of cancer diagnosed within 90 days across AI risk scores. Thus, the incidence of cancer within 90 days in validation test Cohort II was found to be below 12% at an AI score value between 4 and 5 and up to 100% for AI scores from 70 to 90. However, only a limited number of patients in this rather small cohort had high AI scores and for example, there were no patients with AI score between 7 and 8 or between 40 and 45.

Table 3: Bootstrap estimates on grouped risk bins on validation Cohort IB and validation test Cohort II.

Bootstrap estimates of event rate fractiles						
Data partition	Bin	1% fractile	5% fractile	Mean value	95% fractile	99% fractile
Cohort IB	1. Very low (0–30)	0.000%	0.000%	0.352%	1.03%	1.42%
Cohort IB	2. Low (31–50)	0.000%	0.208%	1.49%	3.07%	3.93%
Cohort IB	3. Medium (51–80)	1.22%	1.80%	3.35%	5.13%	5.94%
Cohort IB	4. High (81–90)	1.39%	2.80%	6.31%	10.4%	12.5%
Cohort IB	5. Very high (91+)	18.3%	21.3%	27.9%	34.8%	38.0%
Cohort II	1. Very low (0–30)	0.252%	0.506%	1.49%	2.61%	3.02%
Cohort II	2. Low (31–50)	0.803%	1.20%	2.82%	4.69%	5.81%
Cohort II	3. Medium (51–80)	2.09%	2.91%	4.35%	5.95%	6.66%
Cohort II	4. High (81–90)	4.16%	5.63%	9.55%	13.6%	15.6%
Cohort II	5. Very high (91+)	18.6%	21.0%	27.3%	33.8%	36.9%

Table 4: Performance of the AI score ranging from 0 to 100 in the validation Cohort IB.

AI score	TP, n	FN, n	TN, n	FP, n	Sensitivity, %	Specificity, %	PPV, %	NPV, %
0	55	0	0	987	100.0	0.0	5.3	NA
1	55	0	246	741	100.0	24.9	6.9	100.0
2	51	4	408	579	92.7	41.3	8.1	99.0
3	51	4	531	456	92.7	53.8	10.1	99.3
4	51	4	621	366	92.7	62.9	12.2	99.4
5	46	9	690	297	83.6	69.9	13.4	98.7
6	42	13	732	255	76.4	74.2	14.1	98.3
7	41	14	784	203	74.5	79.4	16.8	98.2
8	39	16	816	171	70.9	82.7	18.6	98.1
9	37	18	849	138	67.3	86.0	21.1	97.9
10	36	19	872	115	65.5	88.3	23.8	97.9
20	27	28	949	38	49.1	96.1	41.5	97.1
30	20	35	965	22	36.4	97.8	47.6	96.5
40	16	39	974	13	29.1	98.7	55.2	96.2
50	13	42	979	8	23.6	99.2	61.9	95.9
60	8	47	982	5	14.5	99.5	61.5	95.4
70	3	52	986	1	5.5	99.9	75.0	95.0
80	2	53	987	0	3.6	100.0	100.0	94.9
90	1	54	987	0	1.8	100.0	100.0	94.8
100	0	55	987	0	0.0	100.0	NA	94.7

FN, false negative; FP, false positive; NPV, negative predictive value; PPV, positive predictive value; TN, true negative; TP, true positive.

of the obtained results. In the current study we performed a validation of the developed model in the “out of time” validation test Cohort II in order to overcome the challenge with overfitting.

Bootstrap estimates performed with the purpose of evaluating the imprecision of the AI methodology, also allows for a comparison of risks between results obtained in the validation Cohort IB and in the validation test Cohort II. For example, a patients’ blood panel result was classified as medium risk corresponding to a mean value risk of 3.35% (5–95% fractile 1.80–5.13) in the validation Cohort IB and of 4.35% (5–95% fractile 2.91–5.95) in the validation test Cohort II.

Presenting data in an AUC format is not optimal, nor is presenting risk scores divided into categories (low, medium, high) as they may be regarded as too imprecise when it comes to clinical decision making in the individual patient. Presenting absolute risk as indicated in Table 5 may be more useful and easier to interpret, as it provide information on the performance of the AI score to predict cancer at a range of different thresholds. In this work, we do not recommend a specific threshold to define when a test is positive or negative, but instead provide the information on risk stratification to be used in clinical decision making in the individual patient. As an example, if an AI score of ≤ 3 was hypothetically considered as the threshold

Table 5: Performance of AI score to predict cancer within 90 days at different thresholds in the validation test Cohort II.

Cut-off	Cohort number, %	Cancer number, %	TP	TN	FP	FN	Sensitivity, %	Specificity, %	PPV, %	NPV, %
Threshold for negative test										
≤1	353(25.8)	4(1.1)	80	349	935	4	95.2	27.2	7.9	98.9
≤2	576(42.1)	7(1.2)	77	569	715	7	91.7	44.3	9.7	98.8
≤3	726(53.1)	15(2.1)	69	711	573	15	82.1	55.4	10.7	97.9
≤4	845(61.8)	20(2.4)	64	825	459	20	76.2	64.3	12.2	97.6
≤5	928(67.8)	27(2.9)	57	901	383	27	67.9	70.2	13.0	97.1
≤6	997(72.9)	30(3.0)	54	967	317	30	64.3	75.3	14.6	97.0
≤7	1,061(77.6)	32(3.0)	52	1,029	255	32	61.9	80.1	16.9	97.0
≤8	1,104(80.7)	32(2.9)	52	1,072	212	32	61.9	83.5	19.7	97.1
≤9	1,140(83.3)	37(3.2)	47	1,103	181	37	56.0	85.9	20.6	96.8
≤10	1,173(85.7)	40(3.4)	44	1,133	151	40	52.4	88.2	22.6	96.6
≤11	1,195(87.4)	42(3.5)	42	1,153	131	42	50.0	89.8	24.3	96.5
≤12	1,215(88.8)	45(3.7)	39	1,170	114	45	46.4	91.1	25.5	96.3
≤13	1,234(90.2)	47(3.8)	37	1,187	97	47	44.0	92.4	27.6	96.2
≤14	1,246(91.1)	48(3.9)	36	1,198	86	48	42.9	93.3	29.5	96.1
≤15	1,257(91.9)	48(3.8)	36	1,209	75	48	42.9	94.2	32.4	96.2
≤16	1,262(92.3)	51(4.0)	33	1,211	73	51	39.3	94.3	31.1	96.0
≤17	1,270(92.8)	51(4.0)	33	1,219	65	51	39.3	94.9	33.7	96.0
≤18	1,277(93.3)	51(4.0)	33	1,226	58	51	39.3	95.5	36.3	96.0
≤19	1,280(93.6)	51(4.0)	33	1,229	55	51	39.3	95.7	37.5	96.0
≤20	1,286(94)	53(4.1)	31	1,233	51	53	36.9	96.0	37.8	95.9
≤25	1,302(95.2)	54(4.1)	30	1,248	36	54	35.7	97.2	45.5	95.9
≤50	1,337(97.7)	63(4.7)	21	1,274	10	63	25.0	99.2	67.7	95.3
≤75	1,362(99.6)	79(5.8)	5	1,283	1	79	6.0	99.9	83.3	94.2
Threshold for positive test										
>1	1,015(74.2)	80(7.9)	80	349	935	4	95.2	27.2	7.9	98.9
>2	792(57.9)	77(9.7)	77	569	715	7	91.7	44.3	9.7	98.8
>3	642(46.9)	69(10.7)	69	711	573	15	82.1	55.4	10.7	97.9
>4	523(38.2)	64(12.2)	64	825	459	20	76.2	64.3	12.2	97.6
>5	440(32.2)	57(13.0)	57	901	383	27	67.9	70.2	13.0	97.1
>6	371(27.1)	54(14.6)	54	967	317	30	64.3	75.3	14.6	97.0
>7	307(22.4)	52(16.9)	52	1,029	255	32	61.9	80.1	16.9	97.0
>8	264(19.3)	52(19.7)	52	1,072	212	32	61.9	83.5	19.7	97.1
>9	228(16.7)	47(20.6)	47	1,103	181	37	56.0	85.9	20.6	96.8
>10	195(14.3)	44(22.6)	44	1,133	151	40	52.4	88.2	22.6	96.6
>11	173(12.6)	42(24.3)	42	1,153	131	42	50.0	89.8	24.3	96.5
>12	153(11.2)	39(25.5)	39	1,170	114	45	46.4	91.1	25.5	96.3
>13	134(9.8)	37(27.6)	37	1,187	97	47	44.0	92.4	27.6	96.2
>14	122(8.9)	36(29.5)	36	1,198	86	48	42.9	93.3	29.5	96.1
>15	111(8.1)	36(32.4)	36	1,209	75	48	42.9	94.2	32.4	96.2
>16	106(7.7)	33(31.1)	33	1,211	73	51	39.3	94.3	31.1	96.0
>17	98(7.2)	33(33.7)	33	1,219	65	51	39.3	94.9	33.7	96.0
>18	91(6.7)	33(36.3)	33	1,226	58	51	39.3	95.5	36.3	96.0
>19	88(6.4)	33(37.5)	33	1,229	55	51	39.3	95.7	37.5	96.0
>20	82(6.0)	31(37.8)	31	1,233	51	53	36.9	96.0	37.8	95.9
>25	66(4.8)	30(45.5)	30	1,248	36	54	35.7	97.2	45.5	95.9
>50	31(2.3)	21(67.7)	21	1,274	10	63	25.0	99.2	67.7	95.3
>75	6(0.4)	5(83.3)	5	1,283	1	79	6.0	99.9	83.3	94.2

FN, false negative; FP, false positive; NPV, negative predictive value; PPV, positive predictive value; TN, true negative; TP, true positive. The absolute risk for having a cancer diagnosed within 90 days at different thresholds according to the AI score is provided in the current Table 5. If choosing a threshold for a negative AI score at ≤5 then the risk in this group was found to be 2.9%. However, for a single patient with a score of ≤2, the risk was 1.2%. In comparison, if a threshold for a positive AI score is set to >5, the risk of being diagnosed with a cancer within 90 days in this group of patients was found to be 13%. However, for a patient with a score of >25, the risk was 45.5%.

Figure 4: ROC curve. Target “cancer within 90 days” with the LR method.

Cohort IA: represents the training cohort. Cohort IB: represents the validation cohort. Cohort II: represents the validation test cohort.

for a negative test, the risk for having a cancer within 90 days would be 1.2% for a patient having a score of 2 and 1.1% at a score of 1. However, a patient having an AI score of eight will be classified as having a positive test and in the current study this patient will have a 19% absolute risk of being diagnosed with cancer within 90 days (Table 5).

Thus, the developed AI model can provide a predictive score, identifying patients with an increased risk of having

cancer in addition to identifying those patients who may face a minor risk of having cancer. In this way, the score may serve as a supplementary source of information for the GPs overall assessment and clinical decision making regarding further diagnostic work up and follow-up strategy for the individual patient.

However, even at low AI risk scores; there is a risk of overlooking patients with cancer. This is not surprising bearing in mind that the patient has already consulted the GP due to symptoms. Because the GP has decided that the patient needs further examinations via the “Suspicion of Hidden Cancer” pathway, it may be assumed that a heightened risk of cancer already exists, based on the GPs choice of referring the patient. This is in accordance with the work of Watson et al. [4] that concludes that blood tests in primary care such as hemoglobin, platelets, serum calcium, liver function tests and inflammatory markers may indicate cancer in patients with non-specific symptoms but cannot rule out the presence of cancer.

To provide a context, we have compared performance of the developed AI model with performance of the immunochemical faecal blood test (iFOBT) used in the Danish screening program for colorectal cancer with an estimated sensitivity of about 79% and PPV in a range of 3–8% [10]. Of those participating in the Danish screening program, 6–7% of citizens present a value above the positive threshold and are offered a colonoscopy, where between 3 and 8% will be diagnosed with colorectal cancer. In comparison, in our developed AI model applied to data from the validation test Cohort II, a patient with a positive AI score of three will have a 10.7% risk of being diagnosed with cancer within 90 days. At an AI score 3 we found a sensitivity of 82% and a specificity of 55% as

Table 6: LR model score from 0–100 in the validation Cohort IB.

LR score	TP, n	FN, n	TN, n	FP, n	Sensitivity, %	Specificity, %	PPV, %	NPV, %
0	55	0	0	987	100.0	0.0	5.3	NA
2	53	2	263	724	96.4	26.6	6.8	99.2
4	47	8	557	430	85.5	56.4	9.9	98.6
6	36	19	743	244	65.5	75.3	12.9	97.5
8	30	25	831	156	54.5	84.2	16.1	97.1
10	28	27	879	108	50.9	89.1	20.6	97.0
20	16	39	967	20	29.1	98.0	44.4	96.1
30	10	45	978	9	18.2	99.1	52.6	95.6
40	7	48	980	7	12.7	99.3	50.0	95.3
50	7	48	982	5	12.7	99.5	58.3	95.3
60	7	48	983	4	12.7	99.6	63.6	95.3
70	5	50	983	4	9.1	99.6	55.6	95.2
80	5	50	983	4	9.1	99.6	55.6	95.2
90	3	52	984	3	5.5	99.7	50.0	95.0
100	0	55	987	0	0.0	100.0	NA	94.7

FN, false negative; FP, false positive; NPV, negative predictive value; PPV, positive predictive value; TN, true negative; TP, true positive.

Table 7: LR model score from 0-10 in the validation test Cohort II.

LR score	TP, n	FN, n	TN, n	FP, n	Sensitivity, %	Specificity, %	PPV, %	NPV, %
0	84	0	0	1,284	100.0	0.0	6.1	NA
2	79	5	308	976	94.0	24.0	7.5	98.4
4	69	15	694	590	82.1	54.0	10.5	97.9
6	60	24	935	349	71.4	72.8	14.7	97.5
8	55	29	1,065	219	65.5	82.9	20.1	97.3
10	44	40	1,137	147	52.4	88.6	23.0	96.6

FN, false negative; FP, false positive; NPV, negative predictive value; PPV, positive predictive value; TN, true negative; TP, true positive.

presented in Table 5. A recent paper using DNS methylation patterns for early detection of cancer yielded sensitivities for stage I cancers of 18% and stage II of 43% [11].

A routine set of blood tests has previously been studied by Naeser et al. [12] but from a different perspective. Naeser et al. analyzed the results from 1,499 patients of which 12.2% were subsequently diagnosed with cancer and discovered that the probability of cancer increased with the number of test results outside the reference range or with specific combinations of abnormal test results. However, just counting the number of tests results outside the reference range is insufficient in this context, as it has been shown that even changes within the normal range may indicate increased risk for cancer. Thus, a reduced P-Albumin concentration and an increased platelet number in blood, increase the risk for cancer in a concentration dependent way.

In addition, a recent study addresses the issue of combining simple blood tests to identify primary care patients with unexpected weight loss for cancer investigation. They found that combinations of simple blood test abnormalities could be used to identify patients with unexpected weight loss who warrant referral for investigation, while people with combinations of normal results could be exempted from referral [13].

Our data shows that both AI and LR models can be used for calculating a predictive score for being diagnosed with cancer within 3 months. This risk can be used by the GP in the overall risk assessment, together with the other information obtained from the anamnesis and objective examination. The GP may recommend a faster investigation for patients with a score corresponding to a high risk while a watchful waiting strategy may be used for patients with a score corresponding to a low risk.

It is intuitively easy to use the personal absolute cancer risk as a percentage rate to prioritize those patients with the highest risk vs. those with a much lower risk, already at the outset of the diagnostic process.

Finally, future research may determine whether LR in itself is sufficiently robust to be used as a risk assessment tool, given that one then does not have to adjust an AI algorithm and there is thus no “requirement” for a standardized set of blood tests, if the initiative is to be scaled to other analytical profiles with a different set of blood tests.

Strengths and limitations of this study

The study cohorts in the current study are well defined. A further strength is the validation of the developed model in the “out of time” validation test Cohort II in order to overcome the challenge with overfitting. In addition, the use of routine laboratory tests available for primary care increases its clinical applicability, avoiding the requirement for laboratory tests often only available in a specialized hospital setting.

Our study has limitations. Firstly, the study population is relatively small. Therefore, the study was performed with a full model without considering gender. However, in future studies, gender specific models will be more relevant. In addition, it is a retrospective study, and a prospective design with a reassessment of the set of blood test would probably strengthen the model. In addition, we did not assess, which single blood test from the analytical profile had the greatest value in detecting cancer and this needs to be done in future studies. Finally, prior to a general clinical implementation it is crucial that the score is further validated to determine its applicability in other populations.

This work did not include demographic risks, such as the lifestyle or social conditions of patients, as its main purpose was a proof of concept study as to whether laboratory tests by themselves contained sufficient valuable intrinsic information to be robust when used for AI data processing to

calculate absolute risk of cancer. This however, is a prerequisite if laboratory test results should be used in combination with other kinds of data to enable even more comprehensive risk prediction models, and thereby ultimately leading to better assessment of the patient's risk of having cancer in a low prevalent setting as the primary care.

Conclusions and perspectives

The current study demonstrates the ability to develop an AI model based on routine laboratory blood tests which is able to provide an easy-to use risk score to predict cancer within 90 days. The use of laboratory tests, widely available for primary care, increases its clinical applicability and the AI risk score may prove to be a valuable tool in the clinical decision-making, supporting the GP triage whether a patient needs faster further investigations or instead adopt a watchful waiting strategy.

The AI score, however, needs to be further external validated to determine its applicability in other populations.

A future improvement of the AI risk score may be obtained by further development and customization regarding the panel of included routine laboratory blood tests. In addition, improvements may be obtained by using gender specific models and by combining laboratory test results and demographic data in future prediction models.

Research funding: This project was financed by the Health Authorities of the Region of Southern Denmark.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests: Christian Hardahl is an employee (Subject Matter Expert) at the company from which SAS Viya was purchased.

Informed consent: This project was performed retrospectively without patient contact in connection with the project after permission from the Danish Board for Patient Safety. The project is performed according to the Danish Health Law § 42. No patient data were analysed outside the legal boarder of the Region of Southern Denmark Health System.

Ethical approval: The project was carried out by permission from the Danish Board for Patient Safety (number 3-3013-2954/1) and Hospital Lillebaelt.

References

1. Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: an overview. *J Thorac Dis* 2019;11:5574–84.
2. Fei Y, Li WQ. Improve artificial neural network for medical analysis, diagnosis and prediction. *J Crit Care* 2017; 40:293.
3. Sud A, Torr B, Jones ME, Broggio J, Scott S, Loveday C, et al. Effect of delays in the 2-week-wait cancer referral pathway during the COVID-19 pandemic on cancer survival in the UK: a modelling study. *Lancet Oncol* 2020;21:1035–44.
4. Watson J, Mounce L, Bailey SE, Cooper SL, Hamilton W. Blood markers for cancer. *BMJ* 2019;367:l5774.
5. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* 2018;359:926–30.
6. Chen X, Gole J, Gore A, He Q, Lu M, Jun M, et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. *Nat Commun* 2020;11:3475.
7. Schneider JL, Layefsky E, Udaltsova N, Levin TR, Corley DA. Validation of an algorithm to identify patients at risk for colorectal cancer based on laboratory test and demographic data in diverse, community-based population. *Clin Gastroenterol Hepatol* 2020; 18:2734–41.e6.
8. Gerhardt W, Keller H. Evaluation of test data from clinical studies. I. Terminology, graphic interpretation, diagnostic strategies, and selection of sample groups. II. Critical review of the concepts of efficiency, receiver operated characteristics (ROC), and likelihood ratios. *Scand J Clin Lab Invest Suppl* 1986;181:1–74.
9. Roelofs R, Shankar V, Recht B, Fridovich-Keil S, Hardt M, Miller J, et al. A meta-analysis of overfitting in machine learning. *Neural Information Processing Systems (NeurIPS)*. Vancouver Convention Center, Vancouver CANADA 2019;32:207979247.
10. Robertson D, Lee J, Boland C, Dominitz J, Giardiello F, Johnson D, et al. Recommendations on fecal immunochemical testing to screen for colorectal neoplasia: a consensus statement by the US Multi-Society Task Force on colorectal cancer. *Gastrointest Endosc* 2017;152. <https://doi.org/10.1038/ajg.2016.492>.
11. Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV, CCGA Consortium. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann Oncol* 2020;31:745–59.
12. Naeser E, Moeller H, Fredberg U, Frystyk J, Peter Vedsted P. Routine blood tests and probability of cancer in patients referred with nonspecific serious symptoms: a cohort study. *BMC Cancer* 2017; 17:817.
13. Nicholson BD, Aveyard P, Koshariis C, Perera R, Hamilton W, Oke J, et al. Combining simple blood tests to identify primary care patients with unexpected weight loss for cancer investigation: clinical risk score development, internal validation, and net benefit analysis. *PLoS Med* 2021;18: e1003728.