Probability of False Prostate Cancer Diagnosis Using a Logistic Function Given Age and f/t PSA Ratio

BU-1523-M

Aldo Crossa
Wittenberg University

Johnny Guzmán
California State University, Long Beach

Andie Hodge
University of the Virgin Islands

Brisa N. Sánchez
University of Texas, El Paso

August 11, 1999

Abstract

Prostate specific antigen (PSA) is a protein of free and complex forms found in the bloodstream of patients suffering from prostatic diseases. The free/total PSA ratio is often used in the detection of prostate cancer. We define a logistic function that assigns the probablity of having cancer given a specific value of the f/t PSA ratio and the age of the patient. Since the levels of PSA in a serum sample are affected by storage conditions over time, we construct a model to show the effect of improper storage of serum samples on the f/t PSA readings. We then use the logistic function and the model of sample decay over time to determine the probability of a false positive given the storage conditions. In addition we provide an analysis of cut-off values, and a function that predicts the time necessary for the diagnosis of a patient to change from negative to positive given his age, f/t PSA Ratio, and a cut-off value.

1 Introduction

Prostate cancer is the second most common death from cancer in men and the most common death for men over seventy-five years old. This disease is very rare in men under the age of 50. In 1995, 244,000 new cases of prostate cancer were reported[4]. The likelihood that an individual will eventually be diagnosed with cancer depends on variables such as race (African-American men have the highest rate of prostate cancer in the world), diet (high fat diets), family's medical history and the exposure to cadmium in the work place, amongst other possibilities[8].

The prostate is a small gland located below the bladder and surrounds the urethra (canal that carries urine from the bladder). The prostate is a male sex gland that produces fluid for semen, which leaves the body through the penis during the male orgasm (i.e., ejaculation). As a patient becomes older, the prostate gland becomes very susceptible to diseases that can be dangerous to a patient's life, for example, prostate cancer.

The growth and function of the prostate gland depends on the male hormone testosterone, produced mainly in the testicles. Testosterone stimulates the development of prostate cancer in the same way kerosene fuels a fire[7]. Hence, as the body produces testosterone, prostate cancer will continue to spread. This disease involves an uncontrolled cell division in the peripheral area of the prostate gland. Once a patient is diagnosed with this form of cancer the tumor is categorized into stages A through D depending on its development. Stage A tumors are early tumors and usually not detected. In Stage B, the tumor is still in the prostate and may be large enough to be detected. In Stage C, the cancer is more advanced. It indicates that the tumor has spread outside the prostate to surrounding areas, but it has not spread to other organs. In Stage D, tumors have spread to the lymph nodes and other nearby organs. Tumors recognized at this stage are in most cases lethal to the patient[7]

Prostate cancer can be detected by a prostate-specific antigen (PSA) test. PSA is a glycoprotein 34,000 Daltons in size that is secreted into the seminal fluid by the epithelial (surface) cells of the prostate gland. The glandular ducts prevent PSA from entering the circulatory system in large concentrations. Initially, PSA is released in its free state but will slowly form stable complexes with α_1-antichymotrypsin (ACT) and α_2-macroglobulin (α2M). In serum samples, the most common forms of PSA are in free and PSA-ACT complex forms [6].

A clinical dilemma of prostate cancer diagnosis requires differentation of prostate glandular clinical symptoms caused by enlargement from Bengin Prostate Hyperplasia (BPH) or various other forms of prostate inflammation and prostatic diseases. This differentiation can be achieved by looking at the ratio of free to total PSA (f/t PSA), where total PSA refers to the protein in all its possible forms (free and complex). Although this ratio is used as the most reliable detector of prostate cancer, there is still a probability that a patient is diagnosed false positive (diagnosed with cancer but in reality doesn't suffer from it). Patients can fall into three categories: those that clearly do not suffer from cancer, those that definitely suffer from cancer, and those that belong in a"gray area", where a solid diagnosis could not be stated. Clearly, the idea of current research is to narrow down this gray area and make the diagnosis as accurate as possible.

The f/t PSA ratio test consists of measuring PSA concentration in a patient's serum. The process for serum extraction involves two steps. The blood is drawn from the patient and allowed to clot. Then the clot is centrifuged to separate the serum. After this, the serum is stored until the time for analysis. The problem arises when the samples are not handled at optimal temperature conditions. If samples are not handled properly then the ratio of f/t PSA can be altered significantly because molecular interactions of PSA with other proteins (such as ACT) are affected. This change of the original f/t ratio, if large enough, can in fact change the doctor's diagnosis of a patient. The result of this is that patients are sometimes subjected to treatment such as surgery, when it is not necessary. Some of the secondary effects of surgery are impotence and incontinence. Also, there are patients that will be diagnosed as healthy when in reality they have the cancer. Clearly this second case is far more critical since it can be lethal to the person.

There are two main objectives in this paper. First we define a logistic function that assigns the probability of having cancer given the f/t PSA ratio and the age of the patient. For this we use data from 4870 patients diagnosed either as "no evidence of malignancy" (NEM) or "prostate cancer" (CaP).[1] All the patients from the sample were in the age range of 45 to 90 years old and had a total PSA concentration from 2-20μg/ml.

Second, we investigate the effect of improper handling of the serum samples on the original PSA levels. We model the change of the f/t ratio as a

[1]Provided by Dr. Robert Veltri and Dr. Craig Miller from UroSciences Group, UroCor, Inc., Oklahoma City, Oklahoma.

function of time, assuming that the conditions are not ideal. We combine this model with the logistic regression to provide a probability function that includes storage times and conditions. Thus, if we know at what conditions the sample was stored and the time it was stored, then we can determine the original f/t ratio and therefore determine the actual probability that the patient does, in fact, suffer from cancer. In addition we provide an analysis of cut-off values and a function that predicts the time needed to change a diagnosis from negative to positive.

2 Methodology

2.1 Data Description

Data from 4870 patients was analyzed using Binary Logistic Regression[2]. This data set was obtained from patients who:

- were between the ages of 45-90 years (mean age of 66 and a standard deviation of 8)
- were diagnosed with prostate cancer (CaP) or with no evidence of malignancy (NEM)
- had no previous history of Prostate Cancer
- total PSA concentration levels were between 2 and 20 μg/ml

The data set contains the age, free PSA and total PSA concentrations, f/t PSA ratio, and the real diagnosis of the patient after biopsy for each patient. All the serum samples were frozen at $-80°C$; therefore we assume that the f/t PSA ratio values given are the true values.

Data on decay rates was obtained from Woodrum [6], and Piironen [3]. These articles gave approximate values for the decay rates of free PSA during clotting and approximations for the rate of change of free PSA during storage at 4°C and 25°C. We use Woodrum's estimates for free PSA decay rate during clotting time, and during serum storage at 25°C; and we used Piironen's estimate for free PSA decay during serum storage at 4°C.

[2]A sample of this data set is found in appendix A

2.2 Fitting a Logistic Model

The data was analyzed using Binary Logistic Regression performed with the statistical software package MiniTab^{TM}. As the binary response variable we used 1 for NEM and 0 for CaP. The variables for the model were f/t PSA Ratio and age. Combining age and f/t PSA Ratio gave the highest odds ratio (716:1) for f/t PSA, as compared to the odds ratio (216:1) when only f/t PSA ratio was used as the variable for the model. The MiniTab^{TM} output estimated the values to be used in the logistic probability function[3]. The p-values for the significance of the coefficients were always less than 0.001.

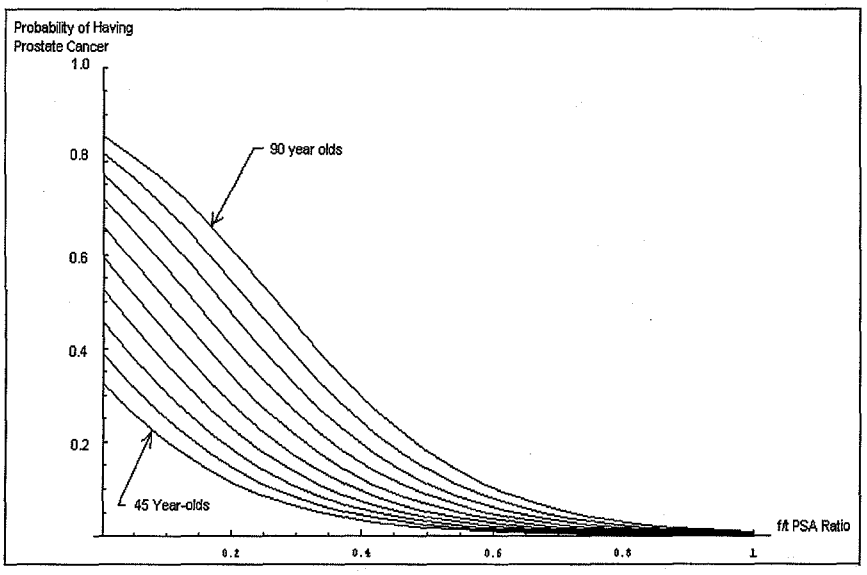

Figure 1: *Probability of Prostate Cancer as a function of f/t PSA Ratio. The ages are held constant for each curve in the graph. Ages are in increments of 5 yrs.*

The resultant probability function has the form,

$$P_{cancer}((f/tPSA), AGE) = \frac{1}{1 + e^{\alpha+\beta(f/tPSA)+\gamma AGE}} \tag{1}$$

where the Logistic fitted coefficients are,

$$\begin{aligned} \alpha &= 3.2210 \\ \beta &= 6.5743 \\ \gamma &= -0.055426 \end{aligned}$$

[3]For samples of these outputs look at appendix B

Therefore, by plugging in the age of the patient, and his f/t ratio at the time of the analysis into equation (1) we get the probability that a patient has prostate cancer. To see estimates of these probabilities in graph form see figure 1.

2.3 Model for f/t PSA Decay

In this section we define the function that predicts the f/t PSA ratio at a given time. We propose an exponential model for the decay of the f/t PSA ratio since the previously mentioned literature gives rates of change of free PSA in percentages. However, since there are two different storage conditions, there are two rates of decay,

λ_1 = Rate of free/total PSA decay during queuing time before centrifugation, depends on temperature T

λ_2 = Rate of free/total PSA decay during serum storage at a given temperature T

Therefore, we consider a piecewise function for how f/t PSA decays. Figure 2 gives a general idea of this function.

Thus,

$$x_c(\tau) = \begin{cases} x_c(0)e^{-\lambda_1 \tau} & \tau \leq c \\ x_c(0)e^{-\lambda_1 c - \lambda_2(\tau - c)} & \tau \geq c \end{cases} \tag{2}$$

represents the f/t PSA ratio at a given time[4] τ; where τ is the time after the blood draw, and c is the queuing time before centrifugation. We assume that the amount of total PSA (complexed and uncomplexed) is constant.

In a realistic setting τ_f, the time between blood draw and the time of analysis, will be greater than c because the analysis is done to the serum, which is obtained after the blood is centrifuged. Since $\tau_f \geq c$, then we use

$$x_c(\tau_f) = x_c(0)e^{-(\lambda_1 c + \lambda_2(\tau_f - c))} \tag{3}$$

to solve for $x_c(0)$, which is the true f/t PSA ratio at the time of blood draw from the patient. This yields,

$$x_c(0) = x_c(\tau_f)e^{\lambda_1 c + \lambda_2(\tau - c)} \tag{4}$$

[4]time is measured in hours

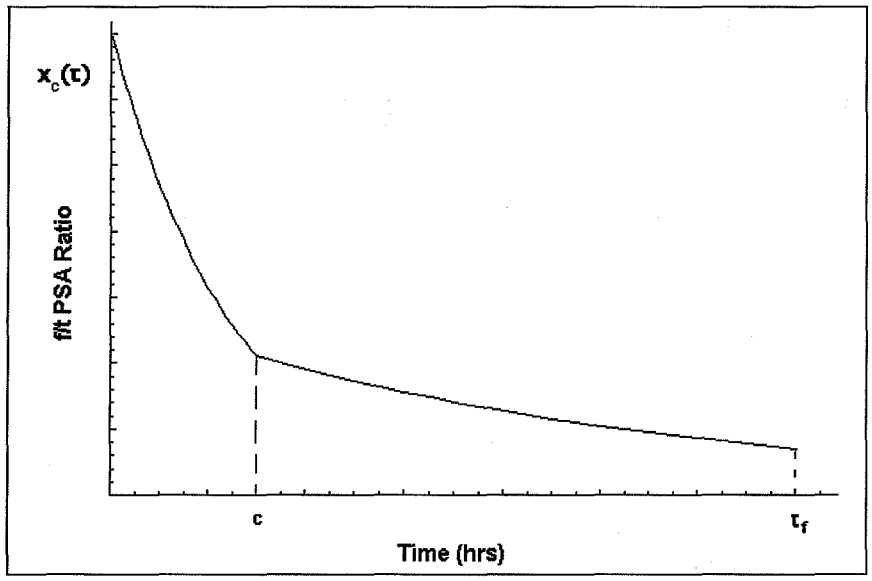

Figure 2: *This figure illustrates how f/t PSA changes before time of centrifugation c, and during the storage time ($\tau_f - c$) before the sample is analyzed.*

2.4 Time Adjusted Logistic Model

In this section we adjust the logistic model so that it considers the decay of f/t PSA Ratio. In order to find the probability that a patient has prostate cancer, given his age, $x_c(\tau_f)$, τ_f, and storage temperature, we need to combine equations (1) and (4). We replace $(f/tPSA)$ in equation (1) with $x_c(0)$ from equation (4). This yields,

$$\begin{aligned} P_{cancer}(x_c(\tau_f), AGE, c, \tau_f) &= \frac{1}{1+e^{\alpha+\beta x_c(0)+\gamma AGE}} \\ &= \frac{1}{1+e^{\alpha+\beta(x_c(\tau_f)e^{\lambda_1 c+\lambda_2(\tau_f-c)})+\gamma AGE}} \end{aligned} \quad (5)$$

Now, we substitute the fitted values α, β, γ and the decay rates from Woodrum [6] and Piioronen [3],

$$\begin{aligned} \lambda_1 &= 0.01 \text{ at } 25^oC \\ \lambda_2 &= 0.0017 \text{ at } 4^oC \\ \lambda_2 &= 0.0005 \text{ at } 25^oC \end{aligned}$$

into equation (5) to attain the cancer probability function,

$$P_{cancer}(x_c(\tau_f), AGE, c, \tau_f) = \frac{1}{1 + e^{3.2210+6.5743(x_c(\tau_f)e^{0.01c+\lambda_2(\tau_f - c)})-0.055426AGE}} \tag{6}$$

where λ_2 depends on the temperature at which the serum samples were stored. Thus, knowing $x_c(\tau_f)$, (τ_f), c and AGE we can calculate the probability of getting prostate cancer.

To illustrate how this probability function is used we give the following example:

Suppose that a practitioner takes a sample from a 61 year old patient at 8:00 am Monday morning, and at 4:00 pm the sample is centrifuged. Suppose that two days later at 5:00 pm the sample was analyzed, and the resulting f/t ratio is 0.3. Then,

$$\begin{aligned} c &= 8\text{hrs.} \\ \tau_f &= 57\text{hrs.} \\ x_8(57) &= 0.3 \end{aligned}$$

Therefore,

$$x_8(0) = x_8(57)e^{-8\lambda_1 - \lambda_2(57-8)}$$

Since we have a recorded f/t ratio of 0.3 at τ_f , we can calculate the initial ratio depending on the conditions the sample was exposed to. Taking the rates of decay (λ_1 and λ_2) determined earlier in the paper then the original ratio should be,

$$x_8(0) = 0.2702 \text{ at storage temperature } 25^oC$$

We can then plug this expression for $x_c(0)$ into the time adjusted probability function to see that,

$$P_{cancer}(x_8(0), 61) = \frac{1}{1 + e^{\alpha+\beta(x_8(0))+\gamma(61)}} \tag{7}$$

is the probability of having cancer at age 61 with a final f/t PSA ratio of 0.3. Therefore,

$$P_{cancer}(x_8(0), 61) = 0.165 \text{ for } 25^oC$$

3 Analysis

3.1 Cut-Off Values

Using the first logistic regression model we calculated the probability that each patient in the data set had cancer. Then, we compared these probabilities to the real diagnosis, and calculated false positive (FP) and false negative (FN) rates given a cut-off value[5].

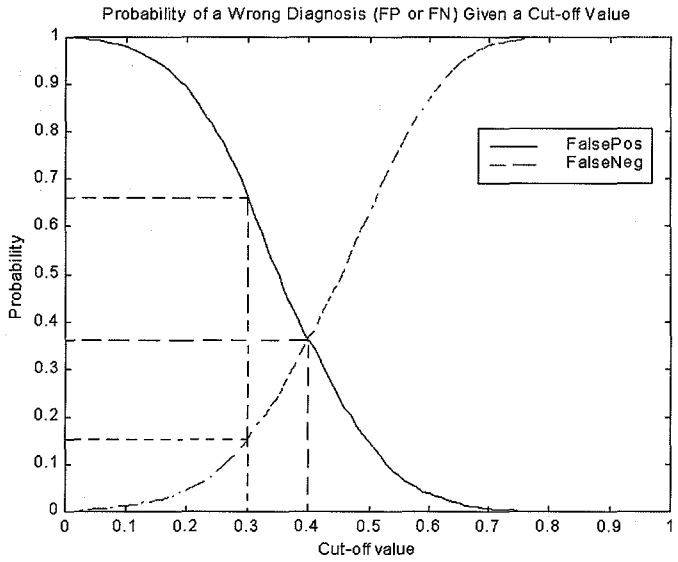

Figure 3: *Probability cut-off values and their corresponding probability of a false diagnosis. The graph shows the probability of diagnosing an FP or a FN when a certain cut-off value is chosen. The cut-off values on the graph were chosen arbitrarily as an example.*

Where,

$$\text{FP probability} = \frac{FP}{TN + FP}$$

and

$$\text{FN probability} = \frac{FN}{TP + FN}$$

[5]See Appendix C for a copy of the MatLab program used for this analysis

Figure 3 shows the results of this numeric comparison that may be useful to decide on an optimal cut-off value. Observe that intrinsically, we are giving the same weight to a FP and to a FN. In real life a FN is more harmful than a FP since the former implies that a person with cancer will not undergo treatment, whereas the latter implies that a person without cancer will undergo treatment. Clearly, the first case (FN) may result in death of a patient whereas FP implies unnecessary treatment. As a realistic example, if a cutoff value of 0.4 is chosen, then from figure 3 the probability of a false positive is the same as the probability of a false negative. Clearly, as we choose a smaller cut-off value, then the probability of FN goes to zero, since we would catch all the cancer cases. However, at the same time, the proportion of FP goes to one, since we would send to treatment all the patients that are cancer free.

Because it is useful to give more weight to a FN than to a FP, we define a K-value as the number of times a FN diagnosis is more important than an FP.

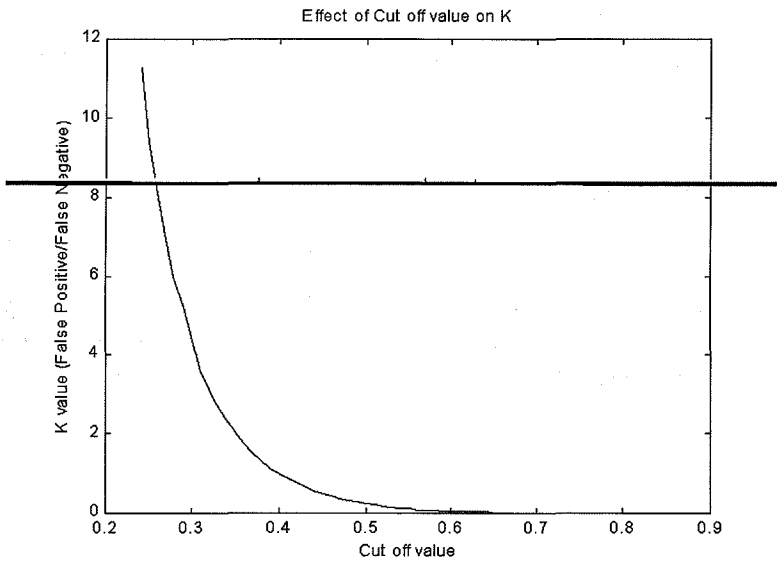

Figure 4: *The K-value represents the ratio of FN/FP probabilities.*

For a fixed K-value there is a cut-off value in which the probability of FP is K times larger than the probability of FN. In figure 4 we see how K decreases as the cut-off value increases. The goal of the cut-off analysis is

to minimize the probabilites of FP and FN diagnosis. K helps to find an appropriate probability cut-off value given the importance of making a FP wrong diagnosis relative to a FN diagnosis.

Once a probability cutoff value, a, has been chosen and since the function that determines the probability of having cancer given the f/t PSA ratio is decreasing (see figure 1), then we can define a f/t PSA ratio cut-off value a^* for a fixed age. Figure 5 is a graphic representation of this statement.

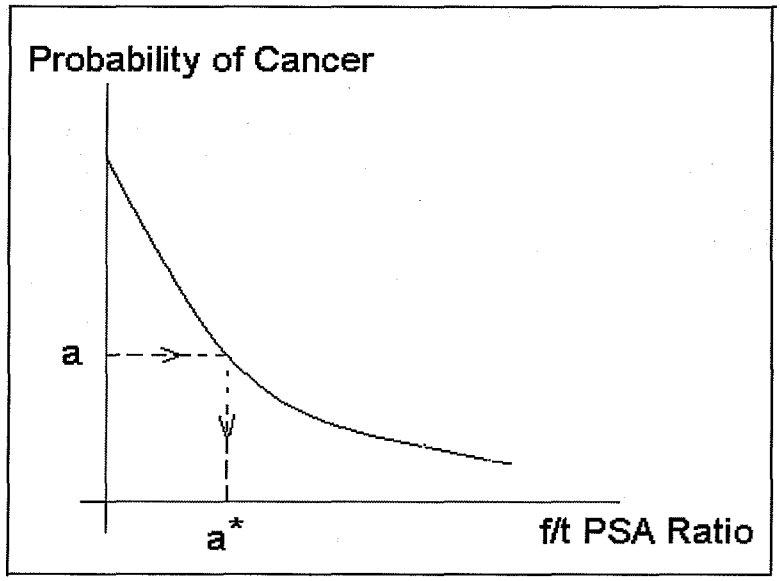

Figure 5: *For a given probability cut-off value, a, we can find the corresponding f/t PSA Ratio cut-off value a^* for a fixed age.*

To solve for the f/t PSA ratio cut off, we solve for a^*,

$$\begin{aligned} P(a^*, AGE) &= a \\ \frac{1}{1+e^{\alpha+\beta a^*+\gamma AGE}} &= a \\ a^* &= \frac{ln(\frac{1-a}{a}) - \alpha - \gamma AGE}{\beta} \end{aligned} \tag{8}$$

Since P_{cancer} is decreasing as $x_c(0)$ increases then,

if $0 \leq x_c(0) \leq a^*$ then $P_{cancer}(x_c(0)) > P(a^*) = a$ (positive diagnosis) and,

if $a^* < x_c(0) \leq 1$ then $P_{cancer}(x_c(0)) < P(a^*) = a$ (negative diagnosis)

3.2 Effect of Storage

Since λ_1 and λ_2 are small, then for small enough τ_f, $x_c(0) \approx x_c(\tau_f)$. Thus we might be tempted to replace $x_c(0)$ with $x_c(\tau_f)$ in the probability function. However, the problem is that this might change the diagnosis. In this section we study at what time after the blood draw is a change in diagnosis possible. In order to do this, we develop a function that predicts the time necessary for a f/t PSA ratio reading to decay to a level where the patient will be diagnosed with cancer, when his original f/t PSA ratio $x_c(0)$ would diagnose no cancer.

Now, if $x_c(0) > a^*$ then $P_{cancer}(x(0)) < a$. Thus the patient will have a negative diagnosis. However, since $x_c(\tau)$ decreases, there exists τ_f large enough such that $x_c(\tau_f) \leq a^*$, which means that if we replace $x_c(0)$ with $x_c(\tau_f)$ then the diagnosis will change from negative to positive.

We find a maximum time τ_m such that if the time τ_f at which the sample was analyzed is greater than τ_m, then the f/t PSA ratio reading $x_c(\tau_f)$ will change enough to alter the diagnosis. (i.e., if $\tau_f < \tau_m$ then the diagnosis will remain negative, otherwise the the results will give an incorrect positive diagnosis).

Recall that

$$x_c(\tau) = \begin{cases} x_c(0)e^{-\lambda_1 \tau} & \tau \leq c \\ x_c(0)e^{-\lambda_1 c - \lambda_2(\tau - c)} & \tau \geq c \end{cases} \tag{9}$$

and notice that τ_m occurs when $x_c(\tau_m) = a^*$ given $x_c(0) > a^*$.

Since $x_c(\tau)$ is a piecewise function, then depending on $x_c(0)$, either $0 \leq \tau_m \leq c$, or $c < \tau_m$, which tells us what part of the function $x_c(\tau)$ to use to solve for τ_m. The borderline case occurs when $x_c(0) = a^* e^{\lambda_1 c}$. Then $x_c(c) = a^* \Rightarrow \tau_m = c$.

This implies that if,

$$a^* < x_c(0) \leq a^* e^{\lambda_1 c} \text{ then we have} \tau_f \leq c$$

whereas if,

$$a^* e^{\lambda_1 c} < x_c(0) \leq 1 \text{ then we have } \tau_f \geq c.$$

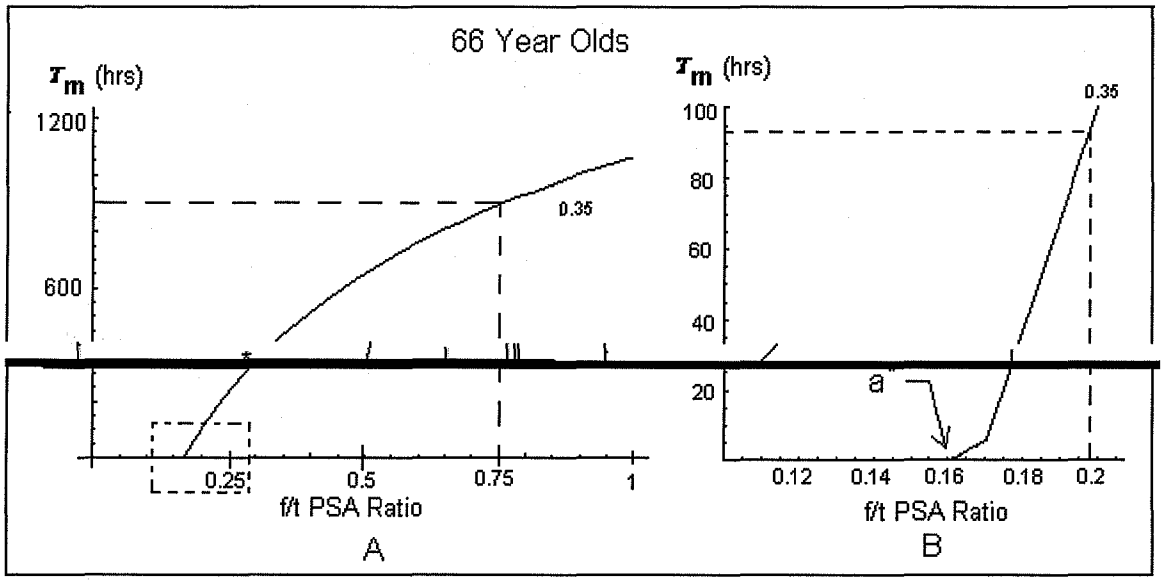

Figure 6: *(A)The graph shows the time it would take a true f/t PSA ratio, from a 66 year old patient, stored at 4° C to decay to the f/t PSA cutoff value. (B) A close up of the area enclosed around the cut-off values.*

Then by using equation (9) and the previous inequalities we have

$$\tau_m(x_c(0)) = \begin{cases} \frac{ln\frac{a^*}{x_c(0)}}{-\lambda_1}, & a^* < x_c(0) \leq a^* e^{\lambda_1 c} \\ \\ \frac{ln\frac{a^*}{x_c(0)} + (\lambda_1 - \lambda_2)c}{\lambda_2}, & a^* e^{\lambda_1 c} < x_c(0) \leq 1. \end{cases} \tag{10}$$

To illustrate the use of this function, we take a probability cut-off value= 0.35, and AGE= 66, and $c = 6$. Then, from equation (8) we know that the f/t PSA cut-off, $a^* = 0.16$. Substituting

$$\begin{aligned} \lambda_1 &= 0.01 \\ \lambda_2 &= 0.0017 \end{aligned}$$

into equation (10) we are able to graph τ_m as a function of $x_c(0)$ to get a graphic representation of the time needed for any $x_c(0)$ (true f/t PSA ratio) of a 66 year old patient to decay to the f/t PSA cut-off value a^*.

Notice that this function increases rapidly. Therefore, if $x_c(0)$ is much larger than the f/t PSA cut-off value, then the time taken for the f/t PSA

ratio to decay will be very large. For example, if the true f/t PSA ratio of the patient is 0.75 then we see from figure 6 that the time needed for the f/t PSA ratio to decay to a^* is approximately 900 hours (or 38 days) after the blood was drawn from the patient.

However, if $x_c(0)$ is close to the f/t PSA cut-off value, then the time for the f/t PSA ratio to decay to a^* is relatively small. For example, take $x_c(0) = 0.2$, then the time needed the f/t PSA ratio to decay to $a^* = 0.16$ is approximately 96 hours (or 4 days). Thus, if the sample was analyzed four days after the blood draw, and $x_c(\tau_f)$ was substituted into the probabability function, instead of $x_c(0)$, the diagnosis will be positive which will be an incorrect diagnosis in this case.

4 Conclusion

In previous literature, cancer diagnosis depended merely on the f/t PSA ratio of a patient. However, in this paper we concentrated on adding age of the patient as an extra parameter. This turns out to be a crucial factor in determining the probability of the patient of having cancer since two patients with the same f/t PSA ratio can have different probabilities of having cancer because of their ages.

Based on these findings, we suggest that practitioners take into account the age of the patient when diagnosing, because the patients will become more susceptible to this type of cancer with time.

Using our time adjusted function with certain parameters we can conclude that the f/t PSA ratio is not affected with time as significantly as we expected. Analysis of critical storage times showed that the diagnosis of a patient can be changed if: 1) the sample is left out room at temperature for unrealistically extensive periods of time (eg. 24 days) or 2) if a given f/t ratio is in the neighbourhood of the cut-off value, then there is a possibility of change in the diagnosis within realistic periods of time (1 or 2 days of storage only). In such a case we would recommend a second and more careful analysis on the individual's serum sample.

The rates λ_1 and λ_2 are critical in determining the time adjusted probability of cancer in a patient. The literature where these values were obtained, had a small sample size and unusual results[6]. This suggests that the rates calculated in these articles may be inaccurate, which will directly affect the

[6]At 4°C the decay rate was greater than at 25°C

accuracy of our model. We encourage future studies to use a more significant sample size to find the rates of f/t ratio that correspond to different storage conditions.

Subsequent papers studying related areas should concentrate on determining other parameters affecting the f/t ratio. Testosterone levels and pH are some other causes for f/t ratio change and could be explored in the future.

Finally, it is important to understand that this paper essentially proposes a methodology to determine the actual f/t ratio of a patient and from it determine the real probability of the patient of suffering from cancer. We suggest that better parameter estimates should be included using the analytical process we followed.

Acknowledgements

This study was supported by the following institutions and grants: National Science Foundation (NSF Grant DMS 9977919); National Security Agency (NSA Grants MDA 9049710074); Presidential Faculty Fellowship Award (NSF Grant DEB 925370) and Presidential Mentoring Award (NSF Grant HRD 9724850) to Carlos Castillo-Chávez; and the office of the provost of Cornell University; Intel Technology for Education 2000 Equipment Grant. Special thanks to Ted Greenwood of the Sloan Foundation; to Carlos Castillo-Chávez for making MTBI possible; to Carlos Moises Hernandez for his guidance throughout the project, Dr. Robert W. Veltri and Craig M. Miller for the extraordinary data they provided us with, and all the MTBI staff and students that motivated us to through this project.

References

[1] Hosmer Jr. D.W., Lemeshow S.: Applied Logistic Regression. John Wiley and Sons Inc., New York, 1989.

[2] Lilja H.: Significance of Different Molecular Forms of Serum PSA. The Free, Noncomplexed Forms of PSA Versus that Complexed to α_1-Antichymotrypsin. Urologic Clinics of North America 20 (4): 681-686, November 1993.

[3] Piironen T., Pettersson K.,Suonpää M., Stenman U., Oesterling J.E., Löuvgren T., Lilja H.: *In Vitro* Stability of Free Prostate-Specific Antigen (PSA) Complexed to α_1-Antichymotrypsin in Blood Samples Urology 48 (6A):81-87, 1996.

[4] U.S. Congress, Office of Technology Assessment, *Costs and Effectiveness of Prostate cancer Screening in elderly Men*, OTA-BP-H-145 (Washington, D.C.: U.S. Government Printing Office, May 1995.

[5] Veltri W.V., Miller M.C.: Free/Total PSA Ratio Improves Differentiation of Benign and Malignant Disease of the Prostate: Critical Analysis of Two Different Test Populations. Urology 53 (4): 736-745, 1999.

[6] Woodrum D., French C., Shammel L. B., Stability of Free Prostate-Specific Antigen in Serum Samples Under a Variety of Sample Collection and Sample Storage Conditions Urology 48 (6A):33-39, 1996.

[7] Tap Holdings Inc. Website: www.prostate.com/Html/diagn.htm

[8] UroSciences Corporation, Oklahoma City, Oklahoma. Website: www.uro.com/caprostate.htm

Appendix A: Data Set

ID	Age	FPSA	PSAT	FrTot	Diagnosis	Binary
UC-17	45	0.22	9.0	0.024	NEM	1
UC-74	45	0.18	4.4	0.041	NEM	1
UC-114	45	0.41	9.0	0.046	NEM	1
UC-985	45	0.52	5.0	0.104	NEM	1
UC-1219	45	0.44	3.8	0.116	NEM	1
UC-1327	45	0.45	3.7	0.122	NEM	1
UC-1859	45	0.48	3.0	0.16	NEM	1
UC-2737	45	0.66	2.3	0.287	NEM	1
UC-3061	45	0.17	4.0	0.043	Cancer	0
UC-3125	45	0.45	9.3	0.048	Cancer	0
UC-3230	45	0.30	5.2	0.058	Cancer	0
UC-29	46	0.10	3.6	0.028	NEM	1
UC-36	46	0.27	9.3	0.029	NEM	1
UC-4558	67	0.44	2.5	0.176	Cancer	0
UC-4564	67	1.10	6.2	0.177	Cancer	0
UC-4575	67	1.00	5.6	0.179	Cancer	0
UC-4595	67	1.20	6.5	0.185	Cancer	0
UC-4654	67	2.80	14.0	0.2	Cancer	0
UC-4719	67	1.00	4.4	0.227	Cancer	0
UC-4748	67	2.20	9.1	0.242	Cancer	0
UC-4767	67	1.10	4.3	0.256	Cancer	0
UC-4782	67	3.20	12.0	0.267	Cancer	0
UC-4825	67	1.70	5.1	0.333	Cancer	0
UC-3	68	0.00	7.0	0	NEM	1
UC-41	68	0.54	17.0	0.032	NEM	1
UC-51	68	0.26	7.2	0.036	NEM	1
UC-91	68	0.26	6.0	0.043	NEM	1
UC-92	68	0.26	6.0	0.043	NEM	1
UC-115	68	0.16	3.5	0.046	NEM	1
UC-117	68	0.52	11.0	0.047	NEM	1
UC-167	68	0.25	4.7	0.053	NEM	1
UC-193	68	0.83	15.0	0.055	NEM	1
UC-206	68	0.25	4.4	0.057	NEM	1
UC-2546	83	1.90	7.9	0.241	NEM	1
UC-2895	83	2.40	5.8	0.414	NEM	1
UC-3018	83	0.57	16.0	0.036	Cancer	0
UC-3038	83	0.74	19.0	0.039	Cancer	0
UC-3040	83	0.21	5.3	0.04	Cancer	0
UC-3269	83	0.61	10.0	0.061	Cancer	0
UC-3618	83	1.20	15.0	0.08	Cancer	0
UC-3656	83	0.62	7.6	0.082	Cancer	0
UC-3754	83	0.50	5.6	0.089	Cancer	0
UC-3864	83	0.60	6.4	0.094	Cancer	0
UC-3919	83	0.98	10.0	0.098	Cancer	0
UC-4129	83	0.44	3.9	0.113	Cancer	0
UC-4182	83	2.00	17.0	0.118	Cancer	0
UC-4193	83	0.83	6.9	0.12	Cancer	0
UC-4422	83	2.10	14.0	0.15	Cancer	0
UC-4502	83	0.84	5.1	0.165	Cancer	0
UC-4545	83	0.78	4.5	0.173	Cancer	0
UC-4744	83	2.10	8.8	0.239	Cancer	0
UC-4763	83	1.70	6.7	0.254	Cancer	0
UC-4849	83	3.00	7.5	0.4	Cancer	0
UC-10	84	0.14	6.7	0.021	NEM	1
UC-40	84	0.11	3.6	0.031	NEM	1

Appendix B: MiniTab^{TM} Output for Logistic Regression

Binary Logistic Regression

```
Link Function:  Logit

Response Information

Variable  Value        Count
Status    1             2961
          0             1909
          Total         4870

     Coef        StDev          Z      P     Ratio     Lower       Predictor
                                                 Odds          95% CI
Predictor         Coef       StDev          Z      P     Ratio     Lower
Upper
Constant        3.2210      0.2534     12.71 0.000
FrTot           6.5743      0.4352     15.11 0.000    716.44    305.30
1681.23
Age          -0.055426    0.003881    -14.28 0.000      0.95      0.94
0.95
All slopes are zero: G = 423.097, DF = 2, P-Value = 0.000     Test that al
Log-Likelihood = -3049.553
Test that all slopes are zero: G = 423.097, DF = 2, P-Value = 0.000

Goodness-of-Fit Tests

Method                  Chi-Square     DF      P
Pearson                   4002.231   3508  0.000
Deviance                  4362.422   3508  0.000
Hosmer-Lemeshow             33.086      8  0.000

Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

                                        Group
Value       1      2      3      4      5      6      7      8      9     10
Total
1
  Obs     153    208    236    271    291    323    357    358    377    387
2961
  Exp   177.3  221.9  247.2  268.5  288.9  307.5  325.7  343.7  367.5  413.0
0
  Obs     334    279    251    216    197    164    130    129    110     99
1909
  Exp   309.7  265.1  239.8  218.5  199.1  179.5  161.3  143.3  119.5   73.0

  Total   487    487    487    487    488    487    487    487    487    486
4870

Measures of Association:
(Between the Response Variable and Predicted Probabilities)
```

Appendix C: MatLab^{TM} Program

```
function A = cutoff(data,mini,max,step)
%function A =cutoff(data,mini,max,step)
%
%Variables
%----------
%data :  Data set used, must be a n x 2 array of values.
% the first column of zeroes and ones, the second of calculated probabilities
%min :  The minimum cut-off value considered
%max :  The maximum cut-off value considered
%step :  The step size between cut-off values
%
%This function computes the False Positive Rate = FP/(number of patients
without cancer)
% = FP/(TN+FP)
%and the False Negative Rate = FN/(number of Patients with cancer)
% = FN/(FN+TP)
data=sortrows(data,1);
k=size(data,1);
totnoncancer=sum(data(:,1)==0);
totcancer=k-totnoncancer;
datanoncancer=data(1:totnoncancer,:);
datacancer=data(totnoncancer+1:k,:);

   x=mini:step:max;
s=length(x)

   for i=1:s
i
surY_cancN = sum(datanoncancer(:,2) >= x(i)); % these go to treatment
without needing it
surN_cancY = sum(datacancer(:,2) < x(i)); % these will not go to
treatment and they need it
FalsePos(i)=surY_cancN/totnoncancer;
FalseNeg(i)=surN_cancY/totcancer;
TotPropFailures(i)=(surY_cancN+surN_cancY)/k;
end;
```

```
   med=sum(FalsePos<FalseNeg)+20;
fp=FalsePos(1:med);
fn=FalseNeg(1:med);
K=fp./fn;
x2=x(1:med);
point=sum(K>10);
x2=x2(point:end);
K=K(point:end);
figure
plot(x2,K)
A=[x2 K];
title('Effect of Cut off value on K');
xlabel('Cut off value')
ylabel('K value (False Positive/False Negative)');
figure
plot(x,FalsePos,'b-',x,FalseNeg,'g--');
legend('FalsePos','FalseNeg');
title('Probability of a Wrong Diagnosis (FP or FN) Given a Cut-off
Value')
xlabel('Cut-off value')
ylabel('Probability')
```