Heliyon 8 (2022) e10932

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Modeling strategies to analyse longitudinal biomarker data: An illustration on predicting immunotherapy non-response in non-small cell lung cancer

Frederik A. van Delft [a], Milou Schuurbiers [b], Mirte Muller [c], Sjaak A. Burgers [c], Huub H. van Rossum [d], Maarten J. IJzerman [a,e,f], Hendrik Koffijberg [a,*,1], Michel M. van den Heuvel [b,1]

[a] Health Technology and Services Research Department, Technical Medical Centre, University of Twente, Enschede, Overijssel, 7522NH, the Netherlands
[b] Department of Respiratory Diseases, Radboud University Medical Center, Nijmegen, Gelderland, 6525GA, the Netherlands
[c] Department of Thoracic Oncology, Netherlands Cancer Institute, Amsterdam, Noord-Holland, 1066CX, the Netherlands
[d] Department of Laboratory Medicine, Netherlands Cancer Institute, Amsterdam, Noord-Holland, 1066CX, the Netherlands
[e] Centre for Cancer Research and Centre for Health Policy, University of Melbourne, Parkville, Melbourne, Victoria, Australia
[f] Peter MacCallum Cancer Centre, Parkville, Melbourne, Victoria, Australia

ARTICLE INFO

Keywords:
NSCLC
Immunotherapy
Response
Serum tumor markers
CYFRA
CEA
CA-125
NSE
SCC

ABSTRACT

Serum tumor markers acquired through a blood draw are known to reflect tumor activity. Their non-invasive nature allows for more frequent testing compared to traditional imaging methods used for response evaluations. Our study aims to compare nine prediction methods to accurately, and with a low false positive rate, predict progressive disease despite treatment (i.e. non-response) using longitudinal tumor biomarker data. Bi-weekly measurements of CYFRA, CA-125, CEA, NSE, and SCC were available from a cohort of 412 advanced stage non-small cell lung cancer (NSCLC) patients treated up to two years with immune checkpoint inhibitors. Serum tumor marker measurements from the first six weeks after treatment initiation were used to predict treatment response at 6 months. Nine models with varying complexity were evaluated in this study, showing how longitudinal biomarker data can be used to predict non-response to immunotherapy in NSCLC patients.

1. Introduction

The introduction of immunotherapy has led to prolonged survival of patients with metastasized non-small cell lung cancer (NSCLC) (Reck et al., 2016). Currently, PD-L1 expression is the best biomarker used to predict immunotherapy response in advanced NSCLC. And at present pembrolizumab is a standard first-line treatment for patients with advanced NSCLC without actionable oncogenic drivers and a PD-L1 expression of >50% (Planchard et al., 2018).

In current practice, tumor dynamics are evaluated by radiological assessment using the Response Evaluation Criteria in Solid Tumors (RECIST) criteria (Eisenhauer et al., 2009). However, these criteria are limited since functional and metabolic changes are not quantified, therefore imaging results, especially early in treatment, may not correspond with treatment effect. Due to uncertainty about tumor response, immunotherapy is often prolonged without patient benefit and with concomitant risk of side effects and unnecessary costs. As a consequence, there is a need for early and accurate assessment of treatment effect, to enable an early decision about treatment continuation. Since the introduction of immunotherapy, multiple other predictive biomarkers have been studied for their ability to predict response, primarily aimed towards patient selection. Biomarkers used in the prediction of immunotherapy response include tumor mutational burden (TMB), exhaled breath condensate, radiomics, or profiling of serum microRNAs (De Vries et al., 2019; Fan et al., 2020; Wei et al., 2018). In the majority of studies, these biomarkers are measured upfront, before treatment is started and predictive validity so far has been limited. However, once treatment is initiated, the tumor dynamically changes over time and therefore the dynamic behavior of biomarkers may be more sensitive to determine tumor response aptly. The longitudinal assessment of biomarkers during the first treatment cycles may be used to predict the probability of progression or can be used for monitoring purposes. In monitoring, changes in a biomarker value are used to determine if a certain event e.g., disease progression occurs.

* Corresponding author.
E-mail address: h.koffijberg@utwente.nl (H. Koffijberg).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.heliyon.2022.e10932
Received 23 May 2022; Received in revised form 9 September 2022; Accepted 29 September 2022

In contrast to prediction models with a biomarker measured at a single point in time, the interpretation of longitudinal biomarker data is challenging as no criteria have been established towards using longitudinal biomarker data in clinical decision making, and clinical use is therefore mainly limited to experience of the individual clinician. A clinical application of longitudinal biomarker measurements is the monitoring for prostate cancer recurrence after prostatectomy or identification of patients at risk of developing castrate resistant prostate cancer after initiation of androgen deprivation therapy. In both cases longitudinal assessment of prostate-specific antigen measurements might help the stratification of high-risk patients and can guide treatment or follow-up (Kim et al., 2016; Tourinho-Barbosa et al., 2018). In advanced NSCLC patients receiving immune checkpoint inhibitors serum tumor marker measurements (Carcinoembryonic antigen (CEA), serum cytokeratin 19 fragment (CYFRA 21.1), cancer antigen 125 (CA-125) and neuro specific enolase (NSE)) can be used to detect early disease progression in the first six months of therapy, since it was shown that these biomarkers are a surrogate for total tumor mass, implying monitoring potential (Molina et al., 2010; Lang et al., 2019; Moritz et al., 2018; Muller et al., 2021).

While serum tumor marker measurements at one time point can be informative, the dynamics of these tumor markers over time will likely provide more information on treatment response. Several methods for the analysis of longitudinal biomarker measurements have been proposed in literature. To date there are no studies comparing models of varying complexity in response prediction using multiple consecutive serum protein measurements. Studies often focus on a single approach, e.g., distinct biomarker patterns, including several consecutive increments, cut-off values, relative changes in biomarker values, and functional principal component analysis (Lund et al., 2014; Sjöström et al., 2001; Sölétormos et al., 2000, Yan et al., 2017; Moritz et al., 2018). Biomarker velocity or doubling time might also provide valuable information (Loeb et al., 2008). Besides these approaches, more complex statistical methods can be used to predict treatment outcomes using longitudinal measurements, e.g., joint-modeling, neural-networks, or landmark analysis (Bull et al., 2020). In joint-models and landmark analysis dynamic changes of covariates over time are incorporated in a survival model. Neural networks are commonly described as a network of neurons in which weights are assigned to the connections between neurons and the neurons contain their own weight, bias term, and activation function. While these methods may each be valuable, guidance on which method to use or studies investigating comparative performance, is lacking. Consequently, in many studies only a single method is applied to predict the outcomes of interest, chosen based on convenience or expertise of the researcher while this may impair best use of the biomarker information and optimal predictive performance (Van Rossum et al., 2021).

This study aims to compare nine analytical methods utilizing longitudinal serum tumor marker measurements in the prediction of immunotherapy effect in advanced NSCLC patients.

2. Methods

This study was based on a previously described patient cohort including patients treated with either nivolumab or pembrolizumab at the Netherlands Cancer Institute (Muller et al., 2021; Schouten et al., 2018). Patients who started treatment between March 2013 and September 2018 were included in the cohort, and follow-up was conducted until January 2019. All patients included in this study were randomly selected for either a training (75%) or a validation (25%) cohort. The training cohort was used to train the prediction models and define prediction thresholds to achieve 95% specificity. Model training focused on specificity since the intended purpose of the prediction models is to inform treatment decisions in order to discontinue treatment early in case of progressive disease (i.e. non-response). For clinical use, a high specificity is required to ensure a low false positive rate and thereby minimizing the chance to falsely withhold treatment. The validation cohort was then used to evaluate and compare the overall model performance. The retrospective collection of data was approved by the local institutional review board and medical ethics committee.

2.1. Tumor marker tests

Serum tumor markers were prospectively measured just prior to start of therapy, at bi-weekly intervals, or at clinical follow-up, which was either every other week for Nivolumab or every three weeks for Pembrolizumab. CA-125, CEA, Cyfra 21.1, and NSE were analyzed on a Roche Cobas 6000 analyzer system. Squamous cell carcinoma antigen (SCC) was performed on a Thermofisher Kryptor immunoassay system. Different prediction models were developed, all based on the same set of biomarkers.

2.2. General model requirements and performance criteria

This study aims to predict response at 6 months after treatment initiation using longitudinal biomarker data obtained in the first six weeks of treatment (Rizvi et al., 2015; Muller et al., 2021). Therefore, the main outcome assessed in this study was non-response at six months after treatment initiation. Prediction and outcome time points were chosen based on currently used RECIST evaluation time points. The first evaluation is performed after six weeks, non-response at six months is chosen to reflect durable clinical benefit. Non-response was defined as progressive disease based on RECIST criteria, clinical progressive disease, or death. Monitoring of response was done through a computed tomography (CT) scan at six weeks, three months, and every three months thereafter. RECIST 1.1 guidelines were used to classify partial response, stable disease, and progressive disease (Eisenhauer et al., 2009).

The baseline biomarker value was determined between 7 days prior to and 1 day after treatment initiation. In case multiple samples were taken during this period, the measurement closest to treatment initiation was selected as the baseline. The 6^{th} week biomarker values were determined between 35 days and 49 days after treatment initiation, with the sample closest to day 42 after treatment initiation selected as the end of the 6^{th} week of therapy. Patients were excluded from the analysis in case required measurements were missing or the number of measurements within the six-week time period was insufficient. Further information on inclusion criteria, handling of missing data, and data transformations is included in Supplemental T 1.

Patients were classified according to the outcome of logical tests, or the class probability derived from a prediction model. Thresholds for both the logical tests and class probability derived form prediction models were derived from a receiver operating characteristics (ROC) curve generated in R statistical software using either the pROC package or a custom function (R Core Team, 2019; Robin et al., 2011). In a previous study Muller et al. developed a prediction model based on the patient cohort described in this study. This prediction model was based on a 50% increase from baseline and a marker dependent minimum value. In that study Muller et al. aimed for a 97.5% specificity and a 20% sensitivity to consider the test as useful (Muller et al., 2021). Since our study uses a more lenient 95% specificity during model training, models are deemed useful when reaching a 20% sensitivity and 95% specificity on the validation results. Furthermore, the diagnostic accuracy of the applied methods was assessed through a bootstrap procedure. In the bootstrap analysis, 1000 samples were randomly selected from the dataset with replacement, the sample size was equal to 50% of the full dataset. The constructed prediction models and prediction threshold were applied to the bootstrap samples. Results obtained through the bootstrap procedure were compared based on the average sensitivity and specificity. The presentation of results was limited to sensitivity and specificity since the logical tests are based on three decision thresholds, resulting in incomplete ROC curves. Therefore it is not possible to calculate the area under the ROC curve.

2.3. Prediction models

Nine prediction models were tested in this study with the handling of the longitudinal data defining the specific differences (Figure 1). Models one to five aim to detect prespecified features in the data, also referred to as logical tests. Models six to nine are based on statistics and machine learning.

Additionally, an upper threshold of relevancy (UTR) and a lower threshold of relevancy (LTR) were applied to methods 2, 3, 4, and 5. If one of the logical tests returned a positive result for a biomarker measurement below the LTR, the positive test result was ignored. A patient was classified as having progressive disease in case two consecutive biomarker measurements exceeded the UTR. Also, the logical test outcome was used in case a biomarker value exceeded the UTR, but the preceding biomarker measurement was below the UTR. All prediction thresholds and thresholds for the LTR and UTR ware chosen to maximize sensitivity at a 95% specificity to ensure a low false positive rate.

Two types of survival models were included in the study. The first model consisted of a Cox proportional hazards model. This model was constructed using three covariates, i.e., the average biomarker value up

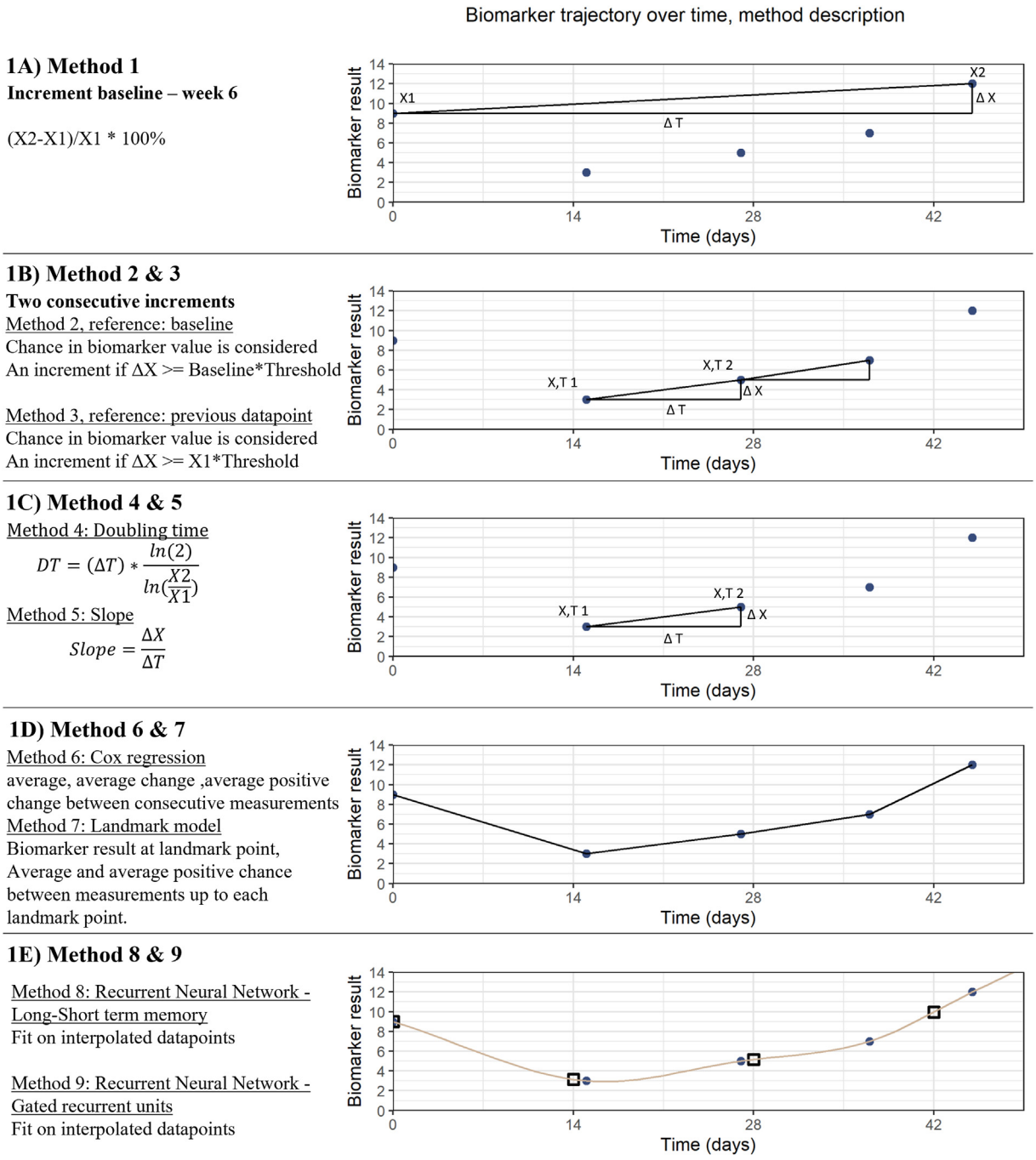

Figure 1. An overview of the methods used in this study. In all windows, the dots represent generated data points used to visualize an exemplary biomarker trajectory. The method based on the increment between baseline and week-six is depicted in 1A. Two consecutive increments, methods 2 and 3 are depicted in 1B. Methods based on the doubling time (method 4) and slope (method 5) are depicted in 1C. Regression based models, methods 6 and 7, are depicted in 1D. Recurrent neural networks, methods 8 and 9, are depicted in window 1E. In 1E, the line through the data points represents a monotone Hermite spline used for interpolation. In this window, the black squares represent the interpolated biomarker results.

to day 49 after treatment initiation, the average change between consecutive measurements, and the average positive change between consecutive measurements. The second model consisted of a landmark model. This model was constructed using the most recent biomarker measurement up to the landmark timepoint, the average biomarker value up to the landmark timepoint, and the average of all positive increments up to the landmark timepoint as covariates. For both models a prediction threshold resulting in a 95% specificity was derived from the ROC curve.

Additionally, two different types of recurrent neural networks (RNN) architectures were included in this study. Multiple RNN architectures are available, with long-short term memory (LSTM) and gated recurrent units (GRU) being the most well known (DiPietro and Hager, 2020). RNNs require time series as input. Since the number and spacing of measurements over time differed per patient a monotone Hermite spline was fit to log transformed data for interpolation. Interpolation results were transformed back using the natural exponential function, thereby removing any negative interpolation results. To aid model training, a second log transformation was performed prior to model training. Both the LSTM and GRU models consisted of 3 layers of 32, 16, and 1 units respectively. The final layer consisted of a fully connected layer with sigmoid activation to allow for binary classification. For both models a dropout and recurrent dropout of 10% were specified for the first two layers.

Supplemental T 1 provides an overview of all methods included in this study and R packages used for model fitting.

3. Results

The patient cohort contained 412 patients of which 307 and 105 were allocated to the training and validation cohort, respectively (Table 1) (Schouten et al., 2020, Schouten et al., 2018).

An overview of the training and validation results is depicted in Table 2. The highest sensitivity achieved on the training data was 32.5% using CYFRA and the method evaluating the increment between baseline and week six, resulting in a specificity of 94.7%. On the validation data the sensitivity and specificity dropped to 24.1% and 93.3%, respectively. On the validation data the highest sensitivity was 38.6% using CYFRA and the RNN-GRU model, resulting in a specificity of 100%. While on the training data this method achieved a sensitivity and specificity of 31.8% and 95.2%, respectively.

The largest decrease in sensitivity between the training and validation data was found for NSE using the Cox model. Showing a decrease of 15.3% between the training (sensitivity 32.0%) and validation (sensitivity 16.7%) results. The largest increase in sensitivity between the training and validation results was 20.1%, and was found for SCC and the method based on the detection of two consecutive increments. However, this increase in sensitivity resulted in a 10.1% decrease in specificity.

Results show that each method exceeded the 20% sensitivity reference value for at least one biomarker in the training set, validation set, and bootstrap average. Validation results show that 17 biomarker and method combinations resulted in a sensitivity >20% while maintaining a 95% specificity. Additionally, 7 biomarker and method combinations resulted in a sensitivity >30%, while maintaining a 95% specificity; CYFRA using two consecutive increments (baseline reference), doubling time, slope, and both RNN models, and NSE using two consecutive increments (both references). This indicated that several biomarker and method combinations can be considered useful, and should be subjected to external validation. Moreover, CYFRA provided the most consistent high sensitivity across methods on the validation data of the included biomarkers.

Results from the individual bootstrap samples and the average sensitivity and specificity obtained by the bootstrap analysis are depicted in Figure 2 and in Supplemental T 2. The results depicted in Figure 2 did not show the correlation between the sensitivity and specificity found in each bootstrap sample, therefore, the correlation and covariance

Table 1. Patient characteristics and description of the training and validation set.

	Training-set	Validation-set	Full cohort
Patients (n (%))	307 (74.5%)	105 (25.5%)	412 (100%)
Mean age (years (SD))	63.7 (9.16)	62.7 (10.1)	63.5 (9.4)
Male sex (n (%))	159 (51.8%)	65 (61.9%)	224 (54.3%)
Treatment			
- Nivolumab (n (%))	272 (88.6%)	100 (95.2%)	372 (90.3%)
- Pembrolizumab (n (%))	35 (11.4%)	5 (4.8%)	40 (9.7%)
Lines of therapy prior to immunotherapy			
- 0 (n (%))	6 (2.0 %)	2 (1.9%)	8 (1.9%)
- 1 (n (%))	237 (77.2%)	80 (76.2%)	317 (76.9%)
- 2 (n (%))	45 (14.7%)	17 (16.2%)	62 (15.0%)
- >2 (n (%))	19 (6.2%)	6 (5.7%)	25 (6.2%)
ECOG performance status at therapy start			
- 0 (n (%))	90 (29.3%)	32 (30.5%)	122 (29.6%)
- 1 (n (%))	183 (59.6%)	56 (53.3%)	239 (580%)
- 2 (n (%))	31 (10.1%)	14 (13.3%)	45 (10.9%)
- 3 (n (%))	2 (0.7%)	2 (1.9%)	4 (1.0%)
- 4 (n (%))	1 (0.3%)	1 (1.0%)	2 (0.5%)
Smoking status			
- Never smoker (n (%))	31 (10.1%)	17 (16.2%)	48 (11.7%)
- Smoker (n (%))	67 (21.8%)	18 (17.1%)	85 (20.6%)
- Former smoker (n (%))	209 (68.1%)	70 (66.7%)	279 (67.7%)
Pack years (years (SD))	33.4 (17.7)	36.3 (20.7)	34.1 (18.5)
Brain metastasis (n (%))	58 (18.9%)	29 (27.6%)	87 (21.1%)
Histology			
- Adenocarcinoma (n (%))	207 (67.4%)	74 (70.5%)	281 (68.2%)
- Squamous (n (%))	68 (22.1%)	19 (18.1%)	87 (21.1%)
- Other	32 (10.4%)	12 (11.4%)	44 (10.7%)
Reason for treatment cessation			
- Progression (n (%))	213 (69.3%)	78 (74.3%)	291 (70.6%)
- irAE (n (%))	26 (8.4%)	12 (11.4%)	38 (9.2%)
- Lost to follow up (n (%))	45 (14.7%)	8 (7.6%)	53 (12.9%)
- Other* (n (%))	23 (7.5%)	7 (6.7%)	30 (7.3%)
Number of patients with PD at 6 months (n (%))	210 (68.4%)	71 (67.7%)	281 (68.2%)
Mean survival after treatment start (days (SD))	232 (198)	255 (225)	238 (206)
Mean duration treatment received in (days (SD))	136 (153)	143 (163)	138 (156)
Patients with biomarker measurements			
- CYFRA (n (%))	306 (99.7%)	103 (98.1%)	409 (99.3%)
- CEA (n (%))	299 (97.4%)	101 (96.2%)	400 (97.1%)
- CA-125 (n (%))	305 (99.3%)	102 (97.1%)	407 (98.8%)
- NSE (n (%))	305 (99.3%)	102 (97.1%)	407 (98.8%)
- SCC (n (%))	258 (84.4%)	80 (76.2%)	338 (82.0%)

*Study end, wish patient, patient's condition, complications.
Carcinoembryonic antigen: CEA, serum cytokeratin 19 fragment: CYFRA 21.1, cancer antigen 125: CA-125, neuro specific enolase: NSE, standard deviation: SD, progressive disease: PD, immune related Adverse Event: irAE.

Table 2. An overview of the sensitivity and specificity calculated for each method and biomarker combination on the training and validation data. All models were trained to achieve a 95% specificity on the training cohort. The highest sensitivity per biomarker is indicated by a black border, sensitivity results >20% and <30% were marked in italics, sensitivity results >30% were marked in bold.

Method	Training data									
	CYFRA		CEA		CA125		NSE		SCC	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
1) Δ baseline-week 6	**0.325**	0.947	*0.200*	0.945	0.183	0.958	0.118	0.945	0.026	0.949
2) 2× increased BL	*0.232*	0.947	*0.281*	0.946	0.154	0.957	*0.279*	0.947	0.180	0.947
3) 2× increased PL	*0.239*	0.952	*0.206*	0.951	0.138	0.950	*0.285*	0.952	0.198	0.956
4) Doubling time	*0.258*	0.952	0.116	0.951	0.170	0.950	*0.285*	0.952	0.081	0.956
5) Slope	**0.313**	0.952	0.116	0.951	*0.239*	0.950	*0.228*	0.952	0.072	0.956
6) Cox model	0.109	0.952	**0.313**	0.951	*0.272*	0.952	**0.320**	0.952	0.143	0.955
7) Landmark model	**0.313**	0.953	**0.313**	0.953	*0.213*	0.953	*0.244*	0.951	0.089	0.962
8) RNN-LSTM	*0.229*	0.952	**0.315**	0.951	*0.217*	0.951	*0.258*	0.952	0.075	0.956
9) RNN-GRU	**0.318**	0.952	*0.255*	0.951	*0.224*	0.951	*0.272*	0.952	0.084	0.956
Method	Validation data									
	CYFRA		CEA		CA125		NSE		SCC	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
1) Δ baseline-week 6	*0.241*	0.933	0.107	0.964	0.103	0.933	0.069	1.000	0.067	1.000
2) 2× increased BL	**0.341**	1.000	0.175	0.926	0.154	1.000	**0.341**	0.963	**0.381**	0.846
3) 2× increased PL	*0.286*	1.000	0.188	0.964	0.174	1.000	**0.347**	0.964	**0.308**	0.875
4) Doubling time	**0.327**	1.000	0.188	1.000	0.196	1.000	*0.286*	1.000	0.115	1.000
5) Slope	**0.347**	1.000	0.167	1.000	*0.283*	0.967	0.163	1.000	0.154	1.000
6) Cox model	0.060	0.929	*0.213*	0.962	*0.277*	1.000	0.167	0.962	0.097	0.824
7) Landmark model	*0.225*	0.923	*0.225*	0.923	*0.233*	0.963	0.132	0.963	0.115	0.900
8) RNN-LSTM	**0.318**	1.000	**0.372**	0.862	*0.262*	0.968	*0.227*	0.966	0.087	1.000
9) RNN-GRU	**0.386**	1.000	*0.256*	0.931	*0.262*	0.968	*0.273*	0.966	0.130	1.000

between the sensitivity and specificity are provided in Supplemental T 3 and 4, respectively. The highest average sensitivity over 1000 bootstrap samples was found for CYFRA using the RNN-GRU, which was 33.6%. While the sensitivity in the training and validation data was 31.8% and 38.6%, respectively. Moreover, the average specificity found for this method and biomarker combination was 96.5%. The lowest average sensitivity found was 3.3% for SCC using the increase in biomarker value between baseline and week six, resulting in an average specificity of 96.2%.

4. Discussion

This is the first study comparing multiple methods using sequential serum tumor markers to predict treatment response in advanced NSCLC patients receiving immune checkpoint inhibitors. The results of the internal validation provided the best estimates for actual model performance. Therefore, the best performing biomarker and method combination was chosen based on the validation results, and was defined as the combination providing the highest sensitivity while maintaining a 95% specificity. Consequently, the RNN-GRU based on CYFRA-data provided the best performance with a sensitivity and specificity of 38.6% and 100%, respectively. The second-best performance was achieved by again CYFRA, now using the slope between consecutive measurements, resulting in a sensitivity and specificity of 34.7% and 100%, respectively. Based on the performance of the best performing model on the validation data, approximately 38% of non-responders can be identified after 6 weeks of therapy. Meaning that other treatment or monitoring choices can be made up to 4.5 months prior to radiological disease progression.

While this novel approach to predict treatment outcomes using sequential serum tumor marker measurements shows promising results, further optimization of models is required to maximize predictive performance. Besides optimization of the prediction models, other clinical endpoints and input sequence lengths should be considered in future studies to optimize clinical benefit. This study is a first step in the development of a clinical prediction model using a novel strategy, therefore more efforts such as external validation of the proposed models in prospective multicenter studies is required to enable adoption of a prediction model in clinical practice. Moreover, results of this study indicate that it might be worthwhile to further evaluate the applicability of the described methods to other disease stages, anti-cancer therapies, or malignancies, than presented in this study.

While combining multiple serum tumor markers in a prediction model is expected to improve the predictive performance, this was not deemed feasible for the logical tests included in this study. The logical tests are based on three thresholds (i.e., a threshold for the test condition, UTR, and LTR), and these thresholds need to be selected for each serum tumor marker included in the model. Resulting in fifteen decision thresholds when combining all five serum tumor markers included in this study.

A broad selection of available methods was made to include models with varying levels of flexibility and complexity. Still, more methods are available (Bull et al., 2020). Overall, our results were indicative of a strong variation in the ability of the chosen prediction methods to extract information from the biomarkers. The RNNs were the most promising, providing the most consistent high sensitivity across most biomarkers. However, less complex methods, e.g., two consecutive increments could outperform the more flexible RNN on some biomarkers, i.e., NSE and SCC. Moreover, using CYFRA, the RNN slightly outperformed two consecutive increments, 38.6% versus 34.1% sensitivity, respectively. The possibly less complicated implementation of a logical test in clinical care, e.g., two consecutive increments, compared to the implementation of a RNN, could result in a greater clinical impact due to more widespread adoption.

While blood samples were assumed to be taken on a bi-weekly basis, some variance in the time interval between measurements was found. To account for this variance in time intervals, measurements taken less than 5 days apart were excluded from the data, as well as measurements performed more than 30 days apart. The constraint on the maximum number of days between measurements did not affect the inclusion of

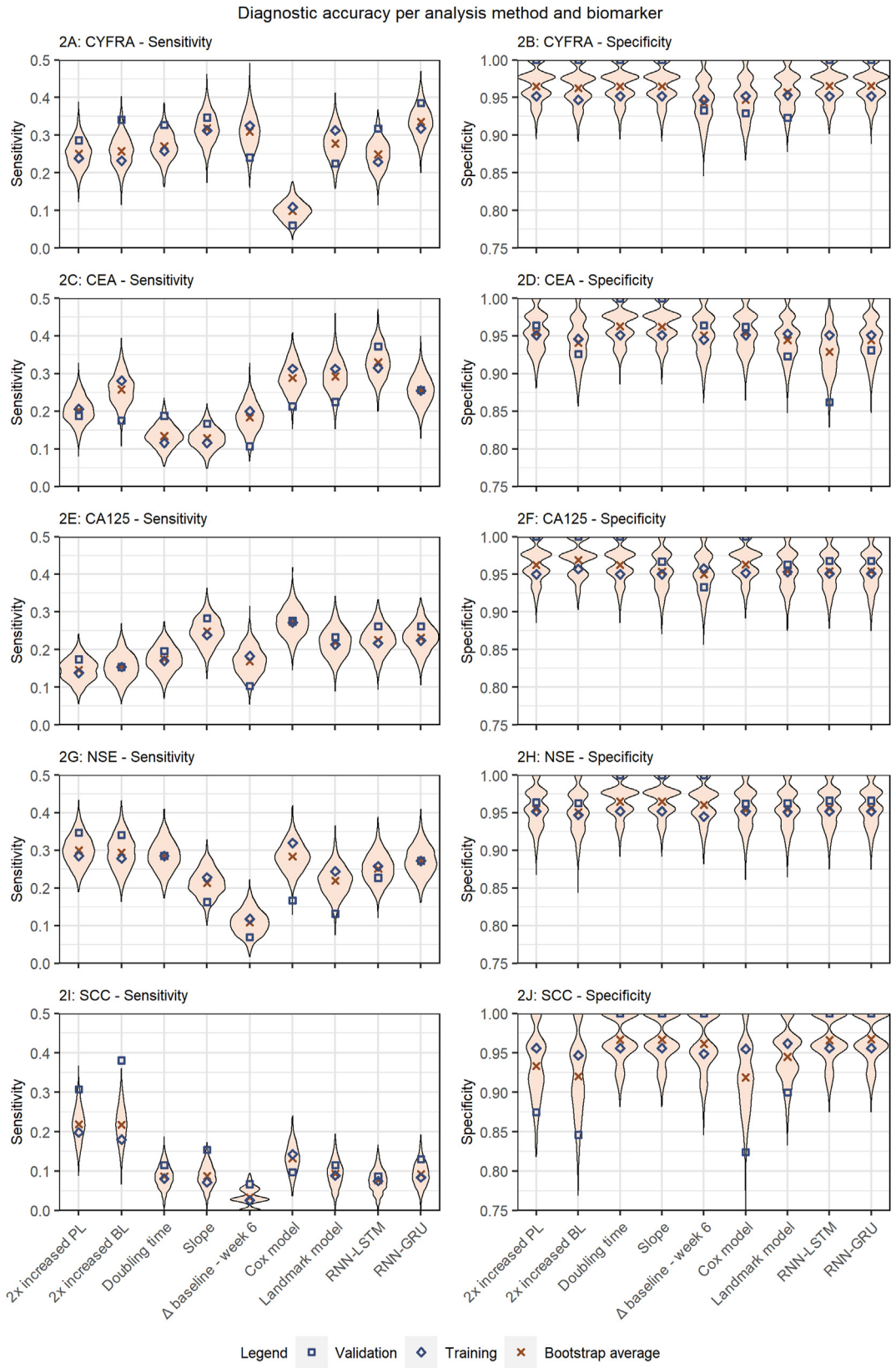

Figure 2. The results per method for CYFRA. CEA. CA-125. NSE. and SCC. The violin plots are used to depict the distribution of the bootstrap results. The "x" marker depicts the bootstrap average. the results obtained on the training and validation set are depicted by the diamond and square marker respectively. The test specificity is depicted on the right side of the multi-window figure (2B: CYFRA, 2D: CEA, 2F: CA-125, 2H: NSE, 2J: SCC). The test sensitivity is depicted on the left side of figure (2A: CYFRA, 2C: CEA, 2E: CA-125, 2G: NSE, 2I: SCC).

patients, since patients with less than three measurements between baseline and day 49 of therapy were excluded from the analysis in a prior step. Three thresholds for the minimum number of days between measurements were assessed with regard to their effect on the sample size. A minimum of five, seven, and ten days were compared. Increasing the minimum from five to ten days excluded an additional 2 to 11 patients depending on the biomarker (Supplemental T 5). This small variation in sample size is unlikely to change the results of this study noticeably. Of the methods included in this study only the RNN required regular time intervals, therefore an interpolation step using Hermite splines was used to transform the data. Except for the method comparing baseline to the measurement at week six, all methods required at least 2 biomarker values between 7 days prior to treatment initiation and 49 days thereafter.

The Cox regression model and the landmark model are both based on covariates which aim to capture the change in biomarker value over time. Therefore, these methods are more dependent on the defined covariates than other methods included in this study. This study did not extensively compare the use of different covariates. However, during model construction the covariates were discussed with a multidisciplinary team to ensure the covariates were able to reflect potentially important data features. A minimal of 10 events per variable (EPV) is generally advised for hazards regression models to result in a statistically valid model (Peduzzi et al., 1995). In this study, the size of the training set was limited to 306 patients or less depending on the tumor marker evaluated. Given that three covariates were used in the regression models, and in all analyses the number of events was much larger than 30 (i.e. the number of events required for 10 EPV in the models with three covariates) the risk of developing statistically invalid models was deemed minimal.

The RNN models included in this study performed well across most biomarkers, on both the training and validation data. Also, the RNN-GRU outperformed all other methods on the validation data using CYFRA measurements as input. Performance of the included RNN models might be limited by the relatively small size of the dataset, especially for SCC which was measured the least frequent. Despite limitations in sample size, other studies in similar settings also report a good performance based on the AUC or other metrics (Kaji et al., 2019, Choi et al., 2016; Ceccarelli et al., 2017; Güler and Übeyli, 2006). While the RNN models performed well, more work in the optimization of the network structure and hyperparameter selection (e.g. number of layers, number of layers, activation functions, and dropout) per biomarker are likely to increase the model performance even further, at least in some biomarkers. However, optimization of neural networks is an elaborate task and not within the scope of this study.

In conclusion, this study showed how models with varying complexity can be used to predict early non-response in immunotherapy treated NSCLC patients based on sequential serum tumor marker measurements. Results show that the performance of prediction methods varies per biomarker, and that for any given biomarker it is worthwhile to compare the performance of different prediction methods. The RNN models presented in this study showed good performance across most biomarkers and should therefore be externally validated. Moreover, most models included in this study performed well using CYFRA measurements as input, indicating that CYFRA provides the most predictive information. With more and more biomarker information routinely collected in clinical practice, and the availability of many types of prediction models, determining the best prediction model is becoming more challenging but also more rewarding.

Declarations

Author contribution statement

Frederik van Delft: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Milou Schuurbiers; Michel M. van den Heuvel; Huub H. van Rossum: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Mirte Muller; Sjaak A. Burgers: Contributed reagents, materials, analysis tools or data; Wrote the paper.

Maarten J. IJzerman: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

Hendrik Koffijberg: Analyzed and interpreted the data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

The authors do not have permission to share data.

Declaration of interest's statement

The authors declare the following conflict of interests

H.H. van Rossum is founder of: Huvaros B.V and SelfSafeSure Bloodcolelctions B.V.

Patents

PCT/NL2016/050315 Method for setting measuring equipment, computer program and measuring equipment.

NL2019/079641 Blood collection device and method for the self-collection of blood by a user.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2022.e10932.

Acknowledgements

None to declare.

References

Bull, L.M., Lunt, M., Martin, G.P., Hyrich, K., Sergeant, J.C., 2020. Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods. Diagn. Progn. Res. 4, 9.

Ceccarelli, F., Sciandrone, M., Perricone, C., Galvan, G., Morelli, F., Vicente, L.N., Leccese, I., Massaro, L., Cipriano, E., Spinelli, F.R., Alessandri, C., Valesini, G., Conti, F., 2017. Prediction of chronic damage in systemic lupus erythematosus by using machine-learning models. PLoS One 12, e0174200.

Choi, E., Schuetz, A., Stewart, W.F., Sun, J., 2016. Using recurrent neural network models for early detection of heart failure onset. J. Am. Med. Inf. Assoc. 24, 361–370.

De Vries, R., Muller, M., Van Der Noort, V., Theelen, W., Schouten, R.D., Hummelink, K., Muller, S.H., Wolf-Lansdorf, M., Dagelet, J.W.F., Monkhorst, K., Maitland-Van Der Zee, A.H., Baas, P., Sterk, P.J., Van Den Heuvel, M.M., 2019. Prediction of response to anti-PD-1 therapy in patients with non-small-cell lung cancer by electronic nose analysis of exhaled breath. Ann. Oncol. 30, 1660–1666.

Dipietro, R., Hager, G.D., 2020. Chapter 21 - deep learning: RNNs and LSTM. In: Zhou, S.K., Rueckert, D., Fichtinger, G. (Eds.), Handbook of Medical Image Computing and Computer Assisted Intervention. Academic Press.

Eisenhauer, E.A., Therasse, P., Bogaerts, J., Schwartz, L.H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., Rubinstein, L., Shankar, L., Dodd, L., Kaplan, R., Lacombe, D., Verweij, J., 2009. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur. J. Cancer 45, 228–247.

Fan, J., Yin, Z., Xu, J., Wu, F., Huang, Q., Yang, L., Jin, Y., Yang, G., 2020. Circulating microRNAs predict the response to anti-PD-1 therapy in non-small cell lung cancer. Genomics 112, 2063–2071.

Güler, İ., Übeyli, E.D., 2006. A recurrent neural network classifier for Doppler ultrasound blood flow signals. Pattern Recogn. Lett. 27, 1560–1571.

Kaji, D.A., Zech, J.R., Kim, J.S., Cho, S.K., Dangayach, N.S., Costa, A.B., Oermann, E.K., 2019. An attention based deep learning model of clinical events in the intensive care unit. PLoS One 14, e0211057.

Kim, Y., Park, Y.H., Lee, J.Y., Choi, I.Y., Yu, H., 2016. Discovery of prostate specific antigen pattern to predict castration resistant prostate cancer of androgen deprivation therapy. BMC Med. Inf. Decis. Making 16, 63.

Lang, D., Horner, A., Brehm, E., Akbari, K., Hergan, B., Langer, K., Asel, C., Scala, M., Kaiser, B., Lamprecht, B., 2019. Early serum tumor marker dynamics predict progression-free and overall survival in single PD-1/PD-L1 inhibitor treated advanced NSCLC-A retrospective cohort study. Lung Cancer 134, 59–65.

Loeb, S., Kettermann, A., Ferrucci, L., Landis, P., Metter, E.J., Carter, H.B., 2008. PSA doubling time versus PSA velocity to predict high-risk prostate cancer: data from the baltimore longitudinal study of aging. Eur. Urol. 54, 1073–1080.

Lund, F., Petersen, P.H., Pedersen, M.F., Hassan, S.O.A., Sölétormos, G., 2014. Criteria to interpret cancer biomarker increments crossing the recommended cut-off compared in a simulation model focusing on false positive signals and tumour detection time. Clin. Chim. Acta 431, 192–197.

Molina, R., Holdenrieder, S., Auge, J.M., Schalhorn, A., Hatz, R., Stieber, P., 2010. Diagnostic relevance of circulating biomarkers in patients with lung cancer. Cancer Biom. 6, 163–178.

Moritz, R., Muller, M., Korse, C.M., Van Den Broek, D., Baas, P., Van Den Noort, V., Ten Hoeve, J.J., Van Den Heuvel, M.M., Van Rossum, H.H., 2018. Diagnostic validation and interpretation of longitudinal circulating biomarkers using a biomarker response characteristic plot. Clin. Chim. Acta 487, 6–14.

Muller, M., Hoogendoorn, R., Moritz, R.J.G., Van Der Noort, V., Lanfermeijer, M., Korse, C.M., Van Den Broek, D., Ten Hoeve, J.J., Baas, P., Van Rossum, H.H., Van Den Heuvel, M.M., 2021. Validation of a clinical blood-based decision aid to guide immunotherapy treatment in patients with non-small cell lung cancer. Tumor Biol. 43, 115–127.

Peduzzi, P., Concato, J., Feinstein, A.R., Holford, T.R., 1995. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J. Clin. Epidemiol. 48, 1503–1510.

Planchard, D., Popat, S., Kerr, K., Novello, S., Smit, E.F., Faivre-Finn, C., Mok, T.S., Reck, M., Van Schil, P.E., Hellmann, M.D., Peters, S., 2018. Metastatic non-small cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann. Oncol. 29, iv192–iv237.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/.

Reck, M., Rodríguez-Abreu, D., Robinson, A.G., Hui, R., Csőszi, T., Fülöp, A., Gottfried, M., Peled, N., Tafreshi, A., Cuffe, S., O'Brien, M., Rao, S., Hotta, K., Leiby, M.A., Lubiniecki, G.M., Shentu, Y., Rangwala, R., Brahmer, J.R., 2016. Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung cancer. N. Engl. J. Med. 375, 1823–1833.

Rizvi, N.A., Hellmann, M.D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J.J., Lee, W., Yuan, J., Wong, P., Ho, T.S., Miller, M.L., Rekhtman, N., Moreira, A.L., Ibrahim, F., Bruggeman, C., Gasmi, B., Zappasodi, R., Maeda, Y., Sander, C., Garon, E.B., Merghoub, T., Wolchok, J.D., Schumacher, T.N., Chan, T.A., 2015. Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. Science 348, 124.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinf.

Schouten, R.D., Egberink, L., Muller, M., De Gooijer, C.J., Van Werkhoven, E., Van Den Heuvel, M.M., Baas, P., 2020. Nivolumab in pre-treated advanced non-small cell lung cancer: long term follow up data from the Dutch expanded access program and routine clinical care. Transl. Lung Cancer Res. 9, 1736–1748.

Schouten, R.D., Muller, M., De Gooijer, C.J., Baas, P., Van Den Heuvel, M., 2018. Real life experience with nivolumab for the treatment of non-small cell lung carcinoma: data from the expanded access program and routine clinical care in a tertiary cancer centre-The Netherlands Cancer Institute. Lung Cancer 126, 210–216.

Sjöström, J., Alfthan, H., Joensuu, H., Stenman, U.H., Lundin, J., Blomqvist, C., 2001. Serum tumour markers CA 15-3, TPA, TPS, hCG β and TATI in the monitoring of chemotherapy response in metastatic breast cancer. Scand. J. Clin. Lab. Investig. 61, 431–441.

Sölétormos, G.R., Hyltoft Petersen, P., Dombernowsky, P., 2000. Progression criteria for cancer antigen 15.3 and carcinoembryonic antigen in metastatic breast cancer compared by computer simulation of marker data. Clin. Chem. 46, 939–949.

Tourinho-Barbosa, R., Srougi, V., Nunes-Silva, I., Baghdadi, M., Rembeyo, G., Eiffel, S.S., Barret, E., Rozet, F., Galiano, M., Cathelineau, X., Sanchez-Salas, R., 2018. Biochemical recurrence after radical prostatectomy: what does it mean? Int. Braz J. Urol. 44, 14–21.

Van Rossum, H.H., Meng, Q.H., Ramanathan, L.V., Holdenrieder, S., 2021. A word of caution on using tumor biomarker reference change values to guide medical decisions and the need for alternatives. Clin. Chem. Lab. Med.

Wei, M., Jin, Q., Hong, L., Matthew, S., Yoganand, B., Ilke, T., Robert James, G., 2018. Radiomic Biomarkers from PET/CT Multi-Modality Fusion Images for the Prediction of Immunotherapy Response in Advanced Non-small Cell Lung Cancer Patients. Proc.SPIE.

Yan, F., Lin, X., Huang, X., 2017. Dynamic prediction of disease progression for leukemia patients by functional principal component analysis of longitudinal expression levels of an oncogene. Ann. Appl. Stat. 11, 1649–1670.