EARLY DIAGNOSIS OF PROSTATE CANCER BY DYNAMIC MODELING AND DATA SCIENCE TOOLS

Oya Hoban, Eylül Kepcen, Özlem Şenel, Yaman Barlas

Industrial Engineering Department
Boğaziçi University
34342 Bebek Istanbul Turkey
+90 (212) 359 6407 / 08
oyahoban@gmail.com, eylulkepcen@gmail.com, senelozlem@hotmail.com, ybarlas@retired.boun.edu.tr

Numerous studies have developed alternative tools to detect prostate cancer in its early stages; however, their scope remains limited because of the strong assumptions they hold, resulting from the limitations in the medical literature. In our project, we study tissue-level dynamics of prostate, and we model the potential tumor presence and dynamics using two methodologies: system dynamics and machine learning (data science). Objective of the study is to come up with an improved diagnosis method supported by two models. We build the dynamic model using stock-flow modeling and simulation to observe the time-dependent dynamics in the prostate. Next, to fill the missing parts of data obtained from the literature, we make use of the dynamic model to produce synthetic data to be used as an input in the machine learning models. Using Python, we build nine different classification models and XGBoost Classifier performs the best among others with an accuracy value of 81.75 and recall value of 87.71. Both models are validated using available real-world data on prostate cancer. Combined outputs from two models provide added information on tumoral status and processes in a given individual. This study can be eventually useful to improve the medical screening procedures towards early diagnosis of prostate cancer.

Keywords: prostate cancer, prostate specific antigen, physiological modeling, dynamic simulation modeling, data science, machine learning

[Word Count: 6884]

1. INTRODUCTION

Prostate cancer is the second most common type of cancer among males worldwide. (Wang, et al., 2022) Prostate cancer stages can be classified according to how advanced and malignant the tumor is. In advanced stages, the survival rate is significantly lower, thus detection of prostate cancer in early stages can increase the probability of survival. (American Cancer Society, 2023) Therefore, early detection/diagnosis of prostate cancer have great importance for both medical doctors and patients.

After a certain age, males are suggested to go into screening for prostate cancer. Late diagnosis and overdiagnosis are two fundamental issues in prostate cancer screening. Late diagnosis may result in advanced and untreatable cancer while overdiagnosis can cause overtreatment of patients. Overtreatment is the case of unnecessarily treating and harming patients who carry clinically insignificant tumors which are expected to stay steady in the lifetime. In this study, we aim to detect the potentially aggressive tumor at the right time. (Loeb, et al., 2014) (NCCN Guidelines, 2022)

We plan to study tumoral characteristics of prostate with a novel approach that combines system dynamics with machine learning. Constructing dynamic models in tissue-level, we focus on the causal relationships between the biological structures of the human body. By modeling prostate characteristics of individuals, we aim to detect the potential tumoral cell growth in an early time.

In this regard, machine learning methods add an extra static information to our study. Adapting several models to detect the tumor presence, we plan to classify the cancerous patients using their prostate-related various characteristics as input. Aim of the current cancer screening procedure is to assist the medical doctors in detecting clinically significant prostate cancer. Objective of this study is coming up with a reliable predictive method supported by two models with the integration of their dynamic and static outputs and to improve performance of the current procedures. With this purpose, an individual level system dynamics model will be built to observe prostate dynamics over time. Simultaneously, multiple machine learning models will be built and tested to predict the cancer presence, and the model with the highest performance accuracy rate will be selected. The co-working of the dynamic and static models is planned to increase the diagnosis performance which may be a significant improvement compared to the current medical practice.

2. BACKGROUND KNOWLEDGE

Prostate cancer occurs from uncontrollable cell proliferation in the prostate gland. (Wang, Zhao, Spring, & DePinho, 2018) (Cannarella, Condorelli, Barbagallo, La Vignera, & Calogero, 2021) The casualties of cancer occurrence are still debatable, but studies show that there is positive correlation between certain characteristics of patients, which are called risk factors, and prostate cancer incidence/mortality. It can be argued that risk factors increase the likelihood of developing cancer and tumor growth rate. In prostate cancer cases, some of the significant risk factors are smoking, genetics, physical activity, sexual activity, increasing age and BMI. (Leitzmann & Rohrmann, 2012) (Pernar, Ebot, Wilson, & Mucci, 2018) Prostate specific antigen (PSA) is produced both in healthy and cancerous prostatic tissues and part of the produced PSA is secreted into blood serum. PSA can be found bound to a protein or free form. The free form of PSA is called free PSA. (Adhyam & Gupta, 2012) There is strong evidence that prostate enlarges with increasing age, and PSA production rate increases. (Zhang, et al., 2013) Cancer cells' walls are more prone to be disrupted, thus they are prone to secrete more PSA into the serum, which supports the claim that increasing PSA levels is one of the prostate cancer indicators. PSA, inside cells, goes into a process called 'proteolytic cleavage', in which PSA is transformed into mature forms, such as free PSA. Because cancer cells proliferate uncontrollably, proteolytic cleavage process completion rate decreases and free PSA portion of total PSA level decreases in people with prostate cancer. It is argued that free PSA to total

PSA ratio is another prostate cancer indicator, negatively correlated with tumoral volume. (Partin, et al., 2002) (Adhyam & Gupta, 2012)

The procedure to prostate cancer detection is that after a certain age, males get PSA level tests and digital rectal examination by medical professionals. Current routine screening procedure can be seen in Figure 1. PSA is produced in cancerous and healthy prostatic cells. If the PSA level of a person is above a certain level, it may be an indicator for prostatic tumor; therefore, he would go into follow-up screening. Follow-up screening comprises Multi-parametric Magnetic Resonance Imaging (mpMRI) and biopsy. The final summary output of mpMRI is called PI-RADS score. It is an indicator of how likely there is a tumor in the imaged area, ranging from 1 to 5. Biopsy output indicates if the sample tissue is cancerous or not, and it is seen as the final determinant of cancer. (NCCN Guidelines, 2022)

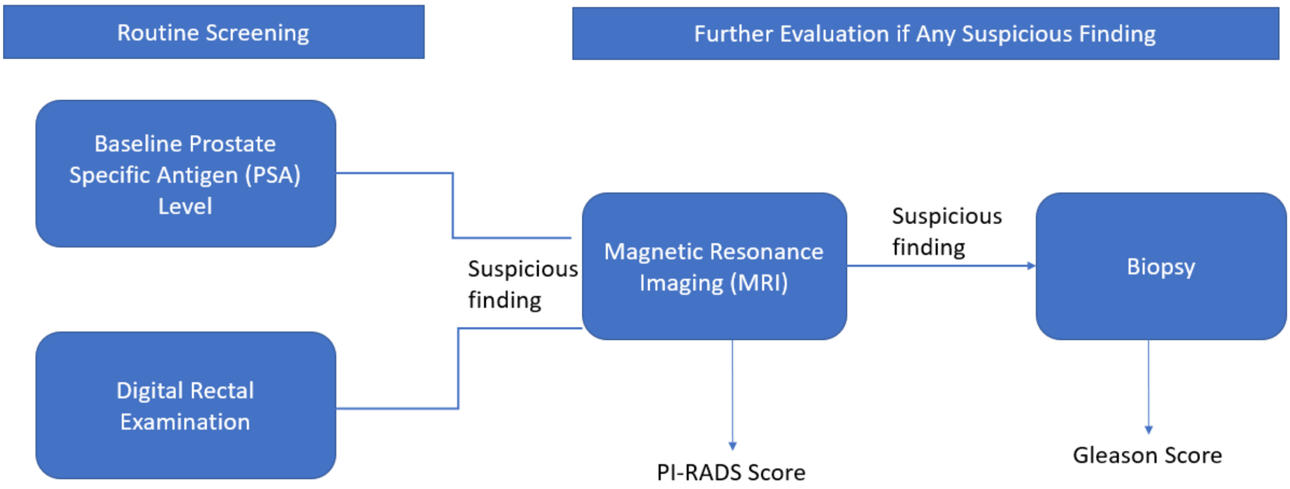

Routine Screening
Further Evaluation if Any Suspicious Finding
Baseline Prostate Specific Antigen (PSA) Level
Digital Rectal Examination
Suspicious finding
Magnetic Resonance Imaging (MRI)
Suspicious finding
Biopsy
PI-RADS Score
Gleason Score

***Figure 1:** Prostate Cancer Screening Procedure (NCCN Guidelines, 2022)*

Since mpMRI and biopsy are not part of routine screening, PSA test output is important for alarming the doctors. Healthy PSA level range can be different for each individual and might depend on multiple parameters: having prostate cancer, prostate enlargement due to increasing age, sexual activity of the person, different diseases such as benign prostatic hyperplasia and prostatitis. (Adhyam & Gupta, 2012) Erroneous results can occur if individualistic factors are not taken into consideration while reviewing PSA test results.

One of the most important factors in prostate cancer detection is the free PSA/Total PSA ratio. Studies show that the free PSA ratio of total PSA tends to be lower in prostate cancer patients. (Adhyam & Gupta, 2012) Detecting cancer before it has become aggressive and untreatable is important. Not incorporating different parameters into cancer detection can result in higher false negative rates, undetected and untreated cancer patients and eventually hazard to the population. (Lila, Ulmert, & Vickers, 2008)

Past mathematical and simulation modeling studies on prostate cancer detection differ in their methodologies. Some of them use PSA levels as an indicator of a tumor, while some focus directly on the potential tumor growth dynamics over years. Using different prostate related parameters, their aim is to predict the tumor presence. Since medical research on cancer epidemiology is still ongoing, certain assumptions were made when developing past models.

Dynamic causal modeling can be useful by generating behaviors of important system variables over time. For this purpose, we first decided to model the prostate tumor dynamics in Vensim using system dynamics approach. Secondly, data science tools such as Support Vector Machines (SVM) and Logistic Regression are highly recommended as classification tools in healthcare decision making. Representing the correlations between parameters, these methodologies are highly used in medical literature for tumor detection.

Machine learning tools are core methodologies to lay out the correlations between parameters that affect PSA levels and detect prostate cancer existence, using individual level data. To understand how different factors affect PSA level, dynamics in the prostate should be observed. However, such tools do not utilize time series data; hence machine learning models have static characteristics and are not adequate to model the dynamics in the prostate. The need for system dynamics modeling arises from this limitation. The system dynamics model adds a dynamic dimension to tumor detection and can be used for foreseeing probable tumor growth which can be negligeable in the imminent time. It can be proposed that foreseeing probable tumor growth can alarm the medical doctors and patients to take precautionary actions to prevent the potential tumor growth.

The relationship between screening output variables, risk factors and prostate cancer's existence need to be analyzed and modeled. The imminent prostate cancer and probable tumor growth detection should be done, individually. Concerning the detection requirements in the short term with static data, machine learning tools can be used. Even though machine learning tools are adequate to lay out the correlations between prostate cancer related factors and predicting cancer existence in the short term; they lack the dynamic dimension. To comprehend the tumor growth, prostate related dynamics, and their effects on PSA; a system dynamics model needs to be built. Prostate related dynamics such as prostate cancer occurrence, tumoral tissue growth, healthy tissue growth, PSA level, free PSA ratio and the risk factors for cancer development are individual specific characteristics. That is why it is a requirement that the prostate dynamics to be modeled individually to comprehend the disease progression and effects of the risk factors.

3. OVERVIEW OF THE SYSTEM DYNAMICS MODEL

The dynamic model was designed to hold an explanatory power about the tissue-level dynamics of prostate. Model has four main stocks centrally (Figure 3). Having interdependencies, these four stocks create six feedback loops in the model. Four of them are balancing loops while the other two are reinforcing loops. Reinforcing loops are the results of birth of prostate and healthy tissue volumes. New tissues being born, their rate of birth increases accordingly. Thus, two reciprocal positive feedbacks create reinforcing loops for each of the tissue stocks. (See Figure 2)

Two of the balancing loops are observed during apoptosis of the tissues. When tissues tend to reproduce, this means an increase in their volume. With volume being larger, apoptosis rates increase accordingly. Hence, these negative feedback loops seek to balance the tumoral and healthy tissue volumes by balancing their population levels and death rates.

Nutrient, the stock which tissues are fed by, are embedded in two negative feedback loops in result of its causal relationships with tumoral and healthy tissues. Tissues are mechanisms that are disposed to grow by consuming nutrient. Their consumption rate increases as their volume increases, which results in a decrease in nutrient levels. Nutrient levels being lower, tissues are forced to consume less nutrient which leads to a decrease in their volumes.

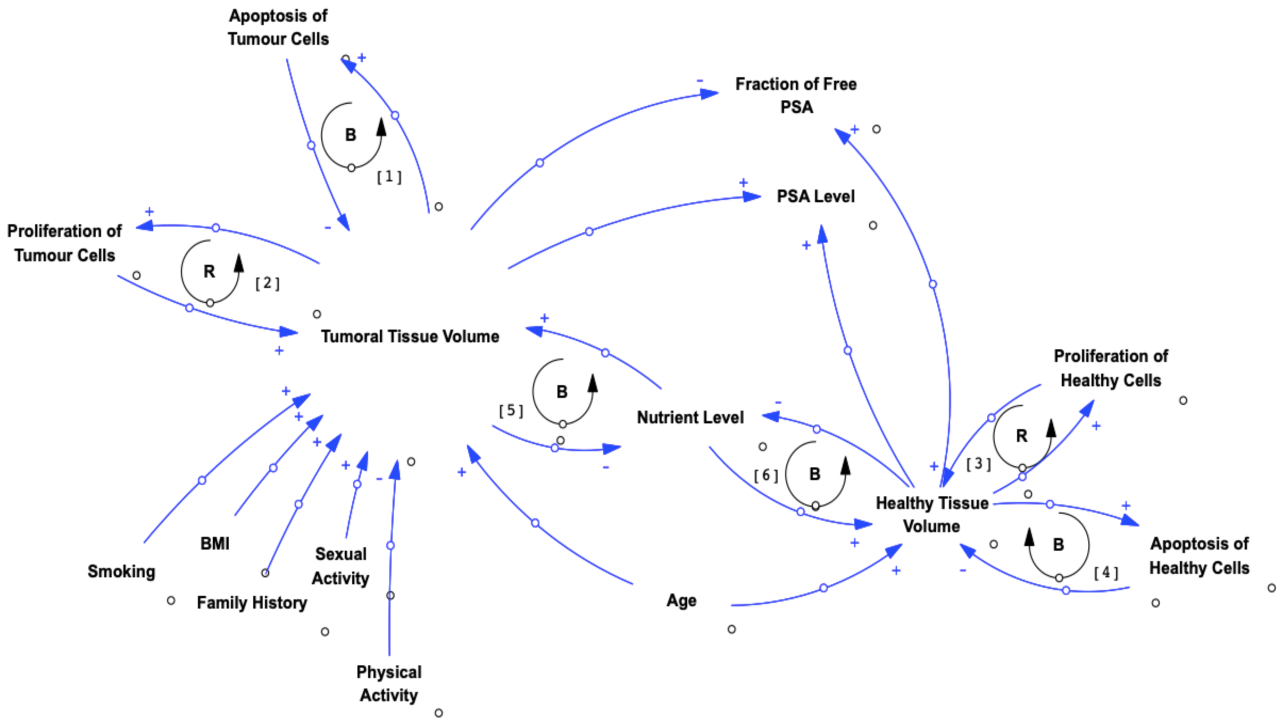
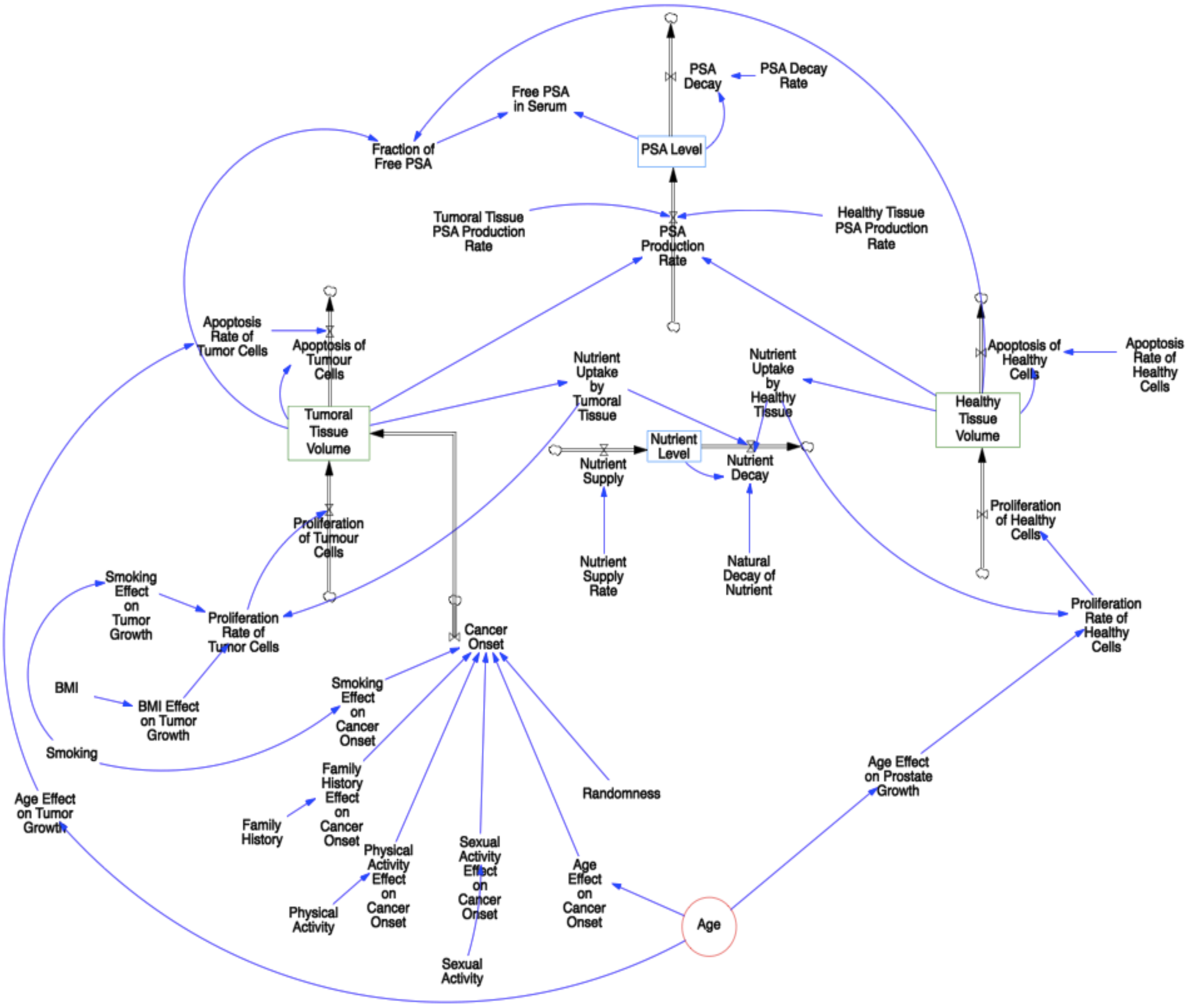


Figure 2: *Causal Loop Diagram*

Model is divided into three fundamental sectors: (1) PSA sector; (2) risk factors sector; (3) prostate tissues sector. Since PSA is not the sole indicator of prostate cancer, it is important to analyze the free PSA level along with its fraction in the serum PSA. Tumoral and healthy tissues are fed by vital nutrients so that they continue to reproduce. Tissues contribute to PSA production, but their rate of production differs by the type of the tissue.

Besides the internal factors mentioned, the model has several external factors, as well (like smoking, BMI…). They are defined to be individual-specific characteristics; thus, their values differ for each individual-specific run. *Age* being an independent stock in the model, it represents the individual's real-time age. Hence, it increases linearly throughout the simulation period, which is ten years as default. A more detailed explanation is provided in the following section.



PSA Decay
PSA Decay Rate
Free PSA in Serum
PSA Level
Fraction of Free PSA
Tumoral Tissue PSA Production Rate
PSA Production Rate
Healthy Tissue PSA Production Rate
Apoptosis Rate of Tumor Cells
Apoptosis of Tumour Cells
Tumoral Tissue Volume
Nutrient Uptake by Tumoral Tissue
Nutrient Uptake by Healthy Tissue
Apoptosis of Healthy Cells
Apoptosis Rate of Healthy Cells
Healthy Tissue Volume
Nutrient Level
Nutrient Supply
Nutrient Decay
Nutrient Supply Rate
Natural Decay of Nutrient
Proliferation of Tumour Cells
Proliferation of Healthy Cells
Smoking Effect on Tumor Growth
Proliferation Rate of Tumor Cells
Cancer Onset
Proliferation Rate of Healthy Cells
BMI
BMI Effect on Tumor Growth
Smoking Effect on Cancer Onset
Smoking
Family History Effect on Cancer Onset
Randomness
Age Effect on Prostate Growth
Age Effect on Tumor Growth
Family History
Physical Activity Effect on Cancer Onset
Sexual Activity Effect on Cancer Onset
Age Effect on Cancer Onset
Age
Physical Activity
Sexual Activity

Figure 3: *Complete Stock Flow Diagram*

4. MODEL DESCRIPTION

Model contains four internal stocks: *Tumoral Tissue Volume, Healthy Tissue Volume, Nutrient Level,* and *PSA Level.* PSA sector includes only *PSA Level* stock, while Prostate Tissues Sector includes remaining three. Table 1 provides the units of the stocks of the model. Tissues are involved into the processes by their PSA production and their interaction with the nutrients.

Table 1*: Units of the Stock Variables of the Model*

Stock	**Unit**
Tumoral Tissue Volume	Liter
Healthy Tissue Volume	Liter

Nutrient Level	Gram
PSA Level	ng/mL

4.1. Assumptions

In the dynamic model, several risk factors are assumed to be affecting the tumoral and healthy tissue dynamics. As mentioned before, some external individual characteristics tend to accelerate the potential tumoral growth. The risk factors included are smoking, sexual activity, physical activity, BMI rate, family history, and age. In the model, we consider these six risk factors as individual-specific characteristics. Thus, they are subject to change in each run. Their values are assumed to be binary. Hence, a risk factor input should be entered as 1 if an individual carries that factor, 0 otherwise. Risk factors are expected to have different levels of effect on the system.

Cancer Onset is located as an inflow to the stock of *Tumoral Tissue Volume*. It is assumed to be the triggering effect which initiates the existence and growth of tumoral tissue. Furthermore, *Randomness* is defined as a variable to explain the random effects on the biological processes in the prostate. For each run, randomness is assumed to have an impact on the tumoral growth initiation. Owing to this, we intend to capture the randomness in human biology, potentially creating an environment that allows rapid cell growth.

Tissues are fed by certain hormones, vital nutrients, vitamins, and especially glucose. (Lorenzo, et al., 2016) Gathering all substances under one name, we called them *nutrient* to prevent unnecessary complexity of the model. Thus, *Nutrient Level* represents all the biological factors which help and accelerate tissue growth.

4.2. Technical Description of the Dynamic Model

Making use of medical literature on prostate cancer, biological interactions and dynamics of prostate tissues are derived. Model is simulated by using Vensim PLE 9.3.4. Causalities of the interactions are built and represented within the model. Thus, we can easily observe the time-dependent dynamic relationships between the variables. To catch the continuous nature of human biology, the time unit of the model is decided to be day, and time step is 0.125. Model is simulated for ten years, which equals to 3650 days.

4.2.1. PSA Sector

$Tumoral\ Tissue\ Volume\ (V_c)$ and $Healthy\ Tissue\ Volume\ (V_h)$ are two main stocks at the center of the system. Together they form the total prostate volume. Since both tumoral and healthy tissues produce serum PSA, they contribute to the inflow rate of $PSA\ Level\ (P_s)$. Healthy tissue produces PSA with a rate of $\alpha_h = 6.25\, ng.mL^{-1}.cm^{-3}.y^{-1}$, (time unit is given as *year* in this equation) while tumoral tissue PSA production rate (α_c) is 15 times of the healthy tissue PSA production rate (α_h). Moreover, PSA has a natural decay rate coefficient of $\gamma = 0.35\ d^{-1}$ (time unit is given as *day* in this equation). (Lorenzo, et al., 2016)

$$P_s = \alpha_h * V_h + \alpha_c * V_c - \gamma * P_s$$

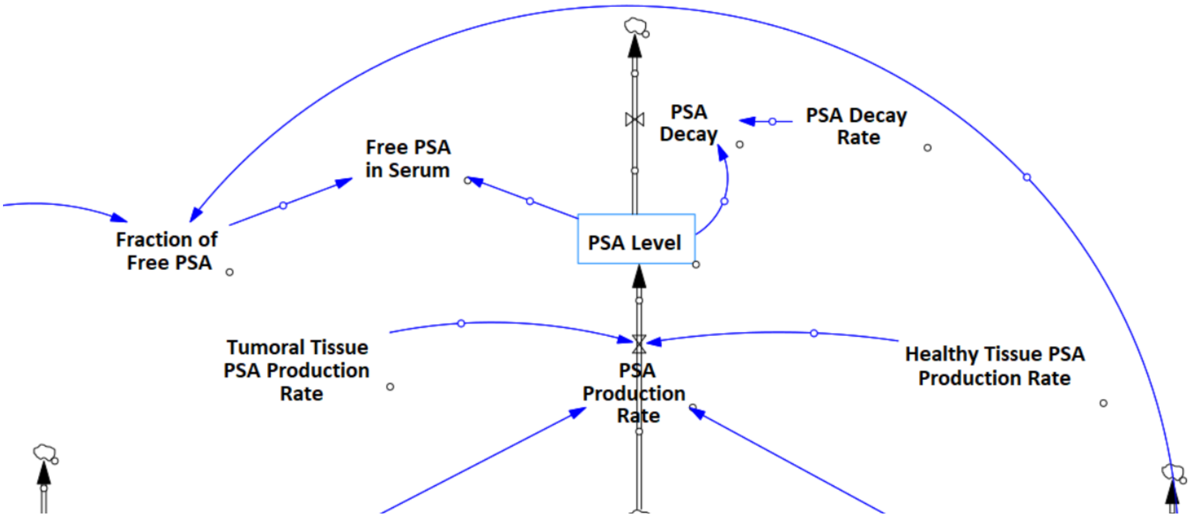


Figure 4: *PSA Sector*

As free PSA ratio is observed to be relatively less in cancer patients than in healthy people; it is important to model the free PSA dynamics along with other PSA parameters. (Adhyam & Gupta, 2012) Healthy prostate tissues tend to produce more inactive PSA which circulates in the blood unboundedly (free PSA), while cancerous cells produce more active PSA circulating in the blood and bound to certain inhibitors such as ACT. (Adhyam & Gupta, 2012) D'Amico et al. provide a set of equations to predict the tumor volume using given values of Gleason grade, total PSA level, and the prostate volume (V_p). (D'Amico, et al., 1997) Based on the causality between free PSA ratio and prostate tissues, we modeled free PSA as directly being affected by tumoral and healthy tissue volumes. Epithelial fraction used in the below equation equals to 0.2 and the PSA leak into serum per cubic centimeter of healthy tissue equals to 0.33, as defined by D'Amico et al. (D'Amico, et al., 1997) Mentioned equation set provided by D'Amico et al. is following:

$$\begin{aligned} &PSA\ from\ healthy\ tissue \\ &\qquad = [(epithelial\ fraction) \\ &\qquad * \left(PSA\ leak\ into\ serum\ per\ cm^3\ of\ healthy\ tissue \right) * (prostate\ volume)] \end{aligned}$$

$$\text{Tumoral Tissue Volume} = \frac{PSA\ from\ tumoral\ tissue}{PSA\ leak\ into\ serum\ per\ cm^3\ of\ tumoral\ tissue}$$

In the dataset, using TNM staging of the patients, we predicted the Gleason grade of each patient. Then, using their prostate volumes and total PSA levels, we estimated the cancer-specific PSA levels along with the tumoral tissue volumes using above equations. Having tumoral tissue volume data, we were able to build a linear regression model. Dependent variable of the model is free PSA ratio, while the independent variables are healthy and tumoral tissue volumes. Deriving the correlation coefficients from the output, we integrated them into the dynamic model. Thus, the variable Fraction of Free PSA (p) is linearly dependent on tumoral and healthy tissue volumes and by multiplying the PSA Level with this fraction, we get the Free PSA Level (P_f).

$$p = 13.7586 - 2.19265 * V_c + 0.18732 * V_h$$

$$P_f = p * P_s$$

4.2.2. Risk Factors Sector

Prostate tissue dynamics are affected by both internal and external factors. Internal factors consist of nutrient level, while external factors consist of certain risk factors such as BMI, smoking, family history, sexual activity, and physical activity. (Perdana, Mochtar, Umbas, & Hamid, 2016) Some of the past studies lack including these risk factors into their models. Since tumor growth cannot be determined only by the internal physiological characteristics, it is important to consider the external factors while building a model. For instance, a 70-year-old male with a family history of cancer has more risk of developing a prostate tumor than a 50 years-old male with a healthy lifestyle.

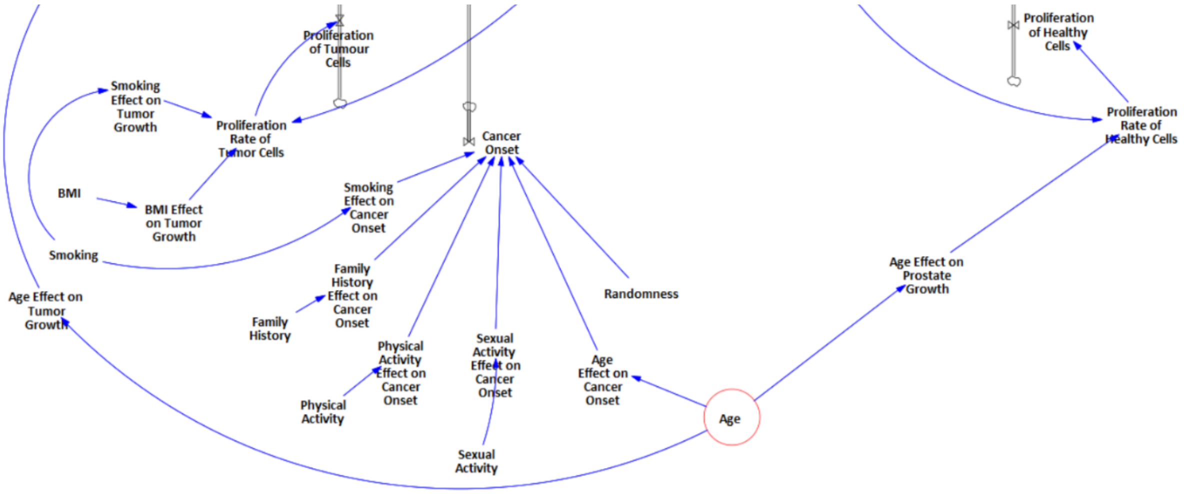

Proliferation of Tumour Cells
Proliferation of Healthy Cells
Smoking Effect on Tumor Growth
Proliferation Rate of Tumor Cells
Cancer Onset
Proliferation Rate of Healthy Cells
BMI
BMI Effect on Tumor Growth
Smoking Effect on Cancer Onset
Smoking
Age Effect on Prostate Growth
Family History Effect on Cancer Onset
Randomness
Age Effect on Tumor Growth
Family History
Physical Activity Effect on Cancer Onset
Sexual Activity Effect on Cancer Onset
Age Effect on Cancer Onset
Physical Activity
Age
Sexual Activity

Figure 5: *Risk Factors Sector*

Risk factor characteristics differ from each other in that not every risk factor directly affects the tumoral tissue growth, but some of them affect the tumor initiation. Thus, we classified the external risk factors according to their impacts on tumor growth and tumor onset. BMI is observed to be affecting proliferation rate of the tumoral cells. (Allott, Masko, &

Freedland, 2013) Smoking increases the risk of developing cancer and increasing the proliferation rate of the tumor. Thus, smoking directly affects both tumoral growth and cancer initiation in the model. An increase in age decreases the apoptosis rate of tumoral cells and increases the proliferation rate of the healthy cells. Thus, age of the patient, which linearly increases throughout the simulation period, affects the tumoral and healthy tissue volumes as well as the cancer onset. Additionally, family history, sexual activity, and physical activity is observed to be influencing the cancer onset, not tumor growth. (Albright, et al., 2015) Even though all risk factors have an influence on the cancer dynamics, their rates are different from one to another. To decide on the correct risk rates, we used estimated relative risk (RR) values from past medical studies on prostate cancer risk factors. Using the relative risks, we calculated the combined probability of the risk factors.

***Table 2**: References for Relative Risks (RR) of the Risk Factors*

Risk Factor	**RR on Cancer Onset**	**RR on Tumor Growth**
Smoking	1.01 - 1.46 (Huncharek, Haddock, Reid, & Kupelnick, 2010)	1.06 - 1.19 (Huncharek, Haddock, Reid, & Kupelnick, 2010)
BMI	-	1.01 - 1.05 (Allott, Masko, & Freedland, 2013)
Family History	2.46 (Albright, et al., 2015)	-
Sexual Activity	1.26 – 1.73 (Perdana, Mochtar, Umbas, & Hamid, 2016)	-
Physical Activity	0.84 - 0.95 (Leitzmann & Rohrmann, 2012)	-

Cancer onset is a variable that determines whether cancer is initiated or not. Starting with a zero-tumor volume, if cancer onset becomes positive, then a negligible volume of tumor is created by the cancer onset and cancer is initiated. There are two main substances that affect the cancer onset value. First is the combined probability of risk factors and second is the randomness of the system. We used base the prostate cancer prevalence as 15.3%. (Huncharek, Haddock, Reid, & Kupelnick, 2010) Randomness represents a random probability which is UNIFORM [0, 1]. If the combined risk probability of the patient exceeds the random probability, then the *Cancer Onset* initiates the cancer and creates a negligibly small volume of tumor.

4.2.3. Prostate Tissues Sector

Three main stocks are located in this core part of the model. These stocks are called $Tumoral\ Tissue\ Volume\ (V_c)$, $Healthy\ Tissue\ Volume\ (V_h)$ and $Nutrient\ Level\ (\sigma)$. They have interdependent relationships with reinforcing and balancing loops present. Proliferation of the cells create a reinforcing loop, while apoptosis of the cells creates a balancing loop. Initial tumor volume for non-cancerous people is in an interval between 0 and 0.01 cc which is a negligible value. Considering the alternative risk scenarios, the initial tumor volume is changed accordingly. Initial healthy epithelial tissue volume is dependent on the age of the patient and changes within 20 cc to 40 cc interval except extreme cases such as benign prostate hyperplasia. Initial PSA values are also dependent on the age and prostate volume of the individual.

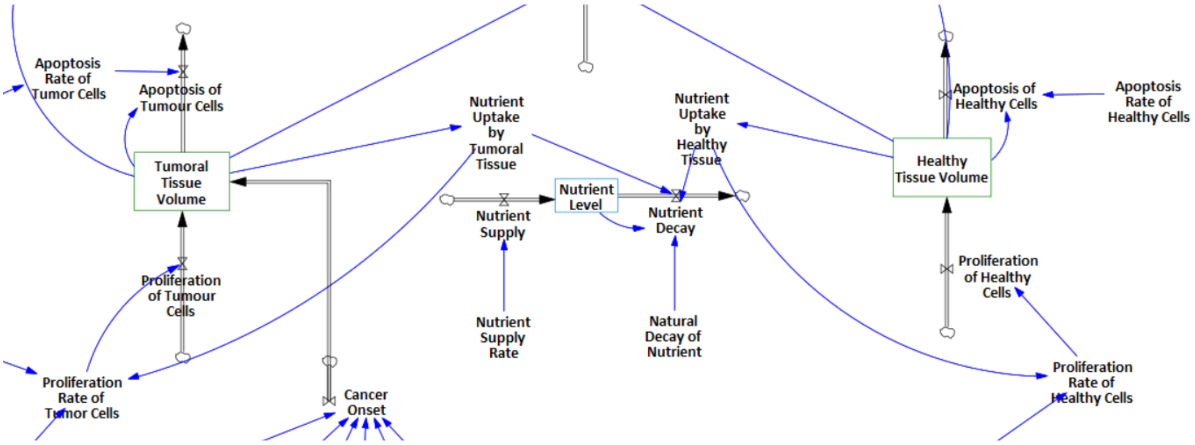

Apoptosis Rate of Tumor Cells
Apoptosis of Tumour Cells
Tumoral Tissue Volume
Proliferation of Tumour Cells
Proliferation Rate of Tumor Cells
Cancer Onset
Nutrient Uptake by Tumoral Tissue
Nutrient Supply
Nutrient Supply Rate
Nutrient Level
Nutrient Uptake by Healthy Tissue
Nutrient Decay
Natural Decay of Nutrient
Apoptosis of Healthy Cells
Apoptosis Rate of Healthy Cells
Healthy Tissue Volume
Proliferation of Healthy Cells
Proliferation Rate of Healthy Cells

***Figure 6**: Prostate Tissues Sector*

Both healthy and tumoral tissues consume nutrient and they contribute to its decay rate, as well. Nutrient uptake rate for the tumoral tissue is $\delta_c = 2.75\ g.L^{-1}.d^{-1}$, and for the healthy tissue is $\delta_h = 2.75\ g.L^{-1}.d^{-1}$. Nutrient decays with a rate coefficient of $\gamma = 0.30\ d^{-1}$. Nutrient supply rate is $s = 3\ g.L^{-1}.d^{-1}$. (Lorenzo, et al., 2016)

$$Nutrient\ Uptake\ by\ Tumoral\ Tissue\ = \delta_c * V_c$$

$$Nutrient\ Uptake\ by\ Healthy\ Tissue\ = \delta_h * V_h$$

$$Natural\ Decay\ of\ Nutrient\ = \gamma * \sigma$$

$$\begin{aligned} &Proliferation\ of\ Tumor\ Cells \\ &\qquad = BMI\ Effect\ on\ Tumor\ Growth * Smoking\ Effect\ on\ Tumor\ Growth \\ &\qquad * Nutrient\ Uptake\ by\ Tumoral\ Tissue \end{aligned}$$

Nutrient uptake is proportional to the volume of the tumoral tissue, and proliferation rate of tumoral cells is proportional to the nutrient uptake. Thus, there is a reinforcing loop which results in an exponential growth of tumoral cells. Moreover, proliferation rate of tumor is affected by both BMI and smoking factors. According to one's BMI value and smoking habits, a relative risk rate is assigned, and they are multiplied with the nutrient uptake rate. Simulating a person with high BMI rate and smoking habit, we observe the proportional growth on the tumoral cells compared to a healthy individual. Also, as a male gets older, his prostate tends to get bigger which is accepted to be normal in certain intervals. If the prostate volume exceeds the normal limits, it is a case called Benign Prostate Hyperplasia (BPH) which is commonly observed among older people. To truly reflect the age effect on the prostate volume, we inserted an age-dependent rate in the proliferation rate of healthy cells.

$$Proliferation\ of\ Healthy\ Cells\ =\ (Nutrient\ Uptake\ by\ Healthy\ Tissue) * (1 + Age\ Effect\ on\ Prostate\ Growth)$$

5. MODEL BEHAVIOR & VALIDATION

Past prostate research studies provide real and estimated data as intervals. Thus, for model calibration, we simulated the system by changing input parameters within the suggested intervals. Having several outputs with various input parameter values, optimal intervals for the model inputs were determined.

The model was tested using structural and behavioral validity tests. For the behavioral validity, we compared model behaviors with real world data under certain scenarios. Those scenarios will be explained in the following. As real-world data, we used the dataset provided in the appendix. We had three fundamental runs for the model validation and showed that the patient-level prostate dynamics are consistent with the real data derived from medical literature. Having previous year PSA and current PSA values for the patients in the data set, we were also able to compare the time-dependent PSA increase rates.
Figure 5 was extracted from another modelling study. In the figure, tumor volume change and serum PSA change over time of a patient can be seen; control line represents patient without an intervention and treatment line represent patient with treatment intervention. Since we don't include treatment effect in our model, control line can be used for validation of our model. Comparing our model and the control line in Figure 5, their behavior under the same baseline conditions match.

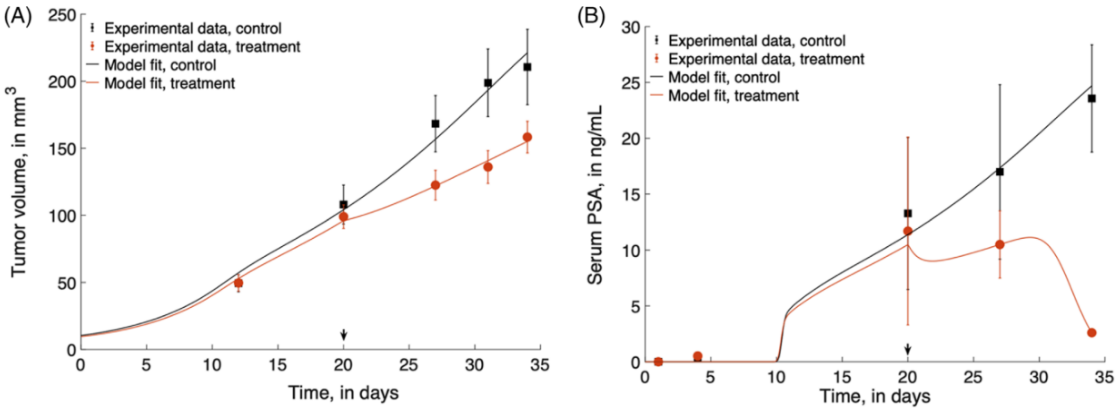


***Figure 5**: (A) Tumor Volume Dynamics and (B) Serum PSA Dynamics Graphs To Be Used For Behavioral Validation (Barnaby, Sorribes, & Jain, 2021)*

5.1. Base Run

Before starting the three scenario runs under different conditions, we present the base run. As a base, we assumed a 55-year-old male with minimal risk factors such as smoking and sexual activity. The subject has no family history in prostate cancer, neither a high BMI rate. Starting with a zero tumoral tissue volume, tumoral proliferation rate stays under control throughout the simulation period of ten years. He doesn't demonstrate an extreme situation in any sectors, for which we can comment that the run is coherent with the real-world behavior. Because of the randomness effect that we added to the model, negligible amounts of tumor may be born but kept under control throughout the simulation.

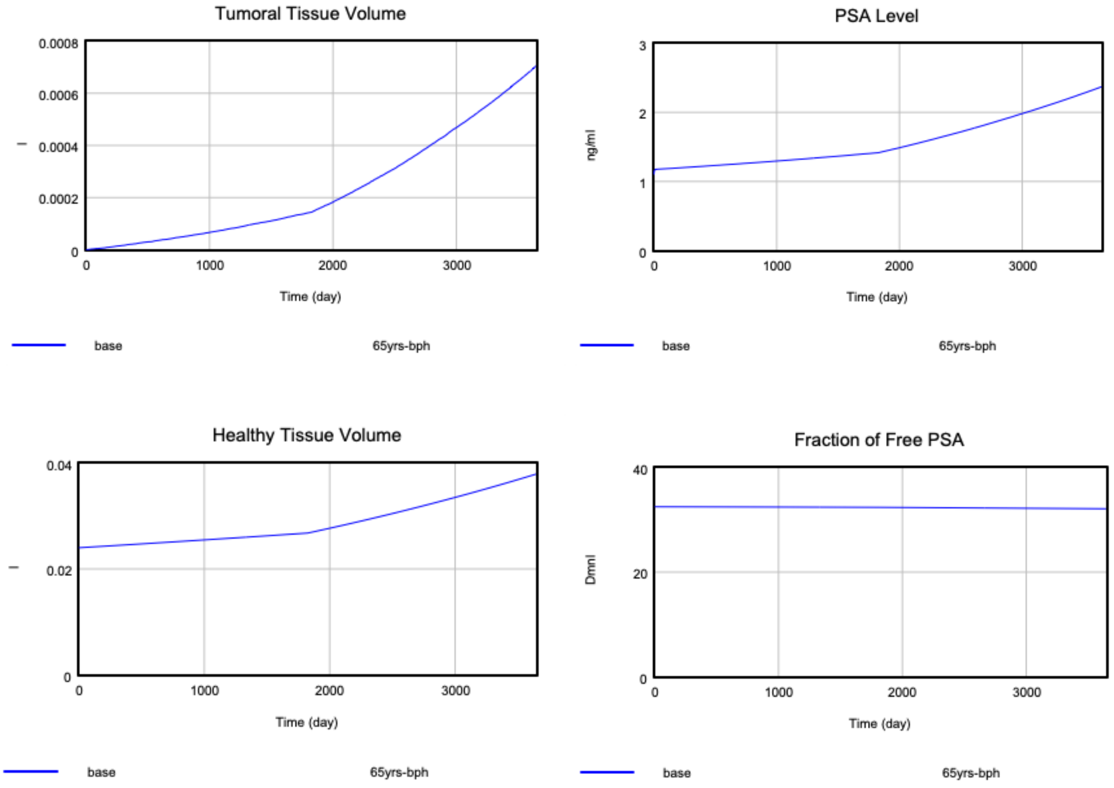


***Figure 6**: Outputs of base scenario*

5.2. Risk Factors Effect

5.2.1. 60-year-old smoker with a *high* prostate volume and BMI rate

This patient is assumed to be a disadvantageous male with older age, high prostate volume, high BMI rate, family history observed, lack of physical activity which results in extreme growth of tumoral tissue. Starting with a negligibly small tumoral tissue volume, tumoral proliferation rate cannot be controlled because of his age and high BMI. Tumor tends to exponentially grow over time and reaches a clinically advanced level during the sixth year of the simulation period. Observing his increasing PSA level, a medical doctor would realize the risk at the sixth year or later. However, simulating his prostate dynamics over years, it is easy to observe potential changes in tissues and PSA levels.

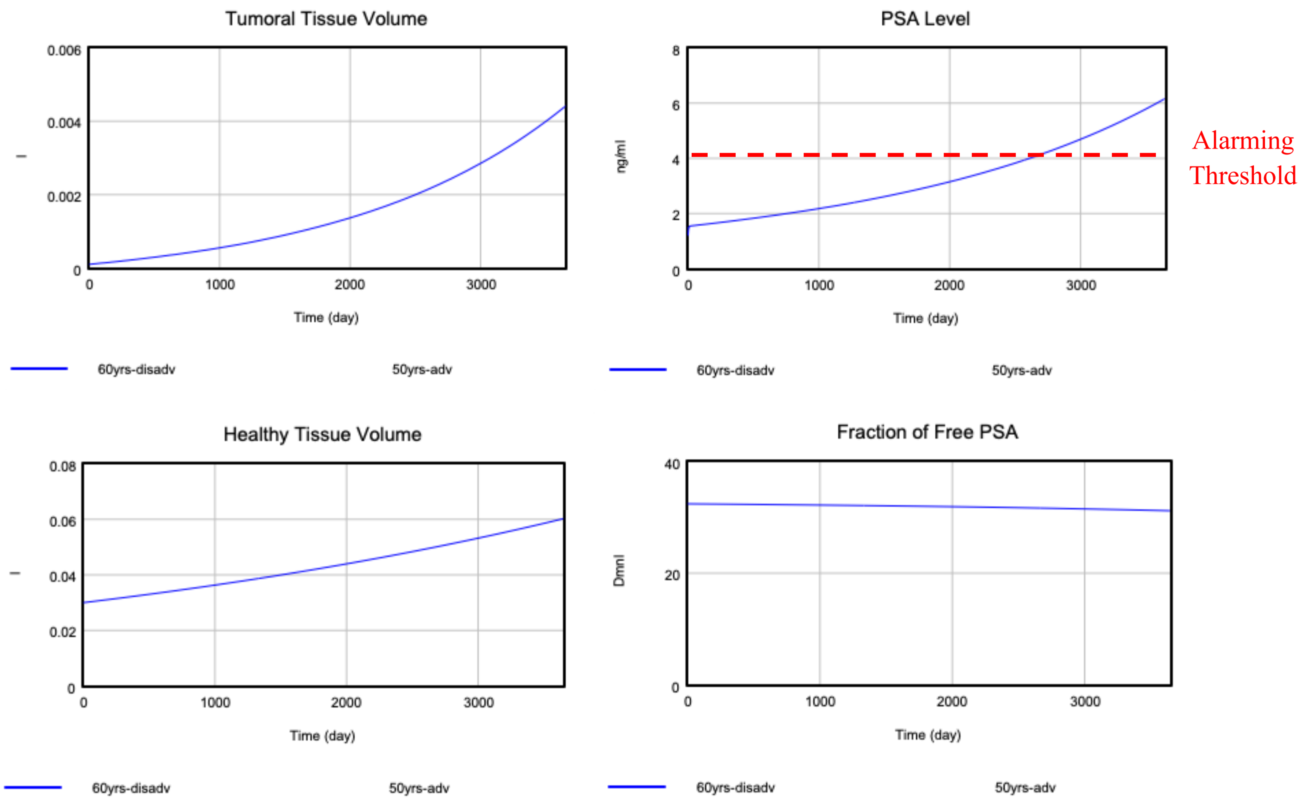


***Figure** 7: Outputs of scenario 1*

5.2.2. 50-year-old non-smoker with a *normal* prostate volume and BMI rate

This patient has a healthy lifestyle and advantageous genetic factors (No history of prostate cancer was observed in the family). Starting with a negligibly small tumoral tissue volume, tumoral proliferation rate stays under control (<0.0008 liters) throughout the simulation period of ten years. We can observe that his healthy prostate easily beats the tumoral cell proliferation. Knowing that a linear increase in PSA over years is normal as long as it doesn't exceed the alarming threshold of 4ng/ml and it doesn't increase exponentially. Starting with a PSA close to 2ng/ml, this patient's PSA value only exceeds 2 ng/ml in the next ten years which is a usual increase by age. Another point to focus is the fraction of free PSA in the serum, since this person doesn't develop a clinically significant tumor over the simulation period, its free PSA fraction tends to stay steady.

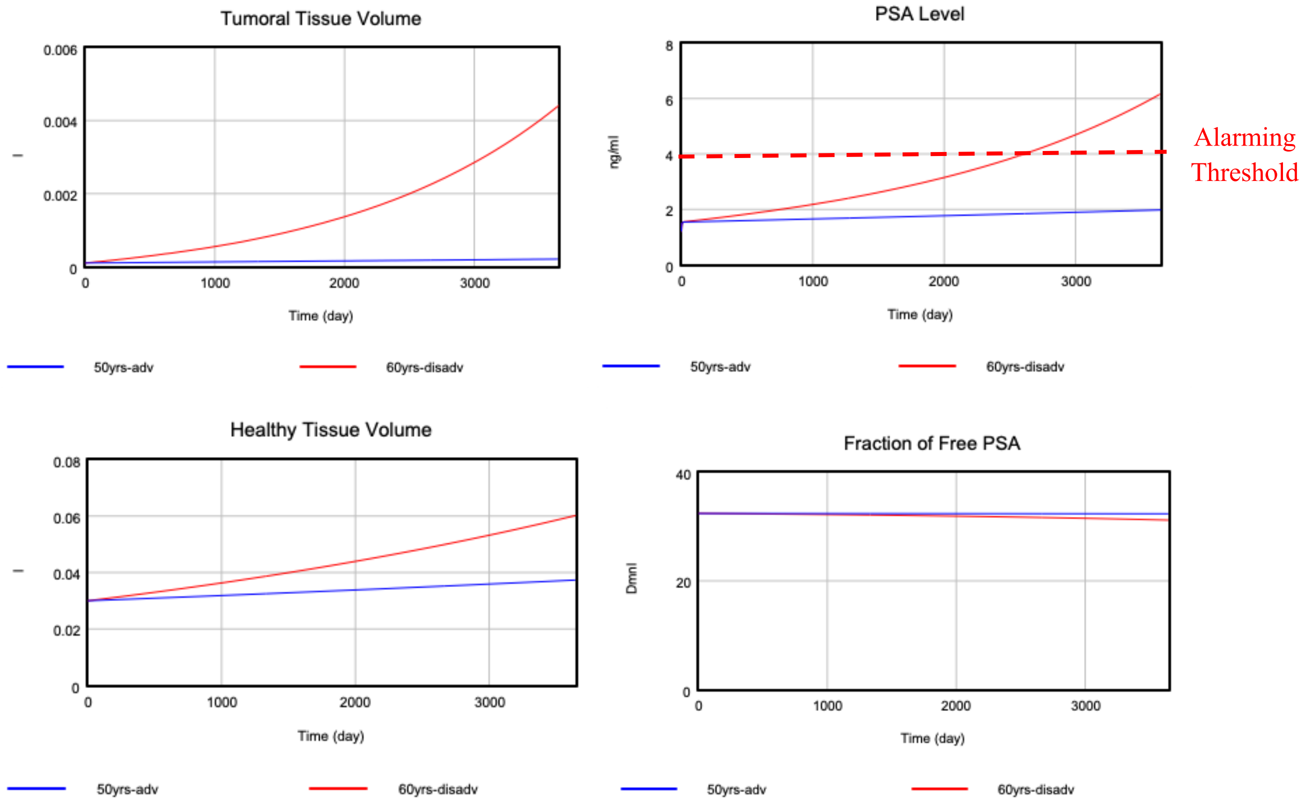


***Figure 10**: Outputs of scenario 2 in comparison with scenario 1*

To truly observe the relative dynamics of the two patients, we compare their scenario runs. It is now more accurate how the first patient cannot control the tumor growth and how the second patient is successful in having control over the tumor. PSA level curves also show a different dynamic of the patients. Exponential growth in PSA is alarming for most of the cases, except for the benign prostate hyperplasia. Through these graphs, we easily observe the effect of risk factors in tumoral tissue growth.

5.3. 65-year-old non-smoker and low BMI with a benign prostate hyperplasia run

Apart from the tumoral growth, prostate volume tends to grow if the patient is diagnosed with benign prostate hyperplasia. In this case, starting with a volume of 45 cc, prostate reaches almost 100 cc at the end of the ten years. For males older than 65, normal range for the prostate volume is 36 cc – 45 cc. (Berges & Oelke, 2011) Even though he does not have a growing tumor, his PSA level exceeds the alarming threshold because of his growing prostate. This case shows how the PSA value may be misleading without analyzing other dynamics in the prostate.

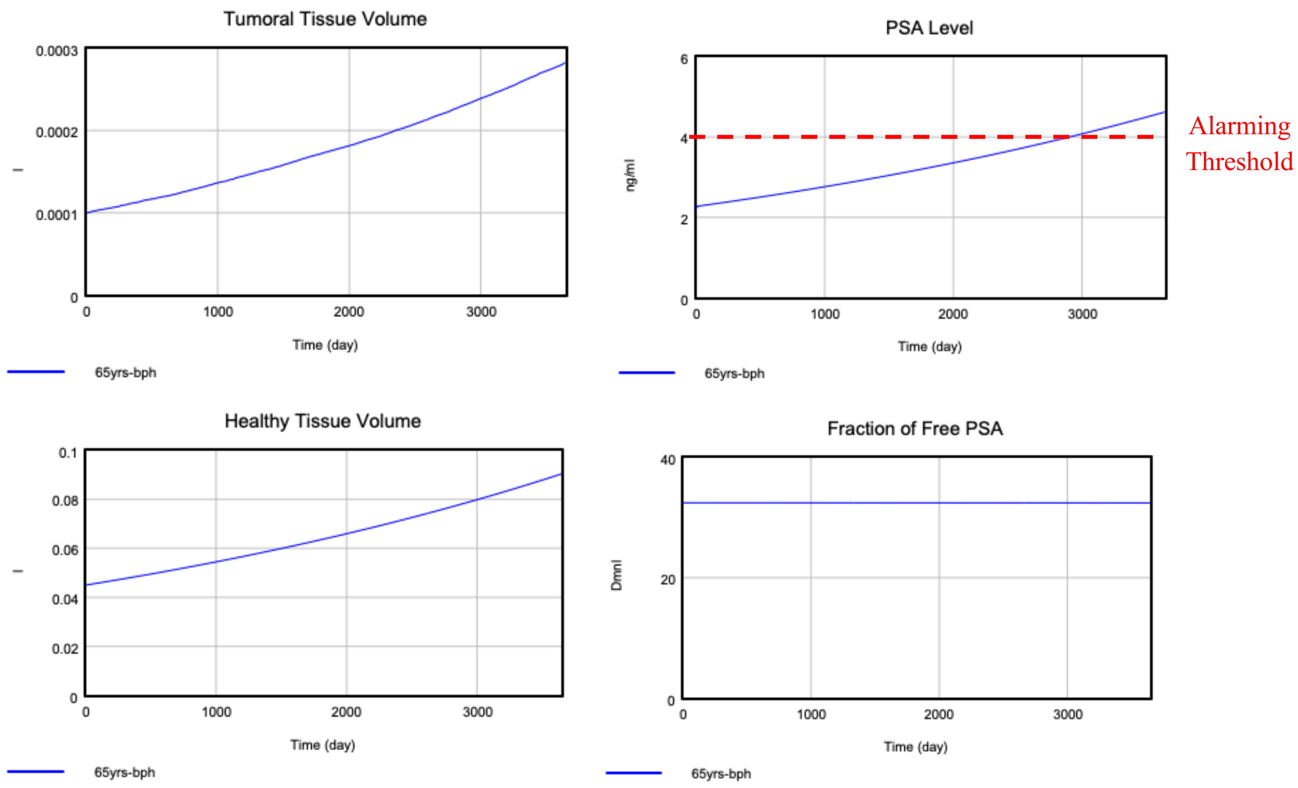


***Figure 11**: Outputs of scenario 3*

Since the model works with considerably small parameter values, a small change in parameters can result in a considerable change in the output. Thus, it is important to address the sensitivity of human physiology in our model, as well. Having the results for the patients, we can observe the changes in output of the patients with relatively different characteristics. Hence, the solution method of the study is consistent with system characteristics and limitations.

Model can be easily implemented to any male, having several prostate-related characteristics as input, we can observe his tissue-level prostate dynamics over periods of time. As long as a person has his prostate related data, the model can sustainably work. Only pre-requisite for the model run is the individual prostate data, screening, and test results. Changing the model period to longer years, we can also analyze lifetime results. Thus, solution is robust considering the potential changes in the parameters and real-life dynamics.

6. MACHINE LEARNING MODEL DESCRIPTION

To be used as inputs for the machine learning models, we gathered three different datasets from previous medical studies and one dataset was provided by Ministry of Health in Turkey. (Klingebiel, et al., 2022) (Nikitina, et al., 2019) (Cebeci & Ozkan, 2021) These datasets contain hundreds of patients' data with certain prostate-related characteristics such as PSA level, free PSA level, previous year's PSA level, prostate volume, and age. To fill the missing clinical characteristics in the data, we used MICE imputation technique which performs best among other techniques such as KNN and Median imputation. Throughout this part of the study, we used Python as the programming language.

Another deficiency of the data is the dominant presence of cancerous patients. For machine learning model to perform well, we needed data of individuals who don't carry a tumor. So that, the model can be trained well to classify the individuals as "cancerous" or "healthy". To this end, we generated hundreds of synthetic data of healthy individuals using system dynamics model. The analysis of the gathered and generated data was coherent with medical literature. This was an important and useful synthesis of the system dynamics model and data science model.

For prediction of prostate cancer, nine different classification algorithms were built. Data was split into two sets: training and validation sets. Five-fold cross validation was performed in the training set to get an insight about overfitting. However, after training the model with the training dataset, the validation set was predicted, and the results of validation set was considered for the comparison and selecting the final model since the results of the validation set didn't carry any data leakage. All parameter changes in the models were made manually in order to parameter optimization.

Here is the list of the classification algorithms implemented:

- Logistic Regression
- Support Vector Machines
- K-Nearest Neighbors
- Naive Bayes
- Decision Trees
- Random Forests
- Gradient Boosting Classifier
- XGBoost Classifier
- LightGBM Classifier

Accuracy, precision, and recall values are indicators for performance of a classification model. Different indices are used for calculations of these indicators. Here are the equations of the performance indicator values:

$$Accuracy = \frac{(False\ Negative\ +\ False\ Positive)}{(True\ negative\ +\ True\ Positive\ +\ False\ Negative\ +\ False\ Positive)}$$

$$\mathrm{Pre}cision = \frac{(True\ Positive)}{(True\ Positive\ +\ False\ Positive)}$$

$$Recall = \frac{(True\ Positive)}{(True\ Positive\ +\ False\ Negative)}$$

As it can be seen from the calculations, lower false positive value indicates higher precision value and lower false negative value indicates higher recall value. In this study, recall value is accepted to be more significant than precision value since false negative detection may result in certain fatalities. (Papageorgiou, et al., 2018) In short, accuracy, false negatives and recall value have more emphasis when analyzing the results of the model and making comparison with other models.

***Table 3**: Comparison of Classification Methods*

For Validation Set	Accuracy	Precision	Recall
K-Nearest Neighbors	72.7506	77.8626	56.9832
Support Vector Machines	73.2648	79.5276	56.4246
Naive Bayes	79.1774	81.4103	70.9497
Decision Tree	75.8355	73.4807	74.3017
Random Forest	81.4910	79.5580	80.4469
Gradient Boosting Classifier	81.2339	77.6042	83.2402
Logistic Regression	80.4627	76.1421	83.7989
LightGBM Classifier	81.4910	79.5580	80.4469
XGBoost Classifier	81.7481	76.2136	87.7095

For comparison, when the accuracy values and recall values considered, XGBoost has the highest accuracy and recall value. Precision value of XGBoost is relatively lower than other models. However, as mentioned before recall value has more significance than precision value for cancer prediction. For conclusion, as it can be seen from Table 3, XGBoost model is chosen as the final model for this study.

6.1. Suggestions

The design can be implemented into the routine prostate cancer screening process. Current screening process can be seen in Figure 12. Suggested model implementations are marked with red arrows in figure 12.

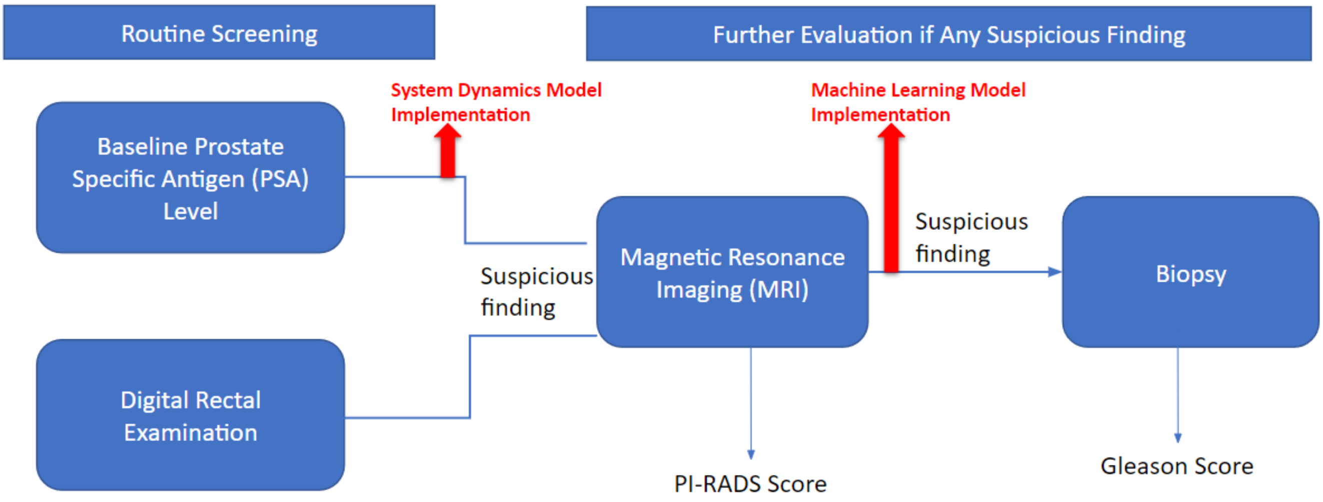

Routine Screening
Further Evaluation if Any Suspicious Finding
System Dynamics Model Implementation
Machine Learning Model Implementation
Baseline Prostate Specific Antigen (PSA) Level
Digital Rectal Examination
Suspicious finding
Magnetic Resonance Imaging (MRI)
Suspicious finding
Biopsy
PI-RADS Score
Gleason Score

***Figure 12**: Prostate Cancer Screening Process*

System dynamics model can be integrated using individual level data and calibrated using the patient's PSA test results to see probable future results. The results from the system dynamics model can alarm the medical doctors and patients to take precautionary actions to prevent the foreseen tumor growth. Machine learning model is implemented by using personal level data, output data of PSA test, and MRI test. Its objective is limited to detecting an existing tumor. The essential part of this process is to detect the tumor which could have been avoided by the standard procedures. The model can alarm the doctors to apply biopsy on the patient, and the biopsy is seen as gold standard in medical field. (NCCN Guidelines, 2022) Machine learning model's aim should be to achieve lower false negative rates on the outcome of MRI, since false positives can be corrected by biopsy in the next steps.

Obtained solutions should be revised with the new screening or test data. Since a change in input parameter may change the output, it is important to consider the future solutions of the model with additional screening and testing results present.

7. CONCLUSION

Tumor detection comes with its challenges, resulting in inadequacy of the screening methods and late diagnosis of cancer patients. In this study, we built system dynamics and machine learning models to detect the potentially aggressive tumor in early stages. We constructed a system dynamics model of the biological mechanisms in prostate to observe the time-dependent pathogenesis of prostate cancer. Building machine learning model as complementary, it was adapted to the study as a prompt classifier. Furthermore, machine learning model was useful for the statistical analysis of the output of prostate cancer screening process, laying out the correlations between the variables, and prediction of the prostate cancer presence. Throughout this study, we intended to develop a reliable predictive method supported by two models with the integration of their dynamic and static characteristics.

We modeled the prostate of a male showing average and normal functioning biological characteristics in tissue-level. PSA being the most significant indicator of prostate cancer, we modeled its relationship with free PSA, as well. Furthermore, we considered the effects of external risk factors on prostate cancer by adapting them in the model. For the machine learning part of the study, we built nine classifying algorithms through Python to find the optimal one, and XGBoost performed best with an accuracy value of 81.75 and recall value of 87.71. System dynamics model was helpful imputing the missing data to be used as input in the ML model.

The system dynamics model was validated through structural and behavioral validity tests. We performed a base run along with the three scenario runs to validate and examine the effects of several risk factors. In base run, a male with normal characteristics is simulated and results are concluded to be coherent with the real-world. Then, we simulated another two runs to demonstrate the effect of the risk factors. First of them is a disadvantageous male carrying each risk factor while second is an advantageous one with healthy lifestyle, carrying no risk factor. In the last run, we indicated the difference between prostate cancer and BPH. Even though PSA seemed to be alarming for prostate cancer, we observed the tumoral volume being steady and concluded that he showed a BPH behavior.

Using the model, we could observe the 10-year-long prostate dynamics in tissue level. The integrated design can help increase the accuracy of prostate cancer detection and assist doctors and patients to take precautions of probable tumor growth, which can reduce the prostate cancer incidences and cancer specific mortality rates. It can be strongly argued that the design will help reducing prostate cancer incidences; however, number of procedures will increase with our additions to the current procedures, which can increase the cost of screening process. As a future work, cost effectiveness analysis can be done on the possible implementation of our design.

Bibliography

Wang, L., Lu, B., He, M., Wang, Y., Wang, Z., & Du, L. (2022). Prostate Cancer Incidence and Mortality: Global Status and Temporal Trends in 89 Countries From 2000 to 2019. *Frontiers in Public Health, 10*, 811044.

Loeb, S., Bjurlin, M. A., Nicholson, J., Tammela, T. L., Penson, D. F., Carter, H. B., . . . & Etzioni, R. (2014). Overdiagnosis and overtreatment of prostate cancer. *European urology, 65*(6), 1046–1055.

Zhang, S. J., Qian, H.-N., Zhao, Y., Sun, K., Wang, H.-Q., Liang, G.-Q., . . . Li, Z. (2013). Relationship between age and prostate size. *Asian Journal of Andrology, 15*(1), 116-120.

Partin, A. W., Hanks, G. E., Klein, E. A., Moul, J. W., Nelson, W. G., & Scher, H. I. (2002). Prostate-specific antigen as a marker of disease activity in prostate cancer. *Oncology (Williston Park), 16*(8), 1024-1038.

Adhyam, M., & Gupta, A. K. (2012). A Review on the Clinical Utility of PSA in Cancer Prostate. *Indian J Surg Oncol, 3*(2), 120-129. Retrieved from National Library of Medicine: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3392481/

Albright, F., Stephenson, R. A., Agarwal, N., Teerlink, C. C., Lowrance, W. T., Farnham, J. M., & Cannon Albright, L. A. (2015). Prostate Cancer Risk Prediction Based on Complete Prostate Cancer Family History. *The Prostate, 75*(4), 390-398.

Allott, E. H., Masko, E. M., & Freedland, S. J. (2013). Obesity and Prostate Cancer: Weighing the Evidence. *European Urology, 63*(5), 800-809.

American Cancer Society. (2023, March 1). *Early Detection, Diagnosis, and Staging*. Retrieved March 19, 2023, from American Cancer Society: https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging/survival-rates.html

Barnaby, J., Sorribes, I., & Jain, H. (2021). Relating prostate-specific antigen leakage with vascular tumor growth in a mathematical model of prostate cancer response to androgen deprivation. *Computational and Systems Oncology, 1*(2), e1014.

Berges, R., & Oelke, M. (2011). Age-stratified normal values for prostate volume, PSA, maximum urinary flow rate, IPSS, and other LUTS/BPH indicators in the German male community-dwelling population aged 50 years or older. *World Journal of Urology, 29*(2), 171-178.

Cannarella, R., Condorelli, R. A., Barbagallo, F., La Vignera, S., & Calogero, A. E. (2021). Endocrinology of the Aging Prostate: Current Concepts. *Front Endocrinol (Lausanne), 12*, 554078.

Cebeci, O. Ö., & Ozkan, A. (2021). An evaluation of factors affecting pain during transrectal ultrasonographic prostate biopsy: a real-life scenario in a retrospective cohort study. *PeerJ, 9*, e12144.

D'Amico, A. V., Chang, H., Holupka, E., Renshaw, A., Desjarden, A., Chen, M., . . . Richie, J. P. (1997). Calculated prostate cancer volume: the optimal predictor of actual cancer volume and pathologic stage. *Urology, 49*(3), 385-391.

Huncharek, M., Haddock, K. S., Reid, R., & Kupelnick, B. (2010). Smoking as a Risk Factor for Prostate Cancer: A Meta-Analysis of 24 Prospective Cohort Studies. *American Journal of Public Health, 100*(4), 693-701.

Klingebiel, M., Arsov, C., Ullrich, T., Quentin, M., Al-Monajjed, R., Mally, D., . . . Schimmöller, L. (2022). Data on the detection of clinically significant prostate cancer by magnetic resonance imaging (MRI)-guided targeted and systematic biopsy. *Data in Brief, 45*, 108683.

Leitzmann, M. F., & Rohrmann, S. (2012). Risk factors for the onset of prostatic cancer: age, location, and behavioral correlates. *Clinical Epidemiology, 4*, 1-11.

Lila, H., Ulmert, D., & Vickers, A. J. (2008). Prostate-specific antigen and prostate cancer: prediction, detection and monitoring. *Nat Rev Cancer, 8*(4), 268-278.

Lorenzo, G., Scott, M. A., Tew, K., Hughes, T. J., Zhange, Y. J., Liu, L., . . . Gomez, H. (2016). Tissue-scale, personalized modeling and simulation of prostate cancer growth. *Proceedings of the National Academy of Sciences of the United States of America, 113*(48), E7663-E7671.

NCCN Guidelines. (2022, February 16). Retrieved October 28, 2022, from National Comprehensive Cancer Network: https://www.nccn.org/guidelines/category_2

Nikitina, A., Sharova, E., Danilenko, S., Selezneva, O., Skorodumova, L., Kanygina, A., . . . Generozov, E. (2019). Data on somatic mutations obtained by whole exome sequencing of FFPE tissue samples from Russian patients with prostate cancer. *Data in Brief, 25*, 104022.

Papageorgiou, V., Apalla, Z., Sotiriou, E., Papageorgiou, C., Lazaridou, E., Vakirlis, S., . . . & Lallas, A. (2018). The limitations of dermoscopy: false-positive and false-negative tumours. *Journal of the European Academy of Dermatology and Venereology, 32*(6), 879-888.

Perdana, N. R., Mochtar, C. A., Umbas, R., & Hamid, A. R. (2016). The Risk Factors of Prostate Cancer and Its Prevention: A Literature Review. *The Indonesian Journal of Internal Medicine, 48*(3), 228-238.

Pernar, C. H., Ebot, E. M., Wilson, K. M., & Mucci, L. A. (2018). The Epidemiology of Prostate Cancer. *Cold Spring Harbor Perspectives in Medicine, 8*(12).

Wang, G., Zhao, D., Spring, D. J., & DePinho, R. A. (2018). Genetics and biology of prostate cancer. *Genes & development, 32*(17-18), 1105-1140.

APPENDIX

	Base Run	Scenario 1	Scenario 2	Scenario 3
Tumoral Tissue Volume	0	0.0001	0.0001	0.0001
Healthy Tissue Volume	0.024	0.03	0.024	0.045
PSA Level	1.1	1.2	1.2	2.2
Age	55	60	50	65
BMI	0	1	0	1
Smoking	1	1	0	0
Family History	0	1	0	0
Physical Activity	0	0	1	1
Sexual Activity	1	0	1	0

***Appendix 1**: Initial values for the scenario runs*