Computers in Biology and Medicine 144 (2022) 105362

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Long short-term memory model – A deep learning approach for medical data with irregularity in cancer predication with tumor markers

Xiaoxing Wu [a,1], Hsin-Yao Wang [b,c,d,1], Peichang Shi [b,h], Rong Sun [e], Xiaolin Wang [e], Zhixiao Luo [e], Fanling Zeng [e], Michael S. Lebowitz [b], Wan-Ying Lin [g], Jang-Jih Lu [b,c], Richard Scherer [b], Olivia Price [b], Ziwei Wang [a,f], Jiming Zhou [b,*], Yonghong Wang [a,e,**]

[a] The First Affiliated Hospital of Chongqing Medical University, Chongqing, 400016, China
[b] 20/20 GeneSystems, Inc, Gaithersburg, MD 20877, USA
[c] Department of Laboratory Medicine, Chang Gung Memorial Hospital at Linkou, Taoyuan City, 33305, Taiwan
[d] PhD Program in Biomedical Engineering, Chang Gung University, Taoyuan City, 33301, Taiwan
[e] Health Management Center, The First Affiliated Hospital of Chongqing Medical University, Chongqing, 400016, China
[f] Department of Gastrointestinal Surgery, The First Affiliated Hospital of Chongqing Medical University, Chongqing, 400016, China
[g] Syu Kang Sport Clinic, Taipei, 11217, Taiwan
[h] University of Maryland, Baltimore County, MD, 20723, USA

A B S T R A C T

Background: Machine learning (ML) has emerged as a superior method for the analysis of large datasets. Application of ML is often hindered by incompleteness of the data which is particularly evident when approaching disease screening data due to varied testing regimens across medical institutions. Here we explored the utility of multiple ML algorithms to predict cancer risk when trained using a large but incomplete real-world dataset of tumor marker (TM) values.

Methods: TM screening data were collected from a large asymptomatic cohort (n = 163,174) at two independent medical centers. The cohort included 785 individuals who were subsequently diagnosed with cancer. Data included levels of up to eight TMs, but for most subjects, only a subset of the biomarkers were tested. In some instances, TM values were available at multiple time points, but intervals between tests varied widely. The data were used to train and test various machine learning models to evaluate their robustness for predicting cancer risk. Multiple methods for data imputation were explored and models were developed for both single time-point as well as time-series data.

Results: The ML algorithm, long short-term memory (LSTM), demonstrated superiority over other models for dealing with irregular medical data. A cancer risk prediction tool was trained and validated for a single time-point test of a TM panel including up to four biomarkers (AUROC = 0.831, 95% CI: 0.827–0.835) which outperformed a single threshold method using the same biomarkers. A second model relying on time series data of up to four time-points for 5 TMs had an AUROC of 0.931.

Conclusions: A cancer risk prediction tool was developed by training a LSTM model using a large but incomplete real-world dataset of TM values. The LSTM model was best able to handle irregular data compared to other ML models. The use of time-series TM data can further improve the predictive performance of LSTM models even when the intervals between tests vary widely. These risk prediction tools are useful to direct subjects to further screening sooner, resulting in earlier detection of occult tumors.

1. Introduction

Diagnosis of cancer at the early stage is one of the most important factors leading to improved cancer survival rates. The 5-year survival rate for colorectal cancer (CRC) is around 90% for Stage I cases but drops to only 10% for late-stage CRC [1]. The improvement of survival rate due to early diagnosis is better than any state-of-the-art therapies used for treating cancer in later stages [2]. Moreover, earlier diagnosis of cancer also results in reduction of treatment costs and loss of economic productivity [3]. In an attempt to capitalize on the value of early cancer diagnosis, many tools have been developed for screening. The majority of these tools screen for only one type of cancer [4]. By way of example, the fecal occult blood test (FOBT) and colonoscopy are used for CRC screening [5], while low-dose computed tomography (LDCT) is used for lung cancer screening [6], and mammography is used for breast cancer screening [7]. Obviously, these tools are developed to screen for only

* Corresponding author. 15810 Gaither Drive, Suite 235, Gaithersburg, MD 20877, USA.
** Corresponding author. The First Affiliated Hospital of Chongqing Medical University, Chongqing 400016, China.
E-mail addresses: jzhou@2020gene.com (J. Zhou), wyh0232983@163.com (Y. Wang).
[1] These authors contributed equally.

https://doi.org/10.1016/j.compbiomed.2022.105362
Received 21 October 2021; Received in revised form 4 February 2022; Accepted 26 February 2022
Available online 9 March 2022

single types of cancer and their application as yearly comprehensive screening modalities would be extremely inconvenient for the general asymptomatic population. The net benefit of cancer screening by these tools would be largely reduced due to reduced willingness to subject oneself to the proscribed screening regimen and low compliance even subsequent to a clinician's orders. Moreover, these cancer screening tools can result in some adverse effects, such as bowel perforation caused by colonoscopy, or pain caused by mammography. The harm brought by the screening tools also hinder their widespread use as screening tools.

To address the issues raised above, pan-cancer screening modalities have begun to emerge including tools utilizing nucleic acid sequencing technology (e.g. Grail's Galleri) [8] or serum protein tumor markers (TM) analysis (e.g. Thrive's CancerSEEK or 20/20 GeneSystems' OneTest™) [4]' [8]. The idea of pan-cancer screening aims to screen multiple types of cancers so that the efficiency of cancer screening would be significantly improved. Moreover, these liquid biopsy-based tools greatly reduce the possible harms that can result from more invasive forms of screening. Phlebotomy of only several milliliters of blood generally does no harm. TMs have been identified, developed, approved, and used in clinical medicine for decades. Typically, individual TMs are used in the following-up of cancers post therapy [9]. In contrast, some studies demonstrated combining multiple TMs as panels which can be used for cancer screening [10]' [11]' [8]. Such combinations are often less susceptible to measurement artifacts compared to the individual markers [11]. The experience of either analyzing or using TMs has been well established in clinical settings. Moreover, analytical measurement of TMs is significantly less costly than nucleic acid sequencing. Thus, TM testing has become a more practical, cost-effective, and popular cancer screening strategy that is widely used in health check-up centers worldwide, especially in East Asia [12]' [13]. For example, in 2018, a total of 435 million health check-up examinations were conducted in China [14], and TM measurement is one of the most common items performed during these health check-ups. TM panels vary greatly across different locales as does the frequency of testing.

Classical machine learning (ML) algorithms generally require complete data for training and testing and while missing data can be imputed using various approaches, this is complicated when employing real-world datasets which often have many missing values due to variations in the implementation of testing panels by different clinicians at different or even the same clinic. We have previously developed an algorithm to predict cancer risk in a real-world cohort of >27,000 subjects using logistic regression (LR) [15]. However, the prior dataset included complete single time-point biomarker data for all subjects. Improvement of this prior model, including its applicability to a wider geographic population where certain biomarker tests may not be available, requires the ability for the model to deal with missing data and time series data. Cancer is a typical disease developing over time. Cancer development is a dynamic process between developing into a local tumor and being eliminated by the immune system [16]. Many cancers can take years to develop to metastasis from their original lesions [17]. Thus, time series data would be more comprehensive than single time-point data to know cancer better. Traditional machine learning approaches (e.g. LR and random forest (RF)) perform well for regular, complete datasets, but often fail when confronted with real-world datasets that have many missing values and many data points. Recurrent neural networks (RNN), such as LSTM and gated recurrent unit (GRU), have been successfully applied in many situations with missing values [18]. These ML models thus represent promising tools for the handling very large real-world datasets which have irregular data, including single time-point values where not all subjects have all values available and time-series data where the interval between multiple tests may vary. LSTM in particular can be adapted to make use of missing value patterns, time intervals and complex temporal dependencies in irregular univariate and multivariate time series data as it has internal gating mechanisms to avoid the vanishing and exploding gradient calculation (supplement Figure 1) [19]' [20]. Thus, these favorable features make LSTM an appropriate algorithm for dealing with both missing data and lack of regularity in the intervals between sequential tests as found in TM data collected from multiple sites and cohorts constituting real-world datasets.

Since training a model with a lot of missing values could lead to biased estimation and thus impact the model's quality, data imputation is necessary to handle missing values. Methods to deal with missing values can be grouped into three major categories: 1) Missing completely at random (MCAR); 2) Missing at random; 3) Not missing at random. Most data imputation approaches focus on MCAR. There are different data imputation methods, the simplest one involves replacing the missing values with 0, constant values, or mean/median values [21]' [22]. This method of data imputation is easy and fast and generally works well with small datasets, however it may not be very accurate and is not recommended for use when imputing categorical variables [23]. The k-nearest neighbors (k-NN) approach is another common method for data imputation. The algorithm uses feature similarity to predict the missing values [24]. The missing values will be replaced based on how closely the present values match to their counterparts in other samples. It is reported that the k-NN method for imputation can achieve much greater accuracy than simple replacement, however it assumes a relationship exists between the various biomarkers, is sensitive to outliers and is computationally expensive since it requires storage of the entire training dataset in local memory. Another popular imputation method is multivariate imputation by chained equation (MICE) [25]. This type of imputation works by filling the missing data multiple times. Multiple Imputations (MIs) are much better than a single imputation as it measures the uncertainty of the missing values in a better way. The chained equations approach is also very flexible and can handle different variables of different data types (i.e., continuous or binary) as well as complexities such as bounds or survey skip patterns [25]. Deep Learning also affords an approach to data imputation that generally works very well with both categorical and non-numerical features. In this approach Machine Learning models are stored which use Deep Neural Networks to impute the missing values [26].

In this study, we aimed to evaluate the robustness of the TM-based cancer screening models by using data collected from two different geographic locations (Chongqing and Taiwan). While the initial goals were to greatly increase the size of the study cohort and explore applicability of our previous models across different geographic locations, we were confronted with differences in the number, type and frequency of tumor markers tested in the different locales. Indeed, even within a single location not all TMs were measured in all patients. Thus, we realized that a more robust model needed to be developed which would allow for risk prediction from incomplete tumor marker panels. Here we have applied a long short-term memory (LSTM) model to develop a novel algorithm which was then used in cross-external validations. We also explored the opportunity for increased accuracy in prediction by using time-series data.

2. Materials and methods

2.1. Inclusion and exclusion criteria of the datasets

We used the health check-up database obtained from the First Affiliated Hospital of Chongqing Medical University (CHQ, Chongqing) and another database from Chang Gung Memorial Hospital (CGMH, Taiwan) incorporating data collected between May 2001 and December 2019. All participants had one or more cancer biomarkers measured for screening purposes and were asymptomatic at the time of testing. All subjects were continuously followed after the initial health check-up examination by monitoring medical records for more than one year to determine the status of cancer diagnosis. Exclusion criteria included loss to follow-up, no further medical examination within one year, and cancer diagnosis before the analytical measurements of TM. This study was approved by the CHQ Ethics Committee (2020-089).

The following TMs were included in the algorithm development: AFP, CA15-3, CA-125, PSA, SCC, CEA, CYFR21-1 and CA19-9. All TM levels were determined from venipuncture-obtained serum samples by an automated chemiluminescence immunoassay analyzer (Cobas 8000 e602, Roche Diagnostics Inc), except for SCC which was measured using commercially available kits from Abbott Diagnostics, Abbott Park, IL, USA). Clinical reference values for each tumor marker are 25 ng/ml for AFP, 25 U/ml for CA 15-3, 35 U/ml for CA-125, 4 ng/ml for PSA, 1.5 ng/ml for SCC, 10 ng/ml for CEA, 3.3 ng/ml for CYFRA 21-1 and 27 U/ml for CA 19-9. The laboratories from which the data was obtained have passed ISO15189 certification (NO.ML00036).

2.2. Dataset preprocessing

To fully evaluate the effect of missing values and data imputation on algorithm development, we employed two different types of datasets. First, we combined the CGMH and CHQ datasets together, and only kept patients who have no missing values for the biomarkers (male: AFP, CEA, CA199, PSA, CYFRA211 and SCC, female: AFP, CEA, CA199, CA125, CA153, CYFRA211and SCC). This dataset includes a total of only 33,226 patients (Supplement Table 1), representing only 20% of the entire dataset. We also explored the use of the entire dataset of 163,174 individuals but has many missing TM values. Then we used 70% of all our raw dataset with missing biomarkers values to build a model and tested our model using 30% dataset (Supplement Figure 2).

2.3. Data imputation

To evaluate the performance of different imputation methods, we either directly replaced missing values with 0, or used the available packages within the Python programming language to employ replacement by KNN, MICE or deep learning imputation. When training the algorithm, the values were imputed for each of 30 repeats and we took the average of AUC for overall model performance comparison.

2.4. Training and validation of models

Seven popular algorithms are compared based on AUC and sensitivities when specificity is set at 0.8. The methods are: Decision tree (DT), Gradient Boosting trees (GB), K nearest neighbor (KNN), Logistic regression (LR), Long short-term memory model (LSTM), Naïve Bayes (NB), and Random forest (RF), which are popular methods in machine learning field [9].

The LSTM model was implemented using the keras package in Python 3.6, while the other models were implemented using the sklearn package in Python 3.6. The most popular and key parameters are fine-tuned and set as the following: for LSTM, we used one hidden layer with 100 LSTM blocks, and the output layer with a single value prediction with the default sigmoid activation function. The network is trained for 10 epochs and batch size is 64. KNN used 10 neighbors. For tree structured models, we tried different numbers of max depths from 1 to 20, ultimately employing depths of 10 and 5 for DT and RF, respectively, for better performance. For the GB model, the max depth was set to 5, number of estimators was set at 100 and the learning rate was set at 1.0. All the other parameters were set to the defaults.

We split the data into different sets: 70% for training, and 30% for testing. There is no universal agreement that how many training processes need to be done, however based on the Central Limit Theorem (CLT), sample sizes equal or greater than 30 are considered sufficient [27], so we repeated this process 30 times. The average AUC of the 30 repeats is used to measure model performance (Supplement Figure 2).

2.5. Follow-up criteria in CHQ

At CHQ, the diagnosis at health check-up involves an analysis of the combination of TM results with other relevant patient examination findings. The TM results of the subjects were classified as elevated if measured at twice the reference value and were grouped according to the results of other relevant examination items. The subjects in different groups were aggressively followed up as outlined in Fig. 1.

2.6. Time series data from CHQ hospital

Time series data is common in real-world cancer screening; however, it is very rarely employed due to the lack of robust model methods for analysis of irregular time series data with greatly varied intervals between tests. Again, the LSTM model was employed to analyze time-series data. The CHQ dataset included some individuals with time-series data including up to four time points for 300 individuals who did not develop cancer in the subsequent follow up period. However, of the 30 individuals who did develop cancer and for whom more than one time-point was available, 7 cancer patients only had data at time point 1, 10 cancer patients had 2 time points, 9 cancer patients had 3 time points and only 5 cancer patients had all data for 4 time points. In order to balance the data between cancer and non-cancer patients, we randomly chose some non-cancer patients and assigned missing values to some data points, so that cancer patients and non-cancer patients had similar patterns regarding patients with biomarker values at each time point. The LSTM algorithm was applied to the time series data analysis after the data preprocessing. To illustrate the effect of number of time points on model performance, we built 4 LSTM models, 1 model using only 1 time point, 1 model using 2 time points, 1 model using 3 time points and 1 model using all the 4 time points (Supplement Table 2).

2.7. Time-to-diagnosis analysis by using Cox's proportional hazards model

Time-to-event data analysis is widely used in oncology, such as the time from cancer diagnosis or treatment initiation to cancer recurrence or death. The Cox proportional hazards (PH) model allows one to describe the survival time as a function of multiple prognostic factors. All cancer patients from CHQ and CGMH were used for Cox analysis.

The survival probability was calculated from a PH model. We used a k-means clustering algorithm to separate the population into elevated and high-risk groups. The log-rank test was performed to check whether the four groups were significantly different.

2.8. Statistical analysis

We used effect size to compare the patient characteristics between CHQ and CGMH due to the large sample size. We applied a Chi-squared test to analyze the distribution of cancer cases, and Fisher's exact test was used for analysis when case number was less than 5 cases.

3. Results

3.1. Subjects

Data was abstracted from two cohorts of subjects: one at CGMH and the other at CHQ. We have previously used the CGMH dataset to derive a cancer risk prediction model [15]. This dataset consists of 27,938 unique subject records all with complete tumor marker data for AFP, CEA, CA19-9, CYFRA21-1, SCC, PSA (males only), CA-125 (females only) and CA15-3 (females only). The second dataset from CHQ consists of 135,236 unique subject records most of whom were tested for a subset of the tumor markers. For some of the CHQ subjects serial testing data is available covering multiple years of tumor marker testing. In total, there were 163,174 participants between the two datasets. For males, 82% had 2 biomarker tests (AFP, CEA), 68% had 3 biomarker tests (AFP, CEA, PSA), and only 36% had 4 biomarker tests (AFP, CEA, PSA, CA199). For females, 83% had AFP and CEA tests, 51% had AFP, CEA and CA199, and only 34% had AFP, CEA, CA125 and CA199 tests. Of all the

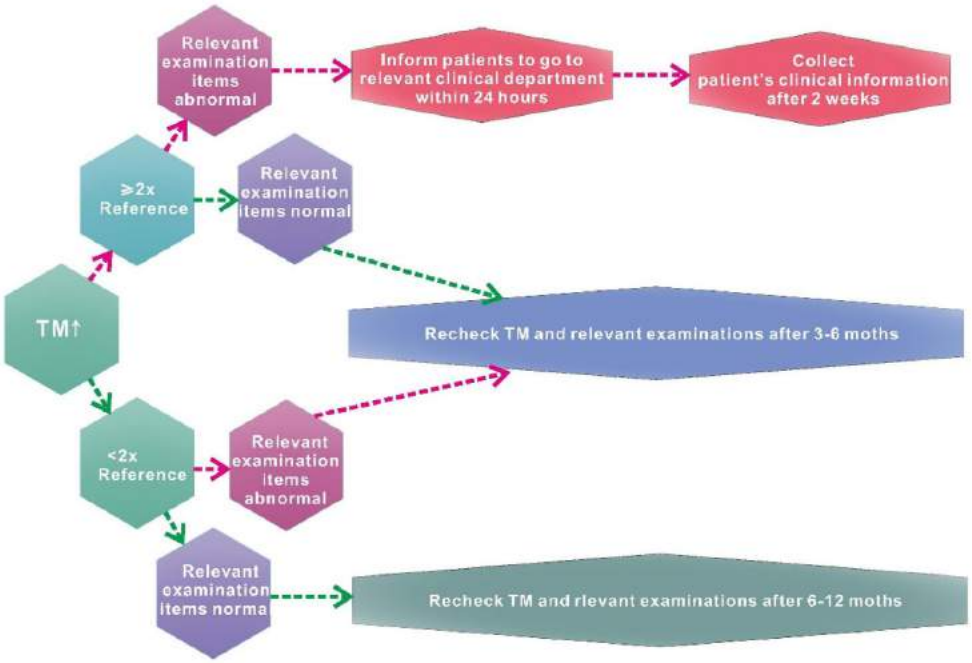

Fig. 1. Decision follow chart for the health checkup population in Chongqing.

participants, 785 (0.481%) were diagnosed with cancer, including 443 (0.321%) in CHQ and 342 (1.22%) in CGMH. We compared the demographic data between the two institutes (Tables 1 and 2). Median was used as the statistic to describe the comparison here to avoid bias due to outliers. The median age was lower in CHQ than in CGMH for both cancer (Table 1(a)) and noncancer groups (Table 1(b)), indicating a younger population for health checkup in CHQ which may account for the lower incidence of cancer in the CHQ population. Regarding TM, several TM presented notable differences between the two institutes. In the cancer (Table 1(a)) and noncancer groups (Table 1(b)), CA125, CA199, and CYFRA211 presented significant higher medians in CHQ. The differences would be larger than the critical differences of the TM. Taking CYFRA211 as the example, the medians were 2.5 ng/ml and 1.63 ng/ml in the cancer group of CHQ and CGMH, respectively. The difference between the medians was 0.87 ng/ml which is 53.37% of 1.63 ng/ml (CGMH medians). The difference could be quite significant given the upper limit of the reference range of CYFRA211 was typically 3.3 ng/ml, but may be related to variations in the types of cancer detected in the populations (Table 2(a)).

The distribution of cancer types, was restricted to some specific types of cancer in CHQ and relatively evenly distributed in CGMH (Table 2 (a)). For CHQ, the top three types of cancer were thyroid cancer, lung

Table 1
Demographic data of the cancer cases (a), and non-cancer cases (b) in Chang Gung Memorial Hospital (CGMH) and Chongqing (CHQ).

(a)

	CGMH					CHQ					
	count	mean	std	median	IQR	count	mean	std	median	IQR	effect size
Age	342	58.6	12.74	58	18	433	52.13	12.99	52	19	1.27
AFP	342	1533	19982	3.25	1.9	380	1327	24834	3.19	2.3	0.97
CA125	156	16.2	18.39	10.66	9	135	22.06	36.96	14.2	10.2	−0.79
CA153	156	10.6	5.31	8.9	5.8	83	12.47	10.98	9	4.99	−0.47
CA199	342	15.3	38.38	7.31	11	177	116.3	1278	11.47	10.5	−2.78
CEA	342	4.51	18.76	1.8	2	390	13.95	181.8	2	1.7	−0.67
CYFRA211	342	2	1.46	1.63	1.3	51	2.77	1.12	2.5	1.71	−0.48
PSA	186	12.4	119.7	1.41	2	159	4.53	17.59	0.75	0.93	0.67
SCC	342	0.66	0.78	0.5	0.5	65	0.89	0.47	0.8	0.5	−0.2

(b)

	CGMH					CHQ					
	count	mean	std	median	IQR	count	mean	std	median	IQR	effect size
Age	27596	48.7	12.01	48	16	134803	44.33	13.67	43	20	0.33
AFP	27596	3.62	6.72	3.05	1.8	117878	4.6	258.2	3.14	2	0
CA125	15160	13.9	48.88	9.53	7.5	26622	16.96	22.37	13.72	9.04	−0.11
CA153	15160	9.66	4.54	8.4	5.6	14462	10.19	5.61	9.1	6.2	−0.1
CA199	27596	9.38	19.53	5.83	9.7	45606	11.96	37.41	9.4	8.63	−0.07
CEA	27596	1.86	5.3	1.5	1.2	122258	2.1	10.27	1.71	1.41	−0.03
CYFRA211	27596	1.49	0.88	1.29	0.9	17090	2.19	1.06	2	1.1	−0.68
PSA	12436	1.3	2.23	0.82	0.8	55240	1.09	1.93	0.78	0.7	0.11
SCC	27596	0.55	0.9	0.3	0.3	11406	0.87	0.5	0.8	0.5	−0.53

Table 2(a)
Cancer cases in Chang Gung Memorial Hospital (CGMH) and Chongqing (CHQ) datasets.

Cancers	CGMH	percentage	CHQ	percentage
Bladder cancer	8	2.34%	1	0.23%
Brain cancer	2	0.58%	0	0.00%
Breast cancer	58	16.96%	51	11.51%
Cervical cancer	19	5.56%	15	3.39%
Colon cancer	31	9.06%	9	2.03%
Duodenum carcinoma.	0	0.00%	1	0.23%
Esophageal cancer	6	1.75%	0	0.00%
Gallbladder carcinoma	1	0.29%	2	0.45%
GIST	1	0.29%	3	0.68%
H&N	9	2.63%	1	0.23%
Intrahepatic bile ducts cancer	1	0.29%	0	0.00%
Kidney cancer	16	4.68%	8	1.81%
Leukemia	12	3.51%	1	0.23%
Liposarcoma	1	0.29%	1	0.23%
Liver cancer	39	11.40%	12	2.71%
Lung cancer	15	4.39%	147	33.18%
Lymphoma	8	2.34%	0	0.00%
Nasal neuroendocrine tumor	1	0.29%	1	0.23%
Ovarian cancer	5	1.46%	5	1.13%
Pancreatic cancer	17	4.97%	3	0.68%
Parotid cancer	1	0.29%	0	0.00%
Prostate cancer	30	8.77%	10	2.26%
Retroperitoneum	1	0.29%	0	0.00%
Skin cancer	11	3.22%	0	0.00%
Stomach cancer	12	3.51%	2	0.45%
Testicle cancer	1	0.29%	0	0.00%
Thyroid cancer	27	7.89%	170	38.37%
Unknown	4	1.17%	0	0.00%
Uterus cancer	5	1.46%	0	0.00%

cancer, and breast cancer. Thyroid cancer cases accounted for 38.37% of all cancer cases, while lung cancer cases accounted for 33.18% and breast cancer cases accounted for 11.51%. The top three cancer types accounted for more than 80% of cancer cases identified at CHQ. Specific cancer types, including urinary tract malignancies (bladder cancer and kidney cancer), gastrointestinal malignancies (esophageal cancer, stomach cancer, hepatocellular carcinoma, colon cancer), hematological malignancies (leukemia and lymphoma), prostate cancer, uterine cancer, skin cancer, and head and neck cancer, were more frequently diagnosed in the CGMH population.

For both CHQ and CGMH groups, the majority of cancer cases (CHQ: 82.2%; CGMH: 54.8%) were in the early stages upon diagnosis. Most of the cancer cases at CHQ were stage 1 (74.72%), which was significantly greater than CGMH (41.72%). Stage 3 and 4 cancer cases in CHQ (3.61%) were significantly less than those in CGMH (45.23%) (Table 2 (b)).

3.2. Cancer screening model development

We first set out to explore which model type and which form of data imputation would yield the best prediction of cancer risk. Towards this end we employed the following machine learning models: decision tree (DT), gradient boosting trees (GB), k-nearest neighbor (kNN), logistic regression (LR), long short-term memory model (LSTM), naïve Bayes (NB), and random forest (RF). We also employed four different methods of data imputation: simple replacement with zero, kNN, MICE and deep learning (Table 3(a)). Algorithms were trained with 70% of the entirety of both datasets from CGMH and CHQ and tested with the remaining 30% of the data (Supplement Figure 2). In training we optimized the AUROC value. Results are presented in Table 3a. Note that of all of the models LSTM and RF achieved the highest AUCs and little effect of the various data imputation methods was noted. Based on the literature, replacement with zero is the easiest approach for imputation but often yields the lowest performance since it does not utilize the relationships among the variables. In our experiment, the results showed that replacement with zero is comparable to other approaches (Table 3(a)). Given its simple utilization, we adopted this method for further analyses. The likely reason for the comparability of this method of imputation with the others is likely due to the inherent imbalanced nature of the real-world dataset. Our cancer rate is about 0.5% and the majority of data belongs to the non-cancer class with a significantly small amount of data belonging to the cancer class. The other imputation methods assume an equal number of samples in each class, thus, the performance of these approaches degrade when the class imbalance grows as in the real-world data. Since in the minority class, very few samples can be used for imputation, the imputed values will lean towards the majority class, which may lead to results not that much different than replacement with zero.

To better understand the performance with missing values, we repeated the experiment limiting the input data to only the samples for which a complete set of TM values were present. We then trained a model on 70% of this. We also used the whole raw dataset to build the model with 70% as training dataset. We subsequently tested the two separate algorithms with the complete data and raw data from the remaining 30% of the datasets, respectively (Supplement Figure 2). Results are reported in Table 3b and indicate that the LSTM model yields the highest AUC and is the most stable to training with missing data vs. complete data.

We explored the contribution of different numbers of TMs to the robustness of the model. Shown in Fig. 2, using only the male data from both CGMH and CHQ, we trained an LSTM model using age and one TM (CEA), two TMs (CEA, AFP), three TMs (CEA, AFP, PSA) or four TMs (CEA, AFP, PSA, CA19-9). Again, we used 70% of the data for training and 30% for testing. Results are presented Table 4 and are compared to an analysis of the same dataset using any single marker above the reference range as the criterion for calling positivity. The comparison to this single threshold approach is made because this approach mimics current practice at health check centers. As would be expected, each additional TM adds to the performance of the model and in all cases greatly exceeds the utility of the single threshold approach (Table 4).

The final LSTM model trained on 70% of the complete combined dataset (CGMH and CHQ), using all male and female data, all biomarkers (AFP, CEA, CA199, PSA, CYFRA211, SCC, CA125, CA153, and imputing missing values by replacement with "0" yielded an AUROC of 0.75.

Table 2(b)
Stages for the cancer cases in Chang Gung Memorial Hospital (CGMH) and Chongqing (CHQ) datasets. *: CGMH excluded stage 4 from dataset.

	CGMH n = 314	percentage	CHQ n = 443	percentage	p value
Stage					<0.001
0	41	13.06%	31	7.00%	
1	131	41.72%	331	74.72%	
2	68	21.66%	62	14.00%	
3	74	23.57%	11	2.48%	
4	*		5	1.13%	
unknown	0	0%	3	0.68%	

Table 3(a)
Data imputation comparison: imputed data vs raw data.

	replace with 0		KNN imputation		MICE imputation		Deep learning imputation	
	AUC	ste	AUC	ste	AUC	ste	AUC	ste
DT	0.67	0.01	0.66	0.01	0.66	0.01	0.66	0.01
GB	0.65	0.01	0.66	0.01	0.64	0.01	0.65	0.01
KNN	0.59	0.00	0.61	0.00	0.56	0.00	0.60	0.00
LR	0.69	0.02	0.69	0.01	0.57	0.02	0.70	0.01
LSTM	0.72	0.00	0.73	0.00	0.68	0.00	0.72	0.00
NB	0.68	0.01	0.69	0.01	0.70	0.00	0.69	0.01
RF	0.71	0.01	0.71	0.00	0.72	0.00	0.71	0.01

Table 3(b)
Model performance comparison.

Miss70%, test missing data 30%						full 70%, full 30%					
	AUC	sde AUC	Sen	Sde Sen	spec		AUC	sde AUC	Sen	sde sen	spec
DT	0.70	0.00	0.40	0.01	0.80	DT	0.66	0.01	0.25	0.02	0.80
GB	0.69	0.01	0.38	0.02	0.80	GB	0.65	0.01	0.29	0.03	0.80
KNN	0.62	0.00	0.16	0.00	0.80	KNN	0.63	0.00	0.15	0.00	0.80
LR	0.70	0.02	0.46	0.02	0.80	LR	0.67	0.01	0.49	0.01	0.80
LSTM	0.74	0.00	0.53	0.01	0.80	LSTM	0.73	0.00	0.51	0.01	0.80
NB	0.72	0.00	0.49	0.01	0.80	NB	0.70	0.00	0.48	0.01	0.80
RF	0.71	0.00	0.48	0.01	0.80	RF	0.73	0.00	0.53	0.01	0.80

DT: Decision tree, GB: Gradient Boosting trees, KNN: K nearest neighbor, LR: Logistic regression, LSTM: Long short-term memory model: NB: Naïve Bayes, RF: Random forest.

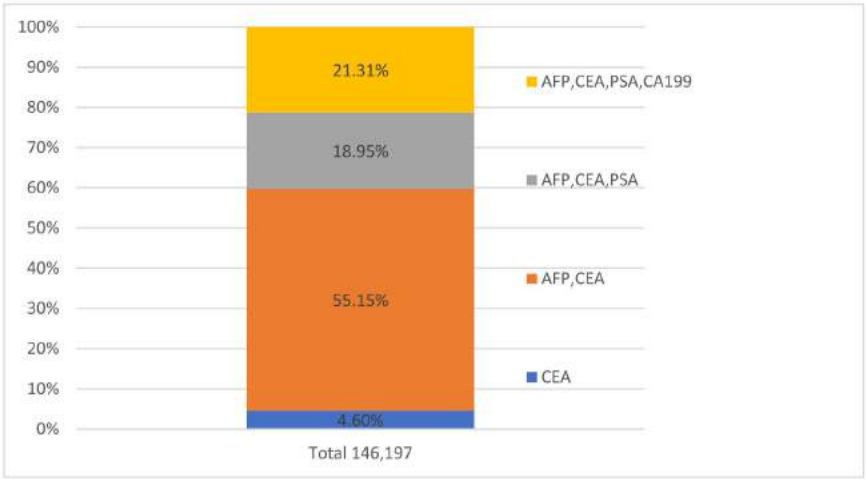

Fig. 2. The percentage of patients with different biomarkers.

Table 4
Performance comparison between LSTM and single threshold.

		LSTM		Single threshold	
Biomarker		Mean	Std	Mean	Std
CEA	AUROC	0.746	0.001	0.533	0.008
	Sensitivity	0.672	0.002	0.097	0.017
	Specificity	0.672	0.002	0.968	0.001
CEA + AFP	AUROC	0.750	0.001	0.544	0.010
	Sensitivity	0.679	0.002	0.124	0.021
	Specificity	0.680	0.002	0.965	0.002
CEA + AFP + PSA	AUROC	0.822	0.002	0.610	0.017
	Sensitivity	0.743	0.003	0.303	0.034
	Specificity	0.744	0.004	0.918	0.004
CEA + AFP + PSA + CA199	AUROC	0.831	0.002	0.628	0.019
	Sensitivity	0.757	0.005	0.354	0.037
	Specificity	0.759	0.005	0.902	0.005

3.3. Time series data model

It is well understood that changes in TM levels over time may be more indicative of the presence of a growing cancer than single time point measurements of the same TMs. As more individuals have been evaluated at annual health check-up appointments, time series data is becoming more available within a real-world cancer screening cohort. As noted above, the LSTM model is well-suited for the analysis of time-series data especially in instances where there is some irregularity in that data (Supplement Figure 1). The dataset we analyzed from CHQ included multi-year TM values (AFP, CEA, PSA, CA199, CA125) from 300 individuals not known to have developed cancer as well as 30 individuals who were diagnosed with cancer during the follow-up period. For the 300 individuals who were not found to have cancer during the follow-up period, complete TM test data was available for four annual test dates. For the 30 individuals who were diagnosed with cancer, 7 cancer patients only had data at time point 1, 10 cancer patients had 2 time points, 9 cancer patients had 3 time points and only 5 cancer patients had all data for 4 time points. In order balance the data between cancer and non-cancer patients, we randomly selected some non-cancer patients and assigned missing values to some data points, so that cancer patients and non-cancer patients had similar patterns regarding patients with biomarker values at each time points. An LSTM algorithm was trained using this time series data analysis after the data preprocessing. To illustrate the effect of number of time points on model performance, we built 4 LSTM algorithms, one using only 1 time point, one using 2 time points, one using 3 time points and one using all 4 time points (Supplement Table 2). The results revealed that the AUROC improved from 0.89 to 0.93 when the number of test point increased

Table 5
Model performance with time series data.

Time points	Biomarkers	AUROC	Sensitivity	Specificity
1	AFP, CEA, PSA, CA199, CA125	0.888	0.833	0.827
2	AFP, CEA, PSA, CA199, CA125	0.908	0.833	0.837
3	AFP, CEA, PSA, CA199, CA125	0.921	0.867	0.863
4	AFP, CEA, PSA, CA199, CA125	0.931	0.867	0.883

from one to four test points (Table 5). The higher AUROC for the one time point model 0.89 developed here vs. 0.75 for the model developed above is due to the use of a more limited dataset (330 vs. 163,174 subjects) and the higher percentage of subjects diagnosed with cancer (9% vs. 0.5%).

3.4. Time-to-diagnosis analysis

We used Cox's proportional hazards algorithm to stratify cancer cases into elevated and high-risk groups in the CGMH and CHQ cohorts, respectively. AFP, CEA and CA199 were the risk factors for cancer diagnosis, $p < 0.001$ (Table 6). CGMH and CHQ use very different criteria for follow-up subsequent to health check examinations. CGMH simply reports resulting TM values to patients and their primary care physicians with no guidance for subsequent follow-up. At CHQ, the health check center is directly involved in follow-up and follows an aggressive scheme as outlined in Fig. 1. Fig. 3a and 3b depict the percentages of subjects diagnosed over time subsequent to a health check visit. It is evident that in CGMH dataset the high-risk group is diagnosed relatively quickly, 50% at ~30 days post visit, while within the elevated risk group the rate of diagnosis is much slower with 50% diagnosed at ~275 days (Fig. 3(b)). The situation at CHQ is very different. Due to the aggressive follow-up, the rate of diagnosis is the same in both the high risk and elevated risk groups, with 50% diagnosed by 25 days post visit and 98% of cases diagnosed by 150 days post visit (Fig. 3(a)). The percentages of early-stage cancers diagnosed in CHQ dataset are significantly higher than CGMH dataset (Table 7). It is evident from this data that the use of TM testing coupled with aggressive follow-up can lead to significantly earlier diagnosis of cancer and at early stages, further supporting the use of machine learning algorithms to better analyze TM results and predict cancer risk.

4. Discussion

In this study, we developed and extensively validated a cancer screening model based on the asymptomatic cohorts whose data was obtained from two independent referral medical centers. Using this complicated real-world data, an LSTM-based cancer screening model was trained which can robustly detect multiple cancer types at the early stages. Moreover, the LSTM algorithm was also advantageous for dealing with time series data of TM panel tests. The time series data enhanced the overall performance of the LSTM models for cancer screening. Additionally, we developed a Cox-regression based model to stratify the cancer cases into different risk groups. The risk groups correlated well to the time-to-cancer diagnosis. In the Cox-regression model, we also found that the intensive follow-up strategy was significantly correlated to early cancer diagnosis. The study could be the first study reporting a tumor markers-based LSTM model being validated by the largest up-to-date real-world data. The robustness and the early cancer diagnosis benefit brought by the approach indicates that a tumor marker-based health check-up combined with an intensive following-up strategy is appropriate to apply to routine health check-up.

Table 6
Characteristics of the risk stratification model used for time-to-cancer diagnosis analysis.

Feature	Coefficients (SE)	HR (95% CI)	p value
Age	0.007358 (0.003806)	1.007385 (0.999898, 1.014927)	0.05
Strategy of follow-up	1.966760 (0.137550)	7.147 (6.228964, 8.201442)	<0.001
AFP	0.000018 (0.000003)	1.000018 (1.000012, 1.000023)	<0.001
CEA	0.008052 (0.002536)	1.008084 (1.003086, 1.013107)	0.001
CA199	0.000203 (0.000061)	1.000203 (1.000084, 1.000322)	<0.001

Using serum tumor markers for cancer screening has raised considerable interests in these years. Wen et al. reported that a TM panel composed of multiple TM can be used as a cancer screening tool in an asymptomatic population [12]. Reference range based single threshold method was used as the interpretative algorithm for the TM results. The diagnostic performance of the TM panel is not optimized due to the fact that the reference ranges of TMs are simply the statistics of a healthy cohort but not designed for cancer screening. Based on the same cohort, Wang et al. harnessed ML algorithms as the interpretative tool for analyzing the TM results [13]. Based on the design, ML algorithms could find a tailored cut-off for a probability score that is derived from multiple TMs. A robust diagnostic performance was independently tested with different cohort in our previous study [4]. Cohen et al. also reported a TM based product (i.e. CancerSEEK) with comparable diagnostic performance [8]. These studies demonstrated robust result by applying ML algorithms in analyzing single time-point TM for cancer screening independently with different cohorts. The irregularity issue of TM data that is commonly encountered in daily screening work has not been well addressed. To improve the TM-based ML model as a practical and useful cancer screening tool in the real world, the study to our best of knowledge is the first study harnessing LSTM as the interpretative tool for the TM-based cancer screening. Our results demonstrated that the LSTM model could cope with data irregularity issue (i.e., missing data and time series data), and thus LSTM is an ideal algorithm for the TM-based cancer screening.

It is well known that missing or irregular sampled data in healthcare is a special challenge since most of the traditional statistical models assume the input data have the similar structure and distributions. Violation of these assumptions can cause learning problems and lead to poor model performance [28]. The general approaches for missing feature problems are handled by imputation, or using more advanced deep learning techniques such as RNN and LSTM models [29]' [30]. Our results showed that for an imbalanced dataset, imputation methods do not lead to better model performance, however they increase the model complexity. On the contrary, an LSTM model yields consistent and robust results with simple replacement of missing values with zero. In the presence of incomplete data, LSTM is robust to missing values by ignoring them when computing statistics in parameter estimation using other non-missing data. Also, in server class imbalanced datasets, the general imputation may impose bias toward the majority classes and lead to poorer performance. This finding could shed light on the strategy to deal with missing feature problems in imbalanced dataset. The age of individuals with cancer detection was significantly older than individuals without cancer. This finding was consistent with the high incidence age of cancers in most cases [31]' [32]. These results suggest that tumor marker screening could benefit most to people over 50 years old. By analyzing the stages of cancer diagnosis, it was found that 592 cases (75.41%) were diagnosed at early stage by screening tumor markers. This was related to the occult onset of most cancers, no characteristic manifestations or clinical manifestations [33]. Therefore, tumor markers, as a means of tumor screening, can be used in the routine health examination of generally healthy people, especially those over 50 years old. The Top 10 incidence rate cancers in our study were thyroid cancer, lung cancer, breast cancer, liver cancer, colorectal cancer, prostate cancer, cervical cancer, renal cell carcinoma, pancreatic cancer, and gastric cancer, which accounted for 94.01% of total cancers in the datasets. But the 2 sites showed different cancer type incidence. In Chongqing, thyroid cancer, lung cancer, breast cancer, cervical cancer and liver cancer are top 5 cancers, and accounted for 89.16% (395 out of 443 cases). But for Taiwan, top 5 cancers are breast cancer, liver cancer, colon cancer, prostate cancer and thyroid cancer, accounted for 54.09% (185 out of 342 cases). Considering the small ethnic differences between the two places, there may be regional differences in cancer [34]. For example, Chongqing has one of the five highest smoking rates in China, and the mortality rate of lung cancer is 9.09/million [35]. In the health

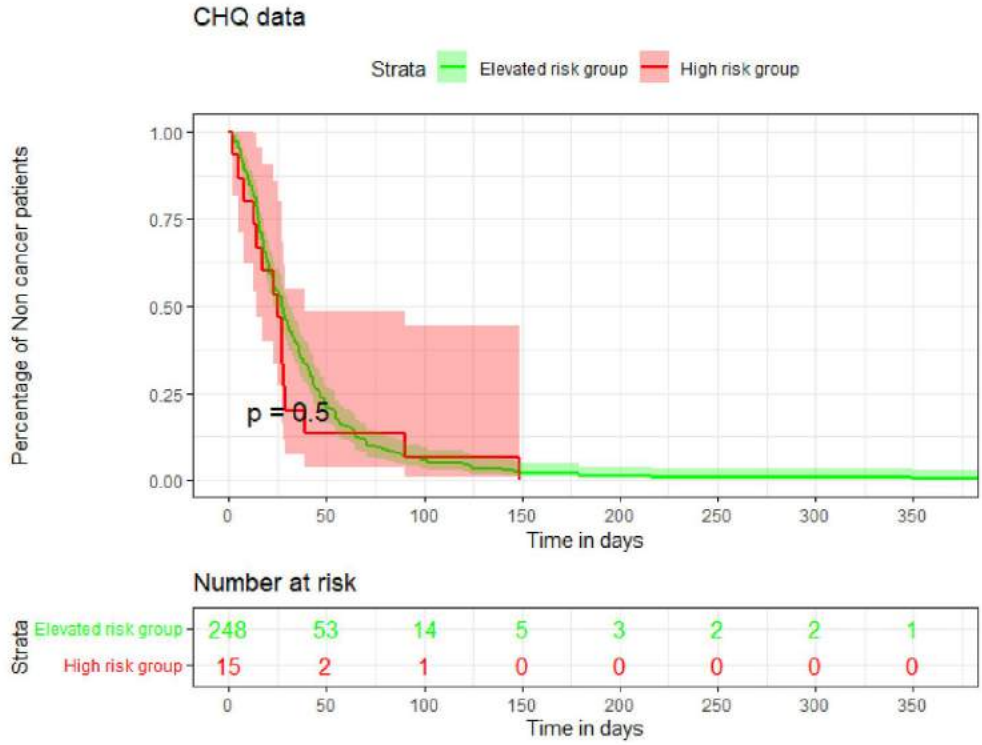

Fig. 3(a). Time-to-cancer diagnosis in different risk groups.

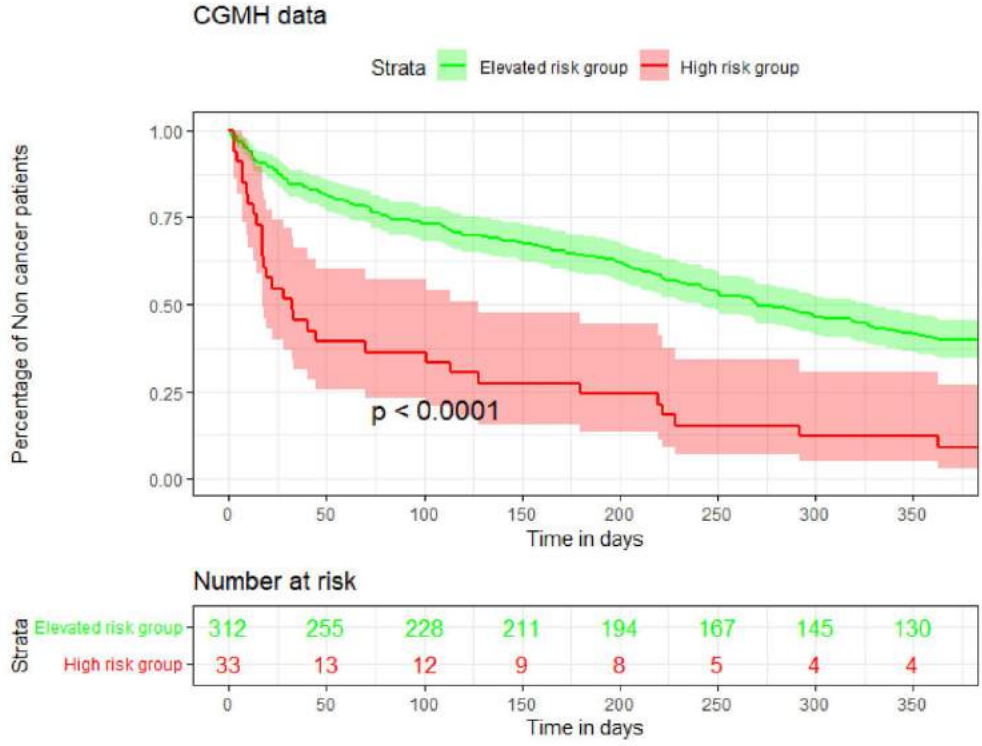

Fig. 3(b). Time-to-cancer diagnosis in different risk groups.

check-up popularization with thin-layer CT and tumor markers, such as CEA, CYFRA211, the detection rate of lung cancers was also significantly increased [36]. Additionally, there are differences in the health check-up examination procedures in the two hospitals. For example, in CHQ, as a routine health check-up examination item, thyroid color B ultrasound had been popularized in the population [37], so the

Table 7
Risk stratification of the cancer cases in Chang Gung Memorial Hospital (CGMH) and Chongqing (CHQ) datasets. *: CGMH excluded stage 4 from dataset.

	CGMH		CHQ	
stage	High risk	Elevated group	High risk	Elevated group
0	0 (0%)	46 (16%)	0 (0%)	25 (9.5%)
1	5 (16%)	133 (45%)	10 (67%)	187 (71%)
2	12 (38%)	56 (19%)	2 (13%)	46 (17%)
3	15 (47%)	58 (20%)	0 (0%)	4 (1.5%)
4	*		2 (13%)	2 (0.8%)

detection rate of thyroid cancer was high and the prognosis of most patients was very good [38].

In analyzing time-to-diagnosis for the cancer cases, the follow-up subsequent to TM results had the most impact on early cancer diagnosis. In the CGMH cohort, the time of 50% cases diagnosed with cancers was corrected to the risk levels assigned by the algorithm. For high-risk group, the time of 50% case diagnosed with cancer was ~30 days post visit, while the elevated group was much later at ~275 days post visit (Fig. 3(b)). Similar findings were reported in a previous study [4]. In the CHQ cohort, the findings were very different; the time of 50% cancer diagnosis for both the high-risk and the elevated risk group were ~25 days, all cancers in high-risk group and 98% of cancers in elevated risk group were diagnosed by 150 days (Fig. 3(a)). Another difference is the stages of cancer diagnosis. In CGMH, 85% cancers diagnosed in the high-risk group were in advanced stages (stage 3 and stage 2, Table 7), while 61% cancers in the elevated risk group were more early stage (stage 0 and stage 1, Table 7). In CHQ, most cancers in the high-risk group (67%) and the elevated risk group (80.5%) were diagnosed at an early stage (Table 7). The different outcomes are due to the intensive follow-up strategy adopted at CHQ for following individuals with abnormal findings that were detected in health checkups (Fig. 1). It is evident that by combining TM screening and clinical follow-ups, health check-up examination can be of significant benefit to patients by detecting cancers earlier before they advance to metastatic disease. This is the first study of semi-quantitative measurement of early cancer diagnosis by TM screening and clinical follow-ups. A prospective study is necessary to clarify the net impact.

Several limitations of this study need to be addressed. First, while the tumor markers are widely used in East Asia, not all tumor markers have been widely employed worldwide. Although powered by AI technique, the irregularity of tumor markers is still the issue that impacts the performance of cancer screening when only one or two tumor markers are tested. Furthermore, our AI models are trained, validated, and independently tested by using data from East Asia only. Direct application of the results to other population in other countries or areas requires further validation in those populations. More bridging studies between different populations are mandatory before broadly applying the AI-added approach. Third, time series data of tumor markers are still much less available than single-point data. Given that time series data theoretically illustrate a more comprehensive view of diseases, more time series data of tumor markers are needed for robustly validating its value. Fourth, we applied LSTM algorithm in analysis of irregular tumor marker data as a proof-of-concept of the approach. Many other AI algorithms that can deal with irregular data were not tested in the study. Further optimization of AI algorithms in analyzing the irregular tumor marker data for cancer screening is worthy of investigation in the future.

5. Conclusion

Based on tumor marker tests, LSTM models can effectively detect cancer earlier than current common medical practice. The performance of the LSTM models are improved by using time series data of tumor marker tests. The LSTM models showed flexibility in dealing with complicated tumor marker tests in real-world datasets, and clinical follow-ups are shown to improve early cancer diagnosis. The approach is readily deployed in routine health check-up examinations.

Funding

Study is funded by National Natural Science Foundation of China (NO.81974385), Chang Gung Memorial Hospital (Linkou), Taiwan (CMRPG3J1791) and 20/20 GeneSysthems, Inc. US (G2021).

Author contributions

Data curation, Xiaoxing Wu, Hsin-Yao Wang, and Peichang Shi; formal analysis, Hsin-Yao Wang, and Peichang Shi; funding acquisition, Jiming Zhou, Yonghong Wang, Ziwei Wang and Jang-Jih Lu; methodology, Xiaoxing Wu, Hsin-Yao Wang, Peichang Shi, Rong Sun, Xiaolin Wang, Zhixiao Luo, Fanling Zeng, Michael Lebowitz, Wan-Ying Lin, Richard Scherer, and Olivia Price; project administration, Jiming Zhou and Yonghong Wu; supervision, Jiming Zhou, Yonghong Wu and Ziwei Wu; writing—original draft, Xiaoxing Wu, Hsin-Yao Wang, Peichang Shi, and Jiming Zhou; writing—review and editing, Xiaoxing Wu, Hsin-Yao Wang, Peichang Shi, Michael Libowitz, Yonghong Wu and Jiming Zhou.

Declaration of competing interest

The 20/20 GeneSystems: INC is now commercializing the algorithms under OneTest brand. Authors Jiming Zhou, Peichang Shi, Michael Libowitz, Richard Scherer, Olivia Price, and Hsin-Yao Wang are affiliated with the company in different capacities. Author Jang-Jih Lu own stocks of 20/20. Authors Xiaoxing Wu, Rong Sun, Zhixiao Luo, Fangling Zeng, Wan-Ying Lin, Yonghong Wu and Ziwei Wang do not have any potential conflicts of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.105362.

References

[1] H. Crooke, M. Kobayashi, B. Mitchell, E. Nwokeji, M. Laurie, S. Kamble, et al., Estimating 1- and 5-year relative survival trends in colorectal cancer (CRC) in the United States: 2004 to 2014, J. Clin. Oncol. 36 (4_suppl) (2018 Feb 1), 587–587.
[2] C. Bettegowda, M. Sausen, R.J. Leary, I. Kinde, Y. Wang, N. Agrawal, et al., Detection of circulating tumor DNA in early- and late-stage human malignancies, Sci. Transl. Med. [Internet] 224 (6) (2014 Feb 19) 926–930, https://doi.org/10.1126/scitranslmed.3007094 [cited 2021 Mar 25];6(224). Available from: https://pubmed.ncbi.nlm.nih.gov/24553385/.
[3] Kakushadze Z, Raghubanshi R, Yu W. Estimating Cost Savings from Early Cancer Diagnosis. [cited 2021 Mar 25]; Available from: www.quantigic.com.
[4] H.-Y. Wang, C.-H. Chen, S. Shi, C.-R. Chung, Y.-H. Wen, M.-H. Wu, et al., Improving multi-tumor biomarker health check-up tests with machine learning algorithms, Cancers (Basel) [Internet] 12 (6) (2020 Jun 1) 1442–1457, https://doi.org/10.3390/cancers12061442 [cited 2021 Mar 25];12(6). Available from: http://www.ncbi.nlm.nih.gov/pubmed/32492934.
[5] H.Y. Wang, T.W. Lin, S.Y.H. Chiu, W.Y. Lin, S Bin Huang, J.C.H. Hsieh, et al., Novel toilet paper-based point-of-care test for the rapid detection of fecal occult blood: instrument validation study, J. Med. Internet Res. [Internet] 22 (8) (2020 Aug 1) 1–16, https://doi.org/10.2196/20261 [cited 2022 Feb 2];22(8). Available from: https://pubmed.ncbi.nlm.nih.gov/32763879/.
[6] P.F. Pinsky, Lung cancer screening with low-dose CT: a world-wide view, J. Med. Internet Res. [Internet] 7 (3) (2018 Jun 1) 234–242 [cited 2022 Feb 2];7(3): 234–42. Available from: https://pubmed.ncbi.nlm.nih.gov/30050762/.
[7] S.S. Nazari, P. Mukherjee, An overview of mammographic density and its association with breast cancer, Breast Cancer [Internet] 25 (3) (2018 May 1) 259–267, https://doi.org/10.1007/S12282-018-0857-5/FIGURES/2 [cited 2022 Feb 2];25(3):259–67. Available from: https://link.springer.com/article/10.1007/s12282-018-0857-5.
[8] 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 [Internet]. Vol. vol. 17. [cited 2021 Mar 25]. Available from: http://science.sciencemag.org/.
[9] Y.J. Tseng, C.E. Huang, C.N. Wen, P.Y. Lai, M.H. Wu, Y.C. Sun, et al., Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data

with machine learning technologies, Int. J. Med. Inform. [Internet] 128 (2019 Aug 1) 79–86 [cited 2021 Mar 25];128:79–86. Available from: https://pubmed.ncbi.nlm.nih.gov/31103449/.

[10] R. Molina, R.M. Marrades, J.M. Augé, J.M. Escudero, N. Viñolas, N. Reguart, et al., Assessment of a combined panel of six serum tumor markers for lung cancer, Am. J. Respir. Crit. Care Med. [Internet] 193 (4) (2016 Feb 15) 427–437, https://doi.org/10.1164/RCCM.201404-0603OC/SUPPL_FILE/DISCLOSURES.PDF [cited 2022 Feb 2];193(4):427–37. Available from: www.atsjournals.org.

[11] J.A. Baron, Screening for cancer with molecular markers: progress comes with potential problems, Nat. Rev. Cancer 12 (5) (2012) 368–371, https://doi.org/10.1038/nrc3260, 125 [Internet]. 2012 Apr 12 [cited 2022 Feb 2];12(5):368–71. Available from: https://www.nature.com/articles/nrc3260.

[12] Y.H. Wen, P.Y. Chang, C.M. Hsu, H.Y. Wang, C.T. Chiu, J.J. Lu, Cancer screening through a multi-analyte serum biomarker panel during health check-up examinations: results from a 12-year experience, Clin. Chim. Acta [Internet] 450 (2015 Oct 23) 273–276, https://doi.org/10.1016/J.CCA.2015.09.004 [cited 2021 Mar 25];450:273–6. Available from: https://pubmed.ncbi.nlm.nih.gov/26344337/.

[13] H.Y. Wang, C.H. Hsieh, C.N. Wen, Y.H. Wen, C.H. Chen, J.J. Lu, Cancers screening in an asymptomatic population by using multiple tumour markers, PLoS One 11 (6) (2016 Jun 1).

[14] Commission NH. Chinese Health Statistics Yearbook, vol. 121, 2019.

[15] H.-Y. Wang, C.-H. Chen, S. Shi, C.-R. Chung, Y.-H. Wen, M.-H. Wu, et al., Improving multi-tumor biomarker health check-up tests with machine learning algorithms, Cancers (Basel) [Internet] 12 (6) (2020), https://doi.org/10.3390/cancers12061442 [cited 2021 Mar 25];12:1442. Available from: www.mdpi.com/journal/cancers.

[16] K.E. De Visser, A. Eichten, L.M. Coussens, Paradoxical roles of the immune system during cancer development, Nat. Rev. Cancer 6 (1) (2006 61) 24–37, https://doi.org/10.1038/nrc1782 [Internet]. 2006 Jan [cited 2022 Feb 3];6(1):24–37. Available from: https://www.nature.com/articles/nrc1782.

[17] B. Vogelstein, K.W. Kinzler, The path to cancer — three strikes and you're out, N. Engl. J. Med. [Internet] 373 (20) (2015 Nov 12) 1895–1898, https://doi.org/10.1056/NEJMP1508811/SUPPL_FILE/NEJMP1508811_DISCLOSURES.PDF [cited 2022 Feb 3];373(20):1895–8. Available from: https://www.nejm.org/doi/full/10.1056/NEJMp1508811.

[18] Z. Che, S. Purushotham, K. Cho, D. Sontag, Y. Liu, Recurrent neural networks for multivariate time series with missing values, Sci. Rep. [Internet] 1 (8) (2018 Dec 1) 6085–6096, https://doi.org/10.1038/s41598-018-24271-9 [cited 2021 Jul 2];8(1):6085. Available from: www.nature.com/scientificreports.

[19] Jo Y, Lee L, Palaskar S. Combining LSTM and Latent Topic Modeling for Mortality Prediction.

[20] R. DiPietro, G.D. Hager, Deep learning: RNNs and LSTM, in: Handbook of Medical Image Computing and Computer Assisted Intervention [Internet], Elsevier, 2019 [cited 2021 Jul 2]. pp. 503–19. Available from: https://jhu.pure.elsevier.com/en/publications/deep-learning-rnns-and-lstm.

[21] T.D. Pigott, A review of methods for missing data, Int. J. Phytoremediation. [Internet] 21 (1) (2001) 353–383 [cited 2021 Jul 2];21(1):353–83. Available from: https://www.tandfonline.com/doi/abs/10.1076/edre.7.4.353.8937.

[22] M.T. Kayembe, S. Jolani, F.E.S. Tan, G.J.P. van Breukelen, Imputation of missing covariate in randomized controlled trials with a continuous outcome: scoping review and new results, Pharm. Stat. [Internet] 19 (6) (2020 Nov 1) 840–860 [cited 2021 Jul 2];19(6):840–60. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/pst.2041.

[23] Z. Zhang, Missing data imputation: focusing on single imputation, Pharm. Stat. [Internet] 4 (1) (2016 Jan 1) 9–17 [cited 2021 Jul 2];4(1):9. Available from:/pmc/articles/PMC4716933/.

[24] (PDF) Review on Missing Value Imputation Techniques in Data Mining [Internet]. [cited 2021 Jul 2]. Available from: https://www.researchgate.net/publication/329625460_Review_on_Missing_Value_Imputation_Techniques_in_Data_Mining.

[25] A. Jadhav, D. Pramod, K. Ramanathan, Comparison of performance of data imputation methods for numeric dataset, Appl. Artif. Intell. [Internet] 33 (10) (2019 Aug 24) 913–933 [cited 2021 Jun 16];33(10):913–33. Available from: https://www.tandfonline.com/doi/abs/10.1080/08839514.2019.1637138.

[26] Y. Duan, Y. Lv, W. Kang, Y. Zhao, A deep learning based approach for traffic data imputation, in: 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 2014. Institute of Electrical and Electronics Engineers Inc.; 2014, 2014, pp. 912–917.

[27] Central Limit Theorem (CLT) Definition [Internet]. [cited 2022 Feb 2]. Available from: https://www.investopedia.com/terms/c/central_limit_theorem.asp.

[28] S. Das, S. Datta, B.B. Chaudhuri, Handling data irregularities in classification: foundations, trends, and future challenges, Pattern Recogn. 81 (2018 Sep 1) 674–693.

[29] C. Xiao, E. Choi, J. Sun, Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review [Internet], in: Journal of the American Medical Informatics Association, vol. 25, Oxford University Press, 2018 [cited 2021 Jul 3]. pp. 1419–28. Available from: https://pubmed.ncbi.nlm.nih.gov/29893864/.

[30] T.A. Lasko, J.C. Denny, M.A. Levy, Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data, PLoS One [Internet] 8 (6) (2013 Jun 24) 66341 [cited 2021 Jul 3];8(6):66341. Available from: www.plosone.org.

[31] Cancer statistics, 2019 | Enhanced Reader [Internet]. [cited 2021 Mar 25]. Available from: chrome-extension://dagcmkpagjlhakfdhnbomgmjdpkdklff/enhanced-reader.html?pdf=https%3A%2F%2Fbrxt.mendeley.com%2Fdocument%2Fcontent%2F095cf861-e97a-3011-8b67-820df1b942aa.

[32] A. Ahmad, Epigenetics in personalized management of lung cancer, in: Advances in Experimental Medicine and Biology, Springer New York LLC, 2016, pp. 111–122.

[33] A.M. Patel, S.G. Peters, Clinical manifestations of lung cancer, Mayo. Clin. Proc. [Internet] 68 (3) (1993) 273–277 [cited 2021 Mar 25];68(3):273–7. Available from: https://pubmed.ncbi.nlm.nih.gov/8474271/.

[34] S. Pilleron, D. Sarfati, M. Janssen-Heijnen, J. Vignat, J. Ferlay, F. Bray, et al., Global Cancer Incidence in Older Adults, 2012 and 2035: A Population-Based Study, 2018.

[35] J.H. Pan Huixian, Zhuoting Li, Ye Zhang, N.J. Wang Rui, A preliminary study on the model of health examination follow up service based on health management, Hosp. Admin. J. Chin. PLA 26 (2019) 756–759.

[36] Q. Yang, P. Zhang, R. Wu, K. Lu, H. Zhou, Identifying the Best Marker Combination in CEA, CA125, CY211, NSE, and SCC for Lung Cancer Screening by Combining ROC Curve and Logistic Regression Analyses: Is it Feasible?, 2018, https://doi.org/10.1155/2018/2082840 [cited 2021 Mar 25]; Available from:.

[37] Operation Skill and Standard Diagnosis and Treatment Are the Basics of Improving the Curative Effect of Thyroid Carcinoma. [Internet]. [cited 2021 Mar 25]. Available from: http://ebhyxbwk.njournal.sdu.edu.cn/EN/10.6040/j.issn.1673-3770.1.2016.01.

[38] K. Bibbins-Domingo, D.C. Grossman, S.J. Curry, M.J. Barry, K.W. Davidson, C. A. Doubeni, et al., Screening for Thyroid Cancer: US Preventive Services Task Force Recommendation Statement, vol. 317, JAMA - Journal of the American Medical Association. American Medical Association, 2017, pp. 1882–1887.