

Deep learning-based identification of patients at increased risk of cancer using routine laboratory markers

Vivek Singh^{1,*}, Shikha Chaganti¹, Matthias Siebert², Soumya Rajesh¹, Andrei Puiu^{3,4}, Raj Gopalan⁵, Jamie Gramz⁶, Dorin Comaniciu¹, and Ali Kamen¹

¹Siemens Healthineers, Digital Technology and Innovation, Princeton, 08540, USA

²Siemens Healthineers, Digital Technology and Innovation, Erlangen, 91052, Germany

³Siemens SRL, Advanta, Brasov, 500007, Romania

⁴Transylvania University of Brasov, Automation and Information Technology, Brasov, 500174, Romania

⁵Siemens Healthineers, Laboratory Diagnostics, Tarrytown, NY 10591, USA.

⁶Siemens Healthineers, Digital and Automation, Malvern, PA 19355, USA.

*vivek-singh@siemens-healthineers.com

ABSTRACT

Early screening for cancer has proven to improve the survival rate and spare patients from intensive and costly treatments due to late diagnosis. Cancer screening in the healthy population involves an initial risk stratification step to determine the screening method and frequency, primarily to optimize resource allocation by targeting screening towards individuals who draw most benefit. For most screening programs, age and clinical risk factors such as family history are part of the initial risk stratification algorithm. In this paper, we focus on developing a blood marker-based risk stratification approach, which could be used to identify patients with elevated cancer risk to be encouraged for taking a diagnostic test or participate in a screening program. We demonstrate that the combination of simple, widely available blood tests, such as complete blood count and complete metabolic panel, could potentially be used to identify patients at risk for colorectal, liver, and lung cancers with areas under the ROC curve of 0.76, 0.85, 0.78, respectively. Furthermore, we hypothesize that such an approach could not only be used as pre-screening risk assessment for individuals but also as population health management tool, for example to better interrogate the cancer risk in certain sub-populations.

Introduction

This paper focuses on the use of multiple biomarkers for the assessment of patients, or identification of otherwise healthy individuals, who are at increased risk of cancer. With the high mortality rate associated with cancer patients, significant research has been conducted to help identify patients at higher risk, starting with identifying medical conditions that increase the risk of cancer, such as diabetes, or genetic predispositions that promote its development¹. Furthermore, various screening procedures have been developed to help facilitate early diagnosis such as the Faecal Immunochemical Test (FIT) and colonoscopy for colorectal cancer (CRC)², mammography for breast cancer³, and low-dose computed tomography (LDCT) for lung cancer⁴. However, cancer screening rates and their uptake remains lower than desired, e.g., in the US⁵. While there are several factors contributing to this low uptake, one of the key factors is the lack of awareness within the general population. This is even more important to address for people who may be at increased risk and would benefit from early and/or regular screening. In other words, there is still a need for convenient tests for early detection of rapidly progressing diseases such as cancer so that intervention can start as early as possible⁶.

Several cancer risk prediction/assessment tools based on demographic, socioeconomic or blood based markers have been developed over the years, and studies have shown that cancer risk assessment algorithms could have an impact in early cancer diagnosis⁷. For instance, the Qcancer 10 year risk algorithm⁸ considers the age, ethnicity, deprivation, body mass index, smoking, alcohol, previous cancer diagnoses, family history of cancer, relevant comorbidities, and medication data for a patient and predicts the cancer risk for 11 types of cancers. Nartowt et al.⁹ reported high concordance in the prediction of CRC into low, medium and high groups using an artificial neural network trained on patient data comprising age, sex, and complete blood count (CBC). ColonFlag¹⁰ can be used to identify individuals at high risk of CRC using specific blood-based markers and refer them to screening procedures such as colonoscopy. More recently, a cell-free DNA-based blood test for the early detection of CRC has been clinically validated in the ECLIPSE study¹¹. Moreover, multi-cancer early detection technologies¹² such as the

Galleri test^{13,14} can identify abnormal methylation patterns in cell-free DNA to detect a cancer signal and predict its origin.

Besides algorithm development, the deployment of the algorithm and communication of the findings play a critical role in acceptance and clinical use of the algorithm¹⁵, and must be taken into account to facilitate screening uptake. To this end, instead of defining a test with a specific set of ingredients catered towards a particular cancer, we propose to use commonly measured blood markers, often obtained during the annual physical exam, and obtain a risk profile for multiple cancers. Furthermore, instead of reporting a risk score, we compute the pre- and post-test odds of a patient at risk of developing cancer over the next 12 months.

A key challenge in developing a model that considers several biomarkers is to deal with a significant degree of missingness in the historical data as not all markers may be obtained at each encounter. Although this issue can be partly alleviated by considering the biomarkers obtained at an annual physical exam, we observed that in real world data, there is still a significant degree of missingness, either due to a lack of awareness, insurance coverage or reimbursement, among other reasons. A standard approach to deal with missingness in input data is to impute the missing values using statistical methods, such as expectation maximization and regression¹⁶. However, the quality of imputed data is limited and can significantly impact the generalization ability of the trained model.

In this work, we address the aforementioned challenges by training a deep learning model, *Deep Profiler*, which takes the age, sex, and commonly obtained blood biomarkers included in CBC and Comprehensive Metabolic Panel (CMP), and outputs a likelihood ratio of a patient to develop cancer over the period of the following 12 months (see Figure 1). The Deep Profiler architecture employs a variational autoencoder (VAE) model that is pre-trained to impute missing data similar to the masked language modeling technique. Subsequently, we train cancer-specific risk prediction models from the shared encoded latent space and compute the likelihood ratio for each patient. We validate the proposed method over screening-relevant cohorts for three different cancers - colorectal, liver, and lung. These are among the top cancers responsible for cancer related mortality rate in the US (<https://seer.cancer.gov/statfacts/html/common.html>, accessed April 30, 2024.).

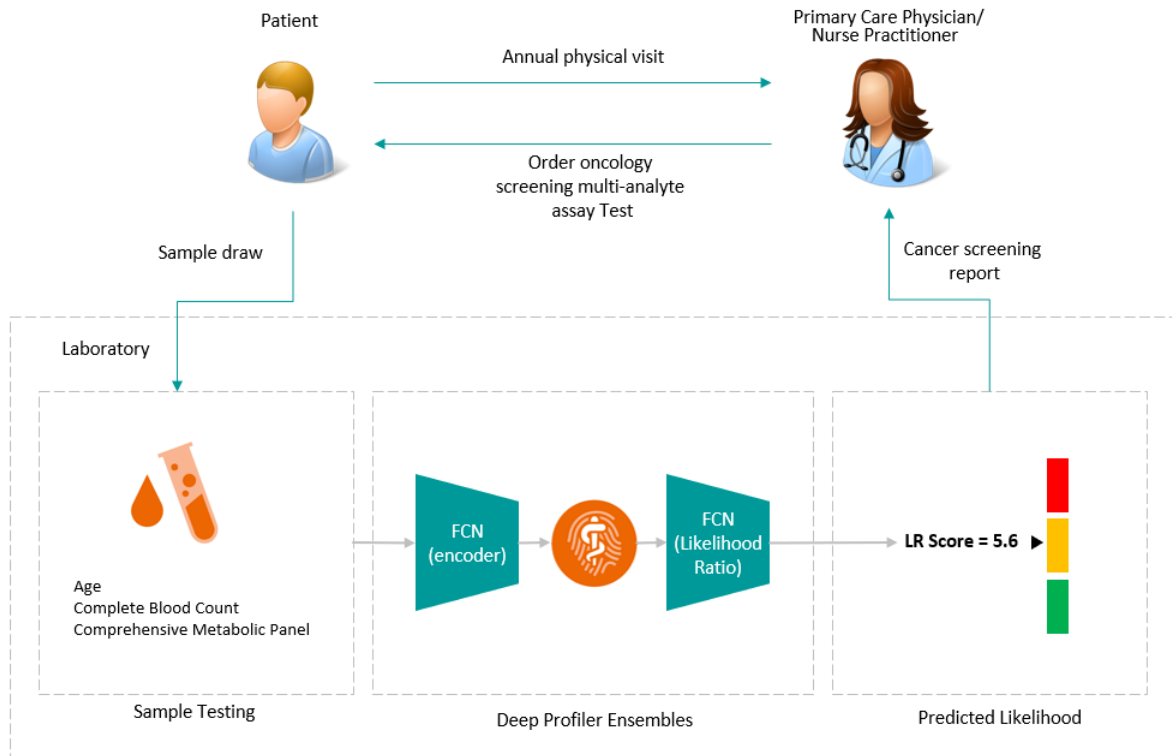


Figure 1. Workflow of using a biomarker-based pre-screening test.

Results

Patient characteristics

To evaluate Deep Profiler across multiple cancer types, we created a cohort of individuals that either have no diagnosis of cancer, or were diagnosed with malignant neoplasms of either colorectal, liver or lung. The validation cohorts each include nearly

5k patients with no cancer diagnosis, and 94, 189 and 224 patients who develop colorectal, liver or lung cancer, respectively. Note that the patients in the validation set were not used during model development and training. The cohorts used for model development include nearly 10k patients with no cancer diagnosis, and 293, 626 and 683 patients who develop colorectal, liver or lung cancer, respectively. Supplementary Table S1 summarizes the statistics of various biomarkers, including age and sex, over the entire cohort covering all three cancer types. Significant differences in the distributions of biomarker measurements in cancer cohorts as compared to control cohorts are indicated.

Likelihood ratio

We train Deep Profiler ensembles to estimate cancer risk for three cancer types: colorectal, liver and lung cancer. Given the biomarker values, the model outputs an array of risk scores (from the ensemble) for each cancer type, which are then utilized to compute the likelihood ratio (LR) of post- to pre-test odds. LRs at increasing risk thresholds, together with Receiver Operating Characteristic (ROC) curves, are shown in Figures 2a, 2b, and 2c for colorectal, liver, and lung cancer, respectively (corresponding precision-recall curves are provided in Supplementary Figure S1).

Since Næser et al. reported that the probability of cancer increased with the number of test results outside the reference range¹⁷, we used the total out-of-reference-range (OoR) markers as a baseline. Indeed, cancer likelihood increases with the total OoR markers. However, the increase is significantly lower as compared to the Deep Profiler models. Similarly, when compared individually with any of the top-5 markers identified by the model for each cancer, Deep Profiler provides a 2-3 fold improvement.

In CRC, age is the only factor considered to determine screening eligibility according to recommendations by the US Preventive Services Task Force (USPSTF)¹⁸. Therefore, we also provide an LR curve for age (normalized between 40 and 85 years) as the single indicator. While CRC shows increased prevalence until the age of about 60, the increase of LR provided by Deep Profiler is significantly higher, with a 4-fold improvement in LR at a threshold greater than 0.8.

Overall, the liver-specific Deep Profiler model performs significantly better as compared to our Deep Profiler models for CRC and lung cancer (LR of ~7.5 vs. ~4.5 at a threshold of 0.8, respectively), which is also reflected in the corresponding ROC curves. Importantly, performance is preserved when assessed in cohort-specific subsets that only include cases for which measurements for at least nine of top-15 markers are available, indicating that the models did not learn patterns related to missing marker measurements.

Relevant biomarkers by cancer type

To gain insights into the laboratory markers with highest impact on the LR, we show cohort-level SHAP summaries for our colorectal, liver, and lung cancer models in Figures 3a, 3b, and 3c, respectively. Since the LR can range from zero to infinity, we passed the computed LR for each sample through a logistic function with mean 5 and scale 0.5. This ensures that outliers, i.e., samples with significantly low or high LR, do not significantly impact the SHAP value distribution.

While only three (age, albumin and hematocrit) of the 15 most important laboratory markers are shared across all cancer types (only age among the five most important markers), there are at least seven markers shared between any two cancer types. In contrast, fine-tuned cancer type-specific prediction models are required to give more weight to markers predictive of only one cancer type, as, in fact, one (neutrophils), three (total protein, WBC and basophils (%)), and three (BUN-creatinine ratio, calcium, and ALT) markers are most important in only colorectal, liver, and lung cancer, respectively.

Several of the biomarkers that are identified as important in our SHAP analysis have also been independently studied and reported. For instance, six of the 12 markers that were integrated into one or both sex-specific colon cancer prediction models by Goshen et al.¹⁹ (ten of which are also available in our cohort) also appear to be most important in our CRC model (RDW, MCV, neutrophils, monocytes, hemoglobin, and AST).

For liver cancer, there are currently no screening recommendations, except for hepatitis B carriers, who are recommended by the American Association for the Study of Liver Diseases (AASLD) to start screening at age 40 and 50 years for men and women, respectively. In our analysis, platelet counts appear as the most important blood marker predictor of liver cancer. In fact, raised platelet counts ($> 400 \times 10^9/l$) indicative of thrombocytosis, or even high-normal platelet counts ($326\text{--}400 \times 10^9/l$) have been reported to be associated with higher cancer incidence, in particular colorectal and lung cancer²⁰. However, patients with high-normal platelet counts had, in general, advanced-stage cancer at diagnosis, which may explain why platelets are particularly important in our liver cancer model, as liver cancer is mostly detected at a late stage, emphasizing the need for improved guidelines concerning screening eligibility.

Of note, the lack of a clear color gradation indicates interactions between different laboratory markers, illustrating the need to train a holistic model by integrating multiple laboratory markers.

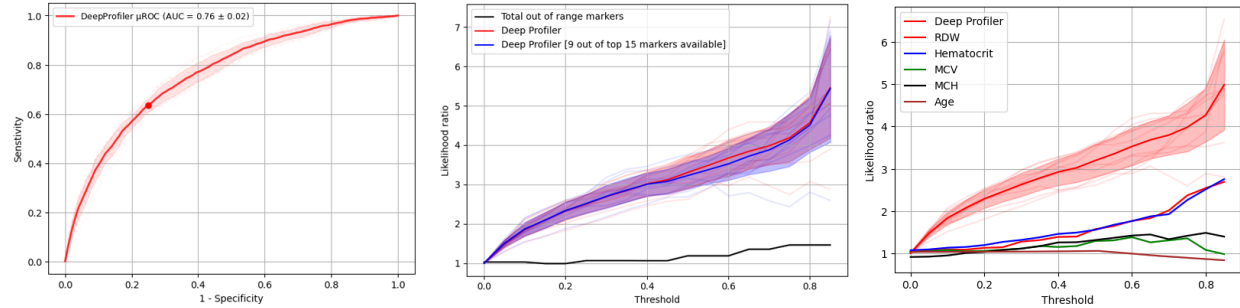
Analysis by comorbidities

For each of the cancer cohorts, we used phecodes²¹ to identify comorbidities that are more likely to be present in the underlying population prior to cancer diagnosis. Relative odds ratios were computed to select comorbidities that are significantly different

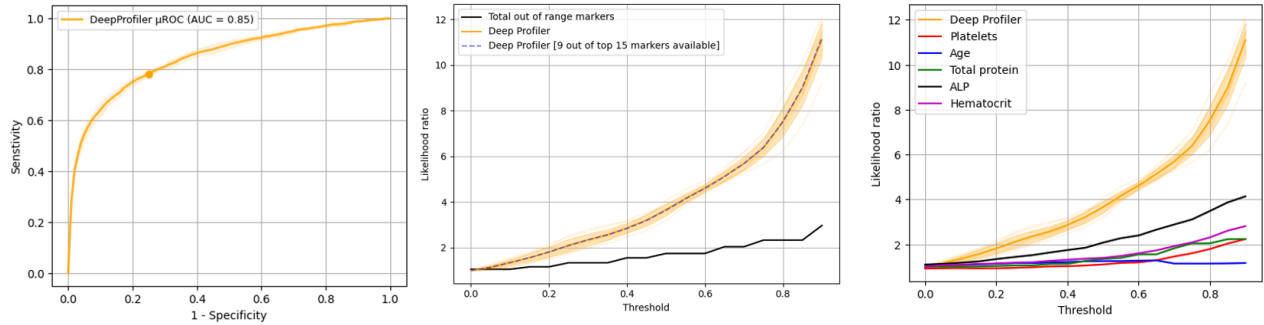
between cancer and control samples using Fisher's exact test (Figures 4a, 4b, and 4c). We found that those are usually the conditions or symptoms that raise suspicion of cancer or other organ dysfunction in a cancer population. For instance, the top three significant comorbidities prior to diagnosis in the CRC cohort are other disorders of the intestine, benign neoplasm of the colon, and hemorrhage of rectum and anus (Figure 4a).

Since the FIT test screens stool for occult blood from the lower intestines, we use the subgroup of patients with blood in stool as a surrogate to evaluate the added value of our model in comparison to an established screening test. Compared to the base prevalence, we see an increase in LR greater than 3-fold, at a risk threshold greater than 0.8 (Figure 4a). Furthermore, we also see added value in subgroups of patients with conditions such as hemorrhage of rectum and anus or gastrointestinal tract.

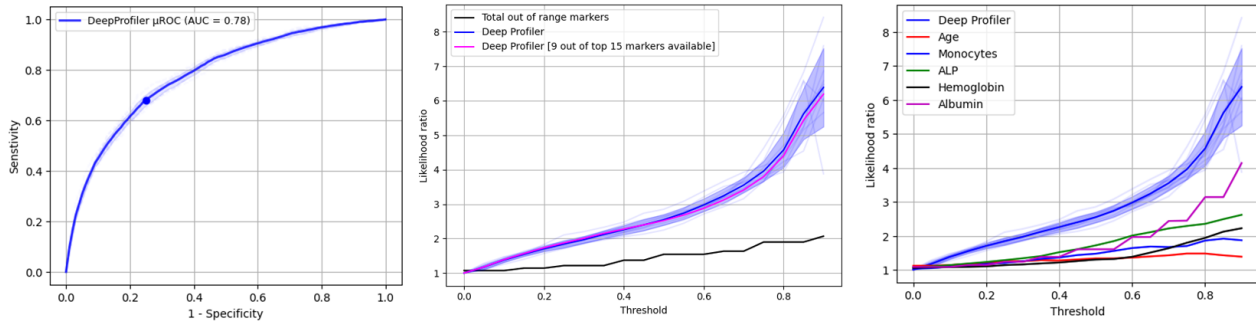
For lung cancer, the USPSTF criteria for screening comprise age and smoking history²². While our lung cohort does not include smoking history, we analyzed the performance of our model on a subgroup of patients having a tobacco use disorder. Notably, our model shows added value in identifying high-risk patients (more than 50 % increase in LR as compared to the base prevalence, at a risk threshold greater than 0.9, Figure 4c).



(a) Colorectal cancer



(b) Liver cancer



(c) Lung cancer

Figure 2. Quantitative performance assessment on (a) colorectal, (b) liver, and (c) lung cancer validation cohorts. ROC curves represent the predictions of the corresponding models (left). Likelihood ratio plots show the likelihood ratios of corresponding models on the full cohort and the subgroup of patients having at least nine out of the top 15 markers available, for increasing risk thresholds (middle). Likelihood ratio curves are compared to those of baseline models that are either based on the total number of OoR markers (middle) or single markers (right), corresponding to the top five markers for each cancer type (cf. Figure 3), including age for CRC. Thin lines depict likelihood ratio curves of base models, while ribbons correspond to one standard deviation of their likelihood ratios. ALP, alkaline phosphatase; MCH, mean corpuscular hemoglobin; MCV, mean corpuscular volume; RDW, red cell distribution width.

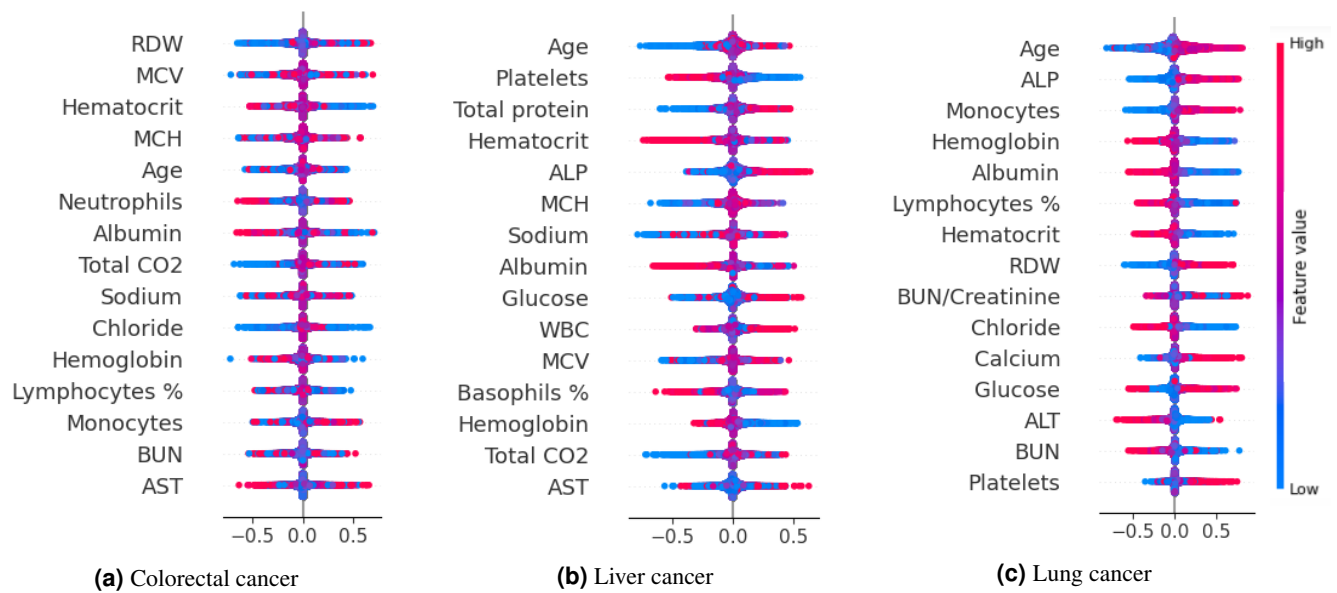


Figure 3. Cohort-level SHAP summaries showing the contribution of the 15 laboratory markers with highest impact on the LR, separately for the (a) colorectal, (b) liver, and (c) lung cancer model. Red and blue points correspond to patients with high and low values of the corresponding laboratory markers, respectively. Laboratory markers are ordered along the y-axis with respect to their overall importance for the prediction, with most important markers at the top. ALP, alkaline phosphatase; AST, aspartate aminotransferase; BUN, blood urea nitrogen; MCH, mean corpuscular hemoglobin; MCV, mean corpuscular volume; RDW, red cell distribution width; WBC, white blood cells.

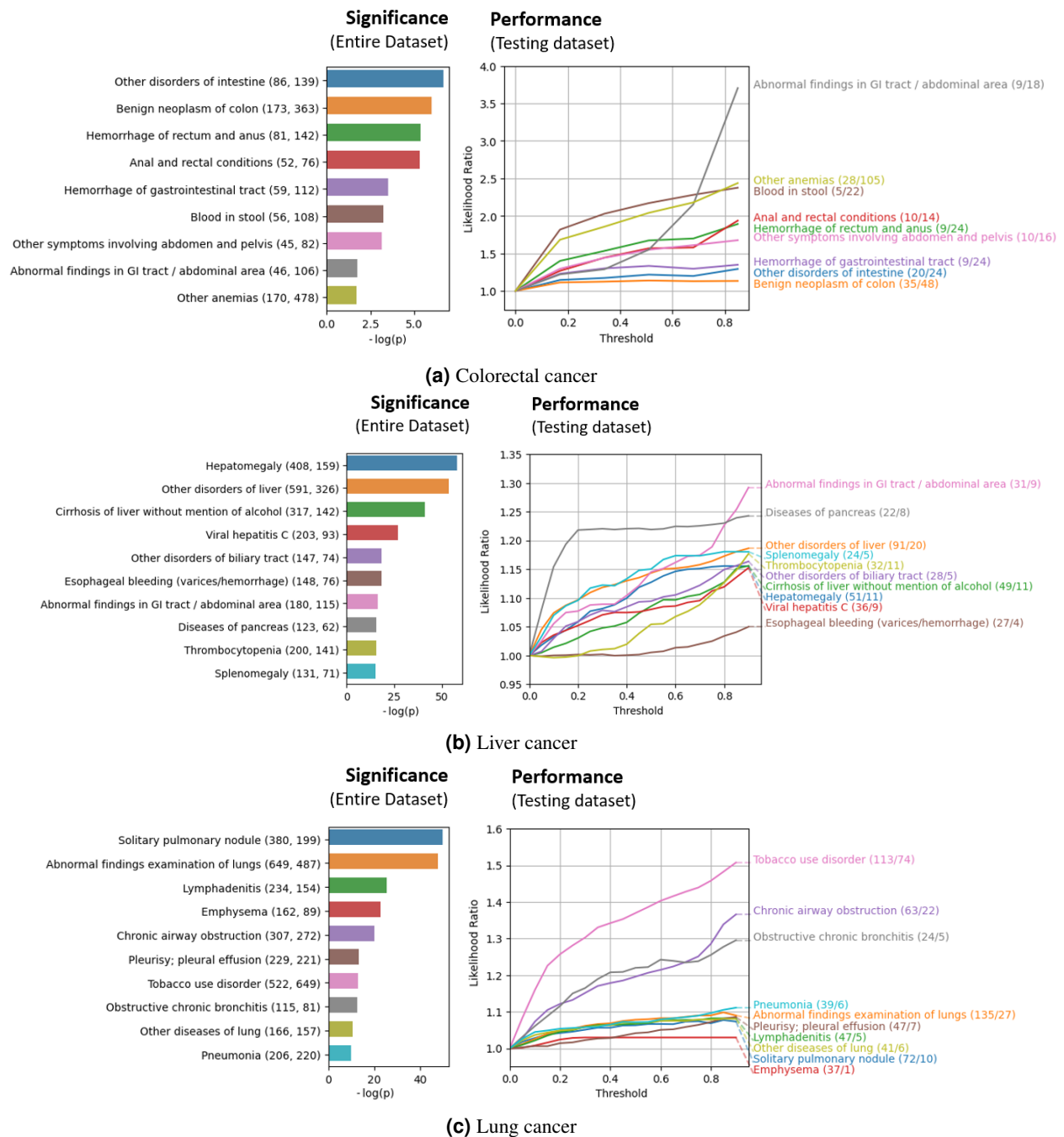


Figure 4. Comorbidity analysis on (a) colorectal, (b) liver, and (c) lung cancer validation cohorts. Likelihood ratio plots show the likelihood ratios of corresponding models on subgroups of patients suffering from the indicated comorbidities, for increasing risk thresholds on the testing dataset. Barplots illustrate the significance of association of the corresponding comorbidity with cancer type (as negative logarithm of the p-value obtained with Fisher's exact test) over the entire dataset. The prevalence of the respective comorbidities in the cancer-positive and control cases is listed with each comorbidity. Note that the base prevalence is different for each comorbidity. GI, gastrointestinal.

Discussion

Here, we report on the development and evaluation of separate deep learning-based risk prediction models for three cancer types based on routine laboratory marker measurements. Although other blood marker-based pan-cancer or cancer-specific risk prediction models have been reported, a direct comparison of performance and markers identified to be important for risk assessment is difficult due to varying (i) study designs, e.g., with respect to the follow-up period until cancer diagnosis (90 days²³ vs. 365 days in our model), (ii) cancer prevalence in training and validation cohorts (25.2% in²⁴ vs. 5.7 % in our lung cohort), (iii) inclusion criteria (e.g., smoking history in lung cancer), (iv) number and availability of markers (e.g., restrict to samples with complete data vs. imputation of missing data), (v) inclusion of non-routine blood markers (e.g., AFP and CA-125¹⁷) as well as patient characteristics other than age and sex²⁵. While the performance is not yet adequate enough for the model to be used as a diagnostic test, we anticipate it to be a valuable tool for recruiting patients into screening programs and, thus, increase screening uptake and efficiency overall, pending additional confirmatory studies. In future work, we also plan to adopt the model to personalize the screening interval after a negative screening result (e.g., based on LDCT).

While the current USPSTF guideline for lung cancer screening only considers age and smoking history for screening eligibility²⁶, a multitude of risk prediction models have been proposed and reported to show improved performance over the USPSTF criteria²⁷. For instance, the PLCO_{m2012} model includes race, body mass index, education, presence of chronic lung disease, personal history of cancer, and family history of lung cancer, in addition to age and smoking history²⁸. Furthermore, a four-marker protein panel (4MP), measuring a precursor form of surfactant protein B (Pro-SFTPB), cancer antigen 125 (CA-125), carcinoembryonic antigen (CEA), and cytokeratin-19 fragment (CYFRA 21-1), has been combined with PLCO_{m2012} to predict lung cancer risk²⁹. However, the required biomarkers are not assessed within routine blood testing, posing a challenge for the test's clinical utility. In contrast, our model is only based on routine laboratory markers and even performs comparable to models that integrate knowledge about smoking status or smoking history^{24,25}.

As cohorts were enriched with patients diagnosed with chronic liver disease, the better performance of our model in liver cancer might originate from an improved differential diagnosis. Interestingly, ALT and AST are not picked up as important markers for liver cancer either, which might also be tied to the model trying to differentiate from chronic liver disease cases in the control set. Analogous, the addition of cases with chronic bowel and lung disease could prove helpful in the colorectal and lung cancer setting, respectively.

The SHAP analysis in Figure 3 depicts the contributions of important laboratory markers to the normalized LR on the cohort level. In addition, we provide the relative contributions of laboratory markers to model prediction on the individual level as waterfall plots in Supplementary Figure S2 for two selected cases: (1) a liver cancer case with only four OoR markers (total CO₂, lymphocytes (%), hemoglobin, and hematocrit, Supplementary Figure S2a), of which total CO₂ and hematocrit are the only 2 of the 15 laboratory markers with highest impact on model prediction (Figure 3b), and a control case with 18 OoR markers (Supplementary Figure S2b), both correctly classified by the model with a high (> 0.8) and low (~ 0) risk, respectively. These example cases illustrate the importance of considering the measurements of a range of markers together to achieve optimal risk assessment.

In this work, we have focused on CBC and CMP blood tests as availability of complete lipid panel blood tests was limited in our cohorts. While lipid panel use is less common in Europe, it is more frequently ordered in the US. In fact, we expect its integration into our risk models to further improve cancer risk assessment by contributing complementary information on physiological or pathological status.

To avoid potential harm resulting from unnecessary follow-up procedures, differential life expectancy, e.g., as a result of competing comorbidities, needs to be taken into account³⁰. This is of particular importance, as chronic diseases may increase the risk of complications from biopsies and cancer treatment. While we have focused on prioritizing patients for screening eligibility, allowing clinicians to discuss with patients who have been predicted to be at increased risk to conduct follow-up screening procedures as early as possible, we plan to investigate different risk grading and threshold approaches that may be applied in dependence of the presence/absence of certain comorbidities.

Similarly, it would be of great value if we could explore the contributions of laboratory markers with respect to cancer stage. Unfortunately, we did not have access to a cohort-linked cancer registry and, therefore, lack information on cancer stage. In the future, we anticipate such an analysis to open up the possibility to develop cancer risk prediction models that also provide an assessment of cancer progress.

Methods

Study selection

We created a cohort of individuals diagnosed with malignant neoplasms of colorectal, liver or lung, or had no confirmed diagnosis of any malignant neoplasm, within the period from 2017 to 2021. We further enriched the cohort by considering additional patients who are diagnosed with chronic kidney disease and/or chronic liver disease but have no personal history

or a novel cancer diagnosis during the aforementioned period. The cohort was created using the Prognos Factor® platform, and the corresponding patient records were obtained from Prognos Health. The patient records include the laboratory marker measurements (blood or urine based) from various encounters as well as CPT³¹ and ICD-10 codes³² from claims. ICD-10 codes of patients were used as surrogate for the diagnosis for selecting patients in this study. All the records obtained correspond to an anonymized healthcare provider in the US.

Screening-based cohorts

To create cohorts for the pre-screening scenario, we utilized the medical claims data to select the patients and their visits that correspond to a screening procedure. Table 1 lists the CPT and ICD-10 codes used for each cancer type. If screening-related codes were not reported for a patient, we used diagnostic procedure codes (listed in Table 1 as well).

Cancer type	Screening		Diagnosis	
	Procedure codes	Encounter codes	Procedure codes	ICD-10 codes
Colorectal	G0105, G0120, G0121, G2204, 45378, 45388, 45330, 45381	Z1211, Z1212, Z1213	45380, 45382, 45384, 45385, 45390	C18, C19, C20
Liver	76700, 76705, 78215	Z1289	47000, 74176, 74177, 74148	C22
Lung	G0296, 71271, X-ray (71045, 71046, 71047, 71048)*	Z122	71250	C34

* Reimbursement as a screening procedure was discontinued in late 2018.

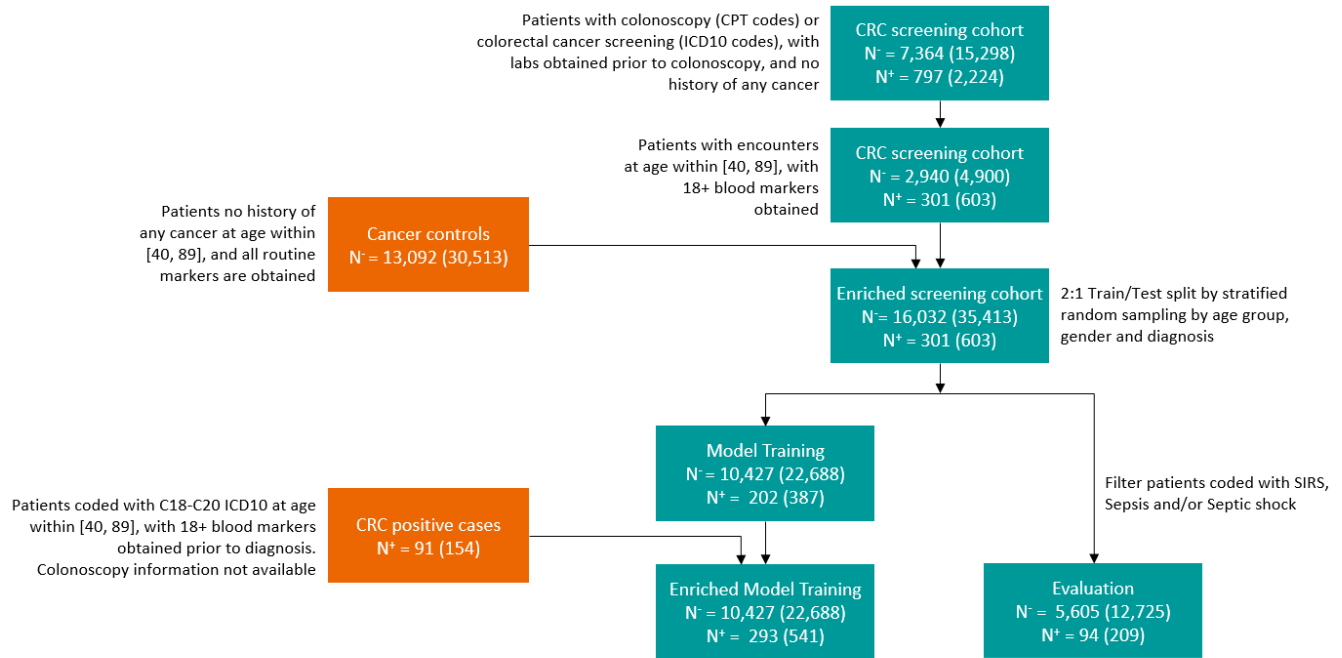
Table 1. Screening and diagnosis relevant CPT and ICD-10 codes used for patient selection and assignment of diagnosis labels for each cancer.

For each patient in the screening cohorts, we next determined if the patient had a positive cancer diagnosis. Although we used ICD-10 codes to select patients for the study from the Prognos Factor® platform, we recognized that the use of ICD-10 codes as confirmed diagnosis may not be sufficiently reliable³³. Therefore, we ensured that at least one additional CPT or ICD-10 code associated with cancer diagnosis and/or therapy (chemotherapy and/or radiation therapy) is present after diagnosis. Figures 5a, 5b, and 5c show the respective consort flow diagrams for each cancer. The screening cohorts for each cancer type as described above correspond to the first cell in the flow diagram for each cancer.

For patients with a positive cancer diagnosis, we only considered the records from visits within 12 months prior to the visit associated with the screening/diagnostic procedure. This ensured that the records all had a cancer diagnosis within the next 12 months. For patients with no cancer, we only considered records for which there are subsequent records available. For each cell in the consort flow diagrams, both the aggregate patient and record counts are shown.

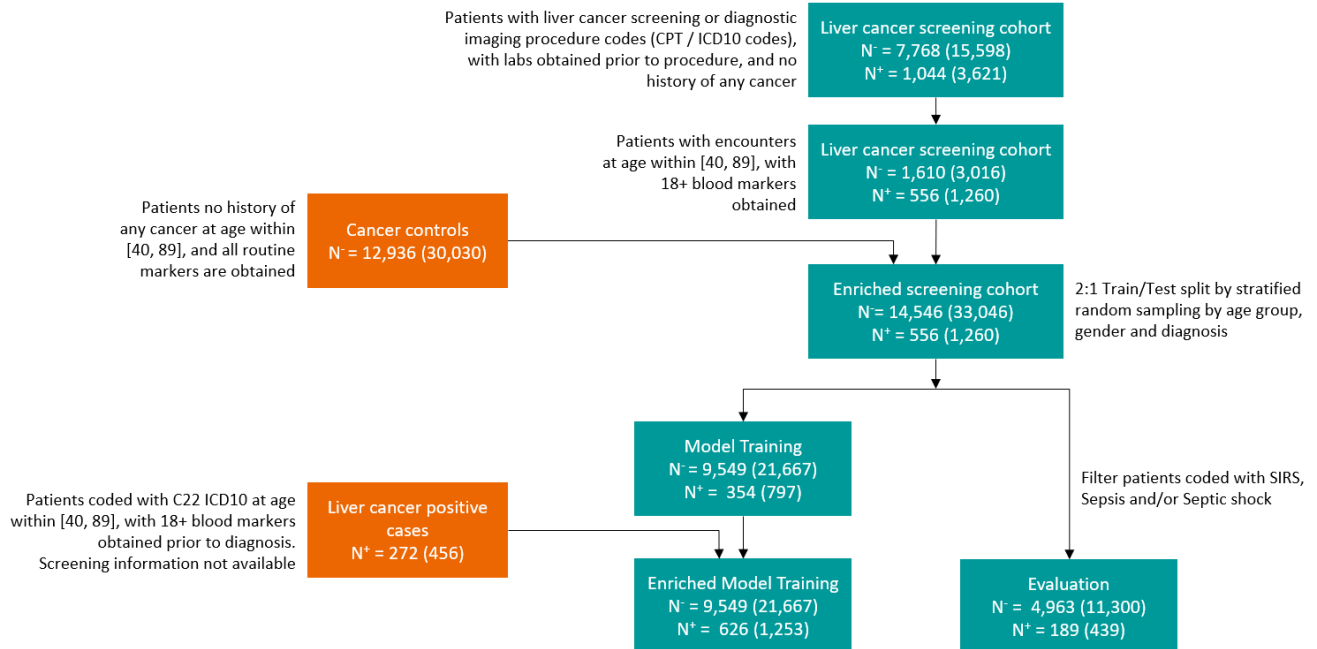
To be consistent with the pre-screening scenario, we next identified the records of all selected patients between the age of 40 and 89, with no prior diagnosis of cancer. The laboratory markers obtained at each visit are tied to the underlying conditions that were diagnosed or monitored and, hence, the number and type of markers varied significantly, e.g., from a single marker to all the markers in CBC and CMP. Thus, we only considered visits with at least 18 markers for our analysis and discarded all other visits. While this approach potentially results in some data loss, it helps avoid the machine learning model to overfit to the specific markers measured. Note that CBC and CMP also include derived markers such as BUN-creatinine ratio which, if missing in the records, are computed using the reported BUN and creatinine values. Similarly, blood-count markers and their percentage markers were computed if either one was not reported.

We subsequently enriched the resulting cohorts by including records of additional patients between the age of 40 and 89 who have no history of cancer and no cancer diagnosis within the next 12 months. While this step enriched the data for model development, more importantly, it also incorporated individuals who are eligible for screening but may not have gone through screening within the period considered in this study. We note that the number of patients (records) that were used to enrich each cancer cohort is not the same in the consort flow diagram. This is primarily due to the fact that there are patients that have gone through a screening procedure for one or more cancers but not all.



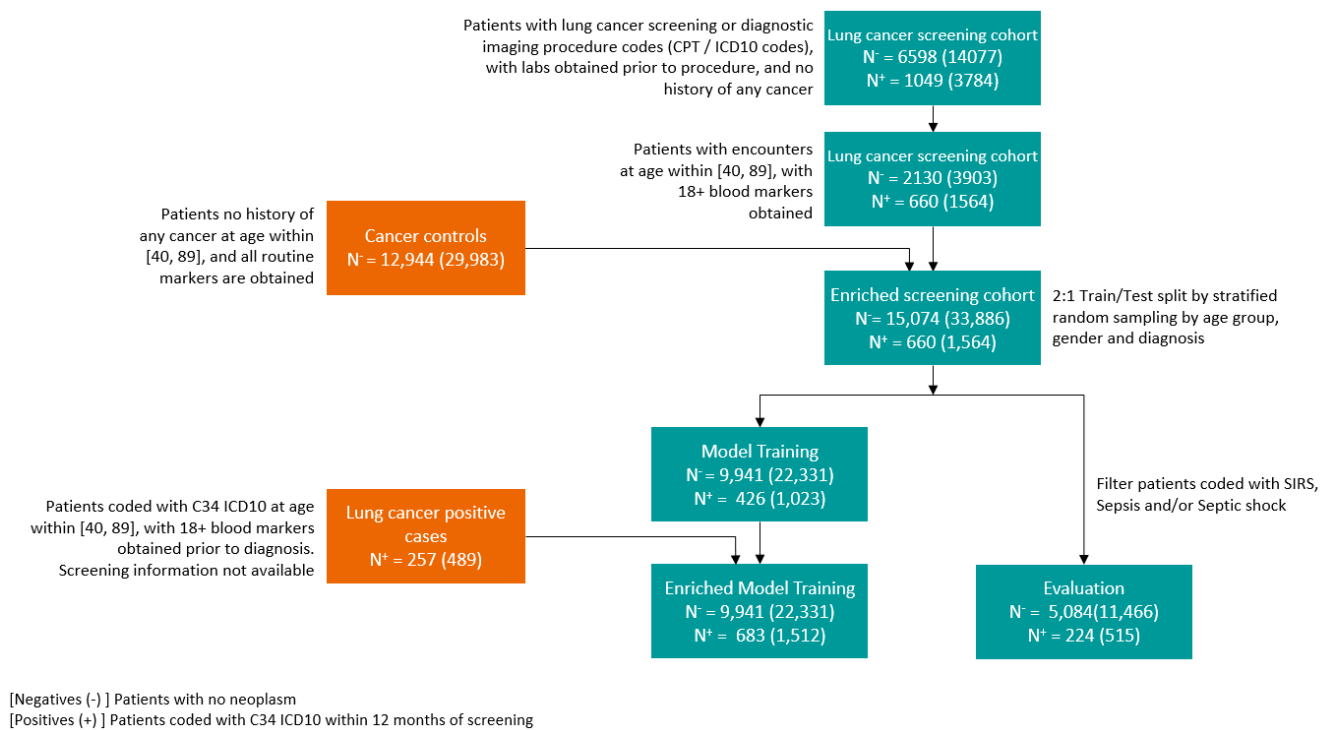
[Negatives (-)] Patients with no neoplasm
[Positives (+)] Patients coded with C18-C20 ICD10 within 12 months of colonoscopy

(a) Colorectal cancer



[Negatives (-)] Patients with no neoplasm
[Positives (+)] Patients coded with C22 ICD10 within 12 months of screening

(b) Liver cancer



(c) Lung cancer

Figure 5. Consort flow diagrams for various cancer cohorts. Each cell depicts the number of cancer positive (N+) and cancer negative (N-) patients. For each entry, both the number of patients and their aggregate encounter count has been reported. SIRS, systemic inflammatory response syndrome.

Data preparation and pre-processing

Given all the patients and their records in the enriched screening cohorts, we next split the data into development and validation cohorts in a 2:1 ratio, stratified by age, sex, and diagnosis. We then performed additional steps to enrich the development cohort that may assist in training a robust model as well as in cleaning the validation cohort to remove any ambiguous records. Specifically, we filter out patient records that correspond to severe/acute infection such as systemic inflammatory response syndrome (SIRS), sepsis, and/or septic shock, since they can result in significant change in CBC and CMP marker levels and may not be relevant in the pre-screening scenario. We also enriched the development cohort by adding newly diagnosed cancer patients whose records do not include a visit associated with any of the screening codes.

Given the development and validation cohorts, we normalized the distributions of biomarker values by subtracting their median and dividing by their inter-quartile distance (IQD). In addition, we transformed the distributions of the following markers to the logarithmic (\log_{10}) scale prior to their normalization: alanine aminotransferase (AST), alkaline phosphatase (ALP), aspartate aminotransferase (AST), bilirubin, creatinine, erythrocyte distribution width (RDW), glucose, leukocytes, lymphocytes, neutrophils, and urea nitrogen. Note that the pre-processing parameters (i.e., median and IQD values) for all markers were only computed on the development cohort, and the same parameter values were then applied to the validation cohort.

Deep Profiler

Given the pre-processed development cohort, we employed a deep neural network called Deep Profiler³⁴ to train models for each of the three cancer scenarios. This method has previously been applied to train models to predict a severity progression risk score for COVID-19 patients based on biomarkers (age and nine blood markers) obtained at the time of admission^{34,35}. Furthermore, it was demonstrated to be robust to marker missingness as well as competitive to other methods such as a logistic regression and boosted forests.

The Deep Profiler model builds on a variational autoencoder architecture³⁶, and consists of three main networks: an encoder for extracting prominent features represented in a latent space, a decoder for reconstructing the input data to ensure data fidelity of the latent feature representation, and, finally, a multi-label classifier network, which is trained to estimate the ordinal risk score. Use of autoencoders to deal with robustness to missing data has been demonstrated to be effective^{37,38}.

In this work, we use the same Deep Profiler architecture for each of the three cancer models, with a symmetric encoder-decoder structure, each comprising of blocks of three fully connected layers with 32 kernels. Each fully connected layer in the encoder is followed by a batch normalization and LeakyReLU (0.2), and each fully connected layer in the decoder is followed by batch normalization and ReLU. Network parameters were learned using the ADAM optimizer with an initial learning rate of 10^{-4} . The training loss is a combination of reconstruction loss (only applied to corresponding input features whose values are not missing), VAE regularization loss and binary cross entropy loss based on the patient's diagnosis.

Prediction uncertainty using Deep Profiler ensembles

While the VAE helps adding robustness to data missingness, undesirable biases may still be present due to the patient population/sampling. To this end, we trained an ensemble of Deep Profiler models. Given the ensemble of models, a confidence interval for the risk score (output of the sigmoid layer) is computed. Each model is trained on different subsets of the data, accounting for variations in the acquisition protocols across hospitals, availability of various biomarkers, measurement methods used to assess the biomarker as well as biomarker measurements such as laboratory values, age range, comorbidities, etc. In this study, for each cancer scenario, we used an ensemble of ten Deep Profiler models.

Estimation of likelihood ratios

Given the patient biomarker data and computed confidence interval over the risk score, the next key step is to present this information such that it can be easily interpreted and acted upon. To this end, we used the risk score and confidence interval to identify the subgroup of patients in the development cohort with similar scores. As a result, we obtained a “similar patient cohort” (a.k.a. patient cohort with similar risk scores) in addition to the development cohort. We then computed the pre- and post-test odds for the patient to have the disease and, subsequently, calculated the likelihood ratio. This can be done separately for each medical condition. Figure 6 below shows one way of presenting this information to the clinician.

Subgroup analysis based on comorbidities

We used phecodes²¹ to analyze comorbidities in each of the cohorts. Phecodes are manually curated groups of ICD-10 codes that are clinically meaningful and relevant to research. They have been widely applied in phenome-wide association studies but, more recently, were also used for rapid electronic health record phenotyping^{39,40}. For the colorectal, liver and lung cancer cohorts, we censored all ICD-10 codes assigned after the date of diagnosis and only mapped the codes assigned before diagnosis to phecodes. This ensured the identification of clinical comorbidities prior to a cancer diagnosis. For control cohorts, we mapped all of the ICD-10 codes recorded to phecodes. Subsequently, we compared the counts of individuals with each

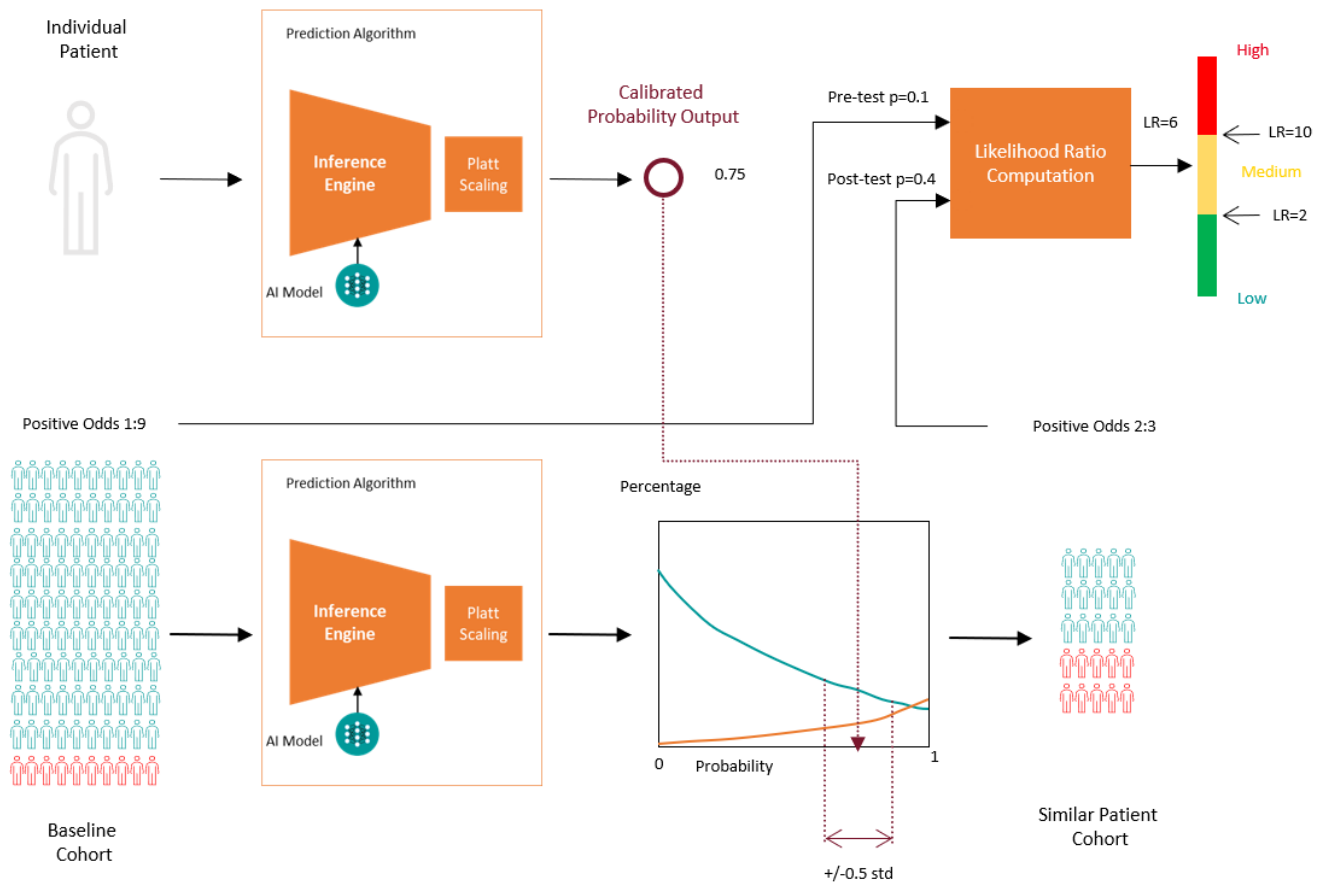


Figure 6. Schematic of the system output interpretation using likelihood ratio.

phecode in the control group against each of the pre-diagnosis cancer cohorts and identified significantly different phecodes, or comorbidities, based on Fisher's exact tests. Finally, we ranked the most significant comorbidities with at least 50 individuals in both the cancer and the control cohort ranked for subgroup analysis and evaluated the respective Deep Profiler models in those subgroups.

Data availability

Entire patient cohort data used in this study was licensed from Prognos Health (prognoshealth.com) via the prognosFACTOR® platform. Prognos Health can be reached at <https://prognoshealth.com/support> or via email at client_support@prognoshealth.com. Kindly reach out to the corresponding author at vivek-singh@siemens-healthineers.com if you are interested in additional details of the patient selection criteria used to obtain the records.

Code availability

The inference algorithms of the models presented in this paper are available from the authors upon request.

References

1. Samadder, N. J. *et al.* Comparison of universal genetic testing vs guideline-directed targeted testing for patients with hereditary cancer syndrome. *JAMA Oncol* **7**, 230–237 (2021).
2. Rex, D. K. *et al.* Colorectal cancer screening: recommendations for physicians and patients from the U.S. Multi-Society Task Force on Colorectal Cancer. *Am J Gastroenterol* **112**, 1016–1030 (2017).
3. Mendes, J. & Matela, N. Breast cancer risk assessment: a review on mammography-based approaches. *J. Imaging* **7** (2021).

4. de Koning, H. J. *et al.* Reduced lung-cancer mortality with volume CT screening in a randomized trial. *N Engl J Med* **382**, 503–513 (2020).
5. Hall, I. J. *et al.* Patterns and trends in cancer screening in the United States. *Prev Chronic Dis* **15**, E97 (2018).
6. Philipson, T. J., Durie, T., Cong, Z. & Fendrick, A. M. The aggregate value of cancer screenings in the United States: full potential value and value considering adherence. *BMC Heal. Serv Res* **23**, 829 (2023).
7. Kostopoulou, O., Arora, K. & Pálfi, B. Using cancer risk algorithms to improve risk estimates and referral decisions. *Commun Med (Lond)* **2**, 2 (2022).
8. Hippisley-Cox, J. & Coupland, C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. *BMJ Open* **5**, e007825 (2015).
9. Nartowt, B. J. *et al.* Robust machine learning for colorectal cancer risk prediction and stratification. *Front Big Data* **3**, 6 (2020).
10. Goshen, R. *et al.* Computer-assisted flagging of individuals at high risk of colorectal cancer in a large health maintenance organization using the ColonFlag test. *JCO Clin Cancer Inf.* **2**, 1–8 (2018).
11. Chung, D. C. *et al.* A cell-free DNA blood-based test for colorectal cancer screening. *N Engl J Med* **390**, 973–983 (2024).
12. Hackshaw, A., Clarke, C. A. & Hartman, A.-R. New genomic technologies for multi-cancer early detection: rethinking the scope of cancer screening. *Cancer Cell* **40**, 109–113 (2022).
13. Klein, E. A. *et al.* Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Ann Oncol* **32**, 1167–1177 (2021).
14. Shao, S. H. *et al.* Multi-cancer early detection test sensitivity for cancers with and without current population-level screening options. *Tumori* **109**, 335–341 (2022).
15. Hamilton, W. Five misconceptions in cancer diagnosis. *Br J Gen Pract* **59**, 441–5, 447; discussion 446 (2009).
16. Heymans, M. W. & Twisk, J. W. R. Handling missing data in clinical research. *J Clin Epidemiol* **151**, 185–188 (2022).
17. Næser, E., Møller, H., Fredberg, U., Frystyk, J. & Vedsted, P. Routine blood tests and probability of cancer in patients referred with non-specific serious symptoms: a cohort study. *BMC Cancer* **17**, 817 (2017).
18. US Preventive Services Task Force *et al.* Screening for colorectal cancer: US Preventive Services Task Force recommendation statement. *JAMA* **325**, 1965–1977 (2021).
19. Goshen, R. *et al.* Predicting the presence of colon cancer in members of a health maintenance organisation by evaluating analytes from standard laboratory records. *Br J Cancer* **116**, 944–950 (2017).
20. Mounce, L. T., Hamilton, W. & Bailey, S. E. Cancer incidence following a high-normal platelet count: cohort study using electronic healthcare records from english primary care. *Br J Gen Pract* **70**, e622–e628 (2020).
21. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation. *JMIR Med Inf.* **7**, e14325 (2019).
22. US Preventive Services Task Force *et al.* Screening for lung cancer: US Preventive Services Task Force recommendation statement. *JAMA* **325**, 962–970 (2021).
23. Soerensen, P. D. *et al.* Using artificial intelligence in a primary care setting to identify patients at risk for cancer: a risk prediction model based on routine laboratory tests. *Clin Chem Lab Med* **60**, 2005–2016 (2022).
24. Flyckt, R. N. H. *et al.* Pulmonologists-level lung cancer detection based on standard blood test results and smoking status using an explainable machine learning approach (2024). [2402.09596](https://doi.org/10.2402/09596).
25. Gould, M. K., Huang, B. Z., Tammemagi, M. C., Kinar, Y. & Shiff, R. Machine learning for early lung cancer identification using routine clinical and laboratory data. *Am J Respir Crit Care Med* **204**, 445–453 (2021).
26. Jonas, D. E. *et al.* Screening for lung cancer with low-dose computed tomography: updated evidence report and systematic review for the US Preventive Services Task Force. *JAMA* **325**, 971–987 (2021).
27. Toumazis, I., Bastani, M., Han, S. S. & Plevritis, S. K. Risk-based lung cancer screening: a systematic review. *Lung Cancer* **147**, 154–186 (2020).
28. Tammemägi, M. C. *et al.* Selection criteria for lung-cancer screening. *N Engl J Med* **368**, 728–736 (2013).
29. Fahrman, J. F. *et al.* Blood-based biomarker panel for personalized lung cancer risk assessment. *J Clin Oncol* **40**, 876–883 (2022).

30. Wu, J. T.-Y., Wakelee, H. A. & Han, S. S. Optimizing lung cancer screening with risk prediction: current challenges and the emerging role of biomarkers. *J Clin Oncol* **41**, 4341–4347 (2023).
31. American Medical Association. *CPT Professional 2024* (American Medical Association, 2023).
32. American Medical Association. *ICD-10-CM 2023: the complete official codebook* (American Medical Association, 2022).
33. Campbell, S. & Giadresco, K. Computer-assisted clinical coding: a narrative review of the literature on its benefits, limitations, implementation and impact on clinical coding professionals. *Heal. Inf Manag.* **49**, 5–18 (2019).
34. Singh, V. *et al.* A deep learning approach for predicting severity of COVID-19 patients using a parsimonious set of laboratory markers. *iScience* **24**, 103523 (2021).
35. Yilmaz, G. *et al.* Concordance and generalization of an AI algorithm with real-world clinical data in the pre-omicron and omicron era. *Heliyon* **10**, e25410 (2024).
36. Pinheiro Cinelli, L., Araújo Marins, M., Barros da Silva, E. A. & Lima Netto, S. *Variational autoencoder*, 111–149 (Springer International Publishing, Cham, 2021).
37. Śmieja, M., Kołomycki, M., Struski, Ł., Juda, M. & Figueiredo, M. A. T. Can auto-encoders help with filling missing data? In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations* (2019).
38. Pereira, R. C., Santos, M. S., Rodrigues, P. P. & Abreu, P. H. Reviewing autoencoders for missing data imputation: technical trends, applications and outcomes. *J. Artif. Intell. Res.* **69**, 1255–1285 (2020).
39. Bastarache, L. Using phecodes for research with the electronic health record: from PheWAS to PheRS. *Annu. Rev Biomed Data Sci* **4**, 1–19 (2021).
40. Kerley, C. I. *et al.* pyPheWAS: a phenome-disease association tool for electronic medical record analysis. *Neuroinformatics* **20**, 483–505 (2022).

Acknowledgements

We would like to thank Dennis Gilbert, Rangarajan Sampath, and Harigovind Singh for valuable discussions about this work. We would like to also thank Manish Chowdhury for assisting us with data preparation during the initial stage of development.

Author contributions statement

V.S., S.C., M.S., R.G., J.G., D.C., and A.K. conceived the model and experiments, V.S. implemented the model, V.S., S.C., M.S., S.R. and A.P. prepared the data, V.S., S.C., and M.S. conducted the experiments, V.S., S.C., M.S., and A.K. analyzed the results, V.S., S.C., M.S., and A.K. wrote the manuscript, which was then reviewed and edited by all authors.

Additional information

Competing interests

V.S., S.C., M.S., J.G., D.C., and A.K. are employed by Siemens Healthineers. A.P. is employed by Siemens SRL. S.R. and R.G. were employed by Siemens Healthineers during the completion of this work.

Disclaimer

The concepts and information presented in this paper are based on research results that are not commercially available. Future commercial availability cannot be guaranteed.

Deep learning-based identification of patients at increased risk of cancer using routine laboratory markers

Vivek Singh^{1,*}, Shikha Chaganti¹, Matthias Siebert², Soumya Rajesh¹, Andrei Puiu^{3,4}, Raj Gopalan⁵, Jamie Gramz⁶, Dorin Comaniciu¹, and Ali Kamen¹

¹Siemens Healthineers, Digital Technology and Innovation, Princeton, 08540, USA

²Siemens Healthineers, Digital Technology and Innovation, Erlangen, 91052, Germany

³Siemens SRL, Advanta, Brasov, 500007, Romania

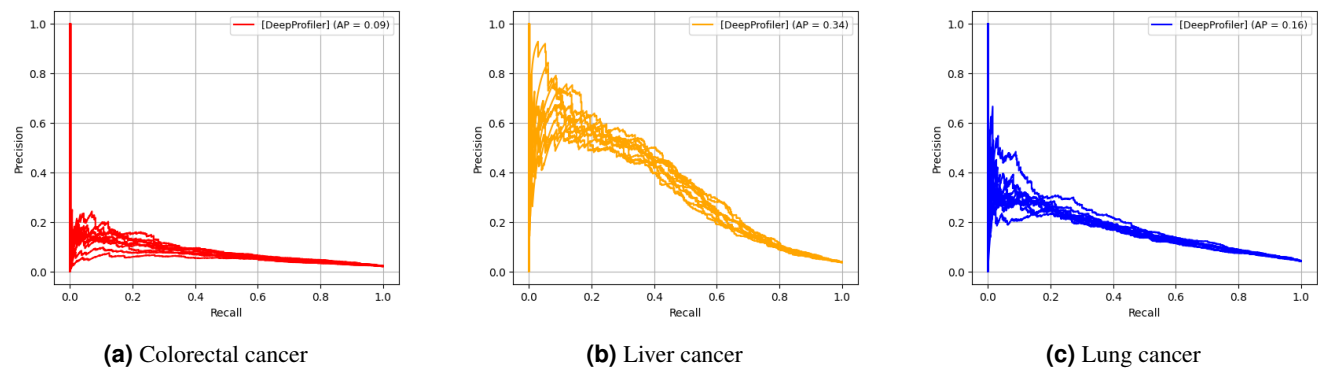
⁴Transylvania University of Brasov, Automation and Information Technology, Brasov, 500174, Romania

⁵Siemens Healthineers, Laboratory Diagnostics, Tarrytown, NY 10591, USA.

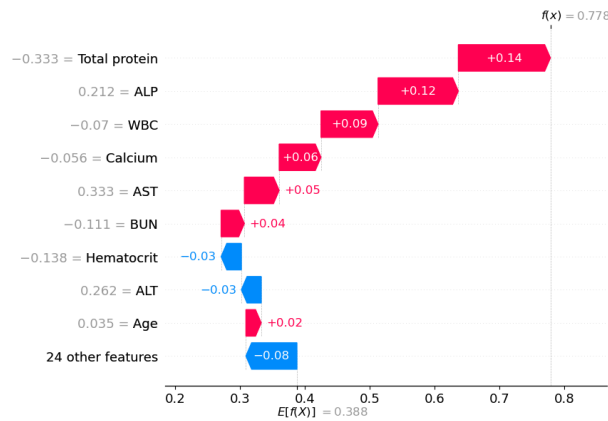
⁶Siemens Healthineers, Digital and Automation, Malvern, PA 19355, USA.

*vivek-singh@siemens-healthineers.com

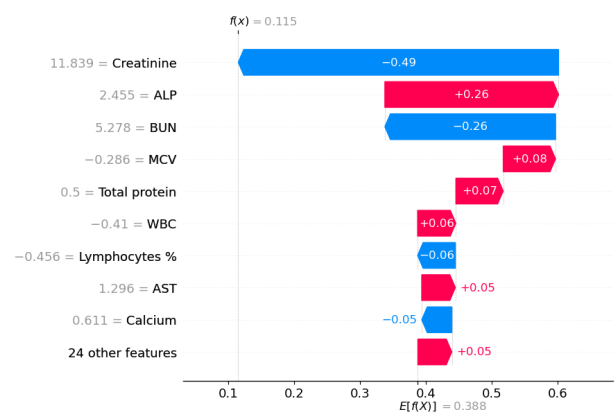
Supplementary Figures



Supplementary Figure S1. Quantitative performance assessment on (a) colorectal, (b) liver, and (c) lung cancer validation cohorts. Precision-recall curves represent the performance of single Deep Profiler models constituting the respective ensemble. The average precision (AP) corresponds to the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight.



(a) Liver cancer case



(b) Control case

Supplementary Figure S2. Waterfall plots of SHAP feature attributions showing the contribution of the nine laboratory markers with highest impact on the normalized LR ($f(x)$) for (a) a liver cancer and (b) a control case. Plots are aligned at the average normalized LR ($E[f(x)]$) across all samples. Samples had to have data available for at least 24 laboratory markers. Normalized laboratory values are listed along with the marker names.

Characteristics	No cancer	Colorectal cancer	Liver cancer	Lung cancer
Encounters	34062	750	1559	1702
Age	61.38 ± 12.28 (17390)	63.11 ± 10.82 (387)*	64.05 ± 11.21 (751)*	67.48 ± 10.03 (752)*
Sex: Male	8473 (17390)	196 (387)	434 (751)*	384 (752)
Albumin (g/dL)	4.5 ± 0.3 (16329)	4.2 ± 0.4 (239)**	4.0 ± 0.6 (631)**	4.2 ± 0.4 (704)**
ALP (U/L)	80.8 ± 45.0 (16790)	96.5 ± 52.4 (311)**	166 ± 169 (704)**	103 ± 77 (713)**
ALT (U/L)	24.6 ± 24.2 (16273)	28.0 ± 24.8 (248)	44.8 ± 74.9 (611)**	21.6 ± 18.1 (708)**
AST (U/L)	24.0 ± 23.4 (16319)	29.4 ± 20.6 (254)**	46.7 ± 54.6 (620)**	24.2 ± 18.6 (698)
Bilirubin (mg/dL)	0.5 ± 0.4 (16745)	0.5 ± 0.3 (313)*	1.0 ± 2.0 (684)**	0.4 ± 0.4 (722)**
BUN (mg/dL)	16.4 ± 8.3 (16681)	16.2 ± 12.0 (284)**	15.9 ± 7.3 (642)*	17.7 ± 9.9 (721)**
BUN-Creatinine Ratio	17.6 ± 5.6 (14545)	17.3 ± 6.0 (264)	18.1 ± 6.0 (599)*	18.0 ± 6.3 (716)
Calcium (mg/dL)	9.4 ± 0.4 (16873)	9.3 ± 0.5 (310)**	9.2 ± 0.6 (689)**	9.3 ± 0.6 (723)**
Chloride (mEq/L)	102 ± 3 (16406)	102 ± 3 (254)*	101 ± 4 (625)	101 ± 4 (716)**
Creatinine (mg/dL)	1.0 ± 0.7 (16666)	1.0 ± 0.7 (285)*	0.9 ± 0.4 (648)**	1.0 ± 0.6 (722)**
Glucose (mg/dL)	109 ± 44 (16371)	120 ± 46 (247)**	126 ± 55 (606)**	119 ± 51 (709)**
Potassium (mEq/L)	4.4 ± 0.4 (16357)	4.4 ± 0.5 (257)	4.4 ± 0.5 (620)	4.4 ± 0.6 (708)**
Sodium (mEq/L)	140 ± 3 (16364)	140 ± 3 (258)	139 ± 4 (627)**	140 ± 4 (713)
Total CO2 (mEq/L)	24.0 ± 2.9 (16348)	23.9 ± 3.2 (252)	24.3 ± 3.2 (622)**	24.1 ± 3.1 (708)
Total protein (g/dL)	7.2 ± 0.5 (16805)	6.9 ± 0.6 (303)**	7.1 ± 0.7 (692)**	7.0 ± 0.6 (721)**
Basophils (cells/mL)	0.0 ± 0.0 (16927)	0.0 ± 0.0 (320)**	0.0 ± 0.0 (660)**	0.0 ± 0.0 (719)**
Basophils % (%)	0.6 ± 0.4 (16944)	0.5 ± 0.4 (325)**	0.5 ± 0.4 (690)**	0.5 ± 0.4 (728)**
Eosinophils (cells/mL)	0.2 ± 0.2 (17162)	0.2 ± 0.2 (359)	0.2 ± 0.2 (700)	0.2 ± 0.2 (728)*
Eosinophil % (%)	2.8 ± 2.4 (17168)	3.1 ± 3.0 (361)	2.7 ± 2.7 (725)**	2.6 ± 3.0 (739)**
Hematocrit (%)	40.9 ± 4.7 (16302)	38.1 ± 5.8 (243)**	37.7 ± 6.0 (621)**	38.0 ± 5.7 (713)**
Hemoglobin (g/dL)	13.5 ± 1.7 (16862)	12.5 ± 2.1 (309)**	12.5 ± 2.2 (675)**	12.5 ± 2.0 (727)**
Hemoglobin A1c %	6.1 ± 1.2 (14112)	6.4 ± 1.5 (140)**	6.5 ± 1.5 (334)**	6.4 ± 1.4 (540)**
Lymphocytes (cells/mL)	2.0 ± 0.8 (16908)	1.7 ± 0.8 (333)**	1.7 ± 1.0 (663)**	1.8 ± 1.4 (714)**
Lymphocytes % (%)	31.7 ± 9.4 (16914)	27.4 ± 11.7 (336)**	26.0 ± 11.8 (691)**	25.8 ± 12.2 (720)**
Neutrophils (cells/mL)	3.7 ± 1.7 (16300)	4.5 ± 3.5 (248)**	4.3 ± 3.1 (577)**	4.7 ± 3.1 (703)**
MCH (pg)	29.7 ± 2.6 (16324)	29.5 ± 3.3 (255)	30.0 ± 3.0 (607)**	29.9 ± 2.5 (707)
MCHC (g/dL)	33.0 ± 1.2 (16341)	33.0 ± 1.5 (259)	33.0 ± 1.5 (589)	32.8 ± 1.2 (708)**
MCV (fL)	89.8 ± 6.3 (16348)	89.3 ± 7.7 (250)	90.5 ± 7.0 (603)	90.9 ± 6.2 (711)**
Monocytes (cells/mL)	0.5 ± 0.2 (16220)	0.6 ± 0.3 (255)**	0.6 ± 0.3 (565)**	0.6 ± 0.3 (702)**
Platelets (cells/mL)	248 ± 72 (16829)	256 ± 118 (314)	223 ± 103 (679)**	259 ± 110 (724)
RDW (%)	13.6 ± 1.4 (16329)	14.5 ± 2.0 (242)**	14.6 ± 2.0 (595)**	14.5 ± 2.1 (715)**
WBC (cells/mL)	6.5 ± 2.3 (17258)	6.9 ± 3.6 (372)	6.9 ± 3.5 (712)	7.4 ± 3.8 (735)

* p-value < 0.01, ** p-value < 0.001, based on Wilcoxon-rank sum tests comparing characteristics of cancer vs. control samples

Supplementary Table S1. Entire cohort characteristics of biomarkers (units) reported as mean and standard error (number of patients for whom the value was recorded). Biomarkers measured in either CMP or CBC are grouped accordingly. ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; BUN, blood urea nitrogen; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; RDW, red cell distribution width; WBC, white blood cells.