**EARLY DETECTION OF AGGRESSIVE CANCER USING LONGITUDINAL BIOMARKER MEASUREMENTS**

**Introduction**

In cancer diagnosis, timing is of the essence. When cancer cells become increasingly heterogeneous, they become more likely to metastasize, invading other parts of the body and rendering effective treatment more difficult. Cancer cells release various biomarkers such as nucleic acids and proteins into the blood [1], many of which can be analyzed to detect cancerous growths at an early stage. A number of biomarkers are released by both healthy and cancerous cells, and when a cancer begins to develop the corresponding biomarker levels deviates from a healthy baseline concentration and grows at a rate proportional to tumor growth [2-3]. To accurately screen patients for the presence of cancer, it is necessary to distinguish between a healthy biomarker concentration and one that is indicative of cancer growth.

One biomarker that is widely used for the purposes of patient screening is prostate-specific antigen (PSA). The clinical standard practice for gauging cancer status using PSA is comparing an individual’s PSA concentration to a universal threshold value of 4 ng/mL [4], which is assumed to be a healthy level across all patients. However, there are several additional factors, primarily age, that can contribute to the rise of a patient’s healthy PSA level beyond a threshold value. Thus, this conventional method retains a high false positive rate. In addition to the variation in the “healthy” range among individuals, there is also variation in an individual’s own biomarker levels among samples [5] as a result of measurement error, circadian rhythms, and other physiological processes that make it difficult to confidently predict cancer onset from thresholding blood biomarker concentrations alone.

Here we present a method employing a longitudinal inspection of blood biomarker levels to screen patients for cancer to enable earlier detection of aggressive cancers with greater accuracy. We first establish a patient’s healthy baseline biomarker level. We then predict whether a patient will have cancer by detecting when biomarker measurements begin to deviate from a healthy baseline, beyond a measurement range attributed to assay noise.

In this study, we aimed to determine the number and frequency of biomarker samples required to classify the presence or absence of tumors within an acceptable (pre-determined) amount of error. We simulated longitudinal data for both healthy and cancer patients, and then normalized biomarker measurements to a patient’s specific baseline using two normalization methods. We then classified the presence or absence of cancer using one of two supervised learning methods. This theoretical study provides us with insight as to how optimal biomarker sampling schedules can be established to improve screening for lethal cancers.

**Materials and Methods**

Blood biomarker measurements were simulated for 200 healthy patients and 200 asymptomatic early-stage cancer patients over a period of 400 weeks. Two methods (mean substraction, standard score (z-score) normalization) were used to identify each patient’s healthy baseline biomarker concentration based on the initial 100 data points. Subsequent measurements were normalized to individual healthy baseline concentrations. Cancer status was then classified based on normalized biomarker levels using 10-fold cross validation and one of two supervised machine learning techniques: 1) *k*-nearest neighbors and 2) classification thresholding.

***Simulation of Longitudinal Biomarker Measurements* (Fig. 1.1)**

For this simulated study, all 400 patients were assumed healthy (without cancer) on day 1. We simulated true (noise-free) biomarker measurements for 200 patients who did not develop cancer during the course of the 400-week study and for 200 patients who eventually developed cancer on day 200 (*t*onset = 200 days). For the healthy patient cohort, the true baseline biomarker concentration  was assumed constant, i.e.,

,

where *ch0* values were randomly selected from a normal distribution with mean 8 ng/mL and standard deviation 1.5 ng/mL.

For patients who eventually developed aggressive cancer, the true biomarker concentration  was assumed to increase according to the Gompertz function,



where  values were also randomly selected from a normal distribution with mean 8 ng/mL and standard deviation 1.5 ng/mL. Tumor growth rate  and decay rate  were randomly selected from uniform distributions with ranges [3.6×10-2, 7.2×10-2] and [5.0×10-3, 1.0×10-2], respectively.

The noise-free measurement, however, is not a realistic model of blood sampling. To account for the various sources of variation and more precisely emulate realistic measurements, noise was introduced to the true values. For this part of the experiment, three different noise models were examined independently to generalize this study to a variety of assays with potentially varying noise.

The *constant error model* is intended for assays with a roughly constant magnitude of noise for each measurement. As such, for each patient, a normal distribution was generated with mean zero and constant standard deviation (representing a numerical noise value). Then, each time point measurement was perturbed using noise values chosen from that distribution The amount of noise is characterized by the value of the standard deviation.

In the *standardized fractional error* model, the noise for each measurement is proportional to the magnitude of that measurement. For each patient, a normal distribution was generated with mean 0 and a constant standard deviation (representing a fractional value). For each measurement, a fraction was selected from the normal distribution, then multiplied by and added to the true measurement. In this model, noise was characterized by the value of the standard deviation.

The *fractional error model* is intended for assays with relatively constant nonzero mean. For each patient, a normal distribution was generated around a nonzero mean (representing a fractional value) and standard deviation of mean×0.5. For each measurement, the mean was multiplied by the magnitude of the measurement, then a value was selected from the newly created distribution, randomly multiplied by 1 or -1, and added to the true measurement. The noise level was characterized by the value of the mean.

***Normalization to Patient-Specific Baselines* (Fig. 2.1)**

Simulated noisy biomarker measurements were normalized using one of two different methods.

(i) Standard score (z-score) normalization

The mean () and standard deviation () of the first 100 days of measurements were computed, and  was assumed to be the patient’s healthy baseline. After the first 100 days, a subsequent measurement  was normalized as follows:

.

(ii) Mean Subtraction

The mean () of the first 100 days of measurements was computed and assumed to be the patient’s healthy baseline. Subsequent measurements  were normalized as follows:

.

***Classification with k-Nearest Neighbors (k-NN)* [7]**

(i) Weighted *k*-NN

Given a patient’s longitudinal biomarker history represented as a multidimensional vector, this algorithm computes the *k* neighbors of minimum Euclidean distance in a set of previously-acquired biomarker measurements belonging to patients who have a known cancer status. The given patient’s cancer status is then classified with the majority label of the *k* nearest neighbors, where labels with smaller Eucliden distance are considered with weight inversely proportional to distance. Mathematically, the distance between the input patient’s history (represented as an *m*-dimensional vector ) and a labeled history (*m*-dimensional vector ) is calculated as

.

For this study, the weighted *k*-NN method with *k* = 5 was applied in two different ways: 1) using fixed observation spans; 2) using dynamically increasing observation spans to mimic a patient’s subsequent visits to provide additional blood samples.

First, we considered the relationship between sampling frequency and observation span, focusing on their relative impacts on classification accuracy. Ten-fold cross validation was then used to assess classification performance. The patients were split into a training set containing 90% of the patients and a testing set containing the remaining 10% of patients, with the ratio of healthy to cancerous patients remaining at 1:1. For each patient in the testing set, the *k* nearest neighbors in the training set were identified, and the test patient was classified according to the status distribution of the neighbors. The following values were calculated:

|  |  |
| --- | --- |
| Accuracy | (# correctly classified) / (total # of signals) |
| Sensitivity (Recall) | (true positive) / (true positive + false negative) |
| Specificity | (true negative) / (true negative + false positive) |
| Precision | (true positive) / (true positive + false positive) |
| F-score | (2 × precision × recall) / (precision + recall) |

In the dynamic *k*-NN approach, a set of 400 patients, 200 cancerous and 200 healthy, was generated. All data was normalized using the standard score and mean subtraction methods. For each patient in the testing set, we began by classifying them using only their first biomarker measurement. If they were classified as healthy, we moved onto their next biomarker measurement and classified them using their first two biomarker measurements. This process was repeated until they were classified as having cancer with a high confidence level or until they reached the end of the 1000 days of observation, at which point they would be assigned a healthy status. The time of classification in days post-onset of cancer was noted. We compared the performance of this dynamic approach using an 80% confidence level threshold and 95% confidence level threshold (Fig. 3.1)

***Classification with Thresholding***

To classify with thresholding, a patient was assigned a cancerous status if their normalized biomarker level crossed a certain threshold value. As in the *k*-NN classification method, we applied thresholding in two different ways: 2) using fixed observation spans, and 2) using dynamically increasing observation spans.

In the fixed length approach, a set of 100 patients, 50 cancerous and 50 healthy, was generated with a certain observation span, 10-day sampling frequency, 15% noise, and a healthy baseline of 10 ± 5 ng/mL. A second set was simulated with no variation in healthy baseline. The sets were normalized using z-score and expanding window methods. We generated a collection of threshold values to test, ranging from a very small threshold (which classified all trajectories as cancerous) to a very large threshold (which classified all trajectories as healthy). Again, the set was split into a 90-10 training-testing set. For each threshold, we classified each patient in the training set using the last observed measurement in the patient’s observation span. If the measurement was larger than the threshold, we classified the patient as cancerous; if it was smaller, we classified the patient as healthy. We then calculated the F-score of the classification using this threshold. Once all thresholds were tested, we selected the threshold that gave the largest F-score, used this “optimal” threshold to classify the patients in the testing set, and calculated the sensitivity, specificity, and F-score of the classification on the testing patients. The threshold selection and classification process was repeated for 10 training-testing splits.

We applied the dynamic approach to the same set of simulated patients used in the *k-*NN dynamic approach. The normalized set was split into a testing and training set ten times, as detailed above. A large range of threshold values was generated. For a given threshold value, we examined the first biomarker measurement for each test patient; if the measurement crossed the threshold, we classified the patient as having cancer and noted the time that the threshold was crossed (time of cancer detection). If the measurement was below the threshold, we examined the patient’s next biomarker measurement. We continued to progress through the patient’s biomarker measurements until either the threshold value was crossed or until we reached the end of the 1000 days of observation, at which point the patient would be classified as healthy. For each threshold value, we calculated the average time the threshold was crossed, and the sensitivity and specificity of classification (Fig 3.2).

**RESULTS**

*k-Nearest Neighbor Classification: Fixed Length Approach*

We examined the performance of the nearest neighbor algorithm in response to changes in three main parameters: (i) the amount of noise in the measurements, (ii) the frequency of measurements, and (iii) the observation span, i.e., the total amount of time over which measurements were taken. The purpose of this analysis was to examine parameters that may be adjusted in the clinic, and to determine clinically feasible noise levels and sampling frequencies that will lead to earliest detection time with highest accuracy. Keeping one of the parameter values fixed, a two-way analysis was performed on the remaining two parameters. The combinations that resulted in a classification accuracy of greater than 80%, greater than 95%, and greater than 99% were plotted, with the x-axis depicting one parameter value and the y-axis depicting the second parameter value (Figs. 1.2, 1.5, 1.6).

With noise fixed at 0%, only two observations were needed for the classification to achieve greater than 99% accuracy, regardless of sampling frequency (Fig. 1.2, A, slope = 1). With noise increased to 15%, the necessary length of observation appeared to grow logarithmically as the frequency of sampling decreased. This relationship was explored in a three-way analysis of sampling frequency, total number of observations, and classification sensitivity and specificity (Fig 1.3). Examining the plot visually, the border between the light yellow colored region (higher sensitivity and specificity) and the darker green colored region (lower sensitivity and specificity) appeared to follow a power function with respect to sampling frequency. The (x,y) pairs corresponding to 95% sensitivity were log-transformed and fitted using least-squares linear regression. The fit was best for higher sampling frequencies; as the time between samples increased beyond 150 days, two observations were sufficient to achieve greater than 95% sensitivity, and the points flattened to a horizontal line (Fig 1.4, A4). Applying the fitted equation to the observation length, which was calculated as the product of the time between samples and the number of observations minus 1, again the fit was most appropriate for higher sampling frequencies (Fig. 1.4, B1).

Classification accuracy was also assessed with two fixed parameters and one varying parameter. With sampling frequency and observation length fixed, increasing noise from 0 % to 10% had little effect on classification performance across all three noise models. As noise increased above 10%, accuracy decreased significantly (Fig. 1.8). With noise and sampling frequency fixed, increasing observation time improved classification accuracy. For all three noise models, once the observation length was long enough, classification accuracy reached its maximum 100%. This occured earliest for the 0-mean percent noise model, followed by the percent noise model, and finally the constant noise model.

Comparing across normalization methods, at low noise levels and with no variation in population baseline, using the z-score normalized, mean-subtracted, or unnormalized measurements produced nearly equivalent results with the nearest neighbor classification. Both the shifting and expanding window autoregressive models had lower sensitivity and specificity values (Fig. 2.2). At higher noise levels and no population baseline variation, the relative results of the autoregressive models improved (Fig. 2.3, 2.4). Upon adding variation to the healthy population baseline, at 5% noise level, the z-score normalized measurements yielded the highest f-scores. As noise level increased, the results across all methods became more similar, and none of the normalization methods clearly outperformed the others.

*Thresholding: Fixed Length Approach*

For z-score normalized measurements, sensitivity, specificity, and the optimal threshold value increase as the observation length increases (Fig. 2.8). For autoregressive normalized measurements, there existed a window of observation spans for which sensitivity, specificity, and optimal threshold values peaked, after which they decreased again (Fig. 2.9). This can likely be explained by the model forecasts eventually “catching up” to the observed values, i.e. the model mistook the abnormally increasing biomarker measurements as an increasing healthy baseline.

*Thresholding: Dynamic Approach*

Selecting an optimal threshold requires optimizing the combination of early detection, sensitivity, and specificity. Smaller thresholds yielded earlier detection time, higher sensitivity, but lower specificity. Larger thresholds yielded higher specificity, but delayed detection time and lower sensitivity. Z-score normalization yielded the best thresholding results, as there existed a range of thresholds for which both sensitivity and specificity were at nearly 100%, and the smallest of these thresholds yielded the earliest detection time (Fig 3.2, A). For all other normalization methods, the sensitivity and specificity curves intersected at a single point, indicating poorer performance and a very limited range of thresholds that could be used to optimally classify patients.

*k-NN versus Thresholding*

With the dynamic approach to *k*-NN, each normalization method had a certain time of detection (fig 3.1, top). For each normalization method, we compared the sensitivity and specificity of classification using 80% confidence *k*-NN to the sensitivity and specificity of classification using thresholding, where the threshold was chosen to give the closest possible detection time to the 80% *k-*NN approach. Across all normalization methods, *k*-NN yielded a sensitivity of 0.11 ± 0.05 higher than thresholding and a specificity of 0.14 ± 0.07 higher than thresholding. Hence, for the simulated set of patients, *k*-NN was the better performing classifier.

**DISCUSSION**

This study developed a framework for making personalized cancer diagnoses tailored to patient-specific baselines by applying supervised learning methods used for anomaly detection [8]. We demonstrated that classification accuracy is dependent on the properties of the biomarker measurements, including noise, frequency of sampling, and observation span. Using the simulated patients, if we desire a given level of accuracy and know any two of the three parameters, we are able to determine the value of the fourth parameter. For example, if we know the expected amount of assay error and want to achieve a certain classification accuracy within a certain detection time, we can calculate how often patients should visit the clinic for blood samples. This could potentially allow for a more efficient allocation of patient and clinician time and hospital resources.

By classifying patients with *k*-NN on normalized measurements, we have combined examinations of patient-specific baselines with examinations of how the trends in the patient’s longitudinal biomarker levels relate to population biomarker trends. (The latter condition is absent in classification by thresholding, which only considers single biomarker measurements when assigning patient status.) Baseline determination allows us to make more personalized diagnoses, and classifying through pattern recognition allows us to better harness the information stored in a patient’s longitudinal biomarker history, leading to earlier detection. Cell culture, mouse model, and patient studies are currently being planned to collect longitudinal biomarker data, and we will then be able to translate the framework developed through this study to real data.

**REFERENCES**

1. Kulasingam V and Diamandis E. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. Nature Clinical Practice Oncology. 2008. 5(10): 588 – 599.

2. Hori SS, Lutz AM, Paulmurugan R, and Gambhir SS. A Model-Based Personalized Cancer Screening Strategy for Detecting Early-Stage Tumors Using Blood-Borne Biomarkers. *Cancer Res*. 2017. 77(10): p. 2570-2584.

3. Hori SS and Gambhir SS. Mathematical Model Identifies Blood Biomarker-Based Early Cancer Detection Strategies and Limitations. *Sci Transl Med*. 2011. 3(109): p. 109ra116.

4. Lin H, McCulloch CE, Turnbull BW, Slate EH, and Clark LC. A Latent Class Mixed Model for Analyzing Biomarker Trajectories with Irregularly Scheduled Observations. Stat Med. 2000. 19(10): 1303-18.

5. Sölétormos G et.al. Biological Variation of Total Prostate-Specific Antigen: A Survey of Published Estimates and Consequences for Clinical Practice. Clinical Chemistry. 2005. 51(8): 1342-1351.

6. Adhikari R, Agrawal RK. An Introductory Study on Time Series Modeling and Forecasting. Lambert Academic Publishing. 2013.

7. Keller JM, Gray M, Givens JA. A Fuzzy *K-*Nearest Neighbor Algorithm. IEEE Transactions on Systems, Man, and Cybernetics. 1985. 15(4): 580-585.

8. Monnet A, Mittal A, Paragios N, Ramesh V. Background Modeling and Subtraction of Dynamic Scenes. Proceedings Ninth IEEE International Conference on Computer Vision. 2003. 2:1305-1312.

[KS1]What factors?

[SH2]Other things to consider:

1. How many healthy patients and cancer patients need to be simulated? We should also increase the population variance to enable larger variation.

2. What happens if you increase *n*h, *n*c, and *d*?

[SH3]What is the assumed doubling time for patients with cancer?

[KS4]1000 days? Yes

[KS5]Motivations? H\_0 ~ G\_0? Adjustable variance, how to determine a single variance?

[SH6]Need reference indicating how these ranges were chosen.

Can also increase variability by making these ranges wider, e.g., what happens if you increase upper limits by up to 1000 fold?

[KS7]Only takes into account frequency WRT days? What about times of day, any other possible factors?

[SH8]n was defined to be the number of patients… use a different variable here (e.g., i).

[KS9]maybe rephrase so emphasizes that it decreases the time from 1000 days to < 1000 days? Sounds like you’re removing the most recent measurements for some reason

[KS10]what exactly does this mean

[KS11]~ normal score

[KS12]arbitrary?

[KS13]Finite distributed lag model

[KS14]Gives covariance of function with itself at pairs of time points

[KS15]In the case of multiple random variables, autocovariance can be expressed as an nxn matrix with entries C\_ij(ts), called an autocovariance matrix, associated with vectors Xt and Xs

[KS16]= E[y\_t y\_t-k] – mu\_y mu\_k

[KS17]At each t?

[KS18]Similarity between observations as a function of lag between them – correlation between the values of ONE variable @ different time points

Autocorrelation between Xs and Xt: Cov(Xs, Xt) / var(Xs)var(xt)

[KS19]How??

[KS20]Unconditional joint distribution does not change when shifted in time – when would it be?

Autocorrelation becomes Cov(Xs, Xt) / var^2

[KS21]What are these values?

[KS22]Arbitrary?

[KS23]How is this produced?

[KS24]Why using k NN??

[KS25]Don’t need the denominator

[KS26]So we *do* have a set of existing patients with known cancer statuses?

[KS27]Testing effect of perturbing noise, sampling frequency, observation span

[KS28]Why? What about day 500?

[KS29]And noise distribution?

[KS30]why are we testing normalization methods at the same time as observation span

[KS31]no distributed lag?

[KS32]abs or positive difference?