11-12-2015

Estimating Likelihood of Having a BRCA Gene Mutation Based on Family History of Cancers and Recommending Optimized Cancer Preventive Actions

Mehrnaz Abdollahian
University of South Florida, mehrnaz@mail.usf.edu

Estimating Likelihood of Having a BRCA Gene Mutation Based on Family History of Cancers and Recommending Optimized Cancer Preventive Actions

by

Mehrnaz Abdollahian

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Industrial Engineering
Department of Industrial and Management Systems Engineering
College of Engineering
University of South Florida

Major Professor: Tapas K. Das, Ph.D.
Getachew Dagne, Ph.D.
Peter J. Fabri, M.D., Ph.D.
Alex Savachkin, Ph.D.
Hui Yang, Ph.D.

Date of Approval:
October 27, 2015

Keywords: Breast Cancer Gene Mutation, Markov Decision Process, Machine Learning Classifiers, Breast and Ovarian Cancers, Robust Optimization

DEDICATION

To the strongest woman in my life, my mom, who taught me the importance of education from early ages. To my dad, who taught me to pursue happiness while achieving my goals. And to my sister who helped me continuously and passionately during all these years.

ACKNOWLEDGMENT

I want to acknowledge first and foremost my major advisor, Dr. Tapas K. Das, for his continues support and guidance. Also, I want to acknowledge the support of Dr. Rebecca Sutphen who generously provided access to data and kept me abreast of the clinician's viewpoint. I want to thank Dr. Peter Fabri for being a significant influencer in my academic life. I sincerely appreciate the contributions of my committee members, Dr. Savachkin, Dr. Dagne, and Dr. Yang, for sharing their knowledge and experience. At the end, my deepest gratitude goes to my parents and my sister for their guidance and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

BRCA1 and BRCA2 are gene mutations that drastically increase chances of developing breast and ovarian cancers, up to 20-fold, for women. A genetic blood test is used to detect BRCA mutations. Though these mutations occur in one of every 400 in the general population (excluding Ashkenazi Jewish ethnicity), they are present in most cases of hereditary breast and ovarian cancer patients. Hence, it is common practice for the physicians to require genetic testing for those that fit the rules as recommended by the National Cancer Comprehensive Network. However, data from the Myriad Laboratory, the only provider of the test until 2013, show that over 70 percent of those tested are negative for BRCA mutations [1]. As there are significant costs and psychological trauma associated with having to go through the test, there is a need for more comprehensive rules for determining who should be tested. Once the presence of BRCA is identified via testing, the next challenge for both mutation carriers and their physicians is to select the most appropriate types and timing of intervention actions. Organizations such as the American Cancer Society suggest drastic intervention actions such as prophylactic surgeries and intense breast screenings. These actions vary significantly in their cost, cancer incidence prevention ability, and can have major side effects potentially resulting in reproduction inability or death. Effectiveness of these intervention actions is also age dependent.

In this dissertation, both an analytical and an optimization framework are presented. The analytical framework uses supervised machine learning models on extended family history of cancers, and personal and medical information from a recent nationwide survey study of women who have been referred for genetic testing for the presence of a BRCA mutation. This framework provides the potential mutation carriers as well as their physician with an estimate of the likelihood of having the mutations. The optimization framework uses a Markov decision process (MDP) model to find cost-optimal and/or quality-adjusted life years (QALYs) optimal intervention strategies for those tested positive for a BRCA mutation. This framework uses a dynamic approach to address this problem. The decisions are made more robust by considering the variation in estimates of the transition probabilities by using a robust version of the MDP model.

This research study delivers an innovative decision support tool that enables physicians and genetic consultants predict the population at high risk of breast and ovarian cancers more accurately. For those identified with presence of the BRCA mutation, the decision support tool offers effective intervention strategies considering either minimizing cost or maximizing QALYs to prevent incidence of cancers.

CHAPTER 1: INTRODUCTION

Early detection of women at high risks of developing breast and ovarian cancers is considered as one of the critical issues in health care. National Cancer Institute (NCI) reported that BRCA1 and BRCA2 gene mutations are accountable for up to 25% and 15% of hereditary breast and ovarian cancer cases, respectively [2]. Based on this report women with BRCA gene mutations have up to 65% and 17% chances of developing breast and ovarian cancers, respectively, compared to 12% and 1.3% chances in the general population, by age 70. BRCA gene mutations may also be associated with increased risks of developing other cancers such as pancreatic, fallopian tube, and peritoneal cancers [2].

BRCA gene mutations are dominant genes that can be passed from an affected parent to their children. It is often assumed that an affected parent has 50% chance of passing the gene to their children regardless of their genders. Given these facts, BRCA gene mutations are mostly seen in families with hereditary breast and ovarian cancers.

There are several tools available to help individuals as well as health practitioners to estimate one's likelihood of having BRCA mutations based on family history of cancers and individual's personal and medical history. National Cancer Comprehensive Network (NCCN) also has guidelines for physicians to refer individuals at high risk of familial breast and ovarian cancers for genetic BRCA testing. A BRCA genetic blood test can detect the

existence of these mutations in an individual. However, this process is costly and based on the Myriad Laboratory report, most people tested are identified with negative test results [1].

1.1 Literature Review

Literature related to finding the likelihood of having a BRCA mutation comprised of empirical and Mendelian models. Stratification by family or personal history is used for empirical models such as Penn [3], Manchester [4], and Frank-Myriad [5]. Whereas, Mendelian models such as BRCAPRO [6] and IBIS [7], are based on Bayesian statistical models given the Mendelian genetic rules of mutation transition.

Empirical models are simple and easy to use. However, for building an accurate empirical model, a large number of samples is needed. Moreover, the rules used for building such models are generalized. For example, family history of cancer is based on the 'total number of cases' in the family instead of individual cases. Mendelian models are more complicated and accurate than empirical models, but they are based on prior probability estimations such as penetrances and allele frequencies which are often underestimated [8] [9]. A more detailed discussion of the models is explained in details in Chapter 2.

In 2004, Marroni et al. compared eight of the existing empirical and Mendelian models for 568 families [9]. The authors concluded that Mendelian models are more accurate overall for predicting the total number of cases with mutations compared to empirical models. However, all the eight models underestimated the likelihood of having mutations in the lower risk population and overestimated individuals at high risk.

In 2008, Antoniou et al. studied the five most recent models for likelihood prediction of BRCA mutations [10]. They concluded that all the models underpredict families with low risk of having a BRCA mutation, which is a large proportion of their study population. Most of these findings, however, are based on populations recruited academically in research centers. Therefore, performances of these models have not been tested yet in the wider population.

In 2011, American BRCA Outcomes and Utilization of Testing (ABOUT) collected nation-wide data from women for whom BRCA testing was requested through a commercial health insurance company [11]. In Chapter 2, the author explains how this data was used to build a novel model using machine learning techniques for estimating likelihood of having a BRCA mutation. Participants of this study have been asked to provide their detailed information of paternal and maternal family histories of cancers as well as their own medical and cancer history information. The proposed approach, free of assumptions of Mendelian models, provide a new and in some cases more accurate way of classifying affected individuals.

The next challenge for individuals tested positive for a BRCA gene mutation is to find the best timing and type of effective intervention actions for preventing breast and ovarian cancers. The appropriate types of actions based on American Cancer Society guidelines are drastic [12]. The screening action recommended is a combination of mammography and magnetic resonance imaging (MRI). Cancer preventive surgeries recommended are prophylactic mastectomy, in which both breasts are removed and prophylactic oophorectomy in which both ovaries are removed. The costs of these actions differ significantly and their

effectiveness changes with time. The open literature offers a handful of simulation driven Markov models that are designed to evaluate ad hoc strategies (sequence of intervention actions) with respect to cost [13] and survival [14]. These simulation-based studies use sensitivity analysis to assess the impact of uncertainties of the input parameters. However, a comprehensive model to select an optimal intervention strategy from the strategy space has not been discussed in the literature. Such a model based tool should have the ability to consider various health state-dependent intervention action choices, their time-dependent impact on health state transition probabilities, and their expected costs and utilities. In this dissertation, the Appendix B presents a published paper entitled *A MDP model for breast and ovarian cancer intervention strategies for BRCA1/2 mutation carriers*. In Chapter 3, Markov decision process (MDP) models capable of determining optimal policies in terms of cost and quality-adjusted life-years (QALYS) is presented. The recommendations by the MDP models can be used as a guideline for both patients and policy holders to make more informed decisions.

Furthermore, to study the effect of estimation errors in measurement of transition probabilities obtained from a simulation study [14], a robust MDP framework is built and tested. The most common way used in the literature for analyzing such uncertainty is sensitivity analysis of specific parameters in the model. However, this ad hoc approach is not fully capable of incorporating the aggregate effect of uncertainties from all the parameter estimations in the model [15]. There have been several papers published in the literature studying the effect of uncertainty in the transition probabilities estimation in a MDP model

[16] [17] [18]. The three main ways to consider uncertainty in MDP models are: a Bayesian approach, an interval approach, and a likelihood based approach. A Bayesian definition of uncertainty proposed by Shapiro et al. [19] is based on a complete knowledge of prior transition probabilities, which often is not the case in practice [17]. Also, an interval approach often leads to a poor statistical representation of uncertainty and very conservative policies. Therefore, in this dissertation a statistically more accurate definition of uncertainty using a likelihood approach proposed by Nilim et al. in 2004 is used [20]. The robust optimal intervention actions are obtained by minimizing/maximizing the expected total cost/reward under the worst-case scenario played by nature. The robust MDP cost-optimal and QALYs-optimal models are presented in more detail in Chapter 4 of this dissertation.

1.2 Research Contributions

The research contributions of the work presented in Chapters 2, 3, and 4 are described next.

Chapter 2 includes the following:

- A more accurate risk estimator model is developed for finding the likelihood of having a BRCA gene mutation given family and personal history of cancers using a gradient boosting model.
- The existing and widely used BRCA likelihood estimation models by genetic consultants are compared with a recent large nation-wide survey data of American women who have undergone BRCA genetic testing.

- Recent and powerful machine learning models such as random forest, regularized logistic regression, and support vector machines are evaluated and compared with the well-known Bayesian and empirical BRCA risk estimators models in the literature.

Chapter 3 includes the following:

- A cost-optimal model is presented for policy holders capable of finding optimal intervention strategies to prevent breast and ovarian cancers for all ages between 30 and 65 for BRCA1 and BRCA2 mutation carriers.
- A QALYs-optimal model is presented capable of finding optimal intervention strategies to prevent breast and ovarian cancers for all ages between 30 and 65 for BRCA1 and BRCA2 mutation carriers.

Chapter 4 includes the following:

- A more reliable optimization model is developed for finding effective cost and QALYs optimal preventive actions for all ages between 30 and 65 for BRCA1 and BRCA2 mutation carriers using a robust MDP model.

1.3 Research Methods

The summary of research methodologies used in this dissertation for Chapters 2, 3, and 4 is explained next.

1.3.1 A Likelihood Estimation of Having a BRCA Mutation

In this section, the models described in Chapters 2, 3, and 4 of this dissertation are summarized. The first model for finding the likelihood of having a BRCA gene mutation

based on a family history of cancers uses a statistical-machine learning framework. This model will provide a guideline for referring an individual for a BRCA genetic testing.

In this dissertation, the state of the art models in the literature are first validated using the dataset from a nationwide study in the United States conducted by American BRCA Outcome and utilization of testing. These models are based on empirical or Mendelian frameworks. Then machine learning classifiers will be built using all the information captured in the survey conducted by ABOUT study on individual's personal and family history of cancers. The models considered in this dissertation consist of gradient boosting model, random forest, support vector machines, and regularized logistic regression. The data have only 9% positive BRCA test results compared to 91% negative results. Since, this data suffers from class imbalance, in order to compare the performance of the models, area under the ROC curve, F1-measure, Mathew correlation coefficient, and the area under the precision-recall curve were used. The machine learning classifiers are explained next.

Gradient boosting model (GBM) is a machine learning classifier which uses an ensemble of weak learners, such as decision trees, iteratively to make a prediction model. In each iteration it gives more weight to misclassified data points. In 2002, the first paper on stochastic gradient boosting was published [21]. It has been widely used since then in the machine learning community and has been recently applied in medical literature [22] [23] [24]. The GBM has more prediction accuracy than decision trees, but it is less interpretable. However, there are measures that can be used to simplify the results of a GBM. Variable importance is one of these tools. Variable importance is proportional to the number of times a variable is

used for splitting weighted by the model improvement. The higher importance corresponds to higher impact on the response variable [25]. The GBM achieves the best outcome results based on the performance criteria as explained in more details in Chapter 2.

Random forest algorithm was developed in 2001 [26]. It is an ensemble of decision trees and uses majority of votes or mode for prediction. It selects a random sample of training data with replacement to build the trees at each iteration. When building a tree, random forest selects a set of covariates at random as splitting candidates for branches. This model is robust to outliers, simple, and can be implemented using parallel computing techniques [26]. It has been used widely in the recent medical literature for classification problems [27] [28] [29]. This method has been implemented for predicting disease risk with an imbalance dataset [28]. The random forest package in R has been reported as the best classifier technique in terms of accuracy when compared with the other 179 classifiers from a wide range of statistical and machine learning methods [30]. More details on parameter tuning and implementation of this method on ABOUT study data are explained in Chapter 2.

Support vector machines (SVMs) are supervised machine learning classifiers. The current version of SVMs are first published in 1995 [31]. SVMs project the input to a higher dimensional space through a linear or non-linear kernel function and find a linear separator for classification problem [32]. This method has a good accuracy and can effectively combine features. SVM is widely used in the classification problems and has been applied in medical literature [32] [33] [34] [35].

Given the training data (x_i, y_i), where $y_i \in (-1, +1)$, the SVM algorithm finds the linear separator as shown in Equation 1.1.

$$w \cdot z + b = 0 \tag{1.1}$$

In Equation 1.1, b is the constant term and z is the projected vector space. The coefficient w is found using:

$$w = \sum_i \alpha_i K(s_i, z) y_i \tag{1.2}$$

In Equation 1.2, α_i is the coefficient, s_i is the support vector, and K(s,z) is the kernel function. This algorithm is discussed in more details in [36]. In Chapter 2 of this dissertation, SVM is applied to the ABOUT study data and is compared with other statistical and machine learning models.

Logistic regression is considered as one of the most popular statistical methods for classification in medical literature [37]. For classification problems with many features, to avoid over-fitting, a regularized version of logistic regression is often used [38]. Logistic regression solves the problem in Equation 1.3.

$$p(y = 1|\theta, x) = \frac{1}{1 + \exp(-\theta^T x)} \tag{1.3}$$

In Equation 1.3, θ represents the parameter of the logistic regression and (x, y), are the data points, in which the binary response variable is $y \in (0, 1)$. A maximum likelihood estimation is used to find the parameters of a logistic regression problem.

A regularized logistic regression adds a constraint to Equation 1.3 that limits the number of parameters that can take non-zero values in the model. Equation 1.3 uses a L1-norm constraint on the number of parameters as shown in Equation 1.4.

$$\min_{\theta} - \sum log(p(y = 1|\theta, x)) + \beta||\theta||_1 \tag{1.4}$$

More details on this method can be found in [38]. In Chapter 2, the implementation of this method on ABOUT study data is explained in more details.

1.3.2 A MDP Model to Find Optimal Cancer Prevention Strategies for BRCA Mutation Carriers

In this section, an optimization framework is implemented to find cost-optimal and quality-adjusted life year-optimal intervention strategies for BRCA gene mutation carriers. A Markov decision process (MDP) is used to model this problem. The state of a BRCA mutation carrier in the MDP model is defined by her age, health status, and prior intervention action history. The value iteration algorithm is utilized to find the optimal intervention strategies for policy makers using cost and for BRCA mutation carriers using quality-adjusted life years. The value iteration algorithm used to solve this MDP problem in Appendix B is described next. The steps for solving a MDP model with a value iteration algorithm are:

Step 1. Set iteration $n = 0$, $\epsilon > 0$, and value at state s, $V_n(s) = 0 \quad \forall s \in S$.

Step 2. For each state, $s \in S$, update the value by using:

$$V_{n+1}(s) = \min_{d \in D_s} \{c(s, d) + (1 - \alpha) \sum_{j \in S} p(j|s, d) V_n(j)\}, \quad \forall s \in \ S, \tag{1.5}$$

where, $c(s,d)$ is the immediate cost of the action d, α is the discount factor, and $p(j|s,d)$ is the transition probability from state s to state j. This process iterates until the convergence is met.

Step 3. Choose the near optimal policy π such that:

$$\pi(s) = \arg\min_{d \in D_s} \{c(s,d) + (1-\alpha) \sum_{j \in s} p(j|s,d) V_n(j)\}, \quad \forall s \in \ S. \tag{1.6}$$

The optimal policies minimize/maximize the total expected cost/reward while avoiding the cancer incidences and death.

1.3.3 A Robust MDP Model to Find Optimal Cancer Prevention Strategies Given Uncertainties in Transition Probabilities

In Chapter 4, performance of the MDP model under transition probability estimation errors was studied using a robust Markov decision process (RMDP) framework. A max-min model was used to obtain robust optimal intervention strategies in the presence of such estimation errors as described below.

$$\prod_{p_s, D_s} := \max_{p \in P_s} \min_{d \in D_s} c(p,d), \forall s \in \ S. \tag{1.7}$$

This RMDP model is solved using a robust dynamic programming algorithm proposed in [17]. The robust dynamic programming algorithm is based on the steps described next.

Step 1. Set iteration $n = 1$ and value of state s, $V_n(s) > 0 \ \forall s \in S$.

Step 2. For all states, $s \in S$, and action, $d \in D$, solve the inner-problem:

$$\sigma_{P_s^d}(V_{nj}) = \max P^T.V_{nj},$$

$$P \geq 0, \quad P.1 = 1, \quad \Sigma F(s,j) \log p(s,j) \geq \beta,$$

where, P is the column vector of transition probabilities, V_{nj} is the column vector of the next states values, $F(s,j)$ is the frequency of visits from state s to state j, $p(s,j)$ is the transition probability from state s to state j, and β is the measure of uncertainty.

Step 3. For all state, $s \in S$, and action, $d \in D$, update the value function by:

$$V_{n+1}(s) = \min_{d \in D_s} \left\{ c(s,d) + (1-\alpha)\sigma_{P_s^d}(V_n) \right\}.$$

This process iterates until the convergence is met.

Step 4. Choose the near optimal policy π such that:

$$\pi(s) = \arg\min_{d \in D_s} \left\{ c(s,d) + (1-\alpha)\sigma_{P_s^d}(V_n) \right\}$$

Robust MDP model solves a a two-layer optimization problem. The inner problem in Step 2 solves the worst-case nature policy based on a chosen uncertainty level. This is an optimization problem which can be solved using a bisection algorithm. In step 4, the optimal policies are chosen to minimize the total expected cost. The more detailed information on

robust dynamic programming employing a likelihood model can be found in Chapter 4. The likelihood function, considered in this dissertation for transition probabilities, is both statistically accurate and numerically tractable [39].

CHAPTER 2: PERFORMANCE OF MACHINE LEARNING MODELS IN PREDICTING PRESENCE OF BRCA MUTATIONS

2.1 Abstract

Accurate prediction of the presence of a BRCA mutation is important as it significantly increases the probability of developing breast and/or ovarian cancers. Existing Mendelian and empirical prediction models may lack satisfactory predictive power. Inclusion of more diverse and/or comprehensive medical and/or family history data may improve performance of BRCA prediction models. A recent study, the American BRCA Outcomes and Utilization of Testing (ABOUT), collected data from a consecutive series of 11,136 individuals requesting BRCA testing through a national commercial health insurer. In this chapter, the ABOUT study dataset is used to examine the power of machine learning models for predicting presence of BRCA mutations. Though machine learning models have received much attention in medical decision making in recent years, they have not been applied to the BRCA prediction problem. In this chapter, the performances of the selected machine learning methods are compared with those from the Mendelian and empirical models widely used by genetic counselors.

Data from the ABOUT study contains variables that are not incorporated into existing predictive models. However, it presents some challenges such as class imbalance (with only 9% positive test results) and missing information. To overcome imbalance, parameters of the

models were tuned by using different class weights and misclassification costs. For missing values, a median/mode imputation method was implemented. Among many machine learning classifiers that have been presented in the literature, gradient boosting model (GBM), random forest (RF), support vector machines (SVM), and regularized logistic regression (RLR) were chosen as, some of the recent comparative studies have found these methods to have relatively better predictive power. For performance evaluation, cross-validation and measurement criteria like the area under the receiver operating characteristic curve (AUC), Matthews correlation coefficient (Phi coefficient), F-measure, and the area under precision-recall plots were used. The performances of the selected machine learning methods were assessed on the ABOUT dataset.

The results show that the GBM outperforms other machine learning methods (SVM, RF, and RLR). Among the currently used clinical models, performance of IBIS is quite comparable to that of GBM. The values of Matthews correlation coefficient, F-measure, AUC, and area under the precision-recall curve for GBM are 0.34, 0.41, 0.76, and 0.29, respectively, and for IBIS are 0.34, 0.39, 0.71, and 0.3, respectively. Family history of BRCA test results collected in the ABOUT study, made a significant difference in the performance of the machine learning models. The performances of the models were assessed again by excluding this information from the features of the data. The performances of all models (with a few exceptions for Myriad) were decreased.

There is significant room for further improvement in the performance of the BRCA prediction models. Experimental results lead us to believe that current data are perhaps missing some of the critical features, which can raise the performance to an acceptable level.

2.2 Introduction

BRCA1 and BRCA2 are gene mutations that drastically increase the chances (up to 20-fold) of developing breast and ovarian cancers for women. A genetic test can detect BRCA1/2 mutations. These gene mutations occur in one for every 400 in the general population other than Ashkenazi Jewish ethnicity for whom it is one in every 50. However, BRCA mutation is present in most patients with hereditary breast and/or ovarian cancers. It is common practice for the physicians to use the rules developed by the National Cancer Comprehensive Network to determine if a BRCA testing is necessary. However, data from the Myriad laboratory (the only provider of the BRCA test until 2013) shows that over 70% of those tested are negative [1]. The high cost of the test and the associated psychological distress warrant a more accurate approach to determining who should be tested.

Existing models in the open literature for predicting presence of BRCA mutation fall under two main categories: empirical and Mendelian. The empirical models use stratification by family or personal history, for example, Myriad tables [1]. Whereas, Mendelian models use statistical predictive tools such as Bayesian statistics. Examples include BRCAPRO [40] and IBIS [41]. Mendelian models generally outperform empirical models since they take into account more detailed history of BRCA mutations and cancers in the pedigree, unlike empirical models. Mendelian models use assumptions about genetic parameters such as

prevalence of BRCA mutations in different ethnicities and races, along with Mendelian rules of gene transmission. However, accurate estimation of the prior probabilities for the parameters presents a limitation for these models. For example, the true prevalence of BRCA mutations is often underestimated due to the shortcomings of genetic testing [8]. Accuracy of the existing Mendelian models suffer from a lack of comprehensive consideration of factors, including race, ethnicity, history of some other types of cancer, previous prophylactic surgeries, and possibility of mutation transmission through the male members of family [42]. Furthermore, most Mendelian models were developed using limited data from academic recruitment studies, and their performances have not yet been fully studied in a wider general population.

In this chapter a more accurate prediction method for BRCA mutations is proposed using a machine learning model (classifier), that is free of prior knowledge/assumptions. The machine learning models considered in this dissertation are: gradient boosting model (GBM), random forest (RF), support vector machines (SVM), and regularized logistic regression (RLR). Though the literature on machine learning presents numerous classification methods, prior comparative studies have suggested that SVM, RF, and GBM tend to perform better [30] [43]. The RLR was included in the study, since logistic regression is used commonly in medical decision making [44]. The chosen machine learning classifiers make use of the expanded family history of other types of cancers besides breast and ovarian, such as prostate, and pancreatic cancers, for up to third degree relatives in the family. The ABOUT survey data presented some challenges such as missing values and class imbalance.

These issues were addressed by using median/mode imputation (for missing values) and by tuning the parameters of models such as class weights and class costs (for imbalance). Other balancing approaches that are examined include SMOTE [45] and propensity scores [46]. The classifiers were tested for various performance measures, including area under the ROC curve (AUC), the area under the precision-recall plot (AUPR), and Matthews correlation coefficient (MCC). MCC ranges between -1 and 1, where 1 indicates perfect correlation between the actual outcome and predicted results (i.e., positive or negative), 0 indicates a random prediction, and -1 shows a negative correlation [47]. The results from the machine learning methods were compared with those from well-known currently used clinical models, including BRCAPRO, IBIS, and Myriad.

2.3 Methodology

In this section, first the ABOUT data set is introduced. Thereafter, the author discusses about the steps for data cleaning and preparation, followed by an outline of the various machine learning models that were implemented.

2.3.1 Data Set

The data from ABOUT study is the first nation-wide patient-reported dataset that includes individuals who had been approved for BRCA testing through one of the largest health insurance company in the U.S. from December, 2011 until December, 2012. A total of 11,136 individuals received a questionnaire (by mail, online, or telephone). The questionnaire is designed to investigate participant's BRCA test outcome, personal medical and surgical history (e.g., breast biopsy and bilateral mastectomy), demographics (age, race and

ethnicity), and family history of cancer (up to third degree relatives, maternal and paternal) including type(s) of cancer, age at diagnosis, and BRCA test results. Among this population study, 3931 individuals responded to the questionnaire and signed the consent form to use their data in the study [11] [48]. Aside from the ABOUT survey data, provider-reported information about each participant's medical and family history of cancers were made available for this research. Some of the information (e.g., history of having triple negative breast cancer) was used to augment participant data. In this study, only women with a risk of having BRCA mutations were considered. Hence, 73 male respondents were excluded. Also, 16 other participants whose responses to the questionnaire were 90% or more incomplete were eliminated from the dataset.

2.3.2 Data Cleaning and Preparation

The ABOUT survey study comprises a high dimensional data set with several hundred features. In order to apply a variety of classifier models, the data needed to be cleaned and prepared. Examples of cleaning include converting descriptive information into categorical variables and gathering additional information through follow up. Data preparation comprised imputing missing values, creating new derived variables, and dealing with class imbalance. Handling missing information is essential for survey data analysis. In this chapter, some of the missing information was retrieved by following up with the participants via email and phone. However, assumptions were made when follow up was not successful. For example, when needed, missing data were replaced by median/mode values of those variables. A critical challenge with the ABOUT data arose from the outcome imbalance with

only 9% positive BRCA test results. To account for the imbalance, some parameters, such as cost and class weights, were tuned while applying some of the classifiers.

Some derived variables were also created. For example, dummy variables were defined for breast cancer diagnosed before age 45 and before age 50, respectively. Other derived variables included a number of first, first and second, first and second and third degree maternal relatives or (separately) paternal relatives with combinations of breast, ovarian, prostate, and pancreatic cancers. The number of maternal or (separately) paternal male relatives with breast cancer was also considered. Age at cancer diagnosis was categorized in 5-year intervals. Separate variables were then considered for the number of first, first and second, first and second and third degree maternal and (separately) paternal relatives with an incidence of one or more cancers in those age intervals. All combined, the prepared data set had a total of 802 variables either original (from the questionnaire) or derived variables.

2.3.3 Statistical and Machine Learning Models

The large number of features and the nature of the problem presents challenges, including correlations between the variables, missing values, and class imbalance. Some of the machine learning methods are well known to overcome these challenges [49]. Gradient boosting model (GBM) can, for example, handle interactions among variables, select important features, and deal with outliers and missing data. These attributes make GBM an attractive choice for the BRCA prediction problem.

For the purpose of performance comparison, some of the well known Mendelian and empirical models, including BRCAPRO, IBIS, and Myriad prevalence tables are used.

2.3.3.1 Machine Learning Models

Gradient Boosting model (GBM) is a non-parametric tool that iteratively combines many weak classifiers that by themselves perform slightly better than chance, to build a strong classifier capable of generating class probabilities. Gradient boosting model (GBM) is a special case of boosting algorithms. At each iteration, GBM builds a new weak model based on the error observed from the whole ensemble models. The flexibility of GBMs in dealing with missing values, correlation, and imbalance makes them a good fit for a variety of classification problems [50]. The GBM method was implemented using the 'dismo' package in R [51].

Random forest (RF) is an ensemble method that uses many independent decision trees for classification. Each tree in random forest is trained on a bootstrap sample of training data using a set of randomly selected features. After a large number of trees have been built, each tree votes for the majority class. Then the votes are used collectively in making decision rules [26]. To implement this method, the 'randomForest' package in R [52] was used. Sample sizes and class weights were tuned to deal with the data imbalance.

Support vector machines (SVM) is a semi-parametric technique developed based on the assumption of linearly separable classes [53]. SVM creates a classification based on a linear combination of the features. It generates weights for covariance based on a transformation of the feature space and then tries to find the best hyperplane that separates the classes. This method is designed for high dimensional data and no assumption is necessary for parametric relationship between the model predictors and outcome. To implement SVM, the 'e1071'

package in R [54] was used. Class weights and cost parameters were tuned to deal with the data imbalance issue.

Logistic regression has been used widely for the classification problems [44]. Regularization is required for cases where a large number of parameters need to be learned. L1-regularized logistic regression has been shown to have good generalization performance in the presence of many irrelevant features [38]. Because of the L1-penalty, the RLR performs both continuous shrinkage and automatic variable selection [55]. For implementing this method, the 'glmnet' package in R [56] [57] was used.

2.3.3.2 State of the Art BRCA Mutation Likelihood Estimators

Myriad mutation prevalence tables [5] present prior probabilities of having a BRCA mutation estimated using ten characteristics of an individual related to age, ethnicity, and family history. The available estimates [5], were obtained by fitting a logistic regression on samples drawn from over 10,000 individuals in Myriads database [58] [59]. It should be noted that Myriad tables consider family history of ovarian cancer and breast cancer only if it was diagnosed before age 50 and these tables have not been updated for several years.

BRCAPRO developed by Parmigiani et al. [6] [60] is a Mendelian model with a Bayesian approach. It uses published BRCA1 and BRCA2 mutation frequencies and cancer specific penetrances to implement Bayesian updating and determine the likelihood of mutation in the pedigree [9]. A 'BayesMendel' package in R [61] is used, which applies the peeling algorithm for analysis of pedigree data in [62]. Several assumptions were made for implementing this model. For instance, if the age of oophorectomy or mastectomy was missing, assumption is

that it was done either at age 40 or at age 60. If the age at second reported breast cancer was missing, it was assumed to be the age of the first breast cancer plus five years. BRCAPRO's query for family member's history of oophorectomy or mastectomy was ignored when such data was not available. Also, when unavailable, marker testing for ER, CK14, CK56, PR, and HER2 inputs were ignored. However, if a participant had a triple negative breast cancer, ER, PR, and HER2 were set to be negatives.

IBIS model developed by Tyrer et al. [7] was the first Mendelian model that utilizes information on endogenous estrogen exposure and history of benign disease using a segregation analysis [63]. IBIS considers information on personal risk factors, BRCA genes, and a hypothetical low penetrance gene carried by 21% of the population [7]. The risk of breast cancer by age 70 for carriers of the hypothetical gene is estimated to be 24% compared to 12% in the general population. IBIS also considers environmental factors such as parity and hormonal factors. Version 6 of the IBIS model [41] was used in this dissertation. In applying the model, the following assumptions were made: a participant is menopausal if she is over 50, premenopausal if less than 45, and perimenopausal between 45 and 50. Hormonal use was ignored due to lack of data. If the age of cancer occurrence was missing for any family member, it was assumed to be either 40 or 60.

2.4 Results

In this section, first, four machine learning models are developed using the ABOUT study dataset. Thereafter, the performances of these models are compared with those from the existing Mendelian and empirical models. The efficacy of the two well-known data bal-

ancing methods: synthetic minority over-sampling technique (SMOTE) [45] and propensity matching score [64] are then tested. For model building and testing, data from 2997 questionnaires of participants with positive or negative results is considered. The responses with either missing or unknown significance are excluded. Finally, the performance analysis of the models is examined by excluding a critical feature (family history of BRCA test result) from the data.

The data was first divided into training and testing data sets containing 70% and 30% of the dataset, respectively. Both data sets were assigned equal proportion of negative to positive results. This was accomplished using 'caret' package in R [65]. Then the variables with near zero variance were removed. This eliminated 139 of the 802 features of the dataset. A 10-fold cross-validation was used for tuning the parameters of the models on the training set. In the rest of this paragraph, tuning of the parameters is explained.

For GBM, there are 5 parameters to tune: loss function distribution, subsampling rate (bag fraction), learning rate (shrinkage), number of trees, and interaction depth. A Bernoulli distribution with a bag fraction of 0.5 was used as recommended in the literature for classification problems [51]. To find the number of trees needed, portioning algorithm is used with 10-fold cross-validation using 'dismo' package in R [51]. The learning parameter was set to 0.01 and the interaction depth was tuned by dividing the data into two sets: training and testing. The parameters for GBM was chosen so as to minimize the area under the ROC curve (AUC).

For RF, two parameters need to be tuned: the number of trees and the number of features needed to grow each tree. The 'caret' package in R was used for this purpose. Sampling by strata was used, i.e., within each class sample size is equal to the number of elements in the class with less frequency. Different estimates for the number of trees was used in the range [500, 1000, 2000, 3000, and 4000].

For implementing SVM, 'e1071' package in R [54] was used. The two parameters need to be tuned are sigma and cost of misclassification for the minority class. A radial kernel and a grid search method to tune the parameters were selected. For implementing RLR no parameter tuning is needed. The 'glmnet' package in R [56] with a 5-fold cross-validation was used.

Table 1 presents the sensitivity and specificity at optimal cut-off points of the machine learning models obtained using the training data set. Among the models, GBM offers the best performance with highest sensitivity 0.71 and second highest specificity of 0.76 (compared to 0.77 for RLR) at the cut-off point of 0.07. GBM selects 160 out of 783 features as influential variables.

Table 5 presents a summary of the top influential features for GBM. The pie chart in Figure 3 shows the aggregate percentage impact of the influential features in different categories: BRCA test results, personal information, hormonal factors, personal cancer information, and family history of cancers on both maternal and paternal sides. The most influential feature (with a 35% impact) is the family BRCA test results. Features related to

maternal and paternal family history of cancers (including breast, ovarian, pancreatic, and prostate cancers) have 24% and 19% impact, respectively.

Table 5 organizes the set of top influential features within each category. Every feature is provided with its percentage of influence, and the top ten features are identified with asterisks. In Table 5 the two most influential features for family history of cancer are: incidence of breast, ovarian, pancreatic, and prostate cancers in the 1st, 2nd, and 3rd degree relatives (including the proband) before age 60 and breast cancer incidence in the first degree relative before age 45.

The performances of the machine learning models were then compared with those from BRCAPRO, IBIS, and Myriad. The models were applied to the testing data set and the following performance measures for comparison were used: AUC, AUPR, F-measure, and MCC. The results are presented in Table 2. Sensitivity- specificity plots are shown in Figure 1. GBM achieves the highest AUC of 0.76 followed by 0.74 for RF and 0.72 for RLR. IBIS and GBM have the highest Matthews correlation coefficient (MCC) of 0.34. GBM and IBIS have the two highest F-measures of 0.41 and 0.39, respectively. GBM and RLR achieve the highest area under the precision-recall (AUPR) graph of 0.29.

Table 4 shows the MCC scores obtained by applying two of the balancing methods: synthetic minority over-sampling technique (SMOTE) [45] and propensity matching score [64]. A 10-fold cross-validation was used on the training set in obtaining the results in Table 4. As the results indicate, applying the balancing methods in this context does not improve the prediction performance of the machine learning models significantly.

The analyses included data both from those who had a known family mutation as well as those who were the first being tested in the family (no a priori knowledge of family mutation). It is well known that the knowledge of BRCA test results in the pedigree has significant influence on the BRCA test outcome. Hence, the data set was further reduced by excluding the participants with known family BRCA test result(s), and the performance assessment of the prediction models were repeated. With this exclusion criteria, the data imbalance increased further to only 5% positive.

Table 3 shows the performance of the models when family history of BRCA testing is excluded from the dataset. Performance of all methods (with a few exceptions for Myriad) decreases compared to those presented in Table 2. This reduction is expected as all the models (except Myriad) make use of the information on BRCA test results in the pedigree; recall that in GBM, it accounted for 35% of the prognostic influence (see Figure 3). IBIS performs better than all other models. Among the machine learning models, GBM performs better. In the absence of family BRCA history, for GBM, paternal family history of cancer accounts for 48% of the prognostic influence. Other influential factors are maternal family history of cancer (23%), personal cancer history (12%), personal information (such as age, race, and history of prophylactic surgeries) (11%), and hormonal factors (6%). Figure 2 shows the sensitivity-specificity plots for the top four models. The data balancing methods were also implemented in this analysis. Like before, balancing does not improve the performance of the models.

2.5 Conclusions

It is estimated that up to 1 million people in the United States carry BRCA gene mutations. The U.S. Preventive Services Task Force (USPSTF) recommends that women with family history associated with an increased risk of BRCA mutations be referred for genetic counseling and consideration of BRCA testing. USPSTF also advises against referring individuals with low risk of hereditary breast and/or ovarian cancer for BRCA testing to avoid potential harm [66] [67]. Hence, it is imperative to have accurate models to predict the likelihood of having a BRCA mutation.

The two main categories of available mutation prediction tools are empirical and Mendelian models. Empirical models such as Myriad tables are easy to use and do not require computer support, but consider only a subset of relevant characteristics. Mendelian models, such as BRCAPRO and IBIS, outperform the empirical models. However, they require input of personal and family history via a computer program and are prone to estimation errors of certain parameters such as penetrance and allele frequencies [9]. A limitation of all empirical and Mendelian models is that they do not fully consider information on other types of cancers such as pancreatic and prostate cancer, which have been linked with having a BRCA mutation [68]. Finally, the data used in developing the models referred here (Myriad, BRCAPRO, and IBIS) were obtained from women with cancer. Therefore, applicability of these models for cancer-free women in the general population is unknown [66].

In this chapter, a set of machine learning based BRCA prediction models were built that benefited from a recent data collected through a study conducted on a national sample of

commercially insured individuals undergoing BRCA testing [11]. In this study, participants provided extensive family cancer history information. The machine learning models that were examined here are gradient boosting model (GBM), random forest, support vector machines, and regularized logistic regression. Among these, GBM attained the highest sensitivity and specificity. Influential factors included family history of breast, ovarian, pancreatic and prostate cancers, as well as male breast cancer.

Results from the machine learning models were compared with those from IBIS, BRCAPRO, and Myriad tables, applied to the same dataset. All the models performed significantly better than chance in discriminating between carriers and non-carriers. Among all the models, GBM had the highest area under the ROC curve. Because this questionnaire data set was imbalanced with only 9% positive BRCA test results, the model performances were assessed using the Matthews correlation coefficient (MCC). MCC performance of IBIS and GBM were comparable and were better than all other models. Though all the models performed significantly better than chance, their MCC values were less than 0.4, which is considered a moderate level of performance. Following the recommendations in the literature on imbalanced dataset [45] [64], two of the data balancing methods, SMOTE and propensity score, were used. These techniques, however, did not significantly improve the MCC scores of the models. Since knowledge of the BRCA test results in the pedigree has a significant influence on the BRCA prediction outcome, the models were tested on the data containing people with no known BRCA history. Models performed lower, suggesting the need for identification of additional features to improve BRCA prediction power.

The study was limited by missing information necessitating assumptions. Nevertheless, the study suggests that the inclusion of additional personal and family cancer history features could improve the BRCA prediction accuracy. Future studies might explore incorporation of other features such as tumor type (e.g., estrogen receptor/progesterone receptor and HER2 status) [10]. Also, further research to develop new methods of class balancing could be beneficial. Finally, evaluation of the usability of the BRCA prediction models in different population settings needs to be done before recommending their routine use [10]. In the end, the IBIS model appears to have reasonable performance characteristics when tested in a recent population of commercially insured individuals undergoing BRCA testing. Given the limitations of this study, performances of the machine learning models do not show a significant improvement over the existing models.

Table 1: Sensitivity and specificity of machine learning models on training data set

Model	Sensitivity	Specificity	Cut-off
GBM	0.71	0.76	0.07
RF	0.64	0.72	0.1
SVM	0.64	0.69	0.08
RLR	0.64	0.77	0.07

Table 2: Performance of machine learning, Mendelian, and empirical models on the test set

Model Category	Model	MCC	AUC	F-measure	AUPR
Machine learning Models	GBM	0.34	0.76	0.41	0.29
	RF	0.32	0.74	0.38	0.25
	RLR	0.32	0.72	0.38	0.29
	SVM	0.3	0.68	0.34	0.23
Mendelian models	BRCAPRO	0.25	0.7	0.3	0.18
	IBIS	0.34	0.71	0.39	0.3
Empirical model	Myriad	0.2	0.57	0.26	0.15

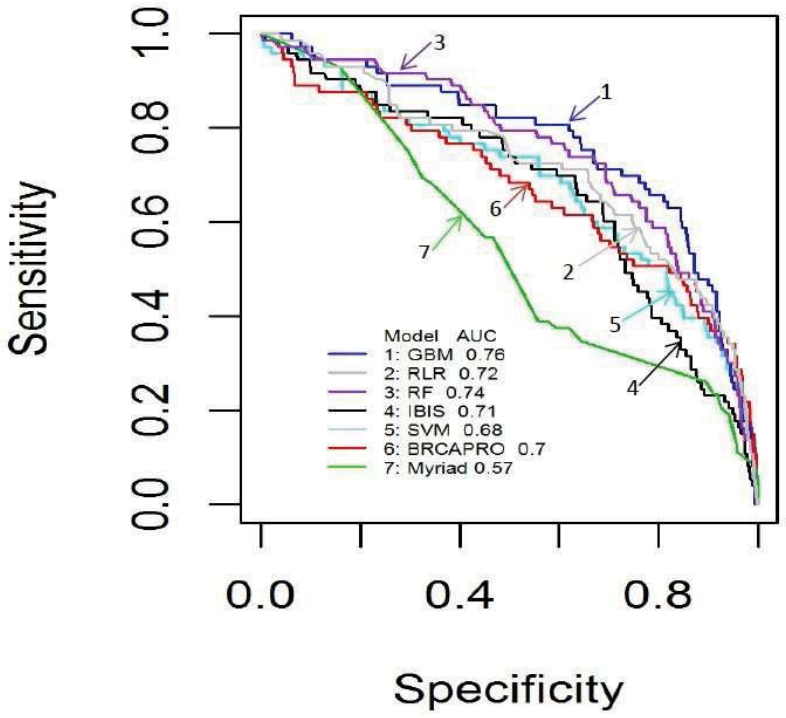

Figure 1: Sensitivity-specificity plot

Table 3: Performance of machine learning, Mendelian, and empirical models on the test set excluding family history of BRCA test

Model Category	Model	MCC	AUC	F-measure	AUPR
Machine learning models	GBM	0.18	0.63	0.29	0.1
	RF	0.14	0.61	0.22	0.09
	RLR	0.1	0.52	0.13	0.07
	SVM	0.14	0.55	0.26	0.08
Mendelian models	BRCAPRO	0.18	0.66	0.26	0.12
	IBIS	0.22	0.65	0.39	0.14
Empirical model	Myriad	0.18	0.63	0.29	0.12

Table 4: Performance of machine learning models after applying balancing methods

Model	MCC-propensity	MCC-SMOTE
GBM	0.27	0.3
RF	0.31	0.34
SVM	0.21	0.24
RLR	0.3	0.3

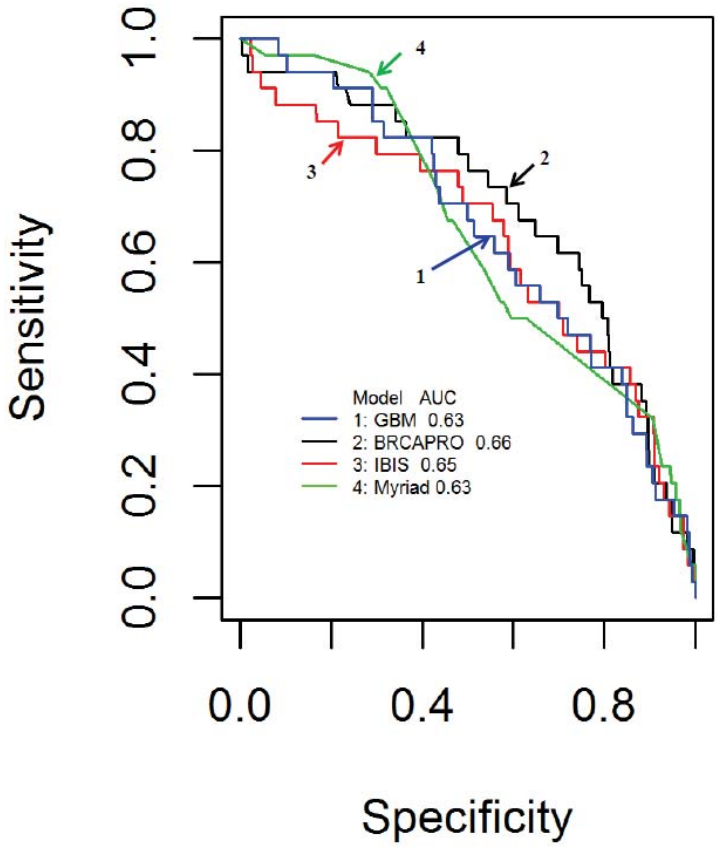

Figure 2: Sensitivity-specificity plot after excluding family history of BRCA test

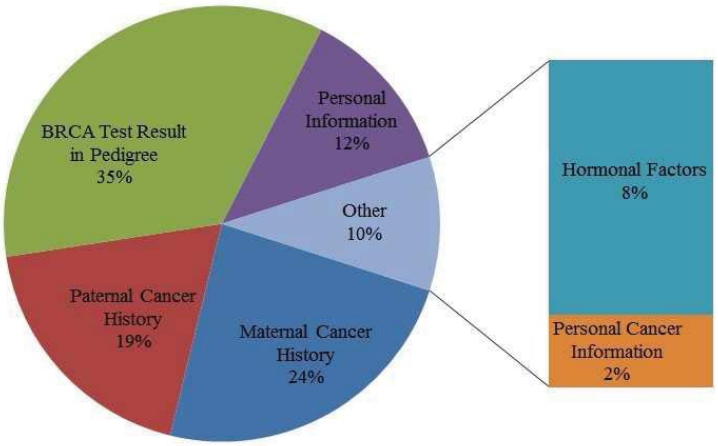

Figure 3: Prognostic factors and their relative influence in the GBM model

Table 5: Summary of the top influential factors and their percentage in the GBM

Personal information
* Age (8%)
* Height and weight (3%)
Race (0.4%)
Ethnicity (0.09%)

Hormonal Factors
* History of Bilateral Mastectomy (2%)
History of bilateral oophorectomy (1%)
Age of first pregnancy (1%)
History of abnormal breast biopsy (0.08%)
Age of start of menstrual (0.04%)

Personal cancer information
Age at first and second breast cancer (0.09%)
Age and History of ovarian cancer (0.08%)
Having a breast cancer before age 45 (0.03%)

Family history cancer information

BRCA test results from first, second, and third degree relatives
* Mother's BRCA test result (20%)
* Sibling's BRCA test result (8%)
* Maternal Aunt and uncle's BRCA test result (3%)

Maternal
* Number of 1st, 2nd, and 3rd degree relatives (including the proband) with breast and ovarian cancers (6%)
* Number of 1st, 2nd, and 3rd degree relatives (including the proband) with breast, ovarian and pancreatic cancers (3%)
Number of 1st , 2nd, and 3rd degree relatives (including the proband) relatives with breast, ovarian, pancreatic, and prostate cancers before age 60 (1%)
Number of 1st degree relatives with breast before age 45 (1%)

Paternal
* Number of 1st, 2nd, and 3rd degree relatives (including the proband) with ovarian cancer (7%)
* Number of 1st, 2nd, and 3rd degree relatives with breast, ovarian , and pancreatic cancers (2%)
Number of 1st and 2nd degree relatives with breast cancer before age 45 (0.07%)
Number of 1st, 2nd, and 3rd degree relatives with breast, ovarian, pancreatic, and prostate cancers before age 60 (0.03%)

* represents the top 10 influential features in the GBM model

CHAPTER 3: A MDP MODEL FOR BREAST AND OVARIAN CANCER INTERVENTION STRATEGIES FOR BRCA1/2 MUTATION CARRIERS

3.1 Abstract

This chapter presents an optimization model that is used to find the cost-optimal and quality-adjusted life year-optimal (QALYs-optimal) intervention strategies for women with BRCA1/2 mutation for ages between 30 to 65. The state of a carrier is defined by her age, health status, and prior intervention actions. Preventive actions considered differ in their costs, major side effects and their cancer prevention abilities. Effectiveness of these actions also depends on the age when they are taken. The surgical actions considered are prophylactic oophorectomy for removing both ovaries, and prophylactic mastectomy for removal of both breasts. Both of these surgical actions cause major side effects. For example, prophylactic oophorectomy increases the risk of heart diseases and osteoporosis. All these considerations make the task of finding optimal intervention strategies a complex decision making problem. At each year starting from age 30, a mutation carrier can make a decision as to whether or not to adopt a screening action. Due to the limitations of data on transition probabilities, the surgical options are considered to be available only at ages 30, 40, and 50. It is assumed that the state of a BRCA carrier at each year depends only on the state of that person a year before. Hence, the intervention decision making problem is modeled as a Markov decision process. The MDP is solved using a value iteration

algorithm. The existing models used in the literature for recommending preventive actions are simulation-based models, which are capable of evaluating a set of policies. A MDP however, is an optimization model that minimizes the total expected cost or maximizes the total expected reward by considering all the possible choices of actions. Details of the MDP model are presented in a recently published paper, *A MDP model for breast and ovarian cancer intervention strategies fro BRCA1/2 mutation carriers*, which can be found in Appendix B. The optimal preventive intervention actions based on cost and QALYs for BRCA1 and BRCA2 mutation carriers suggested by the MDP models can be used as guidelines by both policy makers and individuals.

CHAPTER 4: A ROBUST MDP MODEL UNDER TRANSITION PROBABILITY UNCERTAINTIES FOR BRCA1/2 MUTATION CARRIERS

4.1 Abstract

The Markov decision process (MDP) model, presented in Chapter 3, finds the optimal intervention actions for preventing breast and ovarian cancers for BRCA mutation carriers based on cost or quality-adjusted life years (QALYs). One of the key drivers of the MDP model is transition probabilities. These probabilities were derived from different data sources and simulation studies. Therefore, these probabilities are prone to estimation errors. In this chapter, a robust MDP model (RMDP) capable of dealing with such uncertainty is developed. The robust intervention actions derived from RMDP are presented and compared with those from the MDP model.

4.2 Introduction

The transition probabilities are one of the most important elements of a MDP model. These probabilities explain the stochastic nature of changes in a carrier's health status over time under various interventions. Estimation errors in transition probabilities may thus negatively influence intervention policies obtained from the MDP model. As the transition probabilities were obtained from different public data sources and openly available literature, it was not possible to ascertain the nature of these errors. However, it is essential to develop intervention strategies that are robust to such estimation errors. To accomplish this goal, a

robust Markov decision process model (RMDP) was formulated, from which robust policies were obtained. The RMDP model attempts to optimize the decision criteria assuming that the nature is playing the worst-case scenario. Thus, the RMDP model provides policies that are particularly suited for mutation carriers who are at higher risks of having the worst outcome.

Three different approaches have been presented to the literature that define uncertainty regions for transition probabilities: a Bayesian approach, a polytopes approach, and a statistical likelihood model [17]. The Bayesian approach assumes knowledge of the prior transition probability distributions. Prior probabilities, however, are not always available. The polytopes or interval approach considers uncertainty in a given set. This kind of approach often results in very conservative solutions which may not be statistically accurate. Statistical likelihood models, in most cases, are selected to avoid estimation bias and overly conservative robust policies [17]. The application of the RMDP in medical literature is new but growing. For example, it has been applied for evaluating cost-effectiveness of the fecal immunochemical test screening for colorectal cancer [15]. RMDP has also been used to find optimized medical treatment decisions for patients with type 2 diabetes [69].

This dissertation considers a likelihood statistical model for transition probability uncertainties. A robust dynamic programming approach is used to solve the cost-optimal and QALYs-optimal RMDP models. The formulation of the RMDP is presented next, followed by the results and discussions.

4.3 A RMDP Model Formulation for Finding Optimal Intervention Strategies for BRCA1/2 Mutation Carriers

An RMDP model is built for BRCA1/2 mutation carriers between ages 30 to 65. Model notations used in this section are the same as in Chapter 3. It is assumed that a mutation carrier chooses a preventive action at the beginning of each year. The state of a BRCA mutation carrier, $s \in S$, is defined by her age (a), health status (h), and history of prior intervention actions (i). Intervention actions, $d \in D_s$, are considered to be screening and surgical actions, where D_s denotes the set of actions available in state s. Screening actions are made available yearly and surgical options are made available only at ages 30, 40, and 50. The transition probabilities from state s and action d under uncertainty, $p \in P_s^d$, are assumed to follow a likelihood model by considering the number of times a state is visited under a specified action [70]. A cost-RMDP model objective can be defined as:

$$\min_{d \in D} \max_{p \in P} C(d, p), \tag{4.1}$$

where optimal policies minimize the total expected cost, $C(d, p)$, under the worst-case scenario.

This problem can be solved via a robust dynamic programming algorithm as discussed in [17]. Let $c(s, d)$ be the cost of action d in state s, α be the discount factor, and $\sigma_{P_s^d}(V(n+1)$ defines the worst-case future value. The value of state s in iteration n can be defined as:

$$V_n(s) = \min_{d \in D_s} \left\{ c(s, d) + (1 - \alpha)\sigma_{P_s^d}(V(n+1)) \right\}, \quad \forall s \in S. \tag{4.2a}$$

The optimal policy, $\pi_n^*(s)$, for state s in iteration n is obtained by:

$$\pi_n^*(s) = \arg\min_{d \in D_s} \left\{ c(s,d) + (1-\alpha)\sigma_{P_s^d}(V(n+1)) \right\}, \quad \forall s \in S. \tag{4.2b}$$

In Equation 4.2a, the inner optimization problem presented by $\sigma_{P_s^d}(V(n+1))$ has been solved via a bisection algorithm, as explained in the next section.

4.3.1 A Statistical Likelihood Uncertainty Model

The uncertainty region definition for transition probabilities plays a major role in tractability and conservativeness of the derived policies [69]. Likelihood-based models are considered to be statistically more accurate and less conservative while computationally tractable as discussed in [20]. This uncertainty model is described as follows.

Let N_{sj} denotes the frequency of visits from state s to state j and the transition probability defines by p_{sj}, the log-likelihood model, $L(P)$, for this transition given a Dirichlet distribution can be defined as:

$$L(P) = \sum_{s,j} \log(p_{sj}) N_{sj}, \tag{4.3}$$

where,

$$p_{sj} \geq 0 \quad \forall s, j \in S,$$

$$\sum_j p_{sj} = 1 \quad \forall s \in S.$$

The maximum likelihood β_{max} is obtained by replacing p_{sj} in Equation 4.3 by

$$p_{sj} = N_{sj} / \sum_{k \in S} N_{sk}. \tag{4.4}$$

For an RMDP model with a likelihood uncertainty level, $\beta < \beta_{max}$, the inner optimization problem can be formulated by maximizing the future value as:

$$\max_{p \in P_s^d} \sum_{j \in S} p_{sj} V(j) \tag{4.5}$$

$$L(P) = \sum_{sj} \log(p_{sj}) N_{sj} \geq \beta_s : \quad p_{sj} \geq 0 \quad \forall s, j \in S, \ \sum_j p_{sj} = 1 \quad \forall s \in S,$$

where

$$\beta_s = \beta - \sum_{k \neq s} \sum_j N_{kj} \log(N_{kj} / \sum_{l \in S} N_{kl}).$$

A Lagrangian relaxation method can be implemented for solving this maximization problem. The optimal solution is defined by Lagrangian multipliers λ, μ, and ζ by:

$$p^*_{sj} = \frac{\lambda N_{sj}}{\mu - V(j) - \zeta(j)}. \tag{4.6}$$

As shown in [17], the optimal value of ζ is zero and λ can be written as a function of the μ

$$\lambda(\mu) = (\sum (\frac{N_{sj}}{\mu - V(s)}))^{-1}. \tag{4.7}$$

Hence, in order to find the optimal worst-case scenarios for transition probabilities, the first step is to find μ with a bisection algorithm described below.

Step 1. For a selected level of γ: $\mu_- = V_{max}$ and $\mu_+ = \dfrac{V_{max} - \exp(\beta - \beta_{max})\bar{V}}{1 - \exp(\beta - \beta_{max})}$

$V_{max} = \max(V_j)$, $\bar{V}_s = N^T V$, and $\beta = \sum_{j=1}^{n} N_j \log p_j - 0.5 * \chi^2_{n(n-1),1-\gamma}$

Step 2. While $|\sigma'(\mu_+) - \sigma'(\mu_-)| \leq \delta$:

(a) $\mu = \dfrac{\mu_+ + \mu_-}{2}$

(b) compute σ': $\sigma' = \sum_j N(j) \log \dfrac{\lambda(\mu) N(j)}{\mu - V(j)} - \beta$

(c) if $\sigma' \geq 0$ then $\mu_+ = \mu$ otherwise $\mu_- = \mu$

(d) go to (a)

4.3.2 Robust Dynamic Programming Algorithm

The robust dynamic programming algorithm in [17] is implemented in this Chapter to find the optimal robust intervention strategies for the RMDP model. The algorithm to solve the RMDP model with a likelihood uncertainty is described next.

Step 1. Set $\epsilon \geq 0$, define $V_1 \geq 0$ and iteration $l = 1$

Step 2. Solve the inner problem using the bisection algorithm

$$\sigma_{P_s^d} = \max P^T V \quad \forall s \in S \;\; \forall d \in D_s \tag{4.8}$$

Step 3. For all $s \in S$ and $d \in D_s$ find $V_{l+1}(s)$ using:

$$V_{l+1}(s) = \min_{d \in D_s} \left\{ c(s,d)_+ (1 - \alpha) \sigma_{P_s^d} \right\} \tag{4.9}$$

if $|V_{l+1} - V_l| < \delta$ go to Step 4; else repeat Step 2, where $\delta = \alpha\epsilon/2(1-\alpha)$ and α is the discount factor.

Step 4. Find the optimal policy using:

$$\pi(s) = \arg\min_{d \in D_s} \left\{ c(s,d)_+ (1-\alpha)\sigma_{P_s^d} \right\}, \quad s \in S \tag{4.10}$$

In the next section, the results of the RMDP model are presented and compared with the ones from the MDP model.

4.4 Results

The RMDP model was solved using the robust dynamic programing algorithm. The algorithm was coded in Java and was implemented using an Intel dual core processor with 16 GB RAM. On average, the RMDP took 5 times more than the MDP model to converge. To study the effect of estimation errors on transition probabilities of transient states, the level of uncertainty was altered from 0.1 to 0.5 with an increment of 0.1. Figure 4 displays one-step transition probability changes under different uncertainty levels for a healthy 40 year old BRCA1 mutation carrier who undergone prophylactic mastectomy (PM) and prophylactic oophorectomy (PO) at age 40. At near zero uncertainty, the RMDP transition probabilities are the same as the MDP. As the level of uncertainty increases, the RMDP transition probabilities begin to change. The absorbing state of distant breast cancer has the highest cost (i.e., worst-case scenario). As the level of uncertainty on transition probabilities increases, the probability of ending up in the distant stage of a breast cancer increases and other probabilities decrease.

The RMDP model was solved first by using the cost criteria. For a BRCA1 mutation carrier, with increasing the level of uncertainty, the optimal surgical strategies remain the same, while some screening actions change slightly. This suggests that the MDP model recommendations are robust to the transition probability estimation errors. Whereas, for a BRCA2 mutation carrier, after the level of uncertainty increases to more than 10%, some of the cost-optimal strategies change. The MDP and RMDP with 10% level of uncertainty, cost-optimal intervention strategies are summarized based on the age of a BRCA2 mutation carrier and the prior intervention actions in Table 6 and Table 7, respectively. When the uncertainty level is between 10 to 20 percent, for a healthy 30 year old BRCA2 mutation carrier, the recommendations from both models are the same. However, for a 40 year old BRCA2 mutation carriers with no prior intervention histories, the recommendations differ. For a 40 year old, the MDP model recommends a combination of PM and PO at age 40 as shown in column one of Table 6. Whereas, the RMDP model suggests PM at age 40 but delays PO to age 50 (see Table 7). As, the level of uncertainty increases to 30%, in addition to the previous changes, for a 50 year old with history of PM at age 30, the MDP model advices undergoing PO at age 50. Whereas, the RMDP model recommends yearly screening instead of an additional surgery. The screening actions recommended by the RMDP are only slightly different than the ones suggested by the MDP and can be observed from the tables.

The RMDP model was solved again by considering the states' utilities to find the QALYs-optimal policies. For a BRCA1 mutation carrier, when the uncertainty level is increased to 10%, the recommended actions for a healthy 30 year old with no prior intervention

action changes. The MDP model suggests PO at age 30 followed by PM at age 50 in Table V of the Appendix B but, the RMDP model recommends yearly screening before age 40 followed by a combination of PM and PO at age 40. The RMDP QALYs-optimal policies when the uncertainty level is 10%, are summarized in Table 8. Other strategies remain the same between the MDP and the RMDP model. When the uncertainty level increases to 30%, the RMDP strategies for a healthy mutation carrier with no prior intervention history shifts to no surgery before age 50 and only yearly screening. At the age of 50, the RMDP recommends PM. When the uncertainty reaches 40%, for a healthy mutation carrier with history of PM at age 30, PO is not recommended and substituted with yearly screening and other strategies remain the same. For a BRCA2 mutation carrier, after increasing the uncertainty level to 10%, for a healthy mutation carrier with no prior history, only screening is recommended (see Table 9). After uncertainty level increases to more than 20% for a healthy 40 and 50 year old with prior history of PM at 30, no PO is recommended. The MDP recommended strategies are presented in Table VII of Appendix B.

4.4.1 Assessment of the RMDP Optimal Policies

In this section, some of the RMDP-optimal and MDP-optimal strategies are evaluated and compared. For this purpose, the BRCA Tool [70] health outcome probabilities by age 70 of the recommended strategies is used, similar to Chapter 3. The optimal strategies of the MDP and the RMDP with 30% uncertainty level for a healthy 30 year old BRCA1 mutation carrier with no prior intervention actions are summarized in Table 10. For these optimal policies, Table 12 displays the health outcome probabilities. For cost-optimal strategies,

both RMDP and MDP models suggest a combination of PM and PO by age 30. Therefore, the probabilities reported are the same. However, for QALYs-optimal strategies, the RMDP model has a lower probability of death from other causes of 11% compared to 13% of the RMDP model. However, the RMDP model has higher probabilities for breast and ovarian cancers and lower probability of being healthy. It might be noted that in the RMDP model, death from other causes has the lowest utility for a person i.e., the worst-case scenario. In addition, preventive surgeries, PM and PO, increase the chance of death due to other causes because of their major side effects as explained in Appendix B.

Then, the optimal intervention strategies of MDP and RMDP models are contrasted for a healthy 40 year old BRCA2 mutation carrier with no prior intervention history as listed in Table 11. The health outcome probabilities by age 70 are reported in Table 13. For the cost-optimal strategies, the health outcome probabilities are comparable. However, for QALYs-optimal strategies, the RMDP model has a lower probability of death from other causes of 13% compared to 14% of the MDP model. The RMDP model compared to the MDP model has a higher probability of breast cancer and a lower probability of being healthy by age 70 (see Table 13).

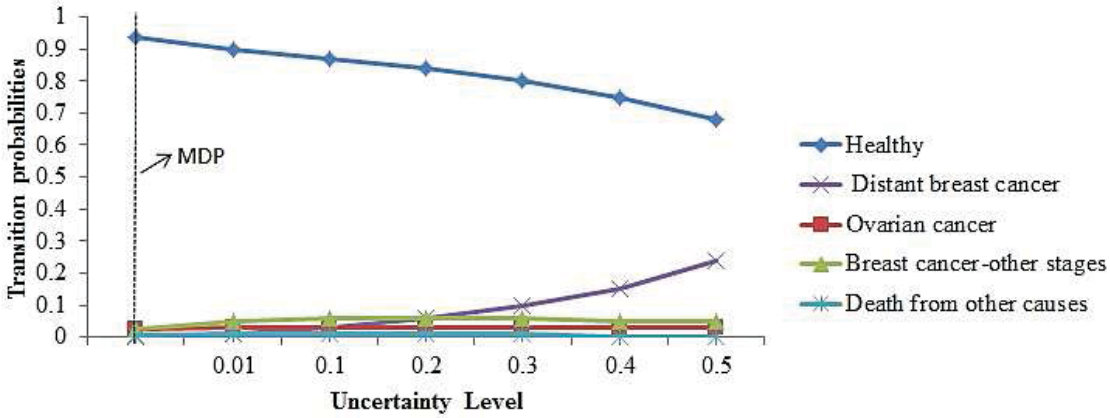

Figure 4: Cost-optimal RMDP one-step transition probabilities for a healthy 40 year old BRCA1 mutation carrier after taking PM+PO-40 action

Table 6: MDP cost-optimal intervention strategies for BRCA2 mutation carriers

Age	None	PO-30	PM-30	PM+PO-30	PO-40	PM-40	PM+PO-40	PM30+PO-40	PM40+PO30	PO-50	PM-50	PM+PO-50	PM30+PO-50	PM40+PO50	PO30+PM40	PO30+PM50
30	PO	NSc	NSc	NSc												
31	Sc	NSc	NSc	NSc												
32	Sc	NSc	NSc	NSc												
33	Sc	NSc	NSc	NSc												
34	Sc	NSc	NSc	NSc												
35	Sc	NSc	NSc	NSc												
36	Sc	NSc	NSc	NSc												
37	Sc	NSc	NSc	NSc												
38	Sc	NSc	NSc	NSc												
39	Sc	NSc	NSc	NSc												
40	**PM+PO**	PM	PO	**NSc**	Sc	NSc	Sc	Sc	NSc							
41	Sc	NSc	NSc	**NSc**	Sc	NSc	Sc	Sc	NSc							
42	Sc	NSc	NSc	**NSc**	Sc	NSc	Sc	Sc	NSc							
43	Sc	NSc	NSc	**NSc**	Sc	NSc	Sc	Sc	NSc							
44	Sc	NSc	NSc	**NSc**	Sc	NSc	Sc	Sc	NSc							
45	Sc	NSc	NSc	**NSc**	Sc	NSc	Sc	Sc	NSc							
46	Sc	Sc	NSc	**NSc**	Sc	NSc	Sc	Sc	NSc							
47	Sc	Sc	NSc	**NSc**	Sc	NSc	Sc	Sc	NSc							
48	Sc	Sc	NSc	**NSc**	Sc	NSc	Sc	Sc	NSc							
49	Sc	Sc	NSc	**NSc**	Sc	NSc	Sc	Sc	NSc							
50	**PM+PO**	PM	PO	**NSc**	PM	**PO**	Sc	Sc	NSc	**Sc**	**Sc**	**Sc**	NSc	**Sc**	Sc	Sc
51	Sc	Sc	NSc	**NSc**	Sc	NSc	Sc	Sc	NSc	**Sc**	**Sc**	**Sc**	NSc	**Sc**	Sc	Sc
52	Sc	Sc	NSc	**NSc**	Sc	NSc	Sc	Sc	NSc	**Sc**	**Sc**	**Sc**	NSc	**Sc**	Sc	Sc
53	Sc	Sc	NSc	**NSc**	Sc	NSc	Sc	Sc	NSc	**Sc**	**Sc**	**Sc**	NSc	**Sc**	Sc	Sc
54	Sc	Sc	NSc	**NSc**	Sc	NSc	Sc	Sc	NSc	**Sc**	**Sc**	**Sc**	NSc	**Sc**	Sc	Sc
55	Sc	Sc	NSc	**NSc**	Sc	NSc	Sc	Sc	NSc	**Sc**	**Sc**	**Sc**	NSc	**Sc**	Sc	Sc
56	Sc	Sc	NSc	**NSc**	Sc	NSc	Sc	Sc	Sc	**Sc**	**Sc**	**Sc**	Sc	**Sc**	Sc	Sc
57	Sc	Sc	NSc	**NSc**	Sc	NSc	Sc	Sc	Sc	**Sc**	**Sc**	**Sc**	Sc	**Sc**	Sc	Sc
58	Sc	Sc	NSc	**NSc**	Sc	NSc	Sc	Sc	Sc	**Sc**	**Sc**	**Sc**	Sc	**Sc**	Sc	Sc
59	Sc	Sc	NSc	**NSc**	Sc	NSc	Sc	Sc	Sc	**Sc**	**Sc**	**Sc**	Sc	**Sc**	Sc	Sc
60	Sc	Sc	NSc	**NSc**	Sc	NSc	Sc	Sc	Sc	**Sc**	**Sc**	**Sc**	Sc	**Sc**	Sc	Sc
61	Sc	Sc	NSc	**NSc**	Sc	NSc	Sc	Sc	Sc	**Sc**	**Sc**	**Sc**	Sc	**Sc**	Sc	Sc
62	Sc	Sc	NSc	**NSc**	Sc	NSc	Sc	Sc	Sc	**Sc**	**Sc**	**Sc**	Sc	**Sc**	Sc	Sc
63	Sc	Sc	NSc	**NSc**	Sc	NSc	Sc	Sc	Sc	**Sc**	**Sc**	**Sc**	Sc	**Sc**	Sc	Sc
64	Sc	Sc	NSc	**NSc**	Sc	NSc	Sc	Sc	Sc	**Sc**	**Sc**	**Sc**	Sc	**Sc**	Sc	Sc
65	Sc	Sc	NSc	**NSc**	Sc	NSc	Sc	Sc	Sc	**Sc**	**Sc**	**Sc**	Sc	**Sc**	Sc	Sc

Table 7: RMDP with 10% uncertainty cost-optimal intervention strategies for BRCA2 mutation carriers

Age	None	PO-30	PM-30	PM+PO-30	PO-40	PM-40	PM+PO-40	PM30+PO-40	PM40+PO30	PO-50	PM-50	PM+PO-50	PM30+PO-50	PM40+PO50	PO30+PM40	PO30+PM50
30	PO	NSc	NSc	NSc												
31	Sc	NSc	NSc	NSc												
32	Sc	NSc	NSc	NSc												
33	Sc	NSc	NSc	NSc												
34	Sc	NSc	NSc	NSc												
35	Sc	NSc	NSc	NSc												
36	Sc	Sc	NSc	NSc												
37	Sc	Sc	NSc	NSc												
38	Sc	Sc	NSc	NSc												
39	Sc	Sc	NSc	NSc												
40	**PM**	PM	PO	NSc	Sc	**Sc**	Sc	Sc	NSc							
41	Sc	Sc	NSc	NSc	Sc	**Sc**	Sc	Sc	NSc							
42	Sc	Sc	NSc	NSc	Sc	**Sc**	Sc	Sc	NSc							
43	Sc	Sc	NSc	NSc	Sc	**Sc**	Sc	Sc	NSc							
44	Sc	Sc	NSc	NSc	Sc	**Sc**	Sc	Sc	NSc							
45	Sc	Sc	NSc	NSc	Sc	**Sc**	Sc	Sc	NSc							
46	Sc	Sc	NSc	NSc	Sc	**Sc**	Sc	Sc	NSc							
47	Sc	Sc	NSc	NSc	Sc	**Sc**	Sc	Sc	NSc							
48	Sc	Sc	NSc	NSc	Sc	**Sc**	Sc	Sc	NSc							
49	Sc	Sc	NSc	NSc	Sc	**Sc**	Sc	Sc	NSc							
50	PM+PO	PM	PO	NSc	PM	**PO**	Sc	Sc	NSc	NSc	Sc	Sc	NSc	**Sc**	Sc	Sc
51	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	NSc	Sc	Sc	NSc	**Sc**	Sc	Sc
52	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	NSc	Sc	Sc	NSc	**Sc**	Sc	Sc
53	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	NSc	Sc	Sc	NSc	**Sc**	Sc	Sc
54	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	NSc	Sc	Sc	NSc	**Sc**	Sc	Sc
55	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	NSc	Sc	Sc	NSc	**Sc**	Sc	Sc
56	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	Sc	**Sc**	Sc	Sc
57	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	Sc	**Sc**	Sc	Sc
58	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	**Sc**	Sc	Sc
59	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	**Sc**	Sc	Sc
60	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	**Sc**	Sc	Sc
61	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	**Sc**	Sc	Sc
62	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	**Sc**	Sc	Sc
63	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	**Sc**	Sc	Sc
64	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	**Sc**	Sc	Sc
65	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	**Sc**	Sc	Sc

Table 8: RMDP with 10% uncertainty QALYs-optimal intervention strategies for BRCA1 mutation carriers

Age	None	PO-30	PM-30	PM+PO-30	PO-40	PM-40	PM+PO-40	PM30+PO-40	PM40+PO30	PO-50	PM-50	PM+PO-50	PM30+PO-50	PM40+PO50	PO30+PM40	PO30+PM50
30	**Sc**	NSc	Sc	Sc												
31	Sc	NSc	Sc	Sc												
32	Sc	NSc	Sc	Sc												
33	Sc	NSc	Sc	Sc												
34	Sc	NSc	Sc	Sc												
35	Sc	NSc	Sc	Sc												
36	Sc	NSc	Sc	Sc												
37	Sc	NSc	Sc	Sc												
38	Sc	NSc	Sc	Sc												
39	Sc	NSc	Sc	Sc												
40	PO+PM	NSc	PO	Sc	NSc	NSc	NSc	NSc	Sc							
41	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc							
42	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc							
43	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc							
44	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc							
45	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc							
46	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc							
47	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc							
48	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc							
49	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc							
50	PM	PM	PO	Sc	PM	NSc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	NSc	Sc	Sc
51	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	NSc	Sc	Sc
52	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	NSc	Sc	Sc
53	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	NSc	Sc	Sc
54	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	NSc	Sc	Sc
55	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	NSc	Sc	Sc
56	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
57	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
58	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
59	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
60	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
61	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
62	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
63	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
64	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
65	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc

4.5 Conclusions

In this chapter, the MDP model presented in Chapter 3 was extended to a robust MDP (RMDP) model by considering uncertainty in transition probabilities. MDP models are powerful decision making tools, but often not used due to their sensitivity to parameter estimation errors [71]. In the open literature, sensitivity analysis was used in Markov chain models for evaluating the effect of uncertainty on the transition probabilities [13] [14]. The transition probabilities of the MDP model were derived from different data sources. Therefore, to better study the total effect of estimation errors on the outcome policies, a RMDP framework was used. This model first finds the confidence regions of the transition probabilities for each state and action, then it optimizes the objective function given the worst-case scenario using those regions [71]. A likelihood definition of uncertainty was used for the RMDP model. The parameters used for this model were the same as in Chapter 3.

Table 9: RMDP with 10% uncertainty QALYs-optimal intervention strategies for BRCA2 mutation carriers

Age	None	PO-30	PM-30	PM+PO-30	PO-40	PM-40	PM+PO-40	PM30+PO-40	PM40+PO30	PO-50	PM-50	PM+PO-50	PM30+PO-50	PM40+PO50	PO30+PM40	PO30+PM50
30	Sc	Sc	NSc	NSc												
31	Sc	Sc	NSc	NSc												
32	Sc	Sc	NSc	NSc												
33	Sc	Sc	NSc	NSc												
34	Sc	Sc	NSc	NSc												
35	Sc	Sc	NSc	NSc												
36	Sc	Sc	NSc	NSc												
37	Sc	Sc	NSc	NSc												
38	Sc	Sc	NSc	NSc												
39	Sc	Sc	NSc	NSc												
40	**Sc**	Sc	PO	NSc	Sc	Sc	Sc	Sc	NSc							
41	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc							
42	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc							
43	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc							
44	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc							
45	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc							
46	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc							
47	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc							
48	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc							
49	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc							
50	Sc	PM	PO	NSc	PM	Sc	Sc	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	NSc
51	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	NSc
52	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	NSc
53	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	NSc
54	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	NSc
55	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	NSc
56	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	NSc
57	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	NSc
58	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	NSc
59	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	NSc
60	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	NSc
61	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	NSc
62	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	NSc
63	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	NSc
64	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	NSc
65	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	NSc

For each experiment, the level of uncertainty was increased by 10%. The cost-optimal and QALYs-optimal strategies were reported.

First, the cost-optimal policies of the RMDP and the MDP model were compared. For a BRCA1 mutation carrier, increasing the level of uncertainty does not significantly change the optimal policies. Therefore, the MDP cost-optimal strategies are robust to changes in transition probabilities. The MDP cost-optimal strategies for a BRCA2 mutation carrier change slightly after increasing the level of uncertainty to 10%. The strategies recommending PM do not change. Therefore, for preventing the worst-case scenario, i.e distant breast cancer, PM is recommended at ages 30, 40, and 50. However, some of the optimal strategies suggesting PO are delayed.

Then, the QALYs-optimal strategies for BRCA1 and BRCA2 mutation carriers were contrasted. For the QALYs models, death from other causes has the lowest utility and hence is considered as the worst-case scenario. In the case of the BRCA1 mutation, the

RMDP optimal strategies changes after increasing the uncertainty level to 10%. The RMDP model for a healthy 30 year old recommends only yearly screening from ages 30 to 40 and delays PO to age 40 and PM to age 50 due to their major side effects. In the case of the BRCA2 mutation, as the level of uncertainty increases, the surgical strategies change to more screening actions similar to BRCA1.

The optimal strategies of the MDP and the RMDP model with 30% uncertainty were selected for a healthy BRCA1 and BRCA2 mutation carriers at ages 30 and 40, respectively. The outcome of the MDP and the RMDP model, with 30% uncertainty level, were contrasted based on the health outcome probabilities by age 70 using the BRCA Tool [70]. For the cost-optimal strategies, the health outcomes for both MDP and the RMDP models are similar. For QALYs-optimal strategies, the RMDP has better outcomes in terms of preventing death from other causes. However, the MDP-recommended strategies have equal or better outcomes for other health states compared to the RMDP model.

Limitations of the RMDP models come from the assumptions made by the original MDP model. For example, surgical actions are still only available at ages 30, 40, and 50. Some of the other shortcomings emerge from the robust MDP model assumptions. For example, the uncertainty for different states are assumed to be uncoupled, which may cause the RMDP strategies to be conservative [72]. Finally, reward values used for solving the MDP models are also prone to estimation errors and can be further considered in the RMDP model. Given, the limitations of the current RMDP model, there is a further room for future research.

Table 10: Comparison of MDP and RMDP optimal strategies for a healthy 30 year old BRCA1 mutation carrier

	Cost-optimal			QALYs-optimal		
Age	**MDP**	**RMDP-10%**	**RMDP-30%**	**MDP**	**RMDP-10%**	**RMDP-30%**
30	**PM+PO**	**PM+PO**	**PM+PO**	**PO**	Sc	Sc
31	NSc	NSc	NSc	NSc	Sc	Sc
32	NSc	NSc	NSc	NSc	Sc	Sc
33	NSc	NSc	NSc	NSc	Sc	Sc
34	NSc	NSc	NSc	NSc	Sc	Sc
35	NSc	NSc	NSc	NSc	Sc	Sc
36	NSc	NSc	NSc	NSc	Sc	Sc
37	NSc	NSc	NSc	NSc	Sc	Sc
38	NSc	NSc	NSc	NSc	Sc	Sc
39	NSc	NSc	NSc	NSc	Sc	Sc
40	NSc	NSc	NSc	NSc	**PM+PO**	Sc
41	NSc	NSc	NSc	NSc	NSc	Sc
42	NSc	NSc	NSc	NSc	NSc	Sc
43	NSc	NSc	NSc	NSc	NSc	Sc
44	NSc	NSc	NSc	NSc	NSc	Sc
45	NSc	NSc	NSc	NSc	NSc	Sc
46	NSc	NSc	NSc	NSc	NSc	Sc
47	NSc	NSc	NSc	NSc	NSc	Sc
48	NSc	NSc	NSc	NSc	NSc	Sc
49	NSc	NSc	NSc	NSc	NSc	Sc
50	NSc	NSc	NSc	**PM**	NSc	**PM**
51	NSc	NSc	NSc	Sc	NSc	Sc
52	NSc	NSc	NSc	Sc	NSc	Sc
53	NSc	NSc	NSc	Sc	NSc	Sc
54	NSc	NSc	NSc	Sc	NSc	Sc
55	NSc	NSc	NSc	Sc	NSc	Sc
56	NSc	NSc	NSc	Sc	NSc	Sc
57	NSc	NSc	**Sc**	Sc	NSc	Sc
58	NSc	NSc	**Sc**	Sc	NSc	Sc
59	NSc	NSc	**Sc**	Sc	NSc	Sc
60	NSc	NSc	**Sc**	Sc	NSc	Sc
61	NSc	NSc	**Sc**	Sc	NSc	Sc
62	NSc	NSc	**Sc**	Sc	NSc	Sc
63	NSc	NSc	**Sc**	Sc	NSc	Sc
64	NSc	NSc	**Sc**	Sc	NSc	Sc
65	NSc	NSc	**Sc**	Sc	NSc	Sc

Table 11: Comparison of MDP and RMDP optimal strategies for a healthy 40 year old BRCA2 mutation carrier

	Cost-optimal		QALYs-optimal	
Age	**MDP**	**RMDP-30%**	**MDP**	**RMDP-30%**
40	**PM+PO**	**PM**	**PM**	**Sc**
41	Sc	Sc	Sc	Sc
42	Sc	Sc	Sc	Sc
43	Sc	Sc	Sc	Sc
44	Sc	Sc	Sc	Sc
45	Sc	Sc	Sc	Sc
46	Sc	Sc	Sc	Sc
47	Sc	Sc	Sc	Sc
48	Sc	Sc	Sc	Sc
49	Sc	Sc	Sc	Sc
50	Sc	**PO**	Sc	Sc
51	Sc	Sc	Sc	Sc
52	Sc	Sc	Sc	Sc
53	Sc	Sc	Sc	Sc
54	Sc	Sc	Sc	Sc
55	Sc	Sc	Sc	Sc
56	Sc	Sc	Sc	Sc
57	Sc	Sc	Sc	Sc
58	Sc	Sc	Sc	Sc
59	Sc	Sc	Sc	Sc
60	Sc	Sc	Sc	Sc
61	Sc	Sc	Sc	Sc
62	Sc	Sc	Sc	Sc
63	Sc	Sc	Sc	Sc
64	Sc	Sc	Sc	Sc
65	Sc	Sc	Sc	Sc

Table 12: Comparison of health outcome probabilities by age 70 for strategies in Table 10 for a healthy 30 year old BRCA1 mutation carrier with no prior intervention history

Outcome measure	**Cost-Optimal**		**QALYs-Optimal**	
Model	**MDP**	**RMDP-30%**	**MDP**	**RMDP-30%**
Strategy	PM+PO-30	PM+PO-30	PO30+PM50	PM-50
Health Outcome				
Death from other causes	0.16	0.16	0.13	0.11
Ovarian cancer	0.07	0.07	0.08	0.3
Breast cancer	0.05	0.05	0.2	0.32
Healthy	0.72	0.72	0.59	0.27

Table 13: Comparison of health outcome probabilities by age 70 for strategies in Table 11 for a healthy 40 year old BRCA2 mutation carrier with no prior intervention history

Outcome measure	**Cost-Optimal**		**QALYs-Optimal**	
Model	**MDP**	**RMDP-30%**	**MDP**	**RMDP-30%**
Strategy	PM+PO-40	PM40-PO50	PM-40	No surgery
Health Outcome				
Death from other causes	0.15	0.14	0.14	0.13
Ovarian cancer	0.03	0.03	0.1	0.1
Breast cancer	0.06	0.07	0.06	0.39
Healthy	0.76	0.76	0.7	0.38

CHAPTER 5: FINAL REMARKS

In this dissertation, a data-driven decision making platform is developed focusing on individuals at high risk of getting breast and ovarian cancers due to their BRCA gene mutations. First a likelihood estimation model is built based on a gradient boosting model. This model is capable of identifying individuals at high risk of having a BRCA mutation based on their family and personal history of cancers. Then, a Markov decision process (MDP) model is formulated to help BRCA mutation carriers and their health providers to find effective intervention actions (based on cost or quality-adjusted life years) to prevent breast and ovarian cancers. Finally, a robust MDP (RMDP) model is presented to study the sensitivity of MDP optimal strategies under uncertainty. In what follows, I present a summary of findings in Chapters 2, 3, and 4 of this dissertation.

In Chapter 2, four machine learning classifiers are used to find the likelihood of having BRCA mutation based on detailed personal and family history of cancer information. The data used for validation of the models emerges from a recent nation-wide survey study (ABOUT) of those who requested BRCA genetic testing through one of the commercial health insurance companies in the United States. This is the first study evaluating existing well-known BRCA risk estimation models using data on general population in the United States. The models considered were gradient boosting model (GBM), random forest, support vector machines, and regularized logistic regression. These models are then compared

and validated using the ABOUT data with some well-known methods in the literature (BRCAPRO, IBIS, Myriad prevalence tables). The GBM model outperforms other existing models as well as other machine learning algorithms based on a selected number of performance criteria such as area under the ROC curve and Matthews correlation coefficient. For the GBM model, history of cancers such as prostate and pancreatic, which were not used in the previous studies, are associated with having a BRCA mutation. The variable importance measure in GBM, finds the history of BRCA testing in the family to be among the most influential factors. However, if this variable is removed from the model, the performance of the machine learning decreases, suggesting the need for additional features. Among the existing models in the literature, IBIS model has a reasonable performance on the test set. The limitations of the GBM model come from the assumptions made due to data incompleteness and imbalance class problems. Features such as estrogen receptor and HER2 status are also not considered in this model.

A genetic consultant or an individual can use the GBM model presented in Chapter 2 to find the likelihood of having the BRCA mutation. If they are at a high risk of having the mutations, they are referred to do a genetic blood test. For those identified with positive mutation results, the next step is to find effective intervention actions to prevent breast and ovarian cancers. The MDP model proposed in Chapter 3 is an attempt to answer this problem. The state of a BRCA mutation carrier is defined by her age, health status, and prior intervention action history. Given the set of screening and surgical intervention actions and cost/utility of these actions in each state, a MDP model is developed. To solve the

MDP model, a value iteration algorithm is used. This framework extends the Markov chain simulation models in the literature from evaluation of several effective policies to offering optimal decision actions. This chapter presents yearly recommendations for BRCA1/2 mutation carriers of ages 30 to 65 with any prior intervention history based on cost/QALYs. The limitations of the MDP models derive from the assumptions made and the data availability on the transition probabilities.

Since transition probabilities play an important role in the solution of a MDP model, in Chapter 4, sensitivity and robustness of the results presented in Chapter 3 are evaluated with the use of a robust MDP (RMDP) platform. A likelihood statistical definition of uncertainty is used for the transition probabilities. The RMDP cost and QALYs optimal intervention strategies for BRCA1 and BRCA2 mutation carriers are reported and compared with the MDP model. The strategies from the MDP model are compared with the ones from the RMDP with respect to health probability outcomes by age 70. The strategies presented by the RMDP can better help those individuals at a high risk of facing the worst-case scenario. The limitations of the RMDP model come from the assumptions of the original MDP models and the RMDP framework as discussed in Chapter 4.

In conclusion, the research presented in this dissertation aims to help individuals and physicians make more informed medical decisions. The more data becomes available through studies such as ABOUT, the more accurate models can be built and the more predictive features can be found to help individuals and their families with difficult and complex medical decisions.

At the end, research on models recommending preventive actions for individuals at high risk of cancers, such as the MDP and the RMDP models, can save millions of lives and millions of dollars in medical treatment expenditure.

REFERENCES

[1] http://www.myriadgroup.com/~/media/Files/Financial\%20Reports/English/Myriad_AR2012.ashx, Myriad Group, AG Annual Report, 2012. Accessed:5/6/2014.

[2] http://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet, BRCA1 and BRCA2: Cancer Risk and Genetic Testing, National Cancer Institute website., 2015. Accessed:8/8/2015.

[3] F. J. Couch, M. L. DeShano, M. A. Blackwood, K. Calzone, J. Stopfer, L. Campeau, A. Ganguly, T. Rebbeck, B. L. Weber, L. Jablon, *et al.*, "Brca1 mutations in women attending clinics that evaluate the risk of breast cancer," *New England Journal of Medicine*, vol. 336, no. 20, pp. 1409–1415, 1997.

[4] D. Evans, D. Eccles, N. Rahman, K. Young, M. Bulman, E. Amir, A. Shenton, A. Howell, and F. Lalloo, "A new scoring system for the chances of identifying a brca1/2 mutation outperforms existing models including brcapro," *Journal of medical genetics*, vol. 41, no. 6, pp. 474–480, 2004.

[5] T. S. Frank, A. M. Deffenbaugh, J. E. Reid, M. Hulick, B. E. Ward, B. Lingenfelter, K. L. Gumpper, T. Scholl, S. V. Tavtigian, D. R. Pruss, *et al.*, "Clinical characteristics of individuals with germline mutations in brca1 and brca2: analysis of 10,000 individuals," *Journal of Clinical Oncology*, vol. 20, no. 6, pp. 1480–1490, 2002.

[6] G. Parmigiani, D. A. Berry, and O. Aguilar, "Determining carrier probabilities for breast cancer–susceptibility genes brca1 and brca2," *The American Journal of Human Genetics*, vol. 62, no. 1, pp. 145–158, 1998.

[7] J. Tyrer, S. W. Duffy, and J. Cuzick, "A breast cancer prediction model incorporating familial and personal risk factors," *Statistics in medicine*, vol. 23, no. 7, pp. 1111–1130, 2004.

[8] H. Kang, R. Williams, J. Leary, C. Ringland, J. Kirk, and R. Ward, "Evaluation of models to predict brca germline mutations," *British journal of cancer*, vol. 95, no. 7, pp. 914–920, 2006.

[9] F. Marroni, P. Aretini, E. DAndrea, M. Caligo, L. Cortesi, A. Viel, E. Ricevuto, M. Montagna, G. Cipollini, S. Ferrari, *et al.*, "Evaluation of widely used models for predicting brca1 and brca2 mutations," *Journal of medical genetics*, vol. 41, no. 4, pp. 278–285, 2004.

[10] A. C. Antoniou, R. Hardy, L. Walker, D. G. Evans, A. Shenton, R. Eeles, S. Shanley, G. Pichert, L. Izatt, S. Rose, *et al.*, "Predicting the likelihood of carrying a brca1 or brca2 mutation: validation of boadicea, brcapro, ibis, myriad and the manchester scoring system using data from uk genetics clinics," *Journal of medical genetics*, vol. 45, no. 7, pp. 425–431, 2008.

[11] J. Armstrong, M. Toscano, N. Kotchko, S. Friedman, M. D. Schwartz, K. S. Virgo, K. Lynch, J. E. Andrews, C. X. A. Loi, J. E. Bauer, *et al.*, "American brca outcomes and utilization of testing (about) study: A pragmatic research model that incorporates personalized medicine/patient-centered outcomes in a real world setting," *Journal of genetic counseling*, pp. 1–11, 2014.

[12] http://www.cancer.org/acs/groups/cid/documents/webcontent/003165-pdf.pdf, Breast Cancer Prevention and Early Detection, American Cancer Society Report, 2015. Accessed:11/10/2015.

[13] K. Anderson, J. Jacobson, D. Heitjan, J. Zivin, D. Hershman, A. Neugut, V. Grann, *et al.*, "Cost-effectiveness of preventive strategies for women with a brca1 or a brca2 mutation," *Annals of internal medicine*, vol. 144, no. 6, p. 397, 2006.

[14] A. W. Kurian, B. M. Sigal, and S. K. Plevritis, "Survival analysis of cancer risk reduction strategies for brca1/2 mutation carriers," *Journal of Clinical Oncology*, vol. 28, no. 2, pp. 222–231, 2010.

[15] J. Goh, M. Bayati, S. A. Zenios, S. Singh, and D. Moore, "Data uncertainty in markov chains: Application to cost-effectiveness analyses of medical innovations," *Working paper*, 2015.

[16] G. N. Iyengar, "Robust dynamic programming," *Mathematics of Operations Research*, vol. 30, no. 2, pp. 257–280, 2005.

[17] A. Nilim and L. El Ghaoui, "Robust control of Markov decision processes with uncertain transition matrices," *Oper. Res.*, vol. 53, pp. 780–798, September-October 2005.

[18] H. Xu and S. Mannor, "Distributionally robust markov decision processes," in *Advances in Neural Information Processing Systems*, pp. 2505–2513, 2010.

[19] A. Shapiro and A. Kleywegt, "Minimax analysis of stochastic problems," *Optimization Methods and Software*, vol. 17, no. 3, pp. 523–542, 2002.

[20] A. Nilim and L. El Ghaoui, *Robust markov decision processes with uncertain transition matrices*. PhD thesis, University of California, Berkeley, 2004.

[21] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[22] K. D. Shetty and S. R. Dalal, "Using information mining of the medical literature to improve drug safety," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 668–674, 2011.

[23] E. J. Atkinson, T. M. Therneau, L. J. Melton, J. J. Camp, S. J. Achenbach, S. Amin, and S. Khosla, "Assessing fracture risk using gradient boosting machine (gbm) models," *Journal of Bone and Mineral Research*, vol. 27, no. 6, pp. 1397–1404, 2012.

[24] Y. Mansiaux and F. Carrat, "Detection of independent associations in a large epidemiologic dataset: a comparison of random forests, boosted regression trees, conventional and penalized logistic regression for identifying independent factors associated with h1n1pdm influenza infections," *BMC medical research methodology*, vol. 14, no. 1, p. 99, 2014.

[25] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *Journal of Animal Ecology*, vol. 77, no. 4, pp. 802–813, 2008.

[26] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[27] F. Yang, H.-z. Wang, H. Mi, W.-w. Cai, *et al.*, "Using random forest for reliable classification and cost-sensitive learning for medical diagnosis," *BMC bioinformatics*, vol. 10, no. Suppl 1, p. S22, 2009.

[28] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC medical informatics and decision making*, vol. 11, no. 1, p. 51, 2011.

[29] T. Shi, D. Seligson, A. S. Belldegrun, A. Palotie, and S. Horvath, "Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma," *Modern Pathology*, vol. 18, no. 4, pp. 547–557, 2005.

[30] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.

[31] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[32] J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana, and A. de Mendonça, "Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests," *BMC research notes*, vol. 4, no. 1, p. 299, 2011.

[33] J. H. Morra, Z. Tu, L. G. Apostolova, A. E. Green, A. W. Toga, and P. M. Thompson, "Comparison of adaboost and support vector machines for detecting alzheimer's disease through automated hippocampal segmentation," *Medical Imaging, IEEE Transactions on*, vol. 29, no. 1, pp. 30–43, 2010.

[34] A. Subasi and M. I. Gursoy, "Eeg signal classification using pca, ica, lda and support vector machines," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8659–8666, 2010.

[35] A. Kharrat, K. Gasmi, M. B. Messaoud, N. Benamrane, and M. Abid, "A hybrid approach for automatic classification of brain mri using genetic algorithm and support vector machine," *Leonardo Journal of Sciences*, vol. 17, no. 1, pp. 71–82, 2010.

[36] P. J. Phillips *et al.*, *Support vector machines applied to face recognition*, vol. 285. Citeseer, 1998.

[37] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," *Journal of biomedical informatics*, vol. 35, no. 5, pp. 352–359, 2002.

[38] S.-I. Lee, H. Lee, P. Abbeel, and A. Y. Ng, "Efficient l˜ 1 regularized logistic regression," in *Proceedings of the National Conference on Artificial Intelligence*, vol. 21, p. 401, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[39] A. Nilim, L. El Ghaou, and V. Duong, "Robust dynamic routing of aircraft under uncertainty," in *Digital Avionics Systems Conference, 2002. Proceedings. The 21st*, vol. 1, pp. 1A5–1, IEEE, 2002.

[40] http://bcb.dfci.harvard.edu/bayesmendel/brcapro.php, BayesMendel Lab, BRCAPRO model, 2013. Accessed:5/6/2014.

[41] http://www.ems-trials.org/riskevaluator/, IBIS Breast Cancer Risk Evaluation Tool, 2014. Accessed:5/6/2014.

[42] J. N. Weitzel, V. I. Lagos, C. A. Cullinane, P. J. Gambol, J. O. Culver, K. R. Blazer, M. R. Palomares, K. J. Lowstuter, and D. J. MacDonald, "Limited family structure and brca gene mutation status in single cases of breast cancer," *Jama*, vol. 297, no. 23, pp. 2587–2595, 2007.

[43] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3446–3453, 2012.

[44] S. C. Bagley, H. White, and B. A. Golomb, "Logistic regression in the medical literature:: Standards for use and reporting, with particular attention to one medical domain," *Journal of clinical epidemiology*, vol. 54, no. 10, pp. 979–985, 2001.

[45] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, no. 1, pp. 321–357, 2002.

[46] R. B. dAgostino, "Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group," *Stat Med*, vol. 17, no. 19, pp. 2265–2281, 1998.

[47] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[48] J. Armstrong, M. Toscano, N. Kotchko, S. Friedman, M. D. Schwartz, K. S. Virgo, K. Lynch, J. E. Andrews, C. X. A. Loi, J. E. Bauer, *et al.*, "Utilization and outcomes of brca genetic testing and counseling in a national commercially insured population: The about study," *JAMA oncology*, pp. 1–10, 2015.

[49] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2005.

[50] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, 2013.

[51] R. J. Hijmans, S. Phillips, J. Leathwick, J. Elith, and M. R. J. Hijmans, "Package dismo," *Circles*, vol. 9, p. 1, 2014.

[52] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[53] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *Intelligent Systems and their Applications, IEEE*, vol. 13, no. 4, pp. 18–28, 1998.

[54] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2012. R package version 1.6-1.

[55] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[56] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[57] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for cox's proportional hazards model via coordinate descent," *Journal of Statistical Software*, vol. 39, no. 5, pp. 1–13, 2011.

[58] J. Powers and J. E. Stopfer, "Risk assessment, genetic counseling, and clinical care for hereditary breast cancer," *Journal of Obstetric, Gynecologic, & Neonatal Nursing*, 2014.

[59] D. A. Berry, E. S. Iversen, D. F. Gudbjartsson, E. H. Hiller, J. E. Garber, B. N. Peshkin, C. Lerman, P. Watson, H. T. Lynch, S. G. Hilsenbeck, *et al.*, "Brcapro validation, sensitivity of genetic testing of brca1/brca2, and prevalence of other breast cancer susceptibility genes," *Journal of Clinical Oncology*, vol. 20, no. 11, pp. 2701–2712, 2002.

[60] D. A. Berry, G. Parmigiani, J. Sanchez, J. Schildkraut, and E. Winer, "Probability of carrying a mutation of breast-ovarian cancer gene brca1 based on family history," *Journal of the National Cancer Institute*, vol. 89, no. 3, pp. 227–237, 1997.

[61] C. Sining, W. Wenyi, W. Broman Karl, A. Katki Hormuzd, P. Giovanni, *et al.*, "Bayesmendel: An r environment for mendelian risk prediction," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, pp. 1–21, 2004.

[62] R. C. Elston and J. Stewart, "A general model for the genetic analysis of pedigree data," *Human heredity*, vol. 21, no. 6, pp. 523–542, 1971.

[63] A. C. Antoniou and D. F. Easton, "Risk prediction models for familial breast cancer," 2006.

[64] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[65] M. K. C. from Jed Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, and T. Cooper, *caret: Classification and Regression Training*, 2013. R package version 5.17-7.

[66] U. S. P. S. T. Force *et al.*, "Genetic risk assessment and brca mutation testing for breast and ovarian cancer susceptibility: Recommendation statement: United states preventive services task force," *The Internet Journal of Oncology*, vol. 3, no. 1, 2004.

[67] V. A. Moyer, "Risk assessment, genetic counseling, and genetic testing for brca-related cancer in women: Us preventive services task force recommendation statement," *Annals of internal medicine*, vol. 160, no. 4, pp. 271–281, 2014.

[68] http://www.cancer.gov/cancertopics/pdq/genetics/breast-and-ovarian/HealthProfessional/page2#Section_1544, National Cancer Institute, Genetics of Breast and Ovarian Cancer, 2014. Accessed:5/6/2014.

[69] Y. Zhang, *Robust Optimal Control for Medical Treatment Decisions.* North Carolina State University, 2014.

[70] A. W. Kurian, D. F. Munoz, P. Rust, E. A. Schackmann, M. Smith, L. Clarke, M. A. Mills, and S. K. Plevritis, "Online tool to guide decisions for brca1/2 mutation carriers," *Journal of Clinical Oncology*, pp. JCO–2011, 2012.

[71] W. Wiesemann, D. Kuhn, and B. Rustem, "Robust markov decision processes," *Mathematics of Operations Research*, vol. 38, no. 1, pp. 153–183, 2013.

[72] S. Mannor, O. Mebel, and H. Xu, "Lightning does not strike twice: Robust mdps with coupled uncertainty," *arXiv preprint arXiv:1206.4643*, 2012.

[73] M. Abdollahian and T. Das, "A mdp model for breast and ovarian cancer intervention strategies for brca1/2 mutation carriers," 2014.

APPENDICES

Appendix A Copyright Permission from the IEEE

Title: A MDP Model for Breast and Ovarian Cancer Intervention Strategies for BRCA1/2 Mutation Carriers
Author: Abdollahian, M.; Das, T.K.
Publication: Biomedical and Health Informatics, IEEE Journal of
Publisher: IEEE
Date: March 2015

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

Appendix B A MDP Model for Breast and Ovarian Cancer Intervention Strategies for BRCA1/2 Mutation Carriers

A MDP Model for Breast and Ovarian Cancer Intervention Strategies for BRCA1/2 Mutation Carriers

Mehrnaz Abdollahian, *Student Member, IEEE*, and Tapas K. Das, *Member, IEEE*

***Abstract*—*Purpose:* Women with BRCA1/2 mutations have higher risk for breast and ovarian cancers. Available intervention actions include prophylactic surgeries and breast screening, which vary significantly in cost, cancer prevention, and in resulting death from other causes. We present a model designed to yield optimal intervention strategies for mutation carriers between the ages of 30 and 65 and any prior intervention history. *Methods:* A Markov decision process (MDP) model is developed that considers yearly state transitions for the mutation carriers and state dependent intervention actions. State is defined as a vector comprising mutation type, health states, prior intervention actions, and age. A discounted value iteration algorithm is used to obtain optimal strategies from the MDP model using both cost and quality-adjusted life years (QALYs) as rewards. *Results:* The results from MDP model show that for 30-year-old women with BRCA1 mutation and no prior intervention history, the cost-optimal strategy is a combination of prophylactic mastectomy (PM) and prophylactic oophorectomy (PO) at age 30 with no screening afterwards. Whereas, the QALYs-optimal strategy suggests PO at age 30 and PM at age 50 with screening afterwards. For BRCA2 mutation carriers at age 30, the cost-optimal strategy is PO at age 30, PM at age 40, and yearly screening only after age 56. Corresponding QALYs-optimal strategy is PM at age 40 with screening. Strategies for all other ages (31 to 65) are obtained and presented. It is also demonstrated that the cost-optimal strategies offer near maximum survival rate and near minimum cancer incidence rates by age 70, when compared to other *ad hoc* strategies.**

***Index Terms*—BRCA1/2 mutations, hereditary breast and ovarian cancer, intervention strategies, Markov decision process (MDP).**

I. Introduction

IT is estimated that more than 300 000 women in the United States carry BRCA1 or BRCA2 gene mutations. These carriers, who make up 5 to 10% of breast cancer patients, have five to 20-fold increased risks of developing breast cancer (56% to 85%) and ovarian cancer (16% to 63%) in their lifetimes [1]. BRCA1/2 mutation carriers may choose to undergo prophylactic mastectomy (PM) and/or prophylactic oophorectomy (PO) to dramatically reduce their risks of breast and ovarian cancers. PM reduces the relative risk of breast cancer in women to 0.1, whereas PO offers a relative risk of 0.6 of breast cancer and 0.04 of ovarian cancer [2]. Most recently, Finch *et al.* [3] reported that PO also reduces overall risk of death by age 70 by 77%. Since the mutation carriers are considered as high risk population, organizations such as American Cancer Society and National Comprehensive Cancer Network suggest annual screening with mammography and magnetic resonance imaging (MRI). MRI offers higher sensitivity in some cases and lower specificity when compared to mammography [4]. Therefore, enhanced cancer screening beginning at a young age using both mammography and MRI with the intent of early detection [5] is considered in this paper. Since the number and type of available drugs for chemoprevention are limited [6] and studies on medications, such as tamoxifen and raloxifene, on reduction of the incidence of invasive breast cancer in BRCA1/2 mutation carriers [7] are limited, we did not consider those in our model.

Mutation carriers and their physicians often struggle with formulating an effective intervention strategy. Usual dilemmas include the choice of preventive surgeries, the age at which to undergo a chosen surgery, and whether or not to screen in any given year [5], [8]. The choice of intervention actions is guided by the following: increased survival, reduced incidence of breast and ovarian cancers, quality-adjusted life years (QALYs), increased probability of death from other causes, and the costs and other undesirable consequences of intervention actions and treatment.

Intervention strategies that include prophylactic oophorectomy are in conflict with the desire of women to have children at older ages. Bilateral salpingo–oophoroctomy causes early menopause, increased heart diseases, higher risk of osteoporosis, and for some women, loss of sexuality and gender identity. Bilateral mastectomy may have a higher psychological impact as it affects the body image [9]. As shown in [10], annual screening with mammography and MRI was linked to the longest quality adjusted survival for women with BRCA1/2 mutation. However, it is expensive and often leads to false positives, thereby increasing anxiety and costs [11].

Finding a good intervention strategy is difficult as the commonly considered strategies have not been assessed comparatively through randomized trials [5]. Most of the studies in the literature also do not fully characterized the patient experience. Moreover, breast cancer inflicts a substantial medical and economic burden to society. Based on the National Cancer Institute report for 2011–2012, $124.6 billion worth of medical care

Manuscript received September 17, 2013; revised November 22, 2013; March 10, 2014; accepted April 12, 2014; Date of publication April 22, 2014; date of current version March 2, 2015.

M. Abdollahian and T. K. Das are with the Department of Industrial and Management Systems Engineering, University of South Florida, Tampa, FL, 33620 USA (e-mail: mehrnaz@mail.usf.edu; das@usf.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2014.2319246

expenditures were made for cancer care in the United States, of which female breast cancer accounted for $16.5 billion and ovarian cancer accounted for $5.1 billion [12]. It was estimated in the American Cancer Society report for 2012 that 39 510 deaths in the U.S. were due to female breast cancer and 15 500 deaths were due to ovarian cancer [12]. It is estimated that a cost-effective intervention strategy for BRCA1/2 mutation carriers may save up to $800 million a year in cancer treatment costs [13].

Open literature offers simulation driven Markov models capable of evaluating given *ad hoc* strategies with regards to cost and survival. To our knowledge, a model-based tool to select an optimal intervention strategy from the set of all possible strategies (strategy space) is not available. Such a model requires the ability to consider the dynamics of variable action choices in different health states, their impact on health state transition probabilities, and their expected costs. Our Markov decision process (MDP) model, as presented in Section II, attempts to fill this gap. See [14] for a discussion on MDP models.

A commonly used method for evaluating decision problems in health care has been "decision tree." But this method is inapplicable in cases where decisions are state dependent and influence the state transition probabilities [15]. The framework of MDP was developed by Bellman (1957), and was extended by Karlin (1955), Howard (1960), Blackwell (1965), to name a few [14]. MDP models have been used in recent years in modeling diverse medical decision making problems including organ transplantation and controlling an epidemic in a closed population [15]. There exist well-known algorithms, such as value iteration, policy iteration, and linear programming, for finding optimal policies from MDP models [14], [15].

Our MDP model uses data from recent medical literature ([5] and [11]) on incidence probabilities of various stages of cancers, utility-weights, availability and cost of interventions, morbidity, and mortality. The model finds optimal strategies by either maximizing the total reward measured in terms of QALYs, or minimizing the total expected cost of intervention and treatment. Our MDP model is capable of identifying optimal intervention actions for any state comprising age, health, and prior intervention status. We considered an expanded and more realistic state space for identifying health and prior intervention status of mutation carriers compared to other models in the open literature. Also in our model, screening option can be turned on or off at any age, whereas in the published literature, the screening decision is made only once for all ages.

II. Markov Decision Process Model for BRCA1/2 Mutation Carriers

We have modeled the yearly state transition process for mutation carriers between ages 30 to 65 as a Markov chain. Subsequently, we have overlayed a decision process model on the Markov chain to form a MDP model. Solution of the MDP model yields the optimal intervention strategy comprising actions at every state. Components of the discrete-time MDP model are decision epoch, state space, decision space, transition probabilities, and reward function, which are described next.

Decision epochs are considered to be the beginning of each new year starting at age 30 till age 65. We define the state of a mutation carrier (s) using a three-tuples as $s = (a, h, i)$, where age $a \in \{30, 31, \ldots, 65\}$, health condition $h \in \{$healthy (hl), breast cancer (bc), ovarian cancer (oc), death from other causes (de)$\}$, and i indicates the intervention status, which is explained later. The breast cancer element (bc) of the health condition is represented by a three-tuple as (ER, size, stage) [11], where ER denotes the estrogen receptor, which can be $+$ or $-$ size indicates the tumor size given by <2 cm or ≥ 2 cm, considered only for the local stage and stage indicates the clinical diagnosis of the tumor as local (l), regional (r), or distant (dt). Hence, there are eight possible conditions of the breast cancer bc element of h, which are $\{(+, <2, l), (+, \geq 2, l), (-, <2, l), (-, \geq 2, l), (+, r), (-, r), (+, dt), (-, dt)\}$. Ovarian cancer ($oc$) element of the health condition h is not further broken into different stages for modeling purposes. Hence, the cardinality of health condition is $|h| = 11$.

The intervention status i of the state of a mutation carrier is given by a two-tuple (sc, su), where sc indicates the screening status and the su indicates the status of preventive surgery. Screening alternatives are $sc \in \{NSc, Sc\}$, where NSc is no screening and Sc is screening with mammography and MRI. Hence, the cardinality of screening status $|sc| = 2$. It is considered that the surgeries can be chosen only at ages 30, 40, or 50, whereas, screening can be chosen at any age. The alternatives for surgery (su) are as follows. For ages 30 up to 39, the surgical status can be $su \in \{$no surgery, PM-30, PO-30, PM+PO-30$\}$, where PM-30 denotes prophylactic mastectomy at age 30, and the other elements are defined similarly. Between the ages 40 and 49, we have $su \in \{$no surgery, PM-30, PM-40, PO-30, PO-40, PM+PO-30, PM+PO-40, PM-30+PO-40, PM-40+PO-30$\}$. Finally, for ages 50 or greater, the intervention status can be $su \in \{$no surgery, PM-30, PM-40, PM-50, PO-30, PO-40, PO-50, PM+PO-30, PM+PO-40, PM+PO-50, PM-30+PO-40, PM-40+PO-30, PM-50+PO-30, PM-50+PO-40, PM-30+PO-50, PM-40+PO-50$\}$. Hence, for ages between 30 to 39 $|su| = 4$, between 40 to 49 $|su| = 9$, and for ages 50 to 65 $|su| = 16$. Therefore, the cardinality of the state space $S = \{s\}$ of a mutation carrier can be given as: $|S| = 10 * 11 * 2 * 4 + 10 * 11 * 2 * 9 + 16 * 11 * 2 * 16 = 8492$.

The available intervention decisions d are considered to be state dependent. At ages 30, 40, and 50, the decisions are either *do nothing* or *conduct surgery and/or screening*. At all intermediate ages (31–39, 41–49, and 51–65) intervention decisions are only either do nothing or start/stop screening. Following are some examples. In state $s_1 =$ (30, *hl*, *NSc*, no surgery), the set of possible decision choices is $D_{s_1} =$ {do nothing, *Sc*, PM, PO, PM+PO, PM+Sc, PO+Sc, PM+PO+Sc}. In state $s_2 =$ (40, *hl*, *NSc*, PM-30), the set of available intervention decisions is $D_{s_2} =$ {do nothing, *Sc*, PO, PO+Sc}. In state $s_3 =$ (52, *hl*, *Sc*, PM-30+PO-40), the set of possible decisions is $D_{s_3} =$ {do nothing, NSc}. In what follows, we characterize the random process that underlie the changes in the state of a mutation carrier.

Let X_a, $a \in \{30, 31, \ldots, 65\}$, denote the state random variable at age a of a mutation carrier. Define a carrier state process $\mathbf{X} = \{X_a : a = 30, 31, \ldots, 65\}$. The probability that

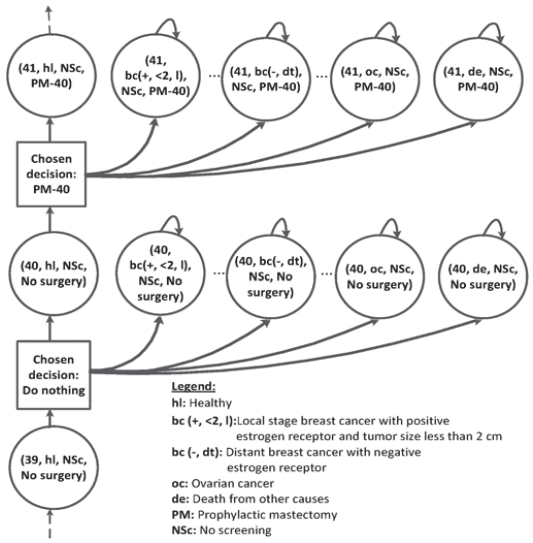

Fig. 1. Sample segment of the one-step state transitions of the MDP model.

a healthy mutation carrier of age a transitions to a particular state at age $a+1$ depends only on her intervention status and health condition at age a and independent of the past. Hence, it can be shown that $P(X_{a+k} = u|X_l, \forall l \leq a; a+k \leq 65) = P(X_{a+k} = u|X_a = v)$, and thus, $\mathbf{X}$ is a Markov chain. Let $\mathcal{D}_a$ denote the decision choice random variable, which is a function of the state random variable X_a. Define $\mathcal{D} = \{\mathcal{D}_a\colon a = 30, 31, \ldots, 65\}$ as the decision process. Then the joint process $(\mathbf{X}, \mathcal{D})$ is a MDP. The structure of the one-step transition probabilities of the Markov chain $\mathbf{X}$ is described next.

Fig. 1 shows a small segment of the one-step transition diagram that captures transitions between ages 39–40 and 40–41 under the decisions to *do nothing* and PM-40, respectively. A BRCA1/2 mutation carrier who is healthy at age 39 and with no history of interventions, can transition at age 40 to any of the health conditions as shown: healthy, breast cancer, ovarian cancer, death from other causes. If the transition state at age 40 is one of the cancer states, which are considered to be absorbing states in our model, the carrier undergoes treatment. If the carrier transitions to healthy state at 40 and decides for PM, the subsequent transitions possible at age 41 are as shown. The probabilities of various transitions depend on the type of mutation, age, and type of chosen intervention. Note that the effect of interventions on the transition probabilities also depend on at what age the intervention was implemented. For example, transition probabilities between ages 40 and 41 would be different if PM was chosen at age 30 or at age 40. It was found in [11] that PM at 30 reduces the incidence of breast cancer at later ages more than PM at 40. This motivated us to supplement the notation of the interventions with age (for example: PM-30, PM+PO-50). Table I exemplifies the one-step transitions in our model by specifying all possible transitions from a particular state of a BRCA2 carrier at age 40 when the action of *NSc* and PO-40 are chosen. The details of how we obtain the probabilities in Table I are given in Table IX of the supplementary web material.

TABLE I
SAMPLE ONE-STEP TRANSITIONS AND CORRESPONDING PROBABILITIES OF A BRCA2 MUTATION CARRIER

Decision	Resulting state at 40	Transition state at 41	Probability
NSc and PO-40	(40, hl, NSc, PO-40)	(41, hl, NSc, PO-40)	0.9386
		(41, bc (+, $<$ 2, l), NSc, PO-40)	0.0107
		(41, bc (+, $\geq$ 2, l), NSc, PO-40)	0.0024
		(41, bc (-, $<$ 2, l), NSc, PO-40)	0.0119
		(41, bc (-, $\geq$ 2, l), NSc, PO-40)	0.003
		(41, bc (+, r), NSc, PO-40)	0.0214
		(41, bc (-, r), NSc, PO-40)	0.0065
		(41, bc (+, dt), NSc, PO-40)	0.003
		(41, bc (-, dt), NSc, PO-40)	0.0012
		(41, oc, NSc, PO-40)	0
		(41, de, NSc, PO-40)	0.0014

III. MDP MODEL SOLUTION

We used the discounted value iteration algorithm [16] for solving our MDP model. Let *n* denote the iteration count, and $V_n(s)$ denote the value of state s in the *nth* iteration. It is shown in (1) how in every iteration the value of the state is updated, where *r(s, d)* is the immediate reward of selecting decision *d* at state *s*, α is the discounting factor, and $p(j|s,d)$ is the one-step transition probability from state s to j under decision d. The value iteration algorithm consists of the following steps.

Step 1. Set $n = 0$ and $V_n(s) = 0\ \forall s \in S$.

Step 2. For each state, $s \in S$, update the value by using:

$$V_{n+1}(s) = \max_{d \in D_s}\{r(s,d) + (1-\alpha)\sum_{j \in S} p(j|s,d)V_n(j)\},$$
$$\forall s \in S. \quad (1)$$

Step 3. Choose strategy π such that:

$$\pi(s) = \arg\max_{d \in D_s}\{r(s,d) + (1-\alpha)\sum_{j \in s} p(j|s,d)V_{n+1}(j)\},$$
$$\forall s \in S. \quad (2)$$

Set $n \leftarrow n+1$.

Step 4. Repeat steps 2 and 3 until $\forall s, |V_{n+1}(s) - V_n(s)| < \theta$, where θ is the chosen threshold value for convergence.

Step 5. Return π as the optimal strategy.

Two main input parameters needed in the implementation of the discounted value iteration algorithm are the one-step transition probabilities ($p(j|s,d)$) and the immediate rewards ($r(s,d)$). In the following subsections, we provide details of how these input parameters are obtained.

A. Computation of One-Step Transition Probabilities

We computed the one-step transition probabilities using the breast and ovarian cancer incidences at specific ages of 30, 40, till 80 (as given in [5]). Since our model considers yearly transitions, we used linear interpolation (as done in [17]) to obtain transition probabilities for intermediate ages. As far as the surgical interventions are concerned, we chose to consider them

only at ages 30, 40, and 50 (similar to [11]). However, the transition probabilities resulting from the surgical interventions that we adopt from [5] are for ages 25, 40, and 50. We used the transition probabilities resulting from surgical interventions at 30 same as that for 25 in [5]. Transition probabilities from healthy (*hl*) to death from other causes (*de*) are calculated using Berkeley mortality database [18]. The mortality rates are reduced by the rates of death caused by breast and ovarian cancers and then further adjusted for the side effects of PO using [11], [18], [19].

Since undergoing PO before age 50 has a two-fold increased risk of cardiovascular disease and a 50% increased risk of osteoporotic hip fracture and dementia, we adjusted the probabilities of death from other causes based on these relative risks [11]. For finding the transition probabilities from a healthy state to a breast cancer state, we used stage and estrogen receptor (ER) proportions from [17]. The probability of transition to the healthy state is found by subtracting from one the total probabilities of transition to other states. When surgeries are implemented at different ages, for example, PM-30 and PO-50, we assumed that the transition probabilities after PO-50 to be the same as PM+PO-30.

We considered that the decision to start or stop screening can be made at any age between ages 30 and 65. Transition probabilities resulting from the screening decision were adopted from [5].

B. Reward: Immediate Cost of Intervention and Treatment Decisions

We considered three types of decisions involving cost: surgery, screening, and cancer treatment. We also considered the cost of death from other causes, which considers only the terminal care cost. Costs of surgeries and screening are assumed to be state independent and their values are adopted from [13] and [21]. For example, cost of PO is considered to be same irrespective of the age it is performed. Cost of cancer treatment in a given stage depends on direct and indirect costs, probability of cancer recurrence, and terminal care cost.

As in [11], we used a stage-based classification. Stage I is defined as local and tumor size less than 2 cm. Stage II is defined as local with tumor size greater than or equal to 2 and some of regional cases. Some of the more advanced regional cases are classified as Stage III. Distant cases are considered stage IV. Since the regional cases can be classified either as Stage II or III, we take the average of the costs as the cost of treatment for regional cases. The cost C_b^k of treating breast cancer in stage $k \in \{I, II, III, IV\}$ is obtained as

$$C_b^k = [C_g^k + C_d^k + C_e^k + C_f^k] * (1 + p_1^k) + p_2^k * C_t \qquad (3)$$

where the cost of diagnosis of cancer C_g^k, includes mammography, diagnostic following initial MRI, and the subsequent follow up MRI, C_d^k is the direct cost of treatment, C_e^k is the indirect cost of lost income during treatment, C_f^k is the cost of follow-up, and C_t is the terminal care cost of the last year of having breast cancer. The probability of recurrence in the next five years, denoted by p_1^k, is obtained from [11], [22], and the probability of death from breast cancer in the next five years, p_2^k, is obtained from [21], [23]. Representative values of these costs are adopted from [13] and [21], which are then adjusted for inflation from 2004 to 2013. The calculated costs are shown in Table II.

TABLE II
UTILITY (FOR QALYS) AND COST VALUES

State	Mean utility weights	Source	Cost ($)	Source
hl	1	[13]	0	
de	0	[13]	34,373	[13]
Disease state treatment				
bc (+, <2, l)	0.86	[20]	86,846	[13]
bc (-, <2, l)	0.86	[20]	91,610	[13]
bc (+, ≥2, l)	0.86	[20]	91,610	[13]
bc (-, ≥2, l)	0.86	[20]	101,172	[13]
bc (+, r)	0.675	[20]	101,172	[13]
bc (-, r)	0.675	[20]	167,804	[13]
bc (+, dt)	0.38	[20]	155,952	[13]
bc (-, dt)	0.38	[20]	254,606	[13]
oc	0.65	[13]	161,619	[13] [21]
Intervention type				
PM	0.76	[13]	13,496	[21]
PO	0.82	[13]	5,518.9	[21]
PM+PO	0.73	[13]	19,015	[21]
Sc	1	[20]	1,667.3	[21]

For stages I and II, we considered that the patients will undergo lumpectomy with axillary dissection, radiotherapy, and endocrine therapy (only for everyone in two patients) as recommended in [24]. Based on [21], we assumed a 12 week long work discontinuity for stages I and II. Stage III and IV treatment considers mastectomy and adjuvant therapy (depending on estrogen receptor (ER) status) [21]. Before age 50, it is estimated that 40% of patients will use bilateral mastectomy and 60% will undergo unilateral mastectomy. After age 50, we consider a 50% chance that patients will choose one of the options [21]. Hence, we used the average cost of unilateral and bilateral mastectomies.

The cost of treatment for ovarian cancer C^o is obtained as

$$C^o = [C_d^o + C_e^o] + p_2^o * C_t^o$$

where C_d^o and C_e^o are the direct and indirect cost of treating ovarian cancer, respectively. p_2^o is probability of death from ovarian cancer in the next five years, and C_t^o is the terminal care cost. All these costs were obtained from [13], [21] and [25] and further adjusted for inflation at rate 3%.

C. Reward: Quality-Adjusted Life Years (QALYs)

We also used QALYs as our reward function. Table II provides the mean utility weights (also referred to in the literature as preference rating) that are used to define the relative value of each disease state. In order to find QALYs for a year of life, one can multiply the utility value of a given state by 1. The utility weights are obtained using time-tradeoff rating and other similar methods discussed in [13] and [20]. Since our decision epochs are at the beginning of each year, the incremental utility of the MDP at each step is considered to be QALY [26]. QALYs-optimal policy maximizes the QALYs for a mutation carrier with no concern for cost.

IV. RESULTS

We solved the MDP model using the value iteration package available in MATLAB [27] using an Intel dual core with 16 GB RAM. We used a 3% discount factor based on an assumed rate

TABLE III
ALL POSSIBLE INTERVENTION STRATEGIES FOR 30 YEAR OLD BRCA1 MUTATION CARRIERS WITH NO PRIOR INTERVENTION

Strategy #	Surgery	Screening
1	No surgery	No
2	No surgery	Yes
3	PM-30	No
4	PM-30	Yes
5	PO-30	No
6	PO-30	Yes
7 (Cost-optimal for BRCA1)	PM30+PO30	No
8	PM30+PO30	Yes
9	PM30+PO40	No
10	PM30+PO40	Yes
11	PM40+PO30	No
12	PM40+PO30	Yes
13	PM30+PO50	No
14	PM30+PO50	Yes
15	PM50+PO30	No
16 (QALYs-optimal for BRCA1)	PM50+PO30	Yes
17	PM-40	No
18	PM-40	Yes
19	PO-40	No
20	PO-40	Yes
21	PM+PO-40	No
22	PM+PO-40	Yes
23	PM40+PO50	No
24	PM40+PO50	Yes
25	PM50+PO40	No
26	PM50+PO40	Yes
27	PM-50	No
28	PM-50	Yes
29	PO-50	No
30	PO-50	Yes
31	PM+PO50	No
32	PM+PO50	Yes

TABLE IV
COMPARISON OF HEALTH OUTCOME PROBABILITIES BY AGE 70 FOR OPTIMAL STRATEGIES FROM DIFFERENT MODELS FOR A BRCA1 MUTATION ZCARRIER

	Anderson et al. [13]	MDP(Cost-optimal)	MDP(QALYs-optimal)	Kurian et al. [11]
Outcome measure	**Cost-effectiveness**	**Cost**	**Quality of life**	**Survival**
Strategy	PM+PO-35	PM+PO-30	PO-30+PM-50	PM25+PO40
Health outcome				
-Death from other causes	0.15	0.16	0.13	0.15
-Ovarian cancer	0.07	0.07	0.08	0.09
-Breast cancer	0.11	0.05	0.2	0.04
-Healthy	0.67	0.72	0.59	0.72

of inflation. The computation time was approximately 20 min. We first solved the model using cost as our reward function. For BRCA1 mutation carriers, Table VI presents the cost-optimal intervention strategies for ages 30 to 65 (rows) and all possible prior intervention history (columns). For example, for a BRCA1 mutation carrier who has undergone PO at age 30, at age 40, PM-40 is recommended as the optimal intervention. Note that, we assumed that all surgical interventions occur at the beginning of the year. Following a surgery, either a screening (*Sc*) or no screening (*NSc*) action is chosen for the rest of the year. Therefore, for the aforementioned scenario, at the beginning of age 40, her prior intervention history changes to PO-30+PM-40. Referring to column PO-30+PM-40, no screening is recommended from 40 to 50, and then yearly screening is recommended for ages 51–65.

For a 30 year old BRCA1 mutation carrier with no prior intervention history, the optimal action is to undergo both PM and PO at age 30 (see row '30' and column 'none'). Now, referring to age 30 and column PM+PO-30, no screening is recommended until age 65.

The cost-optimal intervention strategies for BRCA2 mutation carriers can be found in Table VIII in the supplementary web material. The optimal strategy for a 30 year old with no prior intervention is to undergo PO at age 30. This changes the prior intervention history to column PO-30, which recommends no screening between ages 30–39. At age 40, PM is recommended, which modifies the prior intervention history to PO-30+PM-40. Per this column, no screening is recommended during ages 40–56, followed by yearly screening till 65. If intervention actions are limited to only one surgery at a time, our model recommends PO at age 30 and PM at age 40 for both mutation carriers.

A recent large international prospective study [3] suggests that BRCA1 mutation carriers should undergo prophylactic oophorectomy by age 35 because of the high risk of getting ovarian cancer. This survival-based study also recommends that BRCA2 mutation carriers can wait until age 40 safely because of the lower risk of ovarian cancer. Given this recommendation and the fact that most women are not willing to undergo PO during child bearing ages, we examined a scenario via our cost-based model where PO is not considered before age 40 for BRCA2 mutation carriers. Our model suggests PM at age 30 and PO at age 50. Even when a BRCA2 mutation carrier delays PM to age 40, our model still recommends PO at age 50.

We then solved the MDP model using the QALYs as our reward function. As opposed to the cost-optimal solution that chooses both surgeries at the same time, QALYs-optimal model postpones PM to age 50 for BRCA1 and does not recommend PO at age 30 for BRCA2 mutation carriers. The optimal intervention strategies for a BRCA1 and BRCA2 mutation carrier can be found in Table V and VII. For a 30 year old BRCA1 mutation carrier with no prior intervention, our model suggests PO at age 30 following by PM at age 50. The model recommends yearly screening starting from age 50. For a 30 year old BRCA2 mutation carrier with no prior intervention, the QALYs-model recommends PM-40 and yearly screening only after age 40.

V. ASSESSMENT AND COMPARISON OF OPTIMAL INTERVENTION STRATEGIES

We used an online tool in [28] to assess probabilities of health outcomes by age 70 for several optimal intervention strategies derived from the MDP model, and compared those with outcome probabilities from all the other possible (but nonoptimal) intervention strategies. The outcomes by age 70 are death from other causes, ovarian cancer, breast cancer, and healthy. We constructed a set of intervention strategies for BRCA1 mutation carriers of ages 30, which are shown in Table III. Note that the outcome assessment tool [28] does not allow consideration of partial screening strategies. Hence, for MDP recommended strategies that require partial screening, we have considered their fixed screening variants for comparison purposes.

TABLE V
QALYs-Optimal Intervention Strategies for BRCA1 Mutation Carriers of Ages 30 to 65

	Prior Intervention History																
Age	None	PO-30	PM-30	PM+PO-30	PO-40	PM-40	PM+PO-40	PM-30+PO-40	PM-40+PO-30	PO-50	PM-50	PM+PO-50	PM-30+PO-50	PM-40+PO-50	PO-30+PM-50
30	PO	Sc	Sc	Sc											
31	NSc	NSc	Sc	Sc											
32	NSc	NSc	Sc	Sc											
33	NSc	NSc	Sc	Sc											
34	NSc	NSc	Sc	Sc											
35	NSc	NSc	Sc	Sc											
36	NSc	NSc	Sc	Sc											
37	NSc	NSc	Sc	Sc											
38	NSc	NSc	Sc	Sc											
39	NSc	NSc	Sc	Sc											
40	PM+PO	NSc	PO	Sc	NSc	NSc	NSc	NSc	Sc						
41	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc						
42	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc						
43	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc						
44	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc						
45	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc						
46	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc						
47	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc						
48	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc						
49	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc						
50	PM	PM	PO	Sc	PM	NSc	NSc	NSc	Sc	NSc	NSc	NSc	NSc	NSc	Sc
51	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	NSc	NSc	NSc	NSc	NSc	Sc
52	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	NSc	NSc	NSc	NSc	NSc	Sc
53	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	NSc	NSc	NSc	NSc	NSc	Sc
54	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	NSc	NSc	NSc	NSc	NSc	Sc
55	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	NSc	NSc	NSc	NSc	NSc	Sc
56	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	NSc	NSc	NSc	NSc	NSc	Sc
57	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	NSc	NSc	NSc	NSc	NSc	Sc
58	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	NSc	NSc	NSc	NSc	NSc	Sc
59	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	NSc	Sc
60	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
61	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
62	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
63	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
64	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
65	NSc	NSc	Sc	Sc	NSc	NSc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc

TABLE VI
Cost-Optimal Intervention Strategies for BRCA1 Mutation Carriers of Ages 30 to 65

	Prior Intervention History																
Age	None	PO-30	PM-30	PM+PO-30	PO-40	PM-40	PM+PO-40	PM-30+PO-40	PM-40+PO-30	PO-50	PM-50	PM+PO-50	PM-30+PO-50	PM-40+PO-50	PO-30+PM-40	PO-30+PM-50
30	PM+PO	NSc	NSc	NSc												
31	Sc	NSc	NSc	NSc												
32	Sc	NSc	NSc	NSc												
33	Sc	NSc	NSc	NSc												
34	Sc	NSc	NSc	NSc												
35	Sc	Sc	NSc	NSc												
36	Sc	Sc	NSc	NSc												
37	Sc	Sc	NSc	NSc												
38	Sc	Sc	NSc	NSc												
39	Sc	Sc	NSc	NSc												
40	PM+PO	PM	NSc	NSc	Sc	Sc	Sc	Sc	NSc							
41	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc							
42	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc							
43	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc							
44	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc							
45	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc							
46	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	NSc							
47	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	NSc							
48	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc							
49	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc							
50	PM+PO	PM	PO	NSc	PM	PO	Sc	Sc	NSc	Sc	Sc	Sc	NSc	Sc	Sc	Sc
51	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
52	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
53	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
54	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
55	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
56	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
57	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
58	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
59	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
60	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
61	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
62	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
63	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
64	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc
65	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc	Sc

For all strategies in Table III, we obtained health outcome probabilities by age 70. These are plotted in Fig. 2, for BRCA1 mutation carriers. It may also be noted that the assessment tool [28] does not allow PO before age 35. Since our model considers PO starting at age 30, we assumed that the outcome probabilities of PO-30 are approximated by those of PO-35.

As shown in Fig. 2, the optimal intervention strategy #7 [marked by the vertical line (1)] has the highest probability of being healthy by age 70 (0.72) and the lowest probability of incidence of breast cancer (0.05). It also has a low probability of ovarian cancer (0.07), which is within 1% of the lowest obtained by other strategies, for example, #4 and #5. The optimal strategy #7 has 16% risk of death from other causes compared to, for example, 10% for strategy #2. This increase is very much expected as it is well known that PO increases probability of death from other causes. The QALYs-optimal strategy #16 has a probability of 0.59 for being healthy by age 70 and a probability of 0.08 for ovarian cancer incidence. This strategy also has a low probability (0.13) of death from other causes which compares well with the lowest value of 0.1 offered by strategy #2. The QALYs-optimal strategy has 0.2 probability of breast cancer incidence compared to 0.05 for the cost-optimal strategy (#7). Health outcome results for BRCA2 mutation carriers can be found in Table XI and Fig. 3 provided in supplementary web material section.

VI. Sensitivity Analysis

While examining the effect of normal variations in intervention cost on the cost-optimal strategies, no deviations were observed. Only when we increased the intervention cost by five folds for BRCA1 and ten folds for BRCA2, the optimal strategy for a 50 year old BRCA1/2 mutation carrier changed from PM+PO-50 to PO-50. When the cost increases approached 20 folds, the optimal strategies for both 40 and 50 year olds changed to screening only. Hence, the strategies derived from the MDP model are quite robust to cost estimate variations. When we decreased the utility weights by one standard deviation for BRCA1 mutation carriers, PO-30+PM-40 changed to PM+PO-30 with less frequent screening. For BRCA2 no changes in surgery options were observed with decreased utility. But when we increased the utility weights, lower screening frequencies were recommended. We conducted a sensitivity analysis for the discount factor in the range of 1%–5%. No significant variations in the results were observed.

VII. Discussion

We extended the Markov chain model presented in [13] to a Markov decision model (MDP) by first considering an expanded state space and then by superimposing a state dependent decision process on the Markov chain. The expanded state space for health and treatment status of mutation carriers and the corresponding intervention decision options were developed using information available in [11]. The solution of our MDP model yields cost and QALYs-optimal intervention strategies for healthy BRCA1/2 mutation carriers of age 30 to 65 and with all possible prior intervention status. In our model, the screening decision choice is considered each year as opposed to a one-time decision made at age 25 (as in [11]) or age 35 (as in [13]). In [13], Markov modeling with Monte Carlo simulation

TABLE VII
QALYS-OPTIMAL INTERVENTION STRATEGIES FOR BRCA2 MUTATION CARRIERS OF AGES 30 TO 65

	Prior intervention history																	
Age	None	PO-30	PM-30	PM+PO-30	PO-40	PM-40	PM+PO-40	PM-30+PO-40	PM-40+PO-30	PO-50	PM-50	PM+PO-50	PM-30+PO-50	PM-40+PO-50	PO-30+PM-50
30	NSc	Sc	NSc	NSc											
31	NSc	Sc	NSc	NSc											
32	NSc	Sc	NSc	NSc											
33	NSc	Sc	NSc	NSc											
34	NSc	Sc	NSc	NSc											
35	NSc	Sc	NSc	NSc											
36	NSc	Sc	NSc	NSc											
37	NSc	Sc	NSc	NSc											
38	NSc	Sc	NSc	NSc											
39	NSc	Sc	NSc	NSc											
40	PM	Sc	PO	NSc	Sc	Sc	Sc	Sc	NSc						
41	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc						
42	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc						
43	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc						
44	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc						
45	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc						
46	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc						
47	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc						
48	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc						
49	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc						
50	Sc	PM	PO	NSc	PM	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	NSc
51	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	NSc
52	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	NSc
53	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	NSc
54	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	NSc
55	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	NSc
56	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	NSc
57	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	NSc
58	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	NSc
59	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	NSc
60	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	NSc
61	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	NSc
62	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	NSc
63	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	NSc
64	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	NSc
65	Sc	Sc	NSc	NSc	Sc	Sc	Sc	Sc	NSc	Sc	Sc	Sc	Sc	Sc	NSc

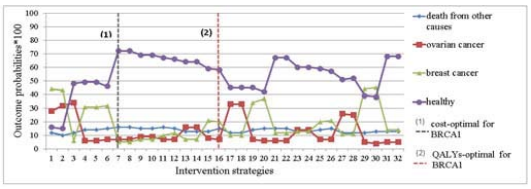

Fig. 2. Health outcome probabilities by age 70 for different intervention strategies (listed in Table III) for 30 year old BRCA1 mutation carriers.

was implemented to find the cost-effective strategies for 35 to 50 year old BRCA1/2 mutation carriers. In [11], a Monte carlo model simulated life histories of a 1980 birth cohort of 1 000 000 female with BRCA1/2 mutations from age 25 until age 100 or death.

In our MDP model, we first used expected present cost to find cost-optimal intervention strategies. True cost of intervention actions, treatment, and the intervention depended transition probabilities from healthy to other states play important roles in yielding the optimal intervention strategies. Since treatment costs are significantly higher than the prevention costs for breast and ovarian cancers, the model attempts to prevent the incidence of cancer while also trying to minimize the cost of intervention decisions. We then used utility weights (that represents the quality adjusted value of a year of life) to find the QALYs-optimal strategies. For this, the model attempts to maximize QALYs for a person with any prior intervention history and any age. The existing simulation based models [11] and [13] are capable of evaluating a given strategy. Whereas, our MDP model can be solved optimally to yield intervention strategies comprising optimal decisions for every state of a healthy mutation carrier. For example, our model has 8492 states and each state has between 1 and 8 possible state dependent decision choices. The value iteration algorithm is able to identify the optimal decision choice for each state. Also, as demonstrated through the use of both cost and QALYs metrics of performance, the MDP model can accommodate a variety of reward measures, for which optimal strategies could vary. Another feature of our model is that it is possible to have an optimal strategy with partial screening. This is an improvement over strategies with fixed screening only that were presented earlier to the literature.

We compared our MDP model with models in [11] and [13] using the health outcome probabilities obtained by the online tool in [28]. As shown in Table IV for a BRCA1 mutation carrier, the MDP cost-optimal intervention strategy, when compared to [11], yields the same probabilities of being healthy (0.72), lower ovarian cancer incidence (0.07), and a slightly higher probability of breast cancer incidence (0.05) and death from other causes (0.16). The MDP cost-optimal strategy has better health outcomes compared to [13] in almost all categories (see Table IV). The MDP QALYs-optimal intervention strategies yields the lowest probability of death from other causes and slightly higher probability of cancer outcomes compared to the other models. A similar comparison for BRCA2 mutation carriers can be found in supplementary web material in Table X.

According to the National Cancer Institute website [29], availability of data on the outcomes of interventions to reduce breast and ovarian risks in BRCA1/2 mutation carriers is limited. There exists uncertainties regarding cancer risk associated with BRCA1/2 mutations [29]. Therefore, management of interventions are done primarily based on expert opinions. We believe that development of model-based decision tools, as presented here, can help both policy makers as well as patients to make more informed intervention decisions. Our model suffers from the following limitations. Since the one-step transition probabilities were adopted from [11], the definition of state vector in our model for a mutation carrier had to be restricted. For example, we could not consider various stages of ovarian cancer and recurrence of breast cancer, as this would require more granular data, than what is currently available, to generate the necessary one-step transition probabilities. Our model considers surgical interventions (PM and PO) only at ages 30, 40, and 50. Once again, limited availability of one-step transition probability data restricted our consideration of surgical interventions in the intermediate ages.

ACKNOWLEDGMENT

The authors would like to express their sincere thanks to Dr. R. Sutphen, MD (Professor of Genetics at University of South Florida) for her guidance on the clinical practices and pointing some key references to us.

REFERENCES

[1] M. Van Roosmalen, L. Verhoef, P. Stalmeier, N. Hoogerbrugge, and W. Van Daal, "Decision analysis of prophylactic surgery or screening

for brca1 mutation carriers: a more prominent role for oophorectomy," *J. Clinical Oncology*, vol. 20, no. 8, pp. 2092–2100, 2002.
[2] V. Grann, J. Jacobson, D. Thomason, D. Hershman, D. Heitjan, and A. Neugut, "Effect of prevention strategies on survival and quality-adjusted survival of women with brca1/2 mutations: an updated decision analysis," *J. Clinical Oncology*, vol. 20, no. 10, pp. 2520–2529, 2002.
[3] T. R. Rebbeck, T. Friebel, H. T. Lynch, S. L. Neuhausen, L. Vant Veer, J. E. Garber, G. R. Evans, S. A. Narod, C. Isaacs, E. Matloff *et al.*, "Bilateral prophylactic mastectomy reduces breast cancer risk in brca1 and brca2 mutation carriers: The prose study group," *J. Clinical Oncology*, vol. 22, no. 6, pp. 1055–1062, 2004.
[4] National Cancer Institute. Genetics of Breast and Ovarian Cancer. (2014). [Online]. Available: http://www.cancer.gov/cancertopics/factsheet/Risk/BRCA
[5] A. Kurian, D. Munoz, P. Rust, E. Schackmann, M. Smith, L. Clarke, M. Mills, and S. Plevritis, "Online tool to guide decisions for brca1/2 mutation carriers," *J. Clinical Oncology*, vol. 30, no. 5, pp. 497–506, 2012.
[6] J. S. Davis and X. Wu, "Current state and future challenges of chemoprevention," *Discovery Med.*, vol. 13, no. 72, pp. 385–390, 2012.
[7] V. A. Moyer, "Risk assessment, genetic counseling, and genetic testing for brca-related cancer in women: US preventive services task force recommendation statement," *Ann. Internal Med.*, 2013.
[8] M. Fatouros, G. Baltoyiannis, and D. Roukos, "The predominant role of surgery in the prevention and new trends in the surgical treatment of women with brca1/2 mutations," *Ann. Surgical Oncology*, vol. 15, no. 1, pp. 21–33, 2008.
[9] D. Evans and A. Howell, "Are we ready for online tools in decision making for brca1/2 mutation carriers?," *J. Clinical Oncology*, vol. 30, no. 5, pp. 471–473, 2012.
[10] V. Grann, P. Patel, J. Jacobson, E. Warner, D. Heitjan, M. Ashby-Thompson, D. Hershman, and A. Neugut, "Comparative effectiveness of screening and prevention strategies among brca1/2-affected mutation carriers," *Breast Cancer Res. Treatment*, vol. 125, no. 3, p. 837, 2011.
[11] A. Kurian, B. Sigal, and S. Plevritis, "Survival analysis of cancer risk reduction strategies for brca1/2 mutation carriers," *J. Clinical Oncology*, vol. 28, no. 2, pp. 222–231, 2010.
[12] American Cancer Society. (2012). Cancer Facts and Figures 2012. [Online]. Available: http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/ document/acspc-031941.pdf
[13] K. Anderson, J. Jacobson, D. Heitjan, J. Zivin, D. Hershman, A. Neugut, and V. Grann, "Cost-effectiveness of preventive strategies for women with a brca1 or a brca2 mutation," *Ann. Internal Med.*, vol. 144, no. 6, p. 397, 2006.
[14] T. Das, A. Gosavi, S. Mahadevan, and N. Marchalleck, "Solving semi-Markov decision problems using average reward reinforcement learning," *Manag. Sci.*, vol. 45, no. 4, pp. 560–574, 1999.
[15] O. Alagoz, H. Hsu, A. Schaefer, and M. Roberts, "Markov decision processes: a tool for sequential decision making under uncertainty," *Med. Decision Making*, vol. 30, no. 4, pp. 474–483, 2010.
[16] A. Gosavi, *Simulation-Based Optimization: Parametric Optimization Techniques and Reinforcement Learning*, vol. 25. New York, NY, USA: Springer, 2003.
[17] D. Easton, D. Ford, and D. Bishop, "Breast and ovarian cancer incidence in brca1-mutation carriers. breast cancer linkage consortium.," *Am. J. Human Genetics*, vol. 56, no. 1, p. 265, 1995.
[18] http://demog.berkeley.edu/~bmd/States/ssa/life.tables/ufper.lt.proj.1x1, University of California, Berkeley: Berkeley Mortality Database, 2014.
[19] http://www.cdc.gov/nchs/data/dvs/MortFinal2006_WorkTable292R.pdf, Centers for Disease Control, National Vital Statistics System, 2008.
[20] R. Pataky, L. Armstrong, S. Chia, A. J. Coldman, C. Kim-Sing, B. McGillivray, J. Scott, C. M. Wilson, and S. Peacock, "Cost-effectiveness of mri for breast cancer screening in brca1/2 mutation carriers," *BMC Cancer*, vol. 13, no. 1, p. 339, 2013.
[21] S. Plevritis, A. Kurian, B. Sigal, B. Daniel, D. Ikeda, F. Stockdale, and A. Garber, "Cost-effectiveness of screening brca1/2 mutation carriers with breast magnetic resonance imaging," *J. Am. Med. Assoc.*, vol. 295, no. 20, pp. 2374–2384, 2006.
[22] http://www.cancer.org/cancer/news/news/study-quantifies-risk-of-breast-cancer-recurrence/, American Cancer Society.
[23] http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-survival-by-stage., American Cancer Society.
[24] M. T. Groot, R. Baltussen, C. A. Uyl-de Groot, B. O. Anderson, and G. N. Hortobágyi, "Costs and health effects of breast cancer interventions in epidemiologically different regions of Africa, North America, and Asia," *Breast J.*, vol. 12, no. s1, pp. S81–S90, 2006.
[25] http://www.ovariancancer.org/about-ovarian-cancer/statistics/., American Cancer Society.
[26] D. Naimark, M. D. Krahn, G. Naglie, D. A. Redelmeier, and A. S. Detsky, "Primer on medical decision analysis: Part 5 working with markov processes," *Med. Decision Making*, vol. 17, no. 2, pp. 152–159, 1997.
[27] MATLAB, version 7.10.0 (R2010a), Natick, MA, USA, The MathWorks Inc., 2010.
[28] http://brcatool.stanford.edu/brca.html/, Stanford University, Decision Tool for Women with BRCA Mutations.
[29] http://www.cancer.gov/cancertopics/pdq/genetics/breast-and-ovarian/HealthProfessional/page2#Section_113, National Cancer Institute, Genetics of Breast and Ovarian Cancer.

Mehrnaz Abdollahian (S'03) is a Ph.D. student in the Department of Industrial and Management Systems Engineering at the University of South Florida, Tampa., USA.

Her research interests include Decision Making, Analytics, and Statistics. She is a student member of INFORMS.

Tapas K Das (M'12) is a Professor and Chair of Industrial and Management Systems Engineering at the University of South Florida, Tampa. His research interest include modeling for decision making in the fields of healthcare, energy, and public health.

He is a Fellow of IIE, and member of INFORMS.