Calibration, Validation, and Analysis of a Combined DES/SD Model of Breast Cancer Screening for Older Women

JEREMY J. TEJADA1
MATT BALLAN2
JULIE IVY3
RUSSELL KING4
JAMES R. WILSON5

Edward P. Fitts Department of Industrial and Systems Engineering,
North Carolina State University,
Campus Box 7906,
Raleigh, NC 27695-7906, USA

1 Member, Institute of Industrial Engineers, and corresponding author. E-mail: jjtejada@ncsu.edu. Telephone: (603) 425-8561.
2 Member, Institute of Industrial Engineers. E-mail: mjballan@ncsu.edu. Telephone: (561) 706-1792.
3 Member, Institute of Industrial Engineers. E-mail: jsivy@ncsu.edu. Telephone: (919) 513-1683.
4 Member, Institute of Industrial Engineers. E-mail: king@ncsu.edu. Telephone: (919) 515-5186.
5 Fellow, Institute of Industrial Engineers. E-mail: jwilson@ncsu.edu. Telephone: (919) 515-6415.
In a follow up article to our two companion articles, we discuss the calibration and validation of our simulation models for screening and treatment of breast cancer in US women aged 65 and older. In the first article we presented a natural history model of breast cancer incidence and progression in randomly sampled individuals from the designated population of older US women, thereby generating a database of these women whose (untreated) cancer history is completely known. In the second article, we discuss the screening-and-treatment simulation, which is composed of a discrete-event simulation (DES) submodel and a system dynamics (SD) submodel that run simultaneously and exchange information to provide a framework for evaluating the complex outcomes of alternative breast cancer screening policies. In this article we discuss a new method for validating a large-scale stochastic multi-response simulation model with respect to numerous disparate sources of noisy real-world data, we discuss our method for calibrating the simulation to historical data during the warm-up period 2001–2010, and we present a comprehensive analysis of the results of our simulations over the time horizon 2012–2020 so as to predict the relative performance of different breast cancer screening policies. We present the most important results and conclusions, including an evidence-based recommendation of screening older women annually up to age 80, and a detailed analysis of the results of using that screening policy in the future.

Keywords: Health care, breast cancer, screening, medical decision making, discrete-event simulation, system dynamics, validation, calibration, combining t-tests

1. Introduction

In a previous sequence of two articles, we developed simulation models for analyzing the occurrence, screening, and treatment of breast cancer in the growing population of US women of age 65 and above. In the first article (Tejada et al., 2013a), we discuss a natural-history simulation of breast cancer incidence and progression in randomly sampled individuals from the designated population of older US women. The natural-history model is a discrete-event simulation (DES) that contains a population growth submodel as well as breast cancer incidence, progression, and survival submodels. The primary output of the natural-history simulation is a database of older women whose (untreated) breast cancer histories are known; and these histories are critical inputs to the screening-and-treatment simulation, the focus of the second article in the sequence (Tejada et al., 2013b). The screening-and-treatment simulation integrates DES and system dynamics (SD) modeling techniques into a single simulation so as to handle the following simultaneously: (i) the screening and treatment decisions and the resulting progression of health states for each individual
in the simulated population; and (ii) the population-level state variables (stocks) and their associated rates of change (flows) that govern the overall operation of the US system for detecting and treating breast cancer.

The focus of the two-article sequence was to develop and exploit a simulation modeling framework for evaluating the effectiveness of breast cancer screening policies for US women who are at least 65 years old. The integrated DES/SD simulation modeling environment provides a flexible tool for evaluating the effectiveness of a wide range of population-level screening policies that can be compared directly on individual women who are representative of the designated population.

In this follow-up to the companion articles on the natural history simulation and the screening-and-treatment simulation, we focus on the calibration, validation, and analysis of the two simulations (Tejada et al., 2012a, Tejada et al., 2012b). In particular we will discuss the following: (i) the formulation of and justification for a new method for validating a large-scale stochastic multi-response simulation model with respect to numerous disparate sources of noisy real-world data; (ii) our method for calibrating the screening-and-treatment simulation to historical data during the warm-up period 2001–2010 using the aforementioned validation procedure; (iii) a comprehensive analysis of the results of our two simulations over the time horizon 2012–2020 so as to predict the relative performance of different breast cancer screening policies for US women of age 65 and above; and (iv) the formulation of recommendations for future work and for future breast cancer screening-and-treatment policies for older women.

The remainder of this article is organized as follows. Section 2 provides background information on, the lack of well-established screening guidelines for women ages 65 and older, the nature of the output of our simulation, and limitations of previous approaches to validation of stochastic multi-response simulations. Section 3 provides a detailed formulation and justification for our new method for validating multi-response simulation models. Section 4 describes our application of the validation procedure presented in Section 3 to calibration of the screening-and-treatment simulation to actual data. This section also contains the results of using the calibration and validation procedures and a discussion of those results. Section 5 presents a detailed set of results from the screening-and-treatment simulation over the time horizon 2012–2020 so as to predict the relative performance of different breast cancer screening policies. This section also includes a comprehensive set of results for the recommended screening policy, annual screening stopping at age 80. Section 6 presents the conclusions and summarizes the article. Tejada (2012) contains a more-detailed discussion of the entire simulation project. The natural history model is fully discussed in Tejada et al. (2012a), and the screening-and-treatment simulation is fully discussed in Tejada et al. (2012b).

2. Background

2.1. Lack of well-established screening guidelines for older women

There are no well-established breast cancer screening guidelines for women at least 65 years old (Mandelblatt et al., 2009; National Cancer Institute, 2009; Nelson et al., 2009; Resnick and McLeskey,
The current guidelines for women of age 65 and above are limited and conflicting. Furthermore, clinical trials for breast cancer screening have generally not included women who are at least 65 years old; and clinicians do not anticipate any clinical trials specific to breast cancer screening in the future (Badgwell et al., 2008; Crivellari et al., 2007; US Preventive Services Task Force, 2009).

Table 1 summarizes current mammography screening guidelines from a variety of sources, including the American Cancer Society (ACS), American Medical Association (AMA), and US Preventive Services Task Force (USPSTF). This table shows that well-defined screening guidelines have not been developed for older woman by any of the public medical authorities.

**Table 1. Mammography Screening Guidelines (Resnick and McLeskey, 2008)**

<table>
<thead>
<tr>
<th>Agency</th>
<th>Middle-Aged Recommendation</th>
<th>Older Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicare/Medicaid Reimbursement</td>
<td>Annual mammography for women &gt;40 yrs (if eligible)</td>
<td>For all women after 70 years old, risk factors associated with breast cancer should be explored every 1–2 years to discuss breast cancer screening and to address the risks and benefits associated with screening, as well as the individual’s comorbidities, life expectancy, health status and quality of life.</td>
</tr>
<tr>
<td>American Academy of Family Physicians</td>
<td>Every 1–2 yrs for women 50–69 yrs</td>
<td></td>
</tr>
<tr>
<td>American College of Obstetricians and Gynecologists</td>
<td>Every 1–2 yrs for women &gt;50 yrs</td>
<td></td>
</tr>
<tr>
<td>American Cancer Society</td>
<td>Yearly for women &gt;40 yrs</td>
<td></td>
</tr>
<tr>
<td>American Medical Association</td>
<td>Yearly after 50 yrs</td>
<td></td>
</tr>
<tr>
<td>Canadian Task Force on Preventive Health Care</td>
<td>Every 1–2 yrs for women 50–69 yrs</td>
<td></td>
</tr>
<tr>
<td>US Preventive Services Task Force</td>
<td>Every 2 yrs for women 50–74 yrs*</td>
<td></td>
</tr>
</tbody>
</table>

*Recently announced change

In December 2009, the USPSTF has announced the following guidelines in their revised recommendation statement (US Preventive Services Task Force, 2009):

1. The USPSTF recommends biennial screening mammography for women aged 50 to 74 years.
2. The USPSTF concludes that the current evidence is insufficient to assess the additional benefits and harms of screening mammography in women 75 years or older.

In particular, the recommendation for screening every 2 years as opposed to screening annually for women between the ages of 50 and 74 has caused considerable controversy. Unlike colorectal cancer screening or a bone marrow biopsy, mammography is non-invasive, and the risk of being diagnosed with breast cancer is known to increase with age. However, annual screening creates numerous false positive exams, which create stressful living situations for patients and can eventually lead to unnecessary biopsies. These types of tradeoffs and controversies highlight the need for a statistically accurate numerical evidence-based decision making tool for evaluating the impact of alternative screening policies.
Nine randomized trials have shown that the benefits of mammography screening for breast cancer include a 20% reduction in the mortality rate; however, eight of those trials did not consider the screening of older patients (Badgwell et al., 2008; Crivellari et al., 2007). Of course, this is complicated by the fact that breast cancer in older patients is believed to be less aggressive (Crivellari et al., 2007; Diab et al., 2000; Downey et al., 2007; Peer et al., 1993); and the benefits of screening may be offset by risks from comorbidities that could lead to other causes of death. These issues formed the need for considerable research regarding effective screening strategies for older women, and these strategies should be able to be dependent upon risk factors, chronological age, and a number of other variables.

2.2. Limitations of previous approaches to validation

Validation is the process of determining the degree to which a model or simulation is an accurate representation of the real world from the perspective of the intended uses of the model or simulation. In many analytical models created in industry and in academic research, validation is often a step that is entirely omitted or lightly touched upon. Many analytical models are extremely complex in nature in involve numerous inputs and outputs of different types, and it can be difficult to bring credibility to the model when attempting to present the structure and results to management or government. In particular, simulation models often have many forms of output, and further statistical analysis may even be done on the simulation output to produce other statistical model output that may need to be validated. Past approaches to multivariate validation are inadequate for the output provided by ours and other complex analytical models.

Multivariate approaches based on Hotelling's $T^2$ statistic (the multivariate generalization of Student's $t$-statistic) are impractical because covariance structure of the noisy real-world data is unknown; moreover, estimating the covariance matrix of the simulation-generated responses would require the number of runs to exceed the number of responses which is not feasible because of the relatively lengthy execution time for each run (15 minutes). Traditional approaches to combining $p$-values from several univariate $t$-tests are invalid when the underlying $t$-tests are not independent of each other; for example, yearly data where the value for one year is clearly dependent on the previous years value. If reasonable data are available for which the model could be validated against, a statistically accurate method for evaluating the ability of the simulation model to capture the behavior of the actual system is a necessary and valuable tool for giving models credibility. The validation procedure developed in the article is intended to provide a framework for performing validation of multi-response analytical models.

2.3. Main performance measures, SEER data, and BCSC data

In this section, we set up all the required notation for individual characteristics, population and sample sizes. We also explain our use of BCSC data as input, and SEER-based statistics statistics for validation of simulation-generated responses.
2.3.1 Simulation generated responses

For \( i = 1, \ldots, k \), the \( i \)th population characteristic \( \mu_i \) refers to the US female 65+ population for a specific calendar year \( CY_i \in \{2001, \ldots, 2020\} \) in the time horizon of the simulation model. Moreover for \( i = 1, \ldots, k \), we let \( N_i \) denote the size of the US female 65+ population at the beginning of year \( CY_i \); and for \( j = 1, \ldots, N_i \), we let \( y_{ij} \) denote the value of the \( i \)th characteristic for the \( j \)th individual in the relevant population. Thus we have

\[
\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij} \quad \text{for} \quad i = 1, \ldots, k; \quad (1.1)
\]

and we let

\[
S_{y_i}^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (y_{ij} - \mu_i)^2 \quad (1.2)
\]

denote the "conventional" definition of the population variance for the \( i \)th characteristic in the context of sampling techniques see Section 2.5 of Cochran (1977).

2.3.2 SEER data

SEER data is almost exclusively used as the data the model is validated against. We assume that SEER data is capturing what actually happened in the population during the years 2001–2011. SEER data was used as input to the simulations in only one case: to determine the survival distribution after treatment was received. However, the Plevritis stage progression model was used to determine the stage of the cancer as a function of tumor size (Plevritis, 2007). Since SEER data was used as input for only one variable, we contend that validating the model against the SEER data is an acceptable approach. In addition, SEER is the most reputable source of cancer data available.

The SEER data set for the \( i \)th characteristic consists of a sample of size \( n_i = f_i N_i \) taken from the relevant population of US females of age 65+; and we let

\[
\{U_{ij} : j = 1, \ldots, n_i\} \quad (1.3)
\]

denote this sample, with sample mean

\[
\bar{U}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} U_{ij} \quad (1.4)
\]

and sample variance

\[
S_{U_i}^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (U_{ij} - \bar{U}_i)^2. \quad (1.5)
\]
If the SEER data set were obtained from the relevant population of $N_i$ US females of age 65+ by simple random sampling (i.e., random sampling without replacement), then the expected value of the sample mean $\overline{U}_i$ would be given by

$$E[\overline{U}_i] = \mu_i \quad \text{for } i = 1, \ldots, k; \quad (1.6)$$

and the variance of $\overline{U}_i$ would be given by

$$Var[\overline{U}_i] = \frac{S^2}{n_i} (1 - f_s) \quad \text{for } i = 1, \ldots, k \quad (1.7)$$

see, for example, Theorems 2.1 and 2.2 of Cochran (1977).

Using Table 1.10 from the SEER Cancer Statistics Review 1975-2009 (Howlander et al., 2012), we determined the SEER database has a total of $n_s = 137,576$ new cases of breast cancer for women 65 and older from the period 2005–2009. Using US Census data (US Census Bureau, 2009), we computed the average population size of US females 65 and older over the period 2005–2009 as $N = 21,982,451$. Thus, we estimate the sampling fraction of the SEER database of 65 and older females with breast cancer to be $f_s = n_s/N = 0.00626$. Note that even if we found some adjustment factor to convert $n_s$ from and incidence-type statistic to a prevalence-type statistic; $f_s$ will still be relatively small, and thus we use the given value of $f_s$ as an approximation for the validation procedure.

Unfortunately, the SEER data set was not obtained by simple random sampling (or any kind of random sampling, for that matter); and thus neither (1.6) nor (1.7) are guaranteed to hold. Moreover, the sample variance (1.5), which under simple random sampling is an unbiased estimator of $Var[\overline{U}_i]$, is not available for all $i$. In the absence of more and better information about the SEER data set, we assume that (1.6) and (1.7) hold to a first approximation for the purposes of validating the simulation output vs. the SEER data set. It must be borne in mind, however, that a significant statistical discrepancy between the results of the simulation and the SEER data set does not necessarily invalidate our simulation, particularly in our unique calibration application.

### 2.3.3 BCSC data

BCSC data served as the primary input to the simulation models (BCSC, 2006). The BCSC data set provides the information on US females aged 65+ that is needed to generate the random initial condition for each run of the simulation. Specifically out of the initial population of $N_0 = 20,582,128$ US females of age 65+ on December 31, 2000, the corresponding BCSC data subset is of size $n_b = 250,509$; and each run of the simulation is initialized with a simple random sample of de-identified individuals taken from this BCSC data subset. The initialization procedure for the simulation is designed to use the proportion $f_s = 0.001$ of the US population of females of age 65+ at the model's starting time, so that each run of the simulation starts with an initial sample of size
\[ n_r = f_r N_0 = 20,582. \]  \hspace{1cm} (1.8)

For the \( i \)th characteristic \( (i = 1, \ldots, k) \) and the \( r \)th run of the simulation \( (r = 1, \ldots, m) \), we let

\[ \{U_{ijr} : j = 1, \ldots, n_r \} \]  \hspace{1cm} (1.9)

denote the corresponding set of characteristic values for the individuals obtained by simple random sampling of the relevant BCSC data subset of US females of age 65+. We would like to assume that (1.9) is a simple random sample of size \( n_r \) from the population of all such characteristic values for the \( N_0 \) US females of age 65+ at the model's starting time; unfortunately some properties of the overall BCSC data set indicate that such an assumption is not strictly true. For example, it is known that at the simulation starting time, women of age 65+ constituted 14.4% of the population of all US women; but the corresponding percentage for the BCSC data set was 24.9%. Nevertheless in the absence of additional information that could be exploited to provide a better method for initializing the simulation model, we assume as a first approximation that each BCSC-based sample of the form (1.9) is a simple random from the underlying population of all such characteristic values for the \( N_0 \) US females of age 65+ at the model's starting time.

2.3.4. Main performance measures

This section presents the performance measures from the screening-and-treatment simulation that are used in the validation procedure. The simulation produces a plethora of results, but the results chosen for validation were done so because: (i) data were available to validate these performance measures against; and (ii) these performance measures are important in determining whether or not the simulation is capturing the behavior of the real US breast cancer screening system accurately. The performance measures can be divided into three categories: population size characteristics, cancer occurrence statistics, and cancer diagnosis statistics.

2.3.4.1. Population size characteristics

Population size characteristics were obtained from US Census data (US Census Bureau, 2009).


Census data were available for the period 2001–2010. The initial population size in 2001 is set to the actual population size as of 1/1/2001, so a comparison for the start of 2001 would have no meaning. The other nine data points are formally validated against simulation-generated observations using our revised \( t \)-statistic method.
2.3.4.2. Cancer-occurrence characteristics

Cancer occurrence characteristics were obtained from SEER data (National Cancer Institute, 2007).

3. US Female 65+ Total Incidence Rates (2005-2009)

[1-3] SEER data were available for the period 2001–2009; however it takes the simulation model until the end of 2004 to "warm up" and produce reasonable values. The five data points for the period 2005–2009 are formally validated against simulation-generated observations using our revised t-statistic method. The volatility in this data make fitting a statistical model to the data very difficult, and testing simulation-generated values against that model would not be a reliable validation technique.


SEER data were available for the period 2001–2009; however it takes the simulation model until the end of 2010 to "warm up" and produce reasonable values for cancer death rates. As expected, there is a longer warm up period needed for cancer death rates than for cancer incidence rates. This leaves us with no data points that can be compared to SEER data with any meaning. However the observed data was relatively stable, which allowed us to fit a linear model to the data, predict values for the period 2011-2020, and test simulation-generated values against that model. The ten model-predicted points are formally validated against simulation-generated observations using our revised t-statistic method.

2.3.4.3. Cancer diagnosis characteristics

Cancer diagnosis characteristics were obtained from SEER data (National Cancer Institute, 2007) and from relevant literature (Gutwein et al., 2011).

1. US Female 65+ Stage Distribution at Diagnosis

Unlike the previous five performance criteria, SEER data for the percentages of women diagnosed in each stage (local, regional, distant) of breast cancer are not available on a yearly basis; instead they are simply overall percentages from the period 2001–2009. Since these values are limited on the range [0,100] because they are percentages, the CI half-lengths generated by the simulation model are very small, on the order of 1%. Thus, this measure was validated from a practical standpoint by simply declaring whether or not the mean simulation values were within ± 10% of the mean data values. There are percentages for local, regional, and distant cancers, so there are three tests for practical significance performed using this data, one for each stage.
2. PPV3 (or Percentage of Benign Biopsies)

The percentage of benign biopsies is not available in SEER or BCSC data. Thus, we used literature (Gutwein et al., 2011) to determine that about 80% of breast biopsies are negative. Since this value is also limited on the range [0,100], the CI half-lengths generated by the simulation model are very small, again on the order of 1%. Thus, this measure was validated from a practical standpoint by simply declaring whether or not the mean simulation value is within ± 10% of the value of 80% derived from literature.

3. Validation Procedure

There are 34 performance measures from the population size and cancer occurrence categories above for which statistical analysis could be used to determine whether or not a significant difference exists between the data and the simulation-generated observations. There a 9 population size estimates, 5 estimates each for invasive incidence, DCIS incidence, and total cancer incidence, and 10 estimates of the breast cancer death rate. The 4 cancer diagnosis characteristics are evaluated from a practical standpoint. In order to determine the set of input parameters which most closely match SEER and US Census data for the period 2001–2011, we develop a measure of overall closeness. The following sections present the development of our validation techniques.

3.1. Validation tests for population-size characteristics

For each \( i \in \{1, ..., 9\} \), we test the population-size validation hypothesis

\[
H_{0i} : E\left[ \bar{X}_i \right] = \mu_i
\]

with the usual one-sample Student's t-statistic

\[
t_i = \frac{\bar{X}_i - \mu_i}{S_{\bar{X}} / \sqrt{m}}.
\]

Sampling fractions need not be considered when measuring the entire population size. This is the performance measure for which the usual one-sample Student's t-statistic can be applied acceptably. Validation of the other performance measures must account for the sampling fractions associated with the data sets used, as well as account for the linear model used in the prediction of breast cancer death rates.

3.2. Validation tests for cancer-occurrence characteristics

3.2.1. Validation tests for cancer incidence rates
From the assumption that (1.9) is a simple random sample, we see that the sample mean

$$\overline{U}_{i, r} = \frac{1}{n_T} \sum_{j=1}^{n_T} U'_{i, jr}$$  \hspace{1cm} (1.12)

has expected value

$$E\left[ \overline{U}_{i, r} \right] = \mu_i$$  \hspace{1cm} (1.13)

and variance

$$Var\left[ \overline{U}_{i, r} \right] = \frac{S_{yi}^2}{n_T} (1 - f_T)$$  \hspace{1cm} (1.14)

for $i = 1, \ldots, k$ and $r = 1, \ldots, m$. Comparing Equations (1.7) and (1.14), we see that

$$Var\left[ \overline{U}_{i} \right] = \left( \frac{n_T}{n_s} \right) \left( \frac{1 - f_S}{1 - f_T} \right) Var\left[ \overline{U}_{i, r} \right] = \left( \frac{n_T}{n_s} \right) \left( \frac{1 - f_S}{1 - f_T} \right) \left( \frac{1 - f_S}{1 - f_T} \right) Var\left[ \overline{U}_{i} \right]$$

$$= \left[ \frac{f_T (1 - f_S)}{f_S (1 - f_T)} \right] Var\left[ \overline{U}_{i, r} \right]$$  \hspace{1cm} for $i = 1, \ldots, k$.  \hspace{1cm} (1.15)

We will exploit Equation (1.15) in formulating test statistics for validating outputs of the simulation against results computed from the SEER data set.

Because samples of the form (1.9) for different runs are independent, we see that the across-run average

$$\overline{U}_i = \frac{1}{m} \sum_{r=1}^{m} \overline{U}_{i, r}$$  \hspace{1cm} (1.16)

has mean $\mu_i$ and variance

$$Var\left[ \overline{U}_i \right] = \frac{Var\left[ \overline{U}_{i, r} \right]}{m} = \frac{S_{yi}^2}{m \cdot n_T} (1 - f_T)$$  \hspace{1cm} (1.17)

for $i = 1, \ldots, k$. We will also exploit Equation (1.17) in formulating test statistics for validating outputs of the simulation.

Now we turn our attention to the simulation model's within-run average output performance measures

$$\left\{ \overline{X}_{i, r} : i = 1, \ldots, k; \ r = 1, \ldots, m \right\}$$  \hspace{1cm} (1.18)
that correspond to the sample average inputs \( \{ \overline{U}_i : i = 1,...,k; \ r = 1,...,m \} \) defined by Equation (1.12).

Because each within-sum sample average output \( \overline{X}_r \) is based on what we assume to be a simple random sample of inputs from the population of US females of age 65+, it is reasonable to postulate that the sample average outputs \( \{ \overline{X}_r : r = 1,...,m \} \) are independent normal random variables with the desired mean \( \mu_i \) for \( i = 1,...,k \). We compute sample mean and variance of the outputs on run \( r \),

\[
\overline{X}_i = \frac{1}{m} \sum_{r=1}^{m} \overline{X}_{ir}
\]

\( (1.19) \)

\[
S^2_{\overline{X}_i} = \frac{1}{m-1} \sum_{r=1}^{m} (\overline{X}_{ir} - \overline{X}_i)^2
\]

\( (1.20) \)

as our unbiased estimators of \( \mu_i \) and \( Var[\overline{X}_r] \), respectively, for \( i = 1,...,k \). Moreover, in view of Equation (1.15) and the fact that each within-run sample average output is based on a simple random sample of model inputs from the population of US females of age 65+, it is reasonable to postulate the following relation between the variance of the SEER-based sample average \( \overline{U}_i \) and the variance of the simulation-based sample average \( \overline{X}_r \):

\[
Var[\overline{U}_i] = \left[ \frac{f_T (1-f_S)}{f_S (1-f_T)} \right] Var[\overline{X}_r] \quad \text{for } i \in \{10,11,...,k\} \text{ and } r = 1,...,m,
\]

\( (1.21) \)

where the population-size characteristics \( \{ N_i : i = 1,...,9 \} \) are known exactly and do not need to be estimated from SEER data. We assume that the \( \{ \overline{U}_i : i = 10,...,k \} \) are also normal random variables, and that \( \overline{U}_i \) is independent of \( \overline{X}_i \) for \( i = 10,...,k \).

### 3.2.2. Validation tests for cancer death rates

For the \( n_D = 10 \) predictions of the annual cancer death rates in the designated population of US females aged 65+ for the time period 2011–2020 and for \( i = 25,...,34 \), let \( \hat{\mu}_i \) represent the linear regression–based prediction of the cancer death rate for year \( 2010 + (i-24) \). As part of the validation of the simulation, we need to evaluate the statistical significance of the discrepancies between the simulation–generated estimates \( \{ \overline{X}_i : i = 25,...,34 \} \) and the regression-based predictions \( \{ \hat{\mu}_i : i = 25,...,34 \} \).

We need to set up some notation for the linear regression used to obtain the prediction \( \hat{\mu}_i \) so that we can estimate \( Var[\hat{\mu}_i - \overline{X}_i] \) for \( i = 25,...,34 \). This regression was based on the historical time series consisting of \( n_H = 9 \) observations of the annual cancer death rate for the period 2001–2009. Let \( D_j \) denote the \( j \)th observed annual cancer death rate (i.e., the response or dependent variable), and let \( A_j \equiv j \) denote the predictor (i.e., the regressor or independent variable) for \( j = 1,...,n_H \). We postulate the classical linear regression model
\[ D_j = \theta_0 + \theta_j A_j + \varepsilon_j , \]  

(1.22)

where the residuals \( \{ \varepsilon_j \} \) in the model (1.22) are assumed to be i.i.d. normal,

\[ \{ \varepsilon_j : j = 1, \ldots, n_H \} \overset{i.i.d.}{\sim} N \left( 0, \sigma^2 \right) . \]  

(1.23)

Using standard regression notation, we let

\[ \overline{D} = \frac{1}{n_H} \sum_{j=1}^{n_H} D_j , \]  

(1.24)

\[ \overline{A} = \frac{1}{n_H} \sum_{j=1}^{n_H} A_j , \]  

(1.25)

\[ S_{AA} \equiv \sum_{j=1}^{n_H} (A_j - \overline{A})^2 , \]  

(1.26)

\[ S_{DD} \equiv \sum_{j=1}^{n_H} (D_j - \overline{D})^2 , \]  

(1.27)

and

\[ S_{AD} \equiv \sum_{j=1}^{n_H} (A_j - \overline{A})(D_j - \overline{D}) , \]  

(1.28)

so that we have the usual least-squares estimates of the regression coefficients,

\[ \hat{\theta}_i = S_{AD} / S_{AA} , \]  

(1.29)

and

\[ \hat{\theta}_0 = \overline{D} - \hat{\theta}_i \overline{A} ; \]  

(1.30)

moreover, the residual mean square is given by

\[ \hat{\sigma}_\varepsilon^2 = \frac{1}{n_H - 2} \sum_{j=1}^{n_H} \left[ D_j - \left( \hat{\theta}_0 + \hat{\theta}_i A_j \right) \right]^2 = \frac{S_{DD} - \hat{\theta}_i S_{AD}}{n_H - 2} \]  

(1.31)

Note that for \( i = 25, \ldots, 34 \), our estimate of population characteristic \( \mu_i \) is

\[ \hat{\mu}_i = \hat{D}_j = \hat{\theta}_0 + \hat{\theta}_i A_j \text{ for } j = n_H + (i - 24) = i - 15 \]  

(1.32)

so that assuming the linear model (1.22) is valid, with \( j = i - 15 \), we have

\[ E[\hat{\mu}_i] = E[\hat{D}_j] = \mu_i \]  

(1.33)

and
\[ \text{Var}[\hat{\mu}_i] = \text{Var}[\hat{D}_j] = \sigma_e^2 \left[ n_i^{-1} + S_{AA}^{-1} (A_j - \overline{A})^2 \right]; \] (1.34)

see, for example, Eq. (3.1.4) of Draper et al. (1998).

Observe that
\[ \overline{A} = \frac{1}{n_H} \sum_{j=1}^{n_H} A_j = \frac{1 + n_H}{2}, \] (1.35)

\[ S_{AA} = \sum_{j=1}^{n_H} \frac{j^2 - (4n_H)^{-1} \left[ n_H (n_H + 1) \right]^2}{6} = \frac{n_H (n_H + 1) (2n_H + 1)}{4}, \] (1.36)

see, for example, Equation 29.2 of Dwight (1961). For \( i = 25, \ldots, 34 \), our estimate of \( \text{Var}[\hat{\mu}_i] \) is given by
\[ \text{Var}[\hat{D}_j] = \hat{\sigma}_e^2 \left[ \frac{1}{n_H} \right] + \left[ \frac{(A_j - \overline{A})^2}{S_{AA}} \right] = \left( \frac{S_{DD} - \hat{\theta}_i S_{AD}}{n_H - 2} \right) \left[ \frac{1}{n_H} \right] + \left( \frac{(j - n_H + 1)^2}{12} \right), \] (1.37)

for \( j = n_H + (i - 24) = i - 15 \)

with the simplification
\[ j - \frac{n_H + 1}{2} = n_H + i - 24 - \frac{n_H + 1}{2} = n_H + 2i - 49, \]

we see that in (1.37),
\[ \left( \frac{A_j - \overline{A}}{S_{AA}} \right)^2 = \left( \frac{n_H + 2i - 49}{2} \right)^2 \left( \frac{n_H (n_H^2 - 1)}{12} \right) = \frac{3(n_H + 2i - 49)^2}{n_H (n_H^2 - 1)}. \] (1.38)

We have that for \( i = 25, \ldots, 34, \)
\[ \text{Var}[\hat{\mu}_i - \overline{X}_i] = \text{Var}[\hat{\mu}] + \text{Var}[\overline{X}_i], \] (1.39)

and this variance is estimated by
\[ \text{Var}[\hat{\mu}_i - \overline{X}_i] = \hat{\sigma}_e^2 \left[ \frac{1}{n_H} + \left( \frac{(A_j - \overline{A})}{S_{AA}} \right)^2 \right] + \left( \frac{S_{X_i}^2}{m} \right) \text{ for } j = n_H + (i - 24) = i - 15, \] (1.40)
where we can exploit (1.37) and (1.38) to obtain convenient computational formulas for \( \mathbb{V} \mathbb{A} \mathbb{R} [\hat{\mu}_i] \). The final two-sample student's t-statistic has the form

\[
t_i = \frac{\hat{\mu}_i - \bar{X}_i}{\sqrt{\frac{\hat{\sigma}^2}{n_H} + \frac{(A_j - \bar{A})^2}{S_{AA}} + \frac{S^2_{\bar{X}_i}}{m}}}^{1/2} \quad \text{for } j = n_H + (i - 24); 
\]

and because there is no reason to suppose that \( \hat{\mu}_i \) and \( \bar{X}_i \) have equal variance, we use an approximation to the degrees of freedom in \( t_i \) in Satterthwaite (1941). Because \( \hat{\sigma}^2 \) has \( n_H - 2 \) degrees of freedom, so does \( \mathbb{V} \mathbb{A} \mathbb{R} [\hat{\mu}_i] \); and of course \( \mathbb{V} \mathbb{A} \mathbb{R} [\bar{X}_i] = \frac{S^2_{\bar{X}_i}}{m} \) has \( m - 1 \) degrees of freedom. Therefore by Satterthwaite's formula, \( t_i \) has approximate degrees of freedom

\[
df_i = \left\lfloor \mathbb{V} \mathbb{A} \mathbb{R} [\hat{\mu}_i] + \mathbb{V} \mathbb{A} \mathbb{R} [\bar{X}_i] \right\rfloor^{1/2} \left\lfloor \frac{\left( \mathbb{V} \mathbb{A} \mathbb{R} [\hat{\mu}_i] \right)^2}{n_H - 2} + \frac{\left( \mathbb{V} \mathbb{A} \mathbb{R} [\bar{X}_i] \right)^2}{m - 1} \right\rfloor 
\]

where by (1.37) and (1.38) we see that

\[
\mathbb{V} \mathbb{A} \mathbb{R} [\hat{\mu}_i] = \left( \frac{S_{DD} - \hat{\theta}_i S_{AD}}{n_H - 2} \right) \left[ \frac{1}{n_H} + \frac{3(n_H + 2i - 49)^2}{n_H(n_H^2 - 1)} \right] \text{ for } i = 25, \ldots, 34, 
\]

and

\[
\mathbb{V} \mathbb{A} \mathbb{R} [\bar{X}_i] = \frac{S^2_{\bar{X}_i}}{m} \text{ for } i = 25, \ldots, 34 
\]

so that the t-statistic (1.41) has approximate degrees of freedom

\[
df_i = \left\lfloor \left( \frac{S_{DD} - \hat{\theta}_i S_{AD}}{n_H - 2} \right) \left[ \frac{1}{n_H} + \frac{3(n_H + 2i - 49)^2}{n_H(n_H^2 - 1)} \right] + \frac{S^2_{\bar{X}_i}}{m} \right\rfloor^{1/2} 
\]

\[
\left\lfloor \left( \frac{S_{DD} - \hat{\theta}_i S_{AD}}{n_H - 2} \right)^2 \left[ \frac{1}{n_H} + \frac{3(n_H + 2i - 49)^2}{n_H(n_H^2 - 1)} \right]^2 + \frac{S^4_{\bar{X}_i}}{m^2(m-1)} \right\rfloor^{1/2} 
\]
3.3. Validation Procedure for Overall Simulation Model

As an overall measure of the discrepancy between the with $k \times 1$ vector
\[ Y \equiv \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_k \end{bmatrix}^T \]  
(1.46)
of responses from the real system and the $k \times 1$ vector
\[ \bar{X} \equiv \begin{bmatrix} \bar{X}_1 & \bar{X}_2 & \cdots & \bar{X}_k \end{bmatrix}^T \]  
(1.47)
of average responses taken over $m$ independent simulation runs, we use the squared length of the $k \times 1$ vector
\[ t \equiv \begin{bmatrix} t_1 & t_2 & \cdots & t_k \end{bmatrix}^T \]  
(1.48)
of standardized discrepancies between the real and simulated responses
\[ D^2 = \| t^2 \| = \sum_{i=1}^{k} t_i^2 = \sum_{i=1}^{k} \frac{(\bar{\mu}_i - \bar{X}_i)}{\tilde{\nu}_i}^2, \]  
(1.49)
where
\[ \bar{\mu}_i = \begin{cases} \mu_i, & \text{for } i \in \{1, \ldots, 9\}, \\ \hat{U}_i, & \text{for } i \in \{10, \ldots, 24\}, \\ \hat{\theta}_u + \hat{\theta}_u \Delta - 15, & \text{for } i \in \{25, \ldots, 34\}, \end{cases} \]  
(1.50)
and
\[ \bar{\nu}_i = \begin{cases} S_X^2 / m, & \text{for } i \in \{1, \ldots, 9\}, \\ S_X^2 \left[ \frac{f_T}{f_S} \left(1 - \frac{f_S}{f_T} \right) + \left\{ \frac{1}{m} \right\} \right], & \text{for } i \in \{10, \ldots, 24\}, \\ S_X^2 \left[ \frac{1}{n_H} \right] + \left( \frac{A_{-15} - A}{S_{A_H}} \right)^2 \left[ \frac{S_X^2}{m} \right], & \text{for } i \in \{25, \ldots, 34\}, \end{cases} \]  
(1.51)
so that for $i = 1, \ldots, k$, we let $\bar{\mu}_i$ denote an estimate of $\mu_i$ derived independently of the simulation model, and we let $\bar{\nu}_i$ denote our estimate of the variance of $\bar{\mu}_i - \bar{X}_i$ with degrees of freedom
\[
\begin{equation}
\text{df}_i = \begin{cases} 
m - 1, & \text{for } i \in \{1, \ldots, 9\}, \\
m - 1, & \text{for } i \in \{10, \ldots, 24\}, \\
\left(\tilde{v}_i\right)^2 \left(\frac{|V\hat{\mu}|^2}{n_H - 2} + \frac{S_n^4}{m^2 (m - 1)}\right), & \text{for } i \in \{25, \ldots, 34\}. 
\end{cases}
\end{equation}
\]

In (1.52), \(\text{Var} \left[ \tilde{\mu}_i \right] \) denotes the usual estimator of the variance of a regression-based prediction,
\[
\text{Var} \left[ \tilde{\mu}_i \right] = \hat{\sigma}^2 \left[ n_H^{-1} + S_A^{-1} \left(A_{i-15} - \bar{A}\right)^2 \right] \quad \text{for } i \in \{25, \ldots, 34\};
\]
see Equations (1.24)–(1.37).

From (1.49)–(1.52), we see that the revised Student's t-statistic for validating the simulation model with respect to the \(i\)th population characteristic can be simply expressed as follows:
\[
t_i = \frac{\tilde{\mu}_i - \bar{X}_i}{\sqrt{\tilde{v}_i}} \quad \text{for } i \in \{1, \ldots, k\};
\]
and under the null hypothesis
\[
H_{0i}: E \left[ \bar{X}_i \right] = E \left[ \tilde{\mu}_i \right],
\]
\(t_i\) has (approximately) Student's t-distribution with \(\text{df}_i\) degrees of freedom.

In vector notation we can write this compactly in a form that closely resembles Hotelling's \(T^2\) statistic.

Let
\[
\tilde{V} = \text{diag} \left(\tilde{V}_1, \ldots, \tilde{V}_k\right) = \begin{bmatrix}
\tilde{V}_1 & 0 & \ldots & 0 \\
0 & \tilde{V}_2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & \tilde{V}_k
\end{bmatrix};
\]
then we see that
\[
D^2 = \left(\mathbf{Y} - \bar{X}\right)^T \tilde{V}^{-1} \left(\mathbf{Y} - \bar{X}\right).
\]
We propose choosing the simulation configuration that minimizes \(D^2\) as the policy which best captures data for the actual system for the 2001–2011. Then, as an overall validation of the selected simulation configuration, one can compute the \(p\)-values associated with the revised \(t_i\) statistics in Equation (1.54), and create a Q-Q plot assessing the extent to which the \(p\)-values look like they have the expected Uniform(0,1) distribution. If \(\{p_g : g = 1, \ldots, L\}\) denotes the \(p\)-values in ascending order so
A Q-Q plot providing a visual assessment of the extent to which the \( p \)-values look like they have the expected Uniform(0,1) distribution is created as follows: plot the points \([p_g, (g - L)^{-1}/L]\) for \( g = 1, \ldots, L \). If the \( p \)-values really came from the Uniform(0,1) distribution, then the plot should look more or less like a straight line. In addition, we feed the resulting \( p \)-values into \( \Phi^{-1}(\bullet) \), the inverse of the standard normal CDF \( \Phi(\bullet) \) to obtain the corresponding \( Z \)-values, combine the \( Z \)-values in an overall Shapiro-Wilk test (Royston, 1993; Royston, 1992; Royston, 1999); and create the corresponding Q-Q plot. Regardless of the final \( p \)-value of the overall Shapiro-Wilk test on the \( Z \)-values

\[
\{Z_i = \Phi^{-1}(p_i): 1, \ldots k\}.
\] (1.58)

Visual inspection of these associated Q-Q plots are the main tools for validation of the selected simulation model. In terms of Q-Q plots, the more the plotted points resemble a straight line the better. Although this is somewhat subjective, this method of evaluating Q-Q plots to validate simulations has been used in many application contexts (Humphrey et al., 2000; Irizarry et al., 2001; Lee et al., 1991), and we are confident we can select the simulation model that best matches historical data for the period 2001–2011 using these Q-Q plots in combination with the overall \( D^2 \) statistic.

At first thought, it may seem appropriate to combine the \( p \)-values using the Fisher chi-squared test, which involves taking the natural log of each \( p \)-value, adding them up, and multiplying by \(-2\). This statistic is chi-squared distributed with \( n - 1 \) degrees of freedom where \( n \) is the number of \( t \)-tests. However, this method is only valid if the \( p \)-values are independent. Performing the von Neumann randomness test (Bartels, 1982; Lada & Wilson, 2006; Lada et al., 2008; Steiger et al., 2005; von Neumann, 1941) on the \( Z_i \)'s from Equation (1.58), we found that for each scenario, this test was failed with a very small level of significance. Thus, using the Fisher chi-squared test is inappropriate for this particular situation where the \( p \)-values are highly correlated. It should be noted that this method can easily be extended to situations where there are a different number \( t \)-tests for each scenario simply dividing the overall effectiveness measure \( D^2 \) by the number of \( t \)-tests.

\[
D_{\text{ADJ}}^2 = \frac{D^2}{k} = \frac{1}{k} \sum_{i=1}^{k} t_i^2.
\] (1.59)

4. Historical Calibration of the Simulation and Validation Results

We developed the validation procedure in Section 3 for our specific simulation modeling application, but the method could be applied to a variety of circumstances, and numerous other forms of statistical tests which generate \( p \)-values could be added while preserving the \( D_{\text{ADJ}}^2 \) statistical measure of overall closeness and the meaning of the aforementioned Uniform and Normal Q-Q plots.

In order to provide a convincing case that this integrated simulation model is able to accurately predict the results of using alternative screening policies for the future period 2012–2020, we developed a
formal method for validating certain performance measures for the period 2001–2009 for which data were available. Since there was no screening policy in place for this period, we cannot impose a policy and then match the results to data by, for example, tuning the adherence submodel. Instead, we must impose several polices and then determine which policy best matches the data, and this "historically calibrated" policy will be in effect from 2001 to 2011. Alternative screening policies are tested for the period 2011–2020. We define a screening policy in terms of a screening interval (which could be different for different women based on their individual attributes) and a stopping age. The screening intervals to be compared for validation are screening every one, two, and three years for all women. Since there was no set stopping age for screening during this time period, we do not want to force every woman to stop screening at some specific age. However, we need to have some control over when women stop screening, and we need to be able to vary the distribution of stopping age. There may be some women who never go for screening after they turn 65, so the absolute minimum stopping age would be 65, but this would simply amount to not screening at all. The oldest possible age in the model is 110; however, it is extremely unlikely any woman reaches this age, and so we set the maximum stopping age at 105 years old. Using these values as the upper and lower limits, we fit seven Beta distributions with modes 70, 75, 80, 85, 90, 95, and 100 (Kuhl et al., 2010). These distributions are defined according to Equation (1.60).

\[
\text{Stopping Age Distribution} = \begin{cases} 
65 + 40 \times \text{Beta}(1.480, 4.360), & \text{if mode} = 70, \\
65 + 40 \times \text{Beta}(2.200, 4.600), & \text{if mode} = 75, \\
65 + 40 \times \text{Beta}(3.117, 4.529), & \text{if mode} = 80, \\
65 + 40 \times \text{Beta}(4.000, 4.000), & \text{if mode} = 85, \\
65 + 40 \times \text{Beta}(4.529, 3.117), & \text{if mode} = 90, \\
65 + 40 \times \text{Beta}(4.600, 2.200), & \text{if mode} = 95, \\
65 + 40 \times \text{Beta}(4.360, 1.480), & \text{if mode} = 100.
\end{cases} 
\]

(1.60)

This yields five screening intervals and seven potential distributions for stopping age, so there are a total of 21 combinations of these two inputs, leaving us with 21 potential screening policies that could be used for 2001–2011. We generated full sets of simulated results for all 21 policies, and then we performed our validation technique on selected performance measures for each of these 21 policies. The best policy was selected, and that policy was used for 2001–2011 regardless of the alternative policy in place from 2012–2020.

We have the following situation: the simulation model produces mean values and confidence interval half-lengths (a measure of the precision of the estimated mean value), but the available data only provide mean values, although we know there is some variation in these reported values. Thus, we develop a formal statistical method for comparing the simulation-generated observations to observed values, and we present a reasonable measure of overall closeness of the simulation results to an observed set of real-world observations. We use this overall measure to determine the screening policy that best matches the
available data. Ten randomly sampled populations were used to generate mean values and CI half-lengths. After experimenting with five simulated populations, we determined that in order to get CI half-lengths of appropriate size, ten simulated populations were necessary. However, as a result, the CI half-lengths for the percentages of women diagnosed in each stage of breast cancer and for the percentage of benign biopsies are very small. Thus, those measures were validated from a practical standpoint by simply declaring whether or not the mean simulation values were within ± 10% of the mean data values.

4.1. Practical significance considerations and suggested modifications

Before moving to the results of implementing this validation procedure, it is important to point out that practical significance should play a role in validation. While some results might be statistically significantly different from the data provided for the actual system, those same results might not be practically significantly different. We illustrate this idea using the output of our model. Initially, we implanted the validation scheme using all 34 performance measures presented in previous sections. However, we quickly realized that the $t$-statistics for the population size were extremely large, while it seemed the population size resulting from the population growth submodel was remarkably close to US Census data for the period 2002–2010. We also note that the population size submodel is unaffected by changes in the screening policy, so the resulting population sizes for the different calibration policies are virtually identical. Thus, we took the population size output from annual screening with a stopping age of 70, and created Table 2 which contains this information along with the Census data for those years, and the computed percentage deviation from the Census data.

<table>
<thead>
<tr>
<th>Year</th>
<th>Census</th>
<th>Simulation</th>
<th>% Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>20687.0</td>
<td>20678.5</td>
<td>-0.04%</td>
</tr>
<tr>
<td>2003</td>
<td>20838.6</td>
<td>20812.2</td>
<td>-0.13%</td>
</tr>
<tr>
<td>2004</td>
<td>20981.8</td>
<td>21002.7</td>
<td>0.10%</td>
</tr>
<tr>
<td>2005</td>
<td>21180.5</td>
<td>21239.8</td>
<td>0.28%</td>
</tr>
<tr>
<td>2006</td>
<td>21418.0</td>
<td>21504.5</td>
<td>0.40%</td>
</tr>
<tr>
<td>2007</td>
<td>21736.9</td>
<td>21834.8</td>
<td>0.45%</td>
</tr>
<tr>
<td>2008</td>
<td>22195.1</td>
<td>22183.6</td>
<td>-0.05%</td>
</tr>
<tr>
<td>2009</td>
<td>22569.1</td>
<td>22602.6</td>
<td>0.15%</td>
</tr>
<tr>
<td>2010</td>
<td>22905.0</td>
<td>23031.3</td>
<td>0.55%</td>
</tr>
</tbody>
</table>

Upon review of Table 2, one can see that the simulation model is tracking the population size of US females 65 and older remarkably well, the largest percentage deviation is 0.50%. The 95% CI half-widths associated with these population size estimates are very small, on the order to $10^2$, leading to extremely large $t$-values even though the output is remarkably close to the actual data. This small half-width issue can lead to misleading $t$-values. Thus, we choose the remove the 9 population size statistics from the
overall statistical validation scheme, and simply concluded that the values of population size are not practically different from the US Census data. This method of studying the results and computing the percentage deviations should be applied to each set of performance measures in conjunction with the statistical validation procedure in order to ensure both statistical and practical significance are accounted for.

4.2. Validation results

Table 3 presents the results from implementing our validation technique. For each of the 21 policies, we give the number of \( t \)-tests passed (of 25), the \( D^2_{ADJ} \) value, the number of practical significance tests passed (of 4), and the von Neumann randomness test \( p \)-value. We also give the Shapiro-Wilk \( p \)-value for the policies, and present the Q-Q plots for the best scenario.

### Table 3. Results from Validation and Calibration Procedure

<table>
<thead>
<tr>
<th>Interval/Measure</th>
<th># of t-tests Passed (95%)</th>
<th>( D^2_{ADJ} )</th>
<th>Practical Tests (±10%)</th>
<th>VN ( p )-value</th>
<th>SW ( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{Interval } = 1 )</td>
<td># Passed (25)</td>
<td>Value</td>
<td>Rank</td>
<td># Passed (4)</td>
<td>( p )-Value</td>
</tr>
<tr>
<td>70</td>
<td>2</td>
<td>156.7</td>
<td>3</td>
<td>2</td>
<td>3.06E-07</td>
</tr>
<tr>
<td>75</td>
<td>5</td>
<td>92.8</td>
<td>1</td>
<td>3</td>
<td>2.93E-07</td>
</tr>
<tr>
<td>80</td>
<td>4</td>
<td>140.8</td>
<td>2</td>
<td>4</td>
<td>2.72E-07</td>
</tr>
<tr>
<td>85</td>
<td>2</td>
<td>210.2</td>
<td>6</td>
<td>4</td>
<td>3.52E-07</td>
</tr>
<tr>
<td>90</td>
<td>1</td>
<td>411.4</td>
<td>19</td>
<td>4</td>
<td>4.29E-07</td>
</tr>
<tr>
<td>95</td>
<td>2</td>
<td>490.4</td>
<td>20</td>
<td>4</td>
<td>3.04E-07</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>590.5</td>
<td>21</td>
<td>4</td>
<td>2.24E-07</td>
</tr>
<tr>
<td>( \text{Interval } = 2 )</td>
<td># Passed (25)</td>
<td>Value</td>
<td>Rank</td>
<td># Passed (4)</td>
<td>( p )-Value</td>
</tr>
<tr>
<td>70</td>
<td>2</td>
<td>263.5</td>
<td>10</td>
<td>0</td>
<td>3.16E-07</td>
</tr>
<tr>
<td>75</td>
<td>2</td>
<td>197.3</td>
<td>4</td>
<td>1</td>
<td>3.13E-07</td>
</tr>
<tr>
<td>80</td>
<td>1</td>
<td>212.8</td>
<td>7</td>
<td>2</td>
<td>2.84E-07</td>
</tr>
<tr>
<td>85</td>
<td>2</td>
<td>207.8</td>
<td>5</td>
<td>3</td>
<td>2.82E-07</td>
</tr>
<tr>
<td>90</td>
<td>0</td>
<td>283.0</td>
<td>13</td>
<td>3</td>
<td>5.01E-07</td>
</tr>
<tr>
<td>95</td>
<td>0</td>
<td>355.9</td>
<td>15</td>
<td>3</td>
<td>7.58E-07</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>370.8</td>
<td>17</td>
<td>3</td>
<td>1.12E-06</td>
</tr>
<tr>
<td>( \text{Interval } = 3 )</td>
<td># Passed (25)</td>
<td>Value</td>
<td>Rank</td>
<td># Passed (4)</td>
<td>( p )-Value</td>
</tr>
<tr>
<td>70</td>
<td>2</td>
<td>370.6</td>
<td>16</td>
<td>0</td>
<td>2.64E-07</td>
</tr>
<tr>
<td>75</td>
<td>3</td>
<td>299.0</td>
<td>14</td>
<td>0</td>
<td>2.35E-07</td>
</tr>
<tr>
<td>80</td>
<td>2</td>
<td>245.9</td>
<td>8</td>
<td>0</td>
<td>2.35E-07</td>
</tr>
<tr>
<td>85</td>
<td>2</td>
<td>260.2</td>
<td>9</td>
<td>0</td>
<td>2.60E-07</td>
</tr>
<tr>
<td>90</td>
<td>2</td>
<td>266.7</td>
<td>11</td>
<td>0</td>
<td>3.12E-07</td>
</tr>
<tr>
<td>95</td>
<td>0</td>
<td>278.8</td>
<td>12</td>
<td>1</td>
<td>1.00E-06</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>379.0</td>
<td>18</td>
<td>1</td>
<td>5.11E-07</td>
</tr>
</tbody>
</table>
Upon examination of the $D^2_{ADJ}$ statistic and the associated Q-Q plots, we believe there is clear evidence that annual screening with a beta distributed stopping age with a mode of 75 matches the data the best. It ranks first in the $D^2_{ADJ}$ statistic, and the associated Uniform and Normal Q-Q plots appear to be the closest to resembling a straight line, although the Uniform Q-Q plot is not close to what we would consider acceptable. The associated Normal Q-Q plot is quite close to resembling a straight line, and this policy also has a Shapiro-Wilk $p$-value that is not significant at the 0.05 level. Thus, we select that policy for use in the periods 2001–2011. In fact, it seems that none of our proposed screening policies could have been deemed "valid" representations of the past. Since none of these policies was actually in place, it makes sense no one policy matches the data reasonably well as we are trying to impose some policy for this time.
period that did not actually exist. With our calibrated screening policy in place for the past, we turn our attention to the results of experimenting with alternative screening policies for future periods.

5. Simulation Results and Discussion

Each time the model is executed, a plethora of results are generated. A full set of results includes several tables with yearly values for each SD level and SD performance measure, a table with yearly averages for individual characteristics of interest, the percentage of women treated, a table of yearly cancer incidence and death rates, graphs describing the stage distribution at diagnosis, graphs of cost and QALYs saved information by category, and graphs with information about methods of detection. Because the full set of results includes a large number of performance measures, we only present the full set of results for the screening policy which saves the most lives for the period 2012–2020. For other policies which are close (or not statistically different in terms of some selected performance measures), we provide the key performance measures only.

5.1. Key performance measures

There are several performance measures of interest when considering the impact of alternative breast cancer screening policies on the designated population as a whole and on the individuals who make up that population. The costs of the screening and treatment processes and the years of life saved are important performance measures for a screening policy, and these outputs can be broken down into categories for even more insight. Ultimately, the goal is to minimize breast cancer deaths, so another key performance measure is the number of deaths caused by breast cancer. The following list describes the key performance measures that are generated by the simulation and how each are categorized.

1. Total Cost
   - By Detection Type (Clinical, Screening)
   - By Procedure Type (Screening, Diagnostic, Biopsy, Treatment)
2. Cost/Life-Year Saved
3. Life-Years Saved
4. Cost/QALY Saved
5. QALYs Saved
   - [2-5] By Detection Type (Clinical, Screening)
6. Cost of False Positive Exams and Benign Biopsies
   - By Procedure Type (Screening, Diagnostic, Benign Biopsy)
7. Number of False Negative Exams Where Detection was Possible
5. Number of Cancer Deaths
6. Stage of Cancer at Diagnosis (Local, Regional, Distant)
7. Type of Detection Percentages (Screening and Clinical)
8. Benign Biopsy Percentage


We choose five of these performance measures as the most important subset, and we ranked each policy in terms of these five performance measures (with rank one denoting the best policy). The five most important performance measures are:

1. Number of breast cancer deaths during and after the year 2012;
2. Number of QALYs saved by screening during the period 2012–2020;
3. Percentage of cancers diagnosed in the distant stage during the period 2012–2020;
4. Cost/QALY saved by screening during the period 2012–2020;
5. The total cost of false positive exams and benign biopsies during the period 2012–2020.

Preventing deaths from breast cancer is the real objective of screening, and that could be considered the most important performance measure. Similarly, the number of QALYs saved is a measure of the amount of life saved and the quality of that life over the entire population, and this is another important performance measure. The cost-effectiveness of screening is a measure of how much we are paying to save one quality-adjusted year of a given woman's life. Cost per QALY saved by screening is something that should be considered, but we argue it is not as important as lives saved; and so long as the policy is reasonably cost-effective, the main objective should be to maximize the lives saved by the policy. The percentage of cancers diagnosed in the distant stage is important because women diagnosed in the distant stage are not likely to live long and their quality of life will be poor. The total cost of false positives is important, but not as important as lives saved or overall cost-effectiveness. The monetary cost of these exams to the population alone can be seen as a problem, but we must also consider the nonmonetary costs for individual women who endure substantial emotional stress because of a false diagnosis of breast cancer. In addition, the more incisions to a woman's breasts, the more difficult is it to identify breast cancers in the future (we did not account for this in our model); and a woman's future adherence to a screening protocol may be adversely affected because of a prior experience where the outcome was a false positive result. For these reasons, we included the total cost of false positive exams and benign biopsies in the top five performance measures; however we argue that these are the least important of the main performance measures.

5.2. Types of screening policies

We have constructed three types of screening policies that can be used for the period 2012–2020: interval screening, risk-based screening, and factor-based screening. Interval screening assigns the same screening interval and stopping age to every woman in the population. Risk-based screening allows for different
screening intervals for both high-risk and low-risk women. Factor-based screening allows for different screening intervals based on one individual breast cancer risk factor, such as breast density, the number of first degree relatives with breast cancer, or some combination of these factors. These individualized risk-based and factor-based policies may be more cost-effective than simply basing the screening interval on age, a practice for which there is no good basis. For the "model calibration" period 2001–2011, the screening policy is set to the interval screening policy that yields results most closely matching SEER data from that time period. The screening policy changes in 2012 to the prescribed policy, which lasts through 2020. These three types of screening policies are essentially three different methods of assigning a screening interval to each woman in the population. However, a screening policy is made up of both a screening interval and a stopping age. Stopping ages are for everyone in the population, and they are not dependent on risk classification or risk factors.

5.2.1. Interval screening

Interval screening is the simplest of the types of screening policies. The screening intervals considered are every one, two, three, four, and five years. Whether using deterministic or stochastic stopping ages, there are seven potential values. This leaves us with 35 combinations of screening interval and stopping age, or in other words 35 possible screening policies of this type.

5.2.2. Risk-based screening

Risk-based screening was suggested by breast cancer advisors, and we use the Barlow one-year risk to determine each woman's overall risk of being diagnosed with breast cancer. The only risk factor that changes over time for each woman in the model is her age, while all other risk factors remain constant. Thus, we recorded each woman's one-year risk upon entering the natural history model and determined the percentiles of that data. We allowed the definition of high risk to vary, so that it can correspond to the top 5%, 10%, 15%, or 20% of women of age 65+ in terms of one-year risk.

Once a definition of high risk is chosen, different screening intervals can be selected for high-risk and low-risk women, with the same options of every one, two, three, four, or five years for each risk classification. In creating optimizations that search through all risk-based policies, we enforce the constraint that the screening interval for low-risk women be greater than or equal to the screening interval for high-risk women, leaving us with 15 combinations of screening intervals. Given that there are 15 combinations of screening intervals, four different definitions of screening, and seven different potential stopping ages, we have 420 risk-based screening policies. An example of a risk-based screening policy is that high-risk women are screened every year, and low-risk women are screened every two years, where by definition high risk corresponds to the top 5% of women of age 65+ in terms of one-year risk; and screening is stopped at the age of 85.
5.2.3. Factor-based screening

Factor-based screening is the most versatile and complex type of screening policy that the integrated screening model can handle. There are eleven risk factors used in the Barlow risk model: five of those factors have three levels, three factors have four levels, two factors have five levels, and one factor has six levels, as shown in Table 4.

**Table 4. Table of Risk-Factors, Their Levels, and Their Categorical Variables**

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Categories and Categorical Variables (43 Total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>65-69, 70-74, 75-79, 80+</td>
</tr>
<tr>
<td></td>
<td>FreqAge7, FreqAge8, FreqAge9, FreqAge10</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Non-Hispanic, Hispanic, Unknown</td>
</tr>
<tr>
<td></td>
<td>FreqEth0, FreqEth1, FreqEth9</td>
</tr>
<tr>
<td>Race</td>
<td>White, Asian, Black, Native American, Other, Unknown</td>
</tr>
<tr>
<td></td>
<td>FreqRace1, FreqRace2, FreqRace3, FreqRace4, FreqRace5, FreqRace9</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>&lt;25, 25-29.99, 30-34.99, &gt;35, Unknown</td>
</tr>
<tr>
<td></td>
<td>FreqBMI1, FreqBMI2, FreqBMI3, FreqBMI4, FreqBMI9</td>
</tr>
<tr>
<td>Age at First Birth</td>
<td>&lt;30, ≥30, No Children, Unknown</td>
</tr>
<tr>
<td></td>
<td>FreqAAFB1, FreqAAFB2, FreqAAFB3, FreqAAFB9</td>
</tr>
<tr>
<td>Previous Breast Procedure</td>
<td>No, Yes, Unknown</td>
</tr>
<tr>
<td></td>
<td>FreqPrevProc1, FreqPrevProc2, FreqPrevProc9</td>
</tr>
<tr>
<td>1st Deg Relatives with BC</td>
<td>None, 1, ≥2, Unknown</td>
</tr>
<tr>
<td></td>
<td>FreqFirstDeg1, FreqFirstDeg2, FreqFirstDeg3, FreqFirstDeg9</td>
</tr>
<tr>
<td>Hormone Therapy Use</td>
<td>No, Yes, Unknown</td>
</tr>
<tr>
<td></td>
<td>FreqHRT1, FreqHRT2, FreqHRT9</td>
</tr>
<tr>
<td>Surgical Menopause</td>
<td>No, Yes, Unknown</td>
</tr>
<tr>
<td></td>
<td>FreqSurgMeno1, FreqSurgMeno2, FreqSurgMeno9</td>
</tr>
<tr>
<td>Result of Last Mammogram</td>
<td>Negative, False Positive, Unknown</td>
</tr>
<tr>
<td></td>
<td>FreqLastMamm1, FreqLastMamm2, FreqLastMamm9</td>
</tr>
<tr>
<td>Breast Density (BI-RADS)</td>
<td>1, 2, 3, 4, Unknown</td>
</tr>
<tr>
<td></td>
<td>FreqDensity1, FreqDensity2, FreqDensity3, FreqDensity4, FreqDensity9</td>
</tr>
</tbody>
</table>

Factor-based screening works in the following way:

1. The user selects the factors on which the screening interval should be based. The number of selected factors can range from only one factor to all eleven risk factors.
2. For those factors that are selected, a screening interval is specified for each level of that factor. For example, age is divided into four levels, so an interval must be specified for each age group.
a. If only one factor is selected, then each woman's screening interval is determined directly by her individual level for that factor.

b. If more than one factor is selected, then there will be a screening interval assigned to each woman for each factor selected; and the smallest such interval (taken over all selected factors) is used as the screening interval for that woman.

As with the other two types of screening policies, the user must specify a stopping age for screening from amongst the seven values. With a little mathematical insight, one can see that if the screening policy is based on all eleven factors and a stopping age, there are 16,329,600 screening policies. If we then allow for the possibility of using any subset of the eleven risk factors, the number of screening policies is so large as to preclude any attempt to optimize system performance by total enumeration. Thus, we must have an intelligent method for searching through these screening policies to find the policy or group of policies that is the most cost-effective.

5.3. Simulation-based comparison of screening policies

Three different objectives were considered through optimization, and to each of these optimization problems we applied policies based on interval screening, risk-based screening, and factor-based screening separately. The results across the different policy types were combined in order to determine the best policy or group of policies for each of the three optimization problems. By an optimization problem, we mean the specification of a performance measure to be optimized as well as the specification of constraints on input variables or on other performance measures (outputs). Before we formulate the three optimization problems, we provide a more comprehensive list of key performance measures. The following three optimization problems were run for all three different types of screening:

1. Objective: Minimize cost/QALY saved for breast cancers detected by mammography during the period 2012–2020. Constraints:
   - All screening intervals (including low and high risk specific intervals) are restricted to 1, 2, 3, 4, or 5 years.
   - The stopping age is restricted to 70, 75, 80, 85, 90, 95, or 100.


   - Total cost/QALY saved during the period 2012–2020 cannot exceed $50,000 (~GDP Per Capita).
   - All screening intervals (including low-and high-risk specific intervals) are restricted to 1, 2, 3, 4, or 5 years.
   - The stopping age is restricted to 70, 75, 80, 85, 90, 95, or 100.
It should be noted that for interval and risk-based screening, we were able to enumerate all possible screening policies and thus conduct a full designed experiment with all input variables at all values for all three optimization problems. However, the number of possible screening policies for factor based screening is over 16 million, and we could not possibly test all possibilities. Thus, we allowed the optimization to run until 5,000 screening policies were tested for each of the three optimizations. While these optimizations are rather simple, extremely complex optimizations could be and may be performed as future work.

After examining the results, we concluded that factor-based screening was unable to compete with risk-based and interval screening in terms of the key performance measures (most likely because we did not find the best group of policies amongst the first 5,000); therefore we focus on the results of risk-based and interval screening. However, the ability to produce results for factor-based screening policies is a key feature of this model and should not be overlooked. The top five policies in terms of the objectives were ranked and placed into tables in order from best to worst. Upon examining the results, we found that optimizations 2 and 3 produced the same set of 5 best policies; and therefore we were left with two sets of 5 best policies—one set for optimization 1, and one set for optimizations 2 and 3.

The best five screening policies in terms of all five performance measures from optimization 1 are:

1. Annual screening for the top 5% in terms of risk, every four years for everyone else, and screening stops at age 70;
2. Annual screening for the top 10% in terms of risk, every two years for everyone else, and screening stops at age 70;
3. Annual screening stops at age 70;
4. Annual screening stops at age 75; and
5. Annual screening stops at age 80.

Ranked results for each performance measure for each policy using optimization 1 are available in Table 5 and Table 6.

Table 5. Top 5 Policies Ranked in Terms of Top 5 Performance Measures for Optimization 1 Part I

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>538.4</td>
<td>5</td>
<td>264.9</td>
<td>5</td>
<td>35.6</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>537.4</td>
<td>4</td>
<td>483.1</td>
<td>4</td>
<td>34.6</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>532.1</td>
<td>3</td>
<td>768.4</td>
<td>3</td>
<td>31</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>522.9</td>
<td>2</td>
<td>1088.8</td>
<td>2</td>
<td>26</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>513.6</td>
<td>1</td>
<td>1238.2</td>
<td>1</td>
<td>21.2</td>
</tr>
</tbody>
</table>
### Table 6. Top 5 Policies Ranked in Terms of Top 5 Performance Measures for Optimization 1 Part II

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>$28,665</td>
<td>1</td>
<td>$1,301,443</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>$29,230</td>
<td>2</td>
<td>$2,691,079</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>$31,974</td>
<td>3</td>
<td>$5,457,012</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>$37,410</td>
<td>4</td>
<td>$8,236,843</td>
<td>14</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>$43,908</td>
<td>5</td>
<td>$9,947,793</td>
<td>13</td>
</tr>
</tbody>
</table>

Upon reviewing Table 5 and Table 6, we can immediately see that the risk-based policies and screening policies with a stopping age of 70 are best in terms of cost-effectiveness and the cost of false positives, but are the worst policies in terms of saving life. The magnitude of the shifts in the performance measures that quantify the saving of life are much too large to consider using one of the most cost-effective policies.

After reviewing the full sets of results for the best policies from optimization 1, it quickly becomes clear that simply minimizing the cost per QALYs saved has detrimental effects on the other performance measures; and our breast cancer advisors stated that those effects could not be tolerated. For example, there was a 15% increase in women diagnosed in the distant stage of breast cancer. This type of shift in the stage distribution was deemed unacceptable. However for the best policy according to optimization 1, the cost-effectiveness of that policy was approximately $29,000 per QALY saved; and this is well below the threshold (GDP per capita) of $50,000 for a cost-effective screening policy. Thus, we decided that a more appropriate method for determining the best screening policy for older US women would be based on optimization problems 2 or 3, wherein an upper bound of $50,000 is imposed on the cost/QALY saved while respectively minimizing deaths due to breast cancer or maximizing QALYs saved. We now move to the results from the best group of policies identified by optimizations 2 and 3, and we note that two of the policies that were the best in terms of all five measures for optimization 1 (policies 4 and 5) are also in the set of the best of policies for optimizations 2 and 3.

The best five screening policies in terms of all five performance measures from optimizations 2 and 3 are:

1. Annual screening stops at age 80;
2. Annual screening stops at age 75;
3. Annual screening for top 10% in terms of risk, every two years for everyone else, and screening stops at age 80;
4. Annual screening for top 5% in terms of risk, every two years for everyone else, and screening stops at age 80; and
5. Biennial screening stops at age 80.
Ranked results for each performance measure for each policy using optimization 1 are available in Table 7 and Table 8.

Table 7. Top 5 Policies Ranked in Terms of Top 5 Performance Measures for Optimization 2 Part I

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>513.6</td>
<td>1</td>
<td>1238.1</td>
<td>1</td>
<td>21.2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>522.9</td>
<td>2</td>
<td>1088.8</td>
<td>4</td>
<td>26</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>525.7</td>
<td>3</td>
<td>814.9</td>
<td>2</td>
<td>25.1</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>527.5</td>
<td>4</td>
<td>803.5</td>
<td>3</td>
<td>25.4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>531.8</td>
<td>5</td>
<td>773.4</td>
<td>5</td>
<td>26.4</td>
</tr>
</tbody>
</table>

Table 8. Top 5 Policies Ranked in Terms of Top 5 Performance Measures for Optimization 2 Part II

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>$43,908</td>
<td>5</td>
<td>$9,947,794</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>$37,410</td>
<td>4</td>
<td>$8,236,844</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>$45,086</td>
<td>3</td>
<td>$5,581,041</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>$45,053</td>
<td>2</td>
<td>$5,361,173</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>$42,843</td>
<td>1</td>
<td>$4,822,807</td>
<td>18</td>
</tr>
</tbody>
</table>

Upon reviewing Table 7 and Table 8, we can see that in terms of cancer deaths, QALYs saved, and percentage of cancers diagnosed in the distant stage, the best screening policy is annual screening with a stopping age of 80. Although this policy ranks third in cost-effectiveness and last in cost of false positives, these are the less important measures; and from a practical standpoint we do not feel the negative impact on life saved cause by using annual screening stopping at age 75 is justified by the increase in cost-effectiveness or by the decrease in cost of false positives. For example, on average there are 9.3 fewer cancer deaths for annual screening stopping at age 80 as compared with stopping age 75; however, we must remember that since we are only simulating 0.1% of the population, this is actually 9,300 fewer cancer deaths, which we consider practically significant.

There are trade-offs between these five main performance measures, and no policy will be best in terms of all five measures. After reviewing the key performance measures for the best policies from optimization 3, we determined that only the top two policies merited further comparison on a statistical basis: annual screening stopping at age 80 and annual screening stopping at age 75. From a practical standpoint, the best policy was superior to the second best in terms of three of the five key performance measures; and, the other two performance measures were the less-important outputs of cost-effectiveness and total cost of false positives. It is important to remember that performance measures that are not rates
or percentages, like the number of breast cancer deaths or the number of QALYs saved, are for 0.1% of the population; therefore these results should be multiplied by 1,000 in order to scale up to the entire population of women 65 and older. Paired Student $t$-tests can be used to identify the policy that is statistically superior to the others in terms of the most important performance criterion. We performed paired-comparison Student $t$-tests using information from each of the ten randomly sampled populations, and across each of the five major performance measures. Table 9 shows the differences across populations for all performance measures, and presents the results of the paired $t$-tests.

Table 9. Results from Paired t-tests on Two Best Policies

| Differences Between Annual Screening Stopping at 80 vs. Annual Screening Stopping at 75 |
|---------------------------------|----------------|----------------|----------------|----------------|
| Population | BC Deaths | QALY Saved | Cost/ QALY Saved | % Distant | False Positive Costs |
| 1 | 10.0 | -36.1 | $12,010 | -6.7 | $1,761,258 |
| 2 | -5.0 | 81.3 | $5,919 | -3.7 | $1,663,328 |
| 3 | 7.0 | 52.0 | $6,902 | -3.4 | $1,869,071 |
| 4 | -49.0 | 265.6 | $3,517 | -6.6 | $1,785,903 |
| 5 | -33.0 | 106.3 | $3,219 | -6.6 | $1,789,784 |
| 6 | 8.0 | 98.8 | $4,295 | -4.3 | $1,695,975 |
| 7 | -8.0 | -21.6 | $7,574 | -3.5 | $1,752,920 |
| 8 | -18.0 | 44.1 | $9,377 | -3.4 | $1,376,040 |
| 9 | -11.0 | 133.9 | $5,846 | -4.3 | $1,728,794 |
| 10 | 6.0 | -18.9 | $6,322 | -5.1 | $1,686,427 |

<table>
<thead>
<tr>
<th>Mean Difference</th>
<th>1.8%</th>
<th>4.0%</th>
<th>14.8%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Deviation</td>
<td>19.5</td>
<td>90.2</td>
<td>$2,693</td>
</tr>
<tr>
<td>95% CI Half-Width</td>
<td>13.9</td>
<td>64.5</td>
<td>$1,927</td>
</tr>
<tr>
<td>95% CI Lower Limit</td>
<td>-23.2</td>
<td>6.1</td>
<td>$4,571</td>
</tr>
<tr>
<td>95% CI Upper Limit</td>
<td>4.6</td>
<td>135.0</td>
<td>$8,425</td>
</tr>
<tr>
<td>Percentage Difference</td>
<td>1.8%</td>
<td>4.0%</td>
<td>14.8%</td>
</tr>
<tr>
<td>Statistical Difference</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>$p$-value</td>
<td>0.165</td>
<td>0.035</td>
<td>3.225E-05</td>
</tr>
</tbody>
</table>

There was not a statistically significant difference in the number of breast cancer deaths, but there was a statistically significant different for all other performance measures. For QALYs saved and the percentage diagnosed in the distant stage, annual screening stopping at age 80 is statistically better. However, for cost-effectiveness and the cost of false positives, annual screening stopping at age 75 is statistically better. As previously discussed, we put more weight on saving lives than on cost-effectiveness and false positives; and in terms of saving lives, we can declare annual screening stopping at age 80 the statistically best policy, and the cost-effectiveness of this policy is under the cost-effectiveness threshold. The only characteristics that could be perceived as negative are the high cost and number of
false positive exams and benign biopsies. However, there would not be a significant increase in these parameters over what they were from 2001–2011; and for that period we tolerated a similar number of false positives. Decreasing false positives comes at the cost of saving lives, which we argue is not worth the trade-off.

5.4. Full set of results for statistically best screening policy

We now present important results for annual screening stopping at age 80, our recommended screening policy. In this article, we present the results for 2012–2020 under our recommended best policy. In Tejada (2012) for most measures we included the results for 2001–2011 under the calibrated policy and the results for 2012–2020 under the calibrated policy (if we continued as we did in the past). This section is also intended to provide a demonstration of the complex and powerful results than can be delivered by simulation modelling applications when the inputs, modelling structure, and output analysis and rigorously developed. Unlike most analytical models, simulation has the ability to account for thousands of complex inputs and outputs, probability distributions of any kind, and interact with users through interfaces, as well as incorporate large-scale optimization and complex statistical analysis. The most comprehensive review of results can be found in Chapter 4 of Tejada (2012). The results presented here are divided up into three sections: cancer incidence and death rates, cancer detection and treatment, and costs and life-years saved.

5.4.1. Results for cancer incidence and death rates

In this section we present the results for cancer incidence rates (both DCIS and invasive cancers) and cancer death rates in Figure 3, Figure 4, and Figure 5. Each graph shows the mean values, 95% upper and lower confidence bands, and if applicable, SEER data for the years 2001–2020. Note that SEER data is only available for 2001–2010 for most cases.

![Invasive Cancer Incidence Rate Per 100,000 Women](image)

**Figure 3.** Invasive Cancer Incidence Rates for Statistically Best Policy
In Figure 3, one can notice the discontinuity that happens starting in the year 2012. This is because the screening policy is the calibrated policy of annual screening stopping at age 75 during the period 2001–2011, and then is changes to annual screening stopping at age 80 starting during the period 2012–2020. Another piece of evidence that the model is valid the fact that it tracks the SEER data remarkably well from during the period 2001–2010, staying in within the 95% CI bands after a short warm-up period.

![DCIS Incidence Rate Per 100,000 Women](image)

**Figure 4.** DCIS Incidence Rates for Statistically Best Policy

Figure 4 shows the same behavior as Figure 3, the same jump in 2012, and the model again track the SEER data reasonably well for the period 2001–2008. Similar conclusions can be made in terms of model validation, and in terms of the effect of changing the screening policy. The scale of this graph allows us to the shift better than the graph of invasive cancer rates.

![Breast Cancer Death Rate Per 100,000 Women](image)

**Figure 5.** Breast Cancer Death Rates for Statistically Best Policy
Figure 5 shows the breast cancer death rates for the entire simulation time horizon 2001–2020. In addition, SEER data are shown for the period 2001–2010, and a linear projection of the same SEER data is shown for the period 2011–2020, the difference shown by contrasting line styles. Breast cancer deaths are the statistic that undergoes the longest warm-up period, and hence our choice to compare our results to the linear projection of the future. One can see there is a slightly negative slope in the SEER data, indicating the breast cancer death rate is dropping over time, which is plausible given advances in screening technology and the additional attention brought to the importance of screening over the last decade. For the first five years starting in 2010, the SEER data is well within the confidence bands, after those five years the continuation of the negative linear trend causes the SEER projection to be less than the simulation estimates. It is impossible to know if that trend will continue for the next 9 years, however, this is best estimate available. It should be noted that were this trend to begin to level off anytime during the period 2011–2015, the SEER data would be well within the confidence limits for the entire future projection.

5.4.2. Results for cancer detection and treatment

In this section we present the results for the method of detection & percentage of benign biopsies, percentage of women treated, and stage distribution at diagnosis in Figure 6, Figure 7, and Figure 8 respectively. Results presented here are for annual screening stopping at age 80 for the period 2012–2020, as well as for the historically calibrated policy for 2001–2011.

![Method of Detection and Benign Biopsy Percentages for Statistically Best Policy](image1)

**Figure 6.** Method of Detection and Benign Biopsy Percentages for Statistically Best Policy

![Biopsy Outcomes Distribution 2012-2020](image2)

**Figure 7.** Percentage of Women Treated for Breast Cancer for Statistically Best Policy
5.4.3. Results for Screening-and-Treatment Costs and Life-Years Saved

In this section we present all the results related to costs and life-years saved. We graph cost by procedure type, costs by method of detection, costs of false positive exams by procedure type, QALYs saved by method of detection, cost per QALY saved by method of detection, life-years saved by method of detection, and cost per life-year saved by method of detection. We present each of these results for both the historically calibrated screening policy and the statistically best screening policy for the years 2012–2020.
Figure 11. False Positive Costs by Procedure Type

Figure 12. QALYs Saved

Figure 13. Cost Per QALY Saved
In addition to the experimentation with screening policies, we also conducted a small experiment with the SD input levels in order to demonstrate the effects of changes in the SD input levels on the other SD levels and outputs. There are two SD inputs with three levels, and two other SD inputs with five levels, yielding a total of 225 \( (3^2 \times 5^2) \) combinations of SD inputs. Given that SD type results, which are typically yearly values of numerous levels, require a large number graphs in our particular case, we chose to simply demonstrate the ability of changes in the SD inputs to affect the output of the simulation. Adherence is a function of the both SD levels and individual DES attributes, so when changes in the SD input levels affect the adherence, this is evidence that the SD and DES submodels are interacting as they were designed to do. In order to demonstrate this, we choose to take our statistically best screening policy and run best-case and worst-case scenarios with the input levels with that same screening policy in effect. This strategy is a type of blocking which removes the effects of the screening policy, and simply shows...
the effects of changing the SD input levels on the other SD levels and model adherence related model outputs. In Section 4.5.4 of Tejada (2012), a series of graphs for both best-case and worst-case scenarios for SD input levels, intermediate SD levels, primary SD levels, and adherence outputs are given, and these could be compared with the graphs of these measures for the default levels of the SD inputs given in Section 4.5.3 of Tejada (2012).

6. Summary and Conclusions

This section presents conclusions, contributions to breast cancer screening policy, contributions to simulation modeling methodology, limitations of the model, and potential future work.

The most cost-effective screening policy for US women at least 65 is annual screening until age 70 for women who are in the top 5% in terms of risk, and a one-time screening at age 68 for all others. However, this optimal policy in terms of cost-effectiveness performs poorly in other areas of performance that are relevant to saving life. After reviewing with our experts the results from this policy and other policies with similar cost-effectiveness, we realized that simply finding the most cost-effective policy was not equivalent to finding the "best" policy to use for women in this age group. Instead, we examined the following alternate objectives while maintaining a reasonable level of cost effectiveness: (a) maximizing QALYs saved, and (b) minimizing the number of breast cancer deaths. We found that from the perspectives of both practical and statistical significance, annual screening for all women from age 65 until age 80 was a superior policy in terms of saving lives; and we presented a full set of results from using this policy. Nevertheless, some policy makers may not judge these performance measures to be the most important; and one of the key features of this model is its ability to evaluate alternate screening policies in terms of almost any relevant performance measure. Thus, others can review full sets of results generated by this model and make decisions about which policy they consider to be the best, and they will have numerical evidence from a statistically validated simulation model to support their perspective.

6.1 Contributions to Breast Cancer Screening Policy

The major contributions of this research to breast cancer screening are: (a) the development of a combined DES/SD simulation tool for the analysis of screening policies and government policies for women at least 65; and (b) comprehensive experimentation with that simulation model to estimate relevant performance measures for screening policies that have the potential to maximize the number of lives saved, to minimize the number of deaths caused by breast cancer, or to maximize system-wide cost effectiveness. We provided a full set of results for such screening policies by determining the proper screening interval and stopping age for individual patients. The simulation can be used by policymakers to inform their decisions about breast cancer screening policy, the number and capacity of breast cancer screening facilities, and the amount of advertising and research for breast cancer. The simulation has the ability to test individualized policies and determine effects on performance.
6.2 Methodological Contributions to Simulation Modeling

The development of the simulation has led to several methodological contributions to simulation modeling. The major contributions are listed below.

1. This simulation can be regarded as a template or a guide for how future combined DES/SD simulation models may be designed for other application domains. In addition, it provides an approach to modeling a complex disease and the screening and treatment of that disease in a population when several disparate performance measures are of key importance.

2. We have identified key performance characteristics to be estimated for breast cancer screening, and we have formulated appropriate point estimators and confidence interval estimators for the expected values of these performance measures.

3. A complete and statistically-based scheme for validating analytical models against actual data. In addition, we outline a procedure for calibrating a model to historical data using this validation scheme.

4. We have formulated a method for sampling a Pearson IV random variable using a transformation of the standard Beta distribution. See Appendix F of Tejada (2012) for a detailed explanation of this procedure and a proof its validity.

6.3 Limitations

As with all mathematical and computer-based models, effective use of the simulation model requires a thorough understanding of the assumptions on which the model is based and of the resulting limitations on the model's applicability in practice. The key limitations of this model are the following.

1. We considered the population of US women at least 65; thus the model should only be applied to screening policies for women at least 65.

2. Add to this list limitations of the validation and calibration developed here.

6.4 Future Work

This modeling methodology as it stands leaves the opportunity for several areas of future work. The most obvious is a version of the model that supports the inclusion of women of all ages and provides the same measures of performance. There is considerable debate about how to screen middle-aged women, and a great deal of literature has focused on those women. Consequently, we choose to inform policy where there was a lack of consensus on the appropriate screening policy as well as a lack of strong evidence supporting what should be done. We could also include a more complicated model of treatment decisions, and a model of survival after treatment that accounts for the type of treatment given. DCIS incidence is accounted for, but our model does not have the ability to account for its progression to invasive cancer, which likely occurs over time with some probability. However, no data or reliable expert opinion was
available to create a model for this progression of DCIS. Along similar lines, once a woman was afflicted with cancer, her survival was computed from survival curves and/or life-tables; the possibilities of curing the disease or recurrence of the disease were not explicitly considered. Future work could be to simulate whether or not the cancer was adequately treated using a treatment decision model; and then based on that information, we could simulate the probability of recurrence. A new risk model for recurrent disease would need to be found in the literature or developed using existing data. Further experimentation with SD model inputs is another possible area of future work. In fact, extending the SD model to include more levels supported by data and creating a less subjective SD model overall could be another potential area of future work. With respect to types of screening policies, we can add as many different screening policies as we want by appropriately extending the model and the user interface.

References


Data collection for this work was supported by the National Cancer Institute-funded Breast Cancer Surveillance Consortium co-operative agreement (U01CA63740, U01CA86076, U01CA86082, U01CA63736, U01CA70013, U01CA69976, U01CA63731, U01CA70040). The collection of cancer data used in this study was supported in part by several state public health departments and cancer registries throughout the U.S. For a full description of these sources, please see: http://www.breastscreening.cancer.gov/work/acknowledgement.html.