Proceedings of the 2021 Winter Simulation Conference
S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, eds.

A SIMULATION MODEL OF BREAST CANCER INCIDENCE, PROGRESSION, DIAGNOSIS AND SURVIVAL IN INDIA

Saumya Gupta
Chandan Mittal
Soham Das
Shaurya Shriyam
Varun Ramamohan

Department of Mechanical Engineering
Indian Institute of Technology Delhi
Hauz Khas
New Delhi 110016, INDIA

Atul Batra

Department of Medical Oncology
All India Institute of Medical Sciences
Ansari Nagar, Aurobindo Marg
New Delhi 110029, INDIA

ABSTRACT

For resource-constrained health systems, it becomes important to evaluate the cost-effectiveness of a breast cancer (BC) screening program prior to its real-world implementation. In this paper, we provide an overview of a new simulation model of BC incidence, progression, and diagnosis developed to maximize the cost-effectiveness of a BC screening program. We describe the development of the modules for BC incidence, tumor growth and diagnosis, and survival estimation of diagnosed patients. Incidence of BC is based on published risk factors and publicly available national cancer registry reports. Survival curves for diagnosed patients based on specific characteristics such as age group and stage of diagnosis are generated using a single published survival curve for the entire population and hazard ratios for each characteristic. Our simulation development approach can provide a model development template for researchers working to develop BC simulation models in other settings with data availability similar to our case.

1 INTRODUCTION & LITERATURE REVIEW

In India, breast cancer (BC) is the most common cancer among women, with an age-adjusted incidence rate (AAR) of 25.8 per 100,000 women, and it is also the leading cause of death from cancer among women in India (Malvia et al. 2017). A key cause for this is late diagnosis, which can be seen from the fact that in India, 45.7% of the patients were diagnosed in advanced stages, compared to 64% of the patients being diagnosed in the local stage itself in the United States (Malvia et al. 2017; ACS 2019).

In this regard, two methods have been suggested to detect BC in early stages: (a) early diagnosis based on symptoms, and (b) screening (WHO 2021). In India, pilot door-to-door BC screening programs were found to be effective, and have yielded the conclusion that with support from the public primary and secondary health infrastructure, screening programs can be a cost-effective option to improve the proportion of cancers detected in the local stage even in low-income settings (Parambil et al. 2019).

In this paper, we describe the development of a novel simulation model of BC incidence, tumor growth and diagnosis, and survival estimation after diagnosis, that we develop to optimize the health and economic impact of a BC screening program in the Indian context. Specifically, using this simulation model, we aim to determine optimal (in terms of maximum cost-effectiveness; i.e., health and economic impact) starting

978-1-6654-3311-2/21/$31.00 ©2021 IEEE

and stopping ages and screening intervals for women using discrete simulation optimization methods such as ranking and selection.

The simulation model that we develop consists primarily of three modules: (a) a BC incidence module; (b) a tumor growth, staging and diagnosis module; and (c) a survival estimation module. Therefore, we focus our literature review on BC simulation models, and specifically in terms of how the above three aspects are handled, where applicable, in the BC simulation literature.

Nearly all published simulation models have been developed for Western contexts. Koleva-Kolarova et al. (2015), in their review of BC simulation models, find that a substantial proportion of the BC simulation literature draws upon seven simulation models. A drawback of using these models to conduct analyses in settings other than those for which they were originally developed is that data for many key parameters specific to the setting under consideration may not be available, and hence the original model parameter estimates may have to be used. We also identified a few simulation models which have been developed outside the framework of these seven simulation models (Ahern et al. 2014; Gray et al. 2017).

With regard to modelling BC incidence, the most common risk factor considered is age. Certain studies modelled the screening of high-risk groups (Román et al. 2019); however, they did not explicitly model risk factor based incidence in a simulated cohort comprising the general population. We identified a few simulation studies where risk factors other than age were included to explicitly model BC incidence (Wu et al. 2013; Tejada et al. 2015).

For tumor growth, two common approaches are use of the exponential distribution for sojourn times or use of the Gompertz growth equation. Most models (Ahern et al. 2014) model tumor growth using the exponential distribution. We identified one study conducted in the Indian context for estimating the cost-effectiveness of BC screening, and the authors have also used the MISCAN Markov model for the natural history of the disease (Okonkwo et al. 2008). A drawback of using Markov chains for modelling tumor progression is that clinical trials that study cancer survival typically publish survival information in the form of survival curves that are rarely exponential, implying that the memoryless property of Markov chains or exponential sojourn times may not be a realistic assumption. Simulation models developed for other types of cost-effectiveness analyses also typically use Markov models for capturing cancer progression (Sun et al. 2019).

The Gompertz growth equation for tumor progression has been used in the studies of Fryback et al. (2006), Mittmann et al. (2015) and Tejada et al. (2015). Tejada et al. (2015) used the Gompertz growth equation to model survival, time of diagnosis and stage assignment using the stage-size relationship from Plevritis et al. (2007). The authors determined lethal tumor sizes and tumor sizes at clinical or mammographic detection on an *a priori* basis, and then determined the time of death and diagnosis using the Gompertz growth equation.

Survival estimation is generally integrated along with tumor growth and disease progression in most BC simulation models, as described for Tejada et al. (2015). We did not identify a study that, similar to our approach, generated survival functions based on risk factors found to be relevant to survival.

Screening is the most common intervention explored for diagnosis of BC in the literature of simulation models. Tejada et al. (2015) incorporate both mammographic and symptomatic (clinical) diagnosis. The authors determined the modal tumor size at clinical detection, and introduced randomness in this tumor size to make the time of detection (estimated from the Gompertz growth equation) stochastic. Further, while Tejada et al. (2015) applied equations from Plevritis et al. (2007) to determine the stage of BC during clinical detection, we applied these equations at every time step to determine the stage of any patient, whether diagnosed or not, based on their tumor size at that point in time.

In summary, the research contributions (RCs) of this work relate to the method of development of the above modules in the presence of limited publicly available information: (RC1) the incidence module was developed based on information from a cross-sectional study conducted to determine risk factors associated with incidence of BC in the Indian capital of New Delhi; (RC2) the annual probability of diagnosis in each stage via methods other than screening (e.g., symptomatic diagnosis) was estimated by combining

estimates of time spent in each stage generated via our tumor growth and cancer stage assignment model in conjunction with cross-sectional study information regarding the proportions of patients diagnosed in each stage; and (RC3) the 'personalized' patient characteristic specific survival estimation module was developed using a single survival curve published for the overall BC population and hazard ratios for survival based on patient characteristics (Gajalakshmi et al. 1997). In particular, our approach towards generating survival functions 'personalized' to individual patient characteristics may in particular be of interest to researchers developing cancer progression models (or other disease progression models using survival curves) in settings where survival data specific to relevant patient characteristics are not available. Similarly, we did not identify another study that described an approach towards estimation of the stage-wise probability of diagnosis in a given time duration similar to our approach.

We now describe the development of the simulation model.

2 SIMULATION MODEL DEVELOPMENT

The simulation is programmed in Python on a computer with 8 gigabytes memory, an Intel *i*7 6^{th} generation processor with a clock speed of 2.5 gigaHertz. The calibration results presented in this section are from 30 replications, each of which required on average 30 minutes of computational runtime. We initialize the model with a population of 100,000 persons (women) aged above 30 years. We do not include persons below the age of 30 years as BC incidence is very low in this age group (NCDIR 2020). The age distribution of the cohort of 100,000 persons at the initialization of the model has been determined based on the census information for women residing in New Delhi (Delhi–Government. 2017). We consider the model applicable for the population of the Indian capital New Delhi, and utilize BC incidence, birth rate and all-cause mortality information specific to New Delhi. The simulation time horizon is fifty years, with the first ten years considered as a warm-up period. We use annual time steps to determine the occurrence of all key events (e.g., births, cancer incidence) in the simulation.

We now describe the disease incidence module.

2.1 Disease Incidence Module

At the start of simulation, no person in the simulated cohort has BC. As the simulation progresses, a set of persons enter the diseased state every year. Given that we geographically situate our model in the New Delhi region of India, age-adjusted incidence rates (AARs, in number of cases per 100,000 persons [women]) were obtained from the incidence data for the Delhi National Capital Region (NCR) published by the National Cancer Registry Programme (NCDIR 2020). The AARs obtained from the NCDIR (2020) report are summarized in Table 1. The proportion of the population in each age group was not reported in the (NCDIR 2020) report; however, this was reported in the previous edition (NCDIR 2016), and hence these parameters were obtained from the 2016 NCRP report. We note here that it is unclear whether the AAR reported in the (NCDIR 2020) report is an annual incidence rate. We assume that this is the case; in the event that it is not, we may overestimate the BC incidence. However, this is an input parameter that can be modified easily by an analyst with incidence data relevant to their analysis.

We also consider other risk factors associated with BC incidence in the Indian context - more specifically, as identified in New Delhi. Multiple cross-sectional studies have been conducted to determine the risk factors associated with BC incidence in the Indian context; however, we identified one relevant study conducted in the Delhi NCR (Pakseresht et al. 2009). Pakseresht et al. (2009) studied sociodemographic and physiological factors associated with BC incidence and find that BC incidence has a significant association (at a 5% level of significance) with BMI, marital status, educational status, occupational status and the number breastfeeding years. The categories associated with each of these risk factors are provided in Table 1, and the odds ratios from Pakseresht et al. (2009) associated with each of these risk factors are also provided in Table 1.

It is likely that the number of breastfeeding years is a proxy factor for parity (the number of children the person has birthed), and hence it is unsurprising that parity, although included by the authors in their study, was not found (marginally) to be a significant risk factor (p-value = 0.06). The rationale underlying the relationship between marital status and employment status and incidence of BC is not entirely clear; however, Pakseresht et al. (2009) discuss that these findings are consistent with other similar studies. We have included these risk factors in our simulation model to be consistent with the policy of considering all statistically significant risk factors for incidence in the simulation model.

Thus, upon instantiation of each person in the simulated cohort, we assign the characteristics listed in Table 1 based on the probabilities in the third column of the table. For example, the 'modal' person would be aged between 30-35 years, have a body mass index (BMI) between 18.5 and 25, live with their spouse, be unemployed, and with their number of breastfeeding years ≤ 6.

The annual risk of incidence of BC, adjusted for each risk factor, is estimated as follows. If p_i denotes the AAR for the i^{th} age group, and m risk factors are considered, with $o_{jk}(j = 1 \text{ to } m,\ k = 1 \text{ to } l_j)$ being the odds ratio for the k^{th} category of the j^{th} risk factor that the person belongs to, then the annual risk of incidence r_i for a person in the simulated cohort is estimated as follows.

$$r_i = p_i \prod_{j=1}^{m} o_{jk} \tag{1}$$

It is important to note that the odds ratios in Table 1 cannot be used as is in equation 1, as applying the odds ratios to p_i would not yield the proportions of persons who are diagnosed with BC that also belong to a particular risk factor category. This is because the corresponding proportions in Table 1 are those associated with the entire cohort considered in the study, and not those associated with persons diagnosed with BC. For example, in Pakseresht et al. (2009), we see that approximately 7% of persons diagnosed with BC also have BMI < 18.5. This implies that our simulation model must also replicate this - that is, the annual probability of incidence r_i must be calibrated so that approximately 7% of persons diagnosed with BC in the simulation have a BMI < 18.5.

We calibrate the r_i by manually varying the odds ratios o_{jk} until the proportions of patients belonging to a risk factor category approach those observed in Pakseresht et al. (2009). The results of the calibration process, including the modified odds ratios for the incidence risk factor characteristics, are provided in Table 2. In all cases, we see that the proportion of BC patients possessing a given risk factor characteristic as observed in Pakseresht et al. (2009) is within the 95% confidence interval (CI) associated with the corresponding estimate from the simulation.

2.2 Tumor Growth, Staging, and Diagnosis Module

We now describe the tumor growth and BC staging module, and then describe the process for estimating the probability of diagnosis in each stage.

At each time step, the probability of a person developing BC is given in (1). If it is determined that a person will develop BC, then the tumor growth process initializes for this person. We assume that tumor begins as a single cell. At every time step, the tumor grows according to the Gompertz tumor growth equation (Norton 1988). This equation expresses the size of the tumor at a given time point in terms of the number of cells. The size of the tumor is estimated using the volume of a single cell, which is assumed to be constant (Plevritis et al. 2007). For two time points t_1 and t_2, where $t_2 > t_1$, the number of cells $N(t_2)$ as a function of $N(t_1)$ is given by the Gompertz tumor growth equation (Norton 1988) as:

$$\begin{aligned} &N(t_2) = N(t_1) \times \exp\left(k(1 - \exp(-b(t_2 - t_1)))\right), \text{ and} \\ &k = \ln\left(\frac{N_{inf}}{N(t_1)}\right) \end{aligned} \tag{2}$$

Table 1: Breast cancer incidence related input parameters.

Patient characteristic		Proportion of population (%)	AAR/Odds ratio
Age Group	30-35 yrs	8.73	16.0
	35 -40 yrs	8.00	32.8
	40-45 yrs	6.61	56.5
	45-50 yrs	5.47	83.4
	50-55 yrs	4.12	115.0
	55-60 yrs	3.18	131.0
	60-65 yrs	3.11	135.0
	65-70 yrs	1.73	155.2
	70-75 yrs	1.24	155.4
	$>$ 75 yrs	1.45	122.2
Body mass index	$<$ 18.5	9.9	1
	18.5-24.99	53.9	1.44
	25-29.99	28.6	1.81
	$\geq$ 30	7.6	4.54
Marital status	Living with spouse	80.2	0.38
	Not living with spouse	19.8	1
Employment status	Employed	8.8	0.28
	Not employed	91.2	1
Number of breastfeeding years	$\leq$ 6 years	53.4	1.91
	$>$ 6 years	46.6	1

Notes: the proportions of the population for the age group characteristic are from the NCDIR (2016) report, and those for the risk factors are from Pakseresht et al. (2009). The parameter estimates associated with the age group patient characteristic in the fourth column are all AARs from NCDIR (2020) report, and those associated with the other patient characteristics are odds ratios.

Here b is the tumor growth rate parameter, with a lognormal distribution (Norton 1988), and N_{inf} is the maximum tumor size, with a mean value of 3.1×10^{12} (Plevritis et al. 2007). Following the approach in Tejada et al. (2015), we make the tumor growth parameter dependent on age by assigning the 25^{th} percentile of the original lognormal distribution as the expected value of b for persons aged 75 years and above, and by assigning the 75^{th} percentile of the original lognormal distribution as the expected value of b for a 25-year old person. For the rest of the cohort (between the ages of 25 years and 75 years), the expected value of b changes linearly between these two values. This is done to incorporate the clinical fact that the rate of BC progression is inversely proportional to the age of the patient, as tumor growth is significantly faster in younger persons when compared to older persons. We assume that the standard deviation of the tumor growth parameter is 0.05 times the mean value per the assumption in Tejada et al. (2015), and then sample b from a lognormal distribution with these mean and standard deviation values for each person in the simulated cohort determined to develop BC. N_{inf} is also assumed to follow a normal distribution with mean and standard deviation per the approach in Tejada et al. (2015) (coefficient of variation = 0.05).

$N(t_2)$ can thus be estimated at any time point t_2 using the k and b values assigned to a particular person, and if t_1 is taken as the time of incidence, and by setting $N(t_1) = 1$. The tumor density D was obtained

Table 2: Breast cancer incidence module: calibration outcomes.

Incidence risk factor		Calibrated odds ratios	Prop. of patients: simulation	Prop. of patients: observed
Body-mass index	< 18.5	0.75	6.76 (0.39)	7.00
	18.5 - 24.99	1.00	50.23 (0.82)	49.60
	25 - 29.99	1.15	30.36 (0.59)	30.40
	≥ 30	1.78	12.64	13.00
Marital status	Living with spouse	1.00	69.58 (1.40)	69.60
	Not living with spouse	1.80	30.42	30.40
Occupational status	Employed	0.37	3.66 (0.20)	3.50
	Not employed	1.00	96.33	96.50
Number of breastfeeding years	≤ 6 years	1.00	63.25 (0.65)	63.70
	> 6 years	0.65	36.74	36.30

Notes: Prop. = proportion; the 'observed' proportion of patients with a given incidence risk factor category are obtained from Pakseresht et al. (2009).

to be as 2.38732×10^5 cells/mm^3 (Plevritis et al. 2007), and thus the tumor volume at time t is given by $v(t) = N(t)/D$. The initial volume v_0 of the tumor (i.e., a single cell), was determined using an estimate of the diameter of a single tumor cell - 2×10^{-2} mm - and assuming a spherical shape of the cell (implying that tumor volume and diameter are related as $v(t) = d(t)^3\pi/6$). Note that while Tejada et al. (2015) use the tumor diameter in determining the BC stage, we perform the stage assignment in a different manner, which we describe below.

2.2.1 Disease Staging and Diagnosis

As the tumor grows at every time step, we update the stage of the cancer at every time step based on the conditional probability of the cancer attaining a particular stage given the size of the tumor. We consider three stages of the cancer: local, regional and distant. We consider this staging system because longitudinal survival data in the Indian context is available for this staging system (Gajalakshmi et al. 1997). Staging systems generally define cancer stages in terms of both tumor size and extent of spread into lymph nodes and other regions; however, Plevritis et al. (2007) determine the cancer stage based solely on the tumor size $v(t)$, and define the probability of being in any stage as follows.

$$P(\text{local} \mid v(t) = v) = \left(\frac{\beta + \gamma(v - v_o)}{\beta + (\eta + \gamma)(v - v_o)}\right)^{\alpha+1} \tag{3}$$

$$P(\text{regional} \mid v(t) = v) = \left(\frac{\eta}{\eta - \omega}\right)\left\{\left(\frac{\beta + \gamma(v - v_o)}{\beta + (\omega + \gamma)(v - v_o)}\right)^{\alpha+1} - \left(\frac{\beta + \gamma(v - v_o)}{\beta + (\eta + \gamma)(v - v_o)}\right)^{\alpha+1}\right\} \tag{4}$$

$$P(\text{distant} \mid v(t) = v) = 1 - P(\text{regional} \mid v(t) = v) - P(\text{local} \mid v(t) = v) \tag{5}$$

In the above equations, $v(t)$ is determined from $N(t)$, which in turn is determined from (2), where t is the time from incidence. The model of disease progression as given in the above equations is based on the work in Plevritis et al. (2007); however, we describe them in the form given in Tejada et al. (2015). We refer the reader to Plevritis et al. (2007) for estimates of the parameters $\alpha, \beta, \gamma, \omega,$ and η.

Upon disease incidence, the cancer is assumed begin in the local stage. Subsequently, at every time step, the probabilities of being in different stages are updated based on the tumor size at that time point. Once the cancer reaches the regional stage, the probability of being in local stage is set to zero. Further calculations of probabilities do not involve the local stage. Similarly, once the cancer reaches the distant stage, it cannot revert to the local or regional stages, and the probabilities for stage assignment are no longer computed for the patient. We now describe how we simulate clinical detection of BC, which can either occur when a patient experiences symptoms, or during a routine health examination that is not part of a targeted screening program.

In order to determine the annual probability of diagnosis by stage for a BC patient, we worked backwards from the proportions of patients diagnosed in each stage, as published in the NCDIR (2020) report. These proportions are 29%, 57%, and 10.30%. It is unclear as to whether the remaining 3.7% are not diagnosed at all or whether their stage is unknown, and because this proportion is relatively small, we ignore it in our subsequent analysis. Using these proportions, we estimate the annual probability of being diagnosed in each stage as follows. If T_l represents the average number of years spent without treatment in, say, the local stage, then we assume that getting diagnosed in this stage can be modeled by a geometric random variable - that is, determining whether a BC patient gets diagnosed in each annual time step represents a trial, and the number of trials until diagnosis in the stage is the geometric random variable. If we denote the overall probabilities of diagnosis in the local, regional, and distant stages as P_l, P_r and P_d, and the corresponding annual probabilities of diagnosis as a_l, a_r and a_d, then the latter set of probabilities can be estimated by solving the following equations:

$$P_l = 0.29 = 1-(1-a_l)^{T_l}; \quad \frac{P_r}{1-P_l} = \frac{0.57}{0.71} = 1-(1-a_r)^{T_r}; \quad \frac{P_d}{1-P_l-P_r} = \frac{0.103}{0.14} = 1-(1-a_d)^{T_d} \tag{6}$$

Note that the left hand sides (LHS) of the equations for estimating a_r and a_l are conditional probabilities - for example, the LHS of the equation for estimating a_r represents the probability that a patient is diagnosed in the regional stage given that they were not diagnosed in the local stage. In order to solve (6), we require the values of T_l, T_r, and T_d. Given the ethical concerns in conducting trials involving untreated cancer patients, we could not find literature regarding the time spent in each stage for untreated BC patients. However, we were able to find a single study published in 1962 that retrospectively estimated the survival of untreated BC patients presenting with distant stage cancer (Bloom et al. 1962). The authors estimate a median survival of 2.7 years for these patients, and hence we assumed that T_d = 2.7 years.

We conducted a separate Monte Carlo simulation of the untreated progression of BC using (3) - (5) for estimating T_l and T_r. A single replication of this simulation involved initializing a BC patient at time $t = 0$ (in the local stage), simulating tumor growth and stage assignment until the cancer attained the distant stage, and recording their times spent in the local and regional states. This process is repeated to estimate T_l and T_r, which are then used to solve (6) to estimate a_l, a_r, and a_d.

The proportions diagnosed in the local, regional, and distant stages, as estimated from the simulation, are 30.24% (standard deviation [SD] = 0.52%), 58.32% (SD = 0.47%), and 11.42% (SD = 0.33%), respectively. We remind readers that the NCDIR (2020) report estimates are 29.0%, 57.0%, and 10.30%, respectively. Given the lack of information regarding uncertainty around the estimates from the NCDIR (2020) report, it is difficult to comment on the statistical significance of the differences between the simulation outcomes and those from the NCDIR (2020) report; however, it is evident the simulation outcomes are reasonably close.

We now describe the implementation of the diagnosis BC via screening alongside clinical diagnosis. At the start of the simulation, the screening program is initialized with parameters such as the time interval

between screening programs (e.g., 1 year, 2 years), the groups of persons eligible for screening (e.g., persons aged between 35-64 years and with BMI $\geq$ 30), the sensitivity and specificity of the screening technique, and the compliance rate (proportion of the eligible persons who voluntarily undergo screening). The screening program itself is envisioned as a drive being conducted by public health authorities at regular time intervals wherein persons (women) are invited to undergo screening at designated facilities.

In the simulation, screening is implemented as follows. As an example, we consider a screening program conducted every 2 years, with persons aged 35-44 being eligible for the program. Under this program, at every alternate time step, a certain proportion of persons in the eligible group (determined by the program compliance parameter) undergo screening. Persons undergo screening until they are no longer eligible; for example, a 41 year old person at the start of the screening program in the simulation could undergo screening twice, at 41 years and 43 years of age. After every positive detection, we assume that a biopsy is done for the confirmation of the diagnosis and staging of the cancer, and hence false positives only incur the costs of screening and biopsy, and not of treatment.

2.3 Survival Estimation Module

Once a person is diagnosed with BC, we determine whether they survive or not at each time step. For undiagnosed patients, all-cause mortality applicable to the general population is applied.

We estimate the survival of persons with a BC diagnosis using survival functions generated based on specific characteristics determined to affect their survival. These 'personalized' survival functions were generated by modifying a single published survival curve constructed for a cohort of BC patients with a mix of these characteristics. A survival function $S_T(t)$ provides the probability of survival at time t. That is, $S_T(t) = P(T \geq t)$, where T is a random variable representing the time to death. Thus if $F_T(t)$ represents the CDF of the time to death random variable (with PDF $f_T(t)$), then $S_T(t) = 1 - F_T(t)$.

We identified only one study in the Indian context that presents a survival curve for BC patients - Gajalakshmi et al. (1997). Gajalakshmi et al. (1997) accessed the survival data of BC patients from the year 1982 to the year 1989, and estimated that the 5-year average survival of BC patients is 47.5%. They also determined that the key factors which determine the survival of BC patients are age at diagnosis, level of education, marital or relationship status, and the stage or clinical extent of disease. The authors provided the hazard ratios corresponding to the different characteristic groups under each risk factor. As discussed in Gajalakshmi et al. (1997), the link between educational status and marital status with BC survival is likely due to the higher socioeconomic class associated with better educational qualifications and marriage, which in turn may mean greater awareness and better access to treatment. From the perspective of calibration of the survival estimation module, we focus primarily on physiological risk factors, which include the age group and cancer stage. Hence, we do not list the information associated with all the risk factors in Table 3, and only provide information regarding the risk factors that we consider in the calibration of the survival estimation module - the age group and cancer stage risk factors.

Using the survival curve published for the entire cohort and the hazard ratios provided in Gajalakshmi et al. (1997), we generated survival functions for each simulated BC patient based on the survival risk factor categories to which they are assigned. The base survival curve was obtained directly from Gajalakshmi et al. (1997) using the software DigitizeIt (www.digitizeit.de). We assume that the published survival curve in applies to a modal person whose characteristics are most common for the survival-related risk factors among the cohort considered in Gajalakshmi et al. (1997). Thus the modal person - from a survival estimation standpoint - is assumed to be aged between 45-54 years, married, not educated, and diagnosed with regional stage BC. The survival function obtained for this modal person is then modified using hazard ratios based on those published in Gajalakshmi et al. (1997) to generate the personalized survival curves for each simulated BC patient with a diagnosis.

Accomplishing this involved assuming the hazard function was related to the risk factor variables via the Cox proportional hazards model (Cox 1972). The hazard function, which models the rate at which the

survival probability is changing with time given that a patient has survived until time t, is the ratio between the probability density function of T and the survival function: $\lambda(t|X_i) = \frac{f_T(t|X_i)}{S_T(t|X_i)}$.

The Cox proportional hazards model is given by the equation below.

$$\lambda(t|X_i) = \lambda_0(t) \times \exp(X_{i1}\beta_1 + \cdots + X_{ip}\beta_p) = \lambda_0(t) \times \exp(X_i^T \beta) \tag{7}$$

In (7), $\lambda(t|X_i)$ represents the hazard function for the i^{th} patient, whose p risk factor characteristics are encoded by the vector $X_i = (X_{i1}, X_{i2}, \cdots, X_{ip})^T$. The coefficients of the X_i, $\beta = (\beta_1, \beta_2, \cdots, \beta_p)$, represent the hazard ratios associated with each risk factor level, and p is the total number of characteristics across all factors determining survival. We note here that p is not the number of risk factors, but rather the total number of risk factor characteristics across all risk factors. For example, if we consider M $(m = 1 - M)$ risk factors and each risk factor had n_m characteristic levels (e.g., the three levels corresponding to tumor stage), then $p = \sum_{m=1}^{M} n_m$. The X_{ij} $(j = 1 - p)$ are binary variables $\in \{0, 1\}$, and are set to 1 if the i^{th} patient is assigned the j^{th} characteristic (i.e., all the other X_{ij} corresponding to that risk factor are set to 0 for this patient).

Patient-specific survival functions are generated using the fact that $S_T(t|X_i) = 1 - F_T(t|X_i)$ and $f_T(t|X_i)$ is the pdf of T. This implies $\lambda(t|X_i)$ can be written as: $\lambda(t|X_i) = -\frac{d}{dt}\log(S_T(t|X_i))$.

Now, if we denote the hazard function representing the modal person (with characteristics given by, say, X_m) as $\lambda_m(t|X_m) = \lambda_0(t) \times \exp(X_m^T \beta)$. Note that we have $S(t|X_M)$ from Gajalakshmi et al. (1997), and hence $\lambda_m(t|X_m)$ at a given time t can be extracted using the following finite difference equation:

$$\lambda(t|X_m) = -\frac{\log(S_T(t|X_m)) - \log(S_T(t - \delta|X_m))}{\delta t} \tag{8}$$

The value of δt is taken as 0.01 time units (years). Note that at $t = 0$, $S_T(t|X_m) = 1$. From the value of $S_T(t|X_m)$, we require the hazard function $\lambda(t|X_i)$ for the i^{th} patient. We obtain this by exploiting the fact that $\lambda(t|X_i) = \lambda(t_0) \times \exp(\beta^T X_i)$, and that we know $\exp(\beta^T X_m)$. This can then be utilized in the following manner to obtain $\lambda(t|X_i)$: $\frac{\lambda(t|X_i)}{\lambda(t|X_m)} = \frac{e^{X_i\beta}}{e^{X_m\beta}}$.

Once $\lambda(t|X_i)$ is estimated as above, we can estimate $S_T(t|X_i)$ by performing the following numerical integration: $S_T(t|X_i) = \exp(-\int_0^t \lambda(t|X_i)dt)$. Note that at $t = 0$, $\log(S_T(t|X_i)) = 0$.

2.4 Survival Module Calibration

A key concern regarding the validation of survival estimation module involves its development based on survival information obtained from a study published in 1997. 5-year mean survival of BC has improved from 47.5% as published in Gajalakshmi et al. (1997) to 66.1% as published in a significantly more recent study (95% CI: 51.5%-80.8%) (Allemani et al. 2018). Therefore validation of the survival estimates generated from the simulation model by comparison with corresponding survival estimates in Gajalakshmi et al. (1997) is not adequate, as both the modeled survival functions as well as calibration targets require updating. Gajalakshmi et al. (1997) publish average five-year survival for each of their risk factor characteristics, each of which can serve as a calibration target. However, because the reasons for the association between educational status and marital status and survival are not fully understood, we chose only the 5-year survival estimates associated with the physiological risk factor characteristics - i.e., age group and tumor stage based characteristics - as calibration targets.

The calibration targets based on age group and clinical stage were estimated using the ratio of the 5-year overall survival published in Allemani et al. (2018) and that from Gajalakshmi et al. (1997). The

5-year mean survival estimate for each age group and tumor stage based risk factor characteristic was multiplied by this ratio; for example, the observed 5-year survival for BC patients aged 35-44 was recorded as 57.4% in Gajalakshmi et al. (1997), yielding an updated calibration target of approximately 79.7%.

Once the calibration targets are estimated, we calibrate the model by varying the relevant hazard ratios (i.e., age group and tumor stage based hazard ratios) manually until the 5-year survival outcomes from the simulation model all attained less than 15% deviation from their calibration targets. However, we placed greater priority on achieving a simulation estimate of the 5-year survival for the entire cohort that was as close as possible to the corresponding calibration target of 66.1%, given that it was the only calibration target obtained directly from external sources. We consider this termination criterion for the calibration process reasonable given that the uncertainty around the overall cohort 5-year survival estimate is considerable - for example, the deviation of the lower limit of its 95% CI is approximately 18.2% from the mean. The hazard ratios from Gajalakshmi et al. (1997), the calibrated hazard ratios, the updated calibration targets, and the simulation survival outcomes are provided in Table 3.

Table 3: Survival estimation module information: hazard ratios and survival outcome comparison.

Survival risk-factor characteristic	Hazard ratios		5-year survival proportion (%)	
	Observed	Calibrated	Calibration target	Simulation estimate
Age at diagnosis (years)				
30-34	1.00	0.47	90.12	94.48 (1.25)
35-44	1.45	2.42	79.70	86.99 (0.68)
45-54	1.67	3.19	69.57	76.31 (0.97)
55-64	2.12	3.93	54.57	62.33 (0.87)
65-74	2.50	4.45	40.27	44.95 (1.52)
≥ 75	4.75	5.80	13.46	14.44 (1.39)
Cancer stage				
Local	1.00	0.36	88.31	81.96 (1.07)
Regional	1.32	1.35	72.76	65.85 (0.72)
Status	3.19	2.90	27.00	25.94 (1.57)
Entire cohort	–		66.10 (51.5 - 80.8)	66.16 (0.62)

Notes: the observed hazard ratios are those provided in Gajalakshmi et al. (1997).

Given that we have presented the simulation modules separately, we briefly discuss the overall set of outcomes generated by the simulation. The outputs of interest from the simulation will depend upon the analysis in question. If the analysis in question is evaluating the cost-effectiveness of a new treatment in comparison to the standard of care, then the relevant outputs would be the medical costs associated with the new treatment and the standard of care, respectively, and the corresponding average life years per BC case associated with each treatment.

From the standpoint of evaluating a screening policy, the key outputs would be the proportion diagnosed in each stage (e.g., the outputs in Section 2.2.1), and the total life years and costs (the sum of the individual survival duration - life years - and medical costs associated with each BC case in the simulated cohort). Thus, an effective screening policy would detect a greater proportion of BC cases in the less advanced (e.g., local) stages, thereby yielding an increase in life years and decrease in costs when compared to a less effective or nonexistent screening policy. These outcomes - total costs and life years - are integrated into a single health economic quantity referred to as the monetary benefit associated with a given screening policy. The difference in the monetary benefits associated with a given screening policy and the status quo

(e.g., no screening) is estimated as the 'net monetary benefit', and a positive net monetary benefit implies that the screening policy under consideration is cost-effective when compared to the status quo. Note that we also consider the willingness of the target population to undergo screening via the incorporation of a compliance parameter in the simulation.

3 DISCUSSION & CONCLUSIONS

In this study, we present a novel simulation model for BC incidence, tumor growth and staging, diagnosis, and survival estimation. We have demonstrated the calibration of each module of the simulation model to external data specific to the Indian context. We develop the model for the Indian context, and hence the approach we have adopted to model incidence, staging, stage-wise clinical diagnosis, and post-diagnosis survival has been driven by the amount and type of data available that is specific to the Indian context. However, as mentioned in Section 1, the approaches we have developed for modeling stage-wise diagnosis and post-diagnosis survival estimation may be useful for researchers/analysts working in settings where data availability is similar to our case.

A limitation of the study involves the lack of age group specific and cancer stage specific survival estimates that are more recently published than those in Gajalakshmi et al. (1997). Even though the National Cancer Registry reports (NCDIR 2020) provide comprehensive cancer incidence information, they do not publish the survival curves that we require for our simulation. Further, a study regarding survival among locally advanced BC in India has been published (Dhanushkodi et al. 2021); however, it does not provide information regarding survival among patients with other BC stages. We also note that our approach towards calibration is a manual approach, which can be computationally inefficient. Therefore, we are exploring the use of simulation optimization approaches to speed up the calibration process.

This simulation model was developed in order to optimize - via discrete simulation optimization methods such as ranking and selection - various aspects of a BC screening program in the Indian context, and thus an immediate avenue of future research involves completing this objective. However, the model may also be adapted for other purposes, such as for evaluating the cost-effectiveness of new treatments or diagnostic interventions.

REFERENCES

ACS 2019. "Breast Cancer Facts and Figures 2019-20, American Cancer Society, Atlanta, GA.". https://bit.ly/3fWD9nI, accessed 8th March 2021.

Ahern, C., Y. T. Shih, W. Dong, G. Parmigiani, and Y. Shen. 2014. "Cost-effectiveness of Alternative Strategies for Integrating MRI into Breast Cancer Screening for Women at High Risk". *British Journal of Cancer* 111(8):1542–1551.

Allemani, C., T. Matsuda, V. Di Carlo, R. Harewood, M. Matz, M. Nikšić, A. Bonaventure, M. Valkov, C. J. Johnson, J. Estève et al. 2018. "Global Surveillance of Trends in Cancer Survival 2000–14 (CONCORD-3): Analysis of Individual Records for 37,513,025 Patients Diagnosed with One of 18 Cancers from 322 Population-based Registries in 71 countries". *The Lancet* 391(10125):1023–1075.

Bloom, H. J. G., W. Richardson, and E. Harries. 1962. "Natural History of Untreated Breast Cancer (1805-1933)". *British Medical Journal* 2(5299):213.

Cox, D. R. 1972. "Regression Models and Life-tables". *Journal of the Royal Statistical Society: Series B (Methodological)* 34(2):187–202.

Delhi–Government. 2017. "Annual Report on Registration of Births and Deaths in Delhi-2017". https://bit.ly/3rXHStg, accessed 24th February, 2021.

Dhanushkodi, M., V. Sridevi, V. Shanta, R. Rama, R. Swaminathan, G. Selvaluxmy, and T. S. Ganesan. 2021. "Locally Advanced Breast Cancer (LABC): Real-World Outcome of Patients From Cancer Institute, Chennai". *JCO Global Oncology* 7:767–781.

Fryback, D. G., N. K. Stout, M. A. Rosenberg, A. Trentham-Dietz, V. Kuruchittham, and P. L. Remington. 2006. "Chapter 7: The Wisconsin Breast Cancer Epidemiology Simulation Model". *JNCI Monographs* 2006(36):37–47.

Gajalakshmi, C., V. Shanta, R. Swaminathan, R. Sankaranarayanan, and R. Black. 1997. "A Population-based Survival Study on Female Breast Cancer in Madras, India". *British Journal of Cancer* 75(5):771–775.

Gray, E., A. Donten, N. Karssemeijer, C. van Gils, D. G. Evans, S. Astley, and K. Payne. 2017. "Evaluation of a Stratified National Breast Screening Program in the United Kingdom: an Early Model-based cost-effectiveness analysis". *Value in Health* 20(8):1100–1109.

Koleva-Kolarova, R. G., Z. Zhan, M. J. Greuter, T. L. Feenstra, and G. H. De Bock. 2015. "Simulation Models in Population Breast Cancer Screening: a Systematic Review". *The Breast* 24(4):354–363.

Malvia, S., S. A. Bagadi, U. S. Dubey, and S. Saxena. 2017. "Epidemiology of Breast Cancer in Indian Women". *Asia-Pacific Journal of Clinical Oncology* 13(4):289–295.

Mittmann, N., N. K. Stout, P. Lee, A. N. Tosteson, A. Trentham-Dietz, O. Alagoz, and M. J. Yaffe. 2015. "Total Cost-effectiveness of Mammography Screening Strategies". *Health Reports* 26(12):16.

NCDIR 2016. "Three-Year Report of Population Based Cancer Registries 2012-2014". https://bit.ly/2U1Fsxg, accessed 20th March 2021.

NCDIR 2020. "Report of National Cancer Registry Programme (ICMR-NCDIR), Bengaluru, India 2020". https://www.ncdirindia.org/All_Reports/Report_2020/default.aspx, accessed 23rd March 2021.

Norton, L. 1988. "A Gompertzian Model of Human Breast Cancer Growth". *Cancer research* 48(24 Part 1):7067–7071.

Okonkwo, Q. L., G. Draisma, A. der Kinderen, M. L. Brown, and H. J. de Koning. 2008. "Breast Cancer Screening Policies in Developing Countries: a Cost-effectiveness Analysis for India". *JNCI: Journal of the National Cancer Institute* 100(18):1290–1300.

Pakseresht, S., G. Ingle, A. Bahadur, V. Ramteke, M. Singh, S. Garg, P. Agarwal et al. 2009. "Risk Factors with Breast Cancer among Women in Delhi". *Indian Journal of Cancer* 46(2):132.

Parambil, N. A., S. Philip, J. P. Tripathy, P. M. Philip, K. Duraisamy, S. Balasubramanian et al. 2019. "Community Engaged Breast Cancer Screening Program in Kannur District, Kerala, India: A Ray of Hope for Early Diagnosis and Treatment". *Indian Journal of Cancer* 56(3):222.

Plevritis, S. K., P. Salzman, B. M. Sigal, and P. W. Glynn. 2007. "A Natural History Model of Stage Progression Applied to Breast Cancer". *Statistics in Medicine* 26(3):581–595.

Román, M., M. Sala, L. Domingo, M. Posso, J. Louro, and X. Castells. 2019. "Personalized Breast Cancer Screening Strategies: A Systematic Review and Quality Assessment". *PLoS ONE* 14(12):e0226352.

Sun, L., Z. Sadique, I. dos Santos-Silva, L. Yang, and R. Legood. 2019. "Cost-effectiveness of Breast Cancer Screening Programme for Women in Rural China". *International Journal of Cancer* 144(10):2596–2604.

Tejada, J. J., J. S. Ivy, J. R. Wilson, M. J. Ballan, K. M. Diehl, and B. C. Yankaskas. 2015. "Combined DES/SD model of Breast Cancer Screening for Older Women, I: Natural-history simulation". *IIE Transactions* 47(6):600–619.

WHO 2021. "Breast Cancer: Prevention and Control, World Health Organization, Geneva, Switzerland". https://www.who.int/cancer/detection/breastcancer/en/, accessed 2nd March 2021.

Wu, Y., M. Yen, C. Yu, and H. Chen. 2013. "Individually Tailored Screening of Breast Cancer with Genes, Tumour Phenotypes, Clinical Attributes, and Conventional Risk Factors". *British Journal of Cancer* 108(11):2241–2249.

AUTHOR BIOGRAPHIES

SAUMYA GUPTA is a senior undergraduate in the Department of Mechanical Engineering at the Indian Institute of Technology Delhi. His email address is ME1160689@mech.iitd.ac.in.

CHANDAN MITTAL is a senior undergraduate in the Department of Mechanical Engineering at the Indian Institute of Technology Delhi. His email address is chandanmittal101@gmail.com.

SOHAM DAS is a PhD student in the Department of Mechanical Engineering at the Indian Institute of Technology Delhi. His email address is 2019MEZ8426@mech.iitd.ac.in.

SHAURYA SHRIYAM is an assistant professor in the Department of Mechanical Engineering at the Indian Institute of Technology Delhi. His email address is shriyam@mech.iitd.ac.in.

ATUL BATRA is an assistant professor in the Department of Medical Oncology at the All India Institute of Medical Sciences at New Delhi. His email address is batraatul85@gmail.com.

VARUN RAMAMOHAN is an assistant professor in the Department of Mechanical Engineering at the Indian Institute of Technology Delhi. His email address is varunr@mech.iitd.ac.in.