Evaluating the Impact of Uncertainty on Risk Prediction: Towards More Robust Prediction Models

Panayiotis Petousis MEng[1], Arash Naeim MD PhD[2], Ali Mosleh PhD[3], William Hsu PhD[1]
[1] Medical Imaging & Informatics, Department of Radiological Sciences and Bioengineering
[2] Department of Medicine, David Geffen School of Medicine
[3] B. John Garrick Institute for the Risk Sciences, Samueli School of Engineering
University of California, Los Angeles, CA

Abstract

Risk prediction models are crucial for assessing the pretest probability of cancer and are applied to stratify patient management strategies. These models are frequently based on multivariate regression analysis, requiring that all risk factors be specified, and do not convey the confidence in their predictions. We present a framework for uncertainty analysis that accounts for variability in input values. Uncertain or missing values are replaced with a range of plausible values. These ranges are used to compute individualized risk confidence intervals. We demonstrate our approach using the Gail model to evaluate the impact of uncertainty on management decisions. Up to 13% of cases (uncertain) had a risk interval that falls within the decision threshold (e.g., 1.67% 5-year absolute risk). A small number of cases changed from low- to high-risk when missing values were present. Our analysis underscores the need for better communication of input assumptions that influence the resulting predictions.

Introduction

An increasing number and variety of patient data being routinely captured have led to new insights into factors that influence the risk of disease such as cancer. Decision tools that aid physicians and patients with assessing these risks in the context of their personal circumstances are one important factor in selecting the appropriate management strategy. A growing number of mathematical models has been developed and validated as tools upon which clinicians can determine whether a patient is considered to have a "high-risk" for cancer and would be suitable candidates for interventions. In breast cancer, models are used to estimate an absolute risk of cancer in women, which influence decisions related to prescribing a risk-reducing pharmacologic intervention (e.g., selective estrogen receptor modulators) or more aggressive screening strategies (e.g., surveillance using breast magnetic resonance imaging). Tamoxifen is one example of a medication investigated for its effectiveness in the prevention of invasive breast cancer for high-risk women. The Breast Cancer Prevention Trial showed that women with a 5-year absolute risk of 1.67% and greater can reduce their risk of invasive cancer by 49% when undergoing chemoprevention compared to taking a placebo [1]. However, use of tamoxifen is not completely without risks and is associated with adverse events such as uterine cancer and blood clotting in the legs or lungs [2]. The purpose of these risk models is to provide physicians and patients with a reasoning tool to weigh the trade-offs between the effects of the intervention with the absolute risks of various health outcomes [3].

While risk prediction models aid in considering potential benefits and costs, these models also have notable limitations. First, models such as the Gail model [4] provide an average risk for a group of women with similar risk factors, not an individual probability of cancer. As such, the interpretation of the predicted risk is unclear for a given individual. Second, uncertainty is an inherent part of risk assessment, given that not all factors related to cancer risk are known or can be measured to the desired precision. Studies have also shown that patient-reported information such as social history and patient outcomes are unreliable [5]–[7]. For clinicians who utilize cancer risk models to make decisions about potential interventions, an understanding about the sensitivity and reliability of self-reported risk factors such as the age of first live birth and family history should be known in the situation that such information is unreliable or missing. For example, heredity information is often complex to elicit from a patient, particularly if she is not completely aware of her siblings' and ancestors' health statuses. Additionally, any risk factor reporting age is often rounded up to the nearest year rather than the true age in months or days. Finally, information that is required to execute the risk model may not be available for a variety of reasons. Missing data are unavoidable in the fast pace, real-world clinical environment. Many models such as the Gail model are a form of a logistic regression model that requires all risk factors to be inputted in order to compute the coefficients for the model or generate an estimated risk. If the patient cannot be subsequently reached to obtain the missing information, data-driven methods such as imputation must be performed to utilize these models. However, the effect of imputation on the validity of risk models has not been thoroughly explored in the medical literature [8], [9]. For instance, datasets often suffer from population

bias such as when the majority of patients are white. In the case of missing data, imputing instances of minority values from unbalanced datasets introduces bias as well as uncertainty in imputed values.

In this work, we present a systematic approach to assess the effect of uncertainty and missing values on risk predictions. Comprehensive assessment of uncertainties in estimated risk metrics requires consideration of uncertainties about the input or parameters of the risk model (parameter uncertainty) as well as uncertainties associated with the form and assumptions of the model (model uncertainty) [10]. The scope of the present work is the treatment of parameter uncertainty. We utilize breast cancer screening as a driving example. Leveraging a large retrospective dataset of women undergoing routine screening, our approach discovers subgroups of similar women from which meaningful value ranges for a given risk factor can be determined. A clustering technique with multiple imputation is used to identify similar patients. Bootstrapping is then used to sample values of similar cases. These values are inputted into the Gail risk model to generate a confidence interval (C.I.) around the absolute risk prediction. We subsequently evaluate the sensitivity of the model to varying inputs. By expressing cancer risk using a C.I., we formalize how uncertainty is expressed, providing additional context to aid physicians in interpreting risk predictions and making management decisions. We analyze the frequency by which uncertainty associated with risk estimates would have potentially changed whether the patient would have been categorized as "high-risk" (e.g., cross the 1.67% risk threshold).

Background

Predicting Absolute Risk of Breast Cancer: The Gail Model

Cancer screening is a large population-based intervention that is at the center of great debate, especially in older patients or for certain cancers such as breast and prostate [4], [11]. Breast cancer screening is particularly contentious. A number of models are in use today to stratify patients into different risk groups [12]. The Gail model is among the earliest and most widely used to estimate absolute risk. The model incorporates age, age at menarche, age at first birth, the number of first-degree relatives with breast cancer, the number of previous breast biopsies, and race in its assessment. The Gail model has been validated in specific cohorts of white American women with specific risk factors but has since been adjusted for individuals of different race and ethnicity. The model calculates the absolute risk of breast cancer by breaking the risk estimation into 3 sub-problems: 1) the estimation of the relative risk using a logistic regression; 2) the estimation of the baseline age-specific breast cancer hazard rate; and 3) the estimation of a long-term probability of developing breast cancer from competing risks, relative risk and the baseline hazard [3]. The Gail model was used to compute risk values for our test population. The probability that a woman at age a with a relative risk $r(t)$ will develop cancer by age $a + t$ can be computed following Equation 1,

$$P(a,\tau,r) = \sum_j \frac{h_{1j}r_j}{h_{1j}r_j + h_{2j}} \frac{S_1(\tau_j - 1)}{S_1(a)} \frac{S_2(\tau_j - 1)}{S_2(a)} \left(1 - exp^{-\Delta_j(h_{1j}r_j + h_{2j})}\right) \quad (1)$$

where j is a defined age interval, h_2 is the risk of death due to other causes (competing hazards), S_2 is the probability of surviving the competing hazards, S_1 is the probability of surviving the death due to breast cancer, t_j is the time at the j-th age interval, a is the baseline age, and t is the time in years between baseline age and predicted age (typically set to 5 years). More information on the implementation of the Gail model can be found in [13]–[16].

Handling Uncertainty in the Data

Simulation-based methods such as Markov Chain Monte Carlo and bootstrapping have certain advantages compared to point estimate imputation methods when dealing with missing or uncertain cases. Even though computationally they are less efficient, they provide a confidence measure in their estimation making them more useful than a point estimate. They simulate possible uncertain values to generate a C.I. that represents the degree of uncertainty. Similarly, this approach can be applied when imputing missing values. Multiple imputation involves the simulation of a user-defined number of complete subsets m which are used to impute missing values. For each missing value, m possible imputed values are generated, reflecting the uncertainty about the true value of the variable. These m imputed values can be used to compute C.I.s [17], [18]. Another class of methods, model-based imputation, refers to estimating the joint distribution among risk factors from which imputed values are generated. To learn such a model, a training set is required to define the joint distribution. Imputation is then performed on a test set with missing values. Finally, clustering-based imputation approaches, identify similar cases from which an imputed value for the missing variable

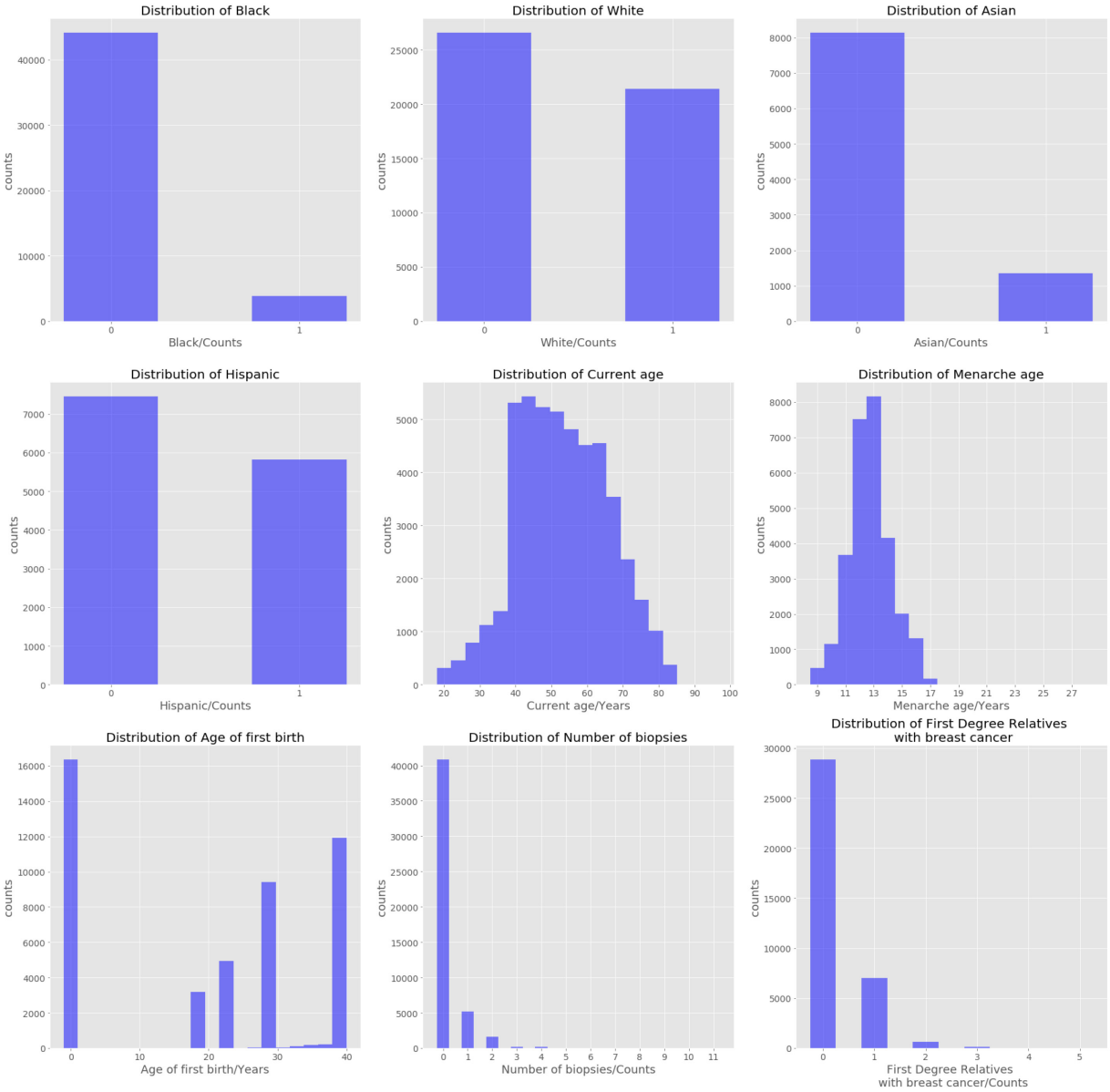

Figure 1: Distributions of each variable for the entire cohort. Race is represented as binary indicator variables.

value is assigned [19]. These approaches are typically implemented using a combination of k-means and k-nearest neighbors (kNN) algorithm [20]. The kNN algorithm is frequently used to cluster cases using variables that do not have missing data from which a set of values from similar cases can be obtained to inform the imputation process.

A significant limitation of these existing methods is the need to utilize complete information for training, which limits the number of cases that can be used. To account for uncertain and imputed values in our breast cancer dataset with the Gail risk model, we propose a multiple clustering imputation methodology that solves the limitations of traditional model-based imputation methods while providing a more informed breast cancer risk representation with C.I.s. Our proposed methodology imputes missing values from cases with complete information using multiple clusters of similar cases. Unlike methods that require complete data, our approach maximizes the use of available data, even ones with missing values. We use bootstrapping to calculate m unique clusters for each case with missing data. Using these m clusters, we generate a range of possible values for missing values, which is used to provide a C.I. of the imputed value. Subsequently, this range of imputed values can be used with risk models to generate a C.I. of risk values.

Variable Name	Variable Type	# Missing (%)
Current age	Continuous	0 (0)
Age of menarche	Continuous	19,572 (21.8)
Age at first child birth	Continuous	1,630 (1.8)
# of 1st degree relatives with breast cancer	Ordinal	11,459 (12.7)
Number of biopsies	Ordinal	1 (<0.1)
White	Indicator variable	0 (0)
Black	Indicator variable	0 (0)
Asian	Indicator variable	0 (0)
Hispanic	Indicator variable	7,127 (7.9)

Table 1: Description of the variables considered by the Gail breast risk model including variable type and percentage of values that are inherently missing in the dataset. The implemented risk model did not adjust for Native Americans; those individuals were excluded from our analysis.

Methods

Dataset

Data on women who underwent breast screening at UCLA was obtained through an institutional review board (IRB)-approved protocol. The dataset consists of 47,980 cases collected during a five-year period. At the time of their breast screening exam, women were asked to complete a questionnaire that collected basic demographics and risk factors related to the Gail model. The purpose of the survey was to obtain all the information necessary to provide a risk estimate using the model and to provide radiologists contextual information about the patient's history. Surveys were typically completed by the patient, largely without assistance from a physician. **Figure 1** depicts the distribution of each variable in the entire cohort. **Table 1** summarizes these variables as well as the number of inherent missing values. The 5-year absolute risk for each case was calculated using the Gail model. The implementation of the Gail risk model that we used as part of this analysis did not adjust for Native Americans. Individuals who self-reported as part of this race category were excluded from our analysis. Individuals with multiple races were not excluded from our analysis.

Overall Approach

Our approach to investigating the influence of uncertainty is illustrated in **Figure 2**. We posit that using C.I.s defined by similar cases for certain input variables can change the interpretation of the absolute risk that is generated by the Gail model. The dataset was randomly split into training (60%) and testing (40%) sets, consisting of 28,788 and 19,192 cases, respectively. Categorical variables such as race were transformed into binary indicator variables, resulting in four variables representing each race and ethnicity categories. Within the testing set, we only considered cases that had complete information, resulting in a total of 1,850 cases. We focused our analysis on this subset of the test set. Missing values were simulated for each case using an unbiased random number generator. The number generation process consists of two random number generators, each producing a value from 0 to 8 (matching the number of input variables). Each random number generator was used to populate a list of 8 elements, corresponding to the number of variables. For elements in each list with the same random integer, the value for the corresponding variable was set as missing.

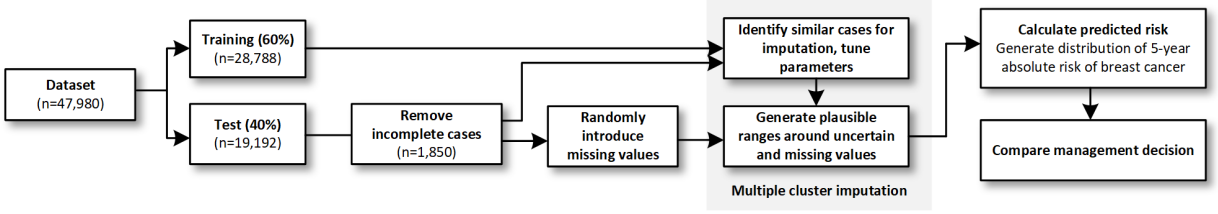

Figure 2: Overall approach. Process by which data collected on women undergoing breast screening were split into a training and test set. The training set was used to perform multiple cluster imputation to generate ranges for uncertain or missing values introduced in the test set. These ranges were used to calculate a range absolute risk scores and interpreted using the current approach of identifying women with a 5-year absolute risk of 1.67% or above as candidates for chemoprevention.

	B	W	A	H	Age	AM	AFLB	Biopsies	FDR_BC
Original data point	0	1	0	0	48	14	39	0	1
Missing Values	0	1	0	0	48	NaN	39	NaN	1
Imputed data point	0	1	0	0	48	(11.7-14.3)	39	0	1

Table 2: Data imputation. An example of an imputed case with the range of possible input values considered for the continuous variables and a point estimate for categorical variables. B: Black, W: White, A: Asian, H: Hispanic, AM: Age at menarche, AFLB: Age at first live birth, Biopsies: Number of biopsies, and FDR_BC: 1st degree relatives with breast cancer. NaN: corresponds to a missing value.

The training set served as a knowledge base of retrospective cases that informs how missing values of the test cases could be imputed. Continuous variables such as current age and age at menarche were varied by ±1 years. The variability introduced into age variables was constrained for two reasons: 1) we hypothesized that the likelihood of a patient getting her age incorrect was small and that the error was more likely due to rounding to the nearest year; and 2) imputation of age would be extremely difficult from the other variables collected. Subsequently, the multiple clustering imputation (MCI) method was used to identify plausible values for categorical variables and generate a range of possible imputed values for the continuous variables. Given the range of imputed values generated using the MCI method, the risk estimate for each case was calculated using the Gail model, yielding a distribution of 5-year absolute cancer risk predictions as well as a 95% C.I. around the median.

Multiple Cluster Imputation (MCI)

Figure 3 illustrates the basic process for generating values using the MCI approach. The algorithm proceeded as follows: for each test case being considered, variables with missing values are identified (**Figure 3-A**). From within the training set, we identify cases that had observed values for the variables that are not missing in the test case. A strength of our approach is maximizing the number of prior cases that are used in this process because our algorithm is able to make use of training cases that have missing information. As such, we will make use of cases that have varying levels of completeness for the variables outside of the ones being imputed. An iterative selection process is performed to generate clusters of cases based on how complete the cases are (**Figure 3-B**). We define a parameter called *tolerance value* to constrain the level of missing information that may exist in a cluster and used for imputation. For example, if the tolerance value is set to 80%, the algorithm would only select cases that have at least 80% of their variables with an observed value outside of the variables that are being imputed. In our case with nine variables, a

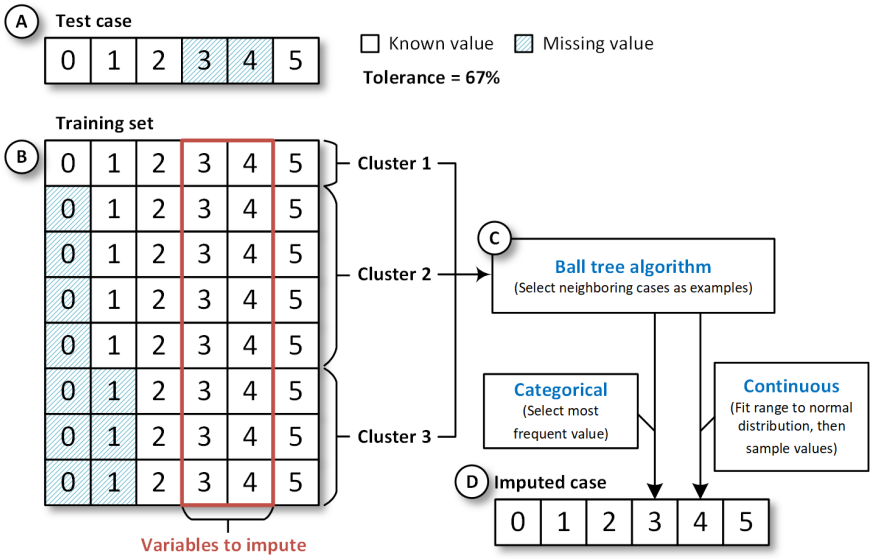

Figure 3: Multiple imputation clustering. Given a test case with two missing values **(A)**, we examine the training set for cases that have values for variables 3 & 4 **(B)**. Within that subset, cases are grouped into clusters based on the percentage of observed variables; all clusters will have observed variables above a predefined tolerance value. A ball tree algorithm **(C)** is used to select the training cases that are most similar to the test case; the range of values defines the permissible values from which the final imputed values are selected **(D)**.

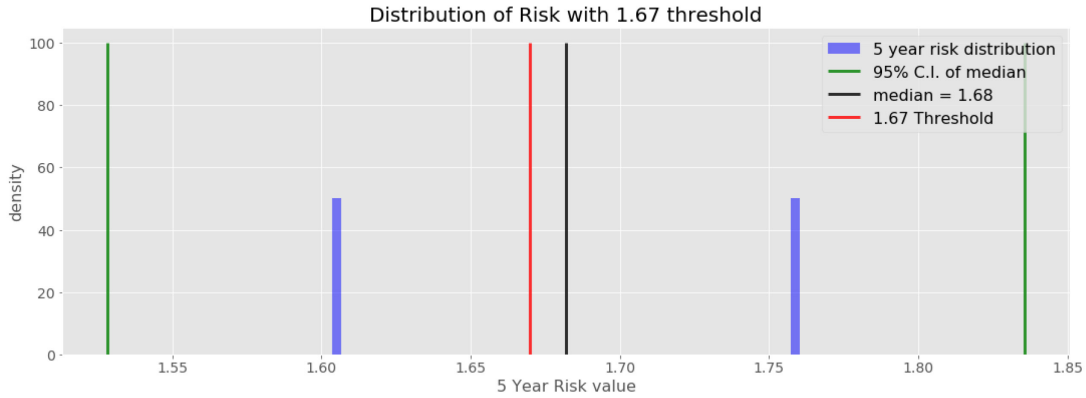

Figure 4: Example distribution of a single case. Distribution of risk predictions for a single patient based on 100 different simulated input. The original risk estimate for the individual is 1.60. The median value is 1.68, with 1.53-1.84 confidence interval.

tolerance of 80% permits only one additional variable to be missing. The entire training dataset is then examined for all possible combinations of variables where only one additional variable is missing. This process is repeated with increasing tolerance values until the value reaches 100%. This bootstrapping process generates multiple clusters of varying levels of completeness, with replacement, from which similar cases can be selected.

A ball tree algorithm [21] is used to select the most similar cases from each cluster compared to the test case (**Figure 3-C**). The ball tree is a binary tree where every node consists of a hypersphere that contains a subset of cases to be searched. The ball tree algorithm used is obtained from scikit-learn [22], and our analysis is implemented in Python. The radius of the hypersphere is user-defined and specified as an input parameter to the algorithm. All cases inside a hypersphere are considered as similar cases. As a space partitioning algorithm, the ball tree efficiently projects points/cases in a multi-dimensional space. The ball tree data structure is a hierarchical binary tree in which each node in the binary tree is split into two clusters with data points added in each cluster based on distance from the centroid of each cluster. In this work, the radius of the hypersphere is adaptive and proportional to the tolerance value: smaller tolerance values are associated with smaller hyperspheres. Once a set of cases have been identified from each cluster, then all the cases are combined into one group. These similar cases are used to impute missing values (**Figure 3-D**). Two types of variables are considered: continuous and ordinal/categorical. Continuous variables are imputed based on the range of values from similar cases in the training set. The minimum and maximum values in these cases define the range of permissible values; we then fit a normal distribution, taking the 50% C.I. of this distribution from the median. Conversely, ordinal and categorical variables are imputed based on the most frequent value for a given variable. **Table 2** provides an example of how imputation was performed on a case with two missing values. Imputed values and the range of imputed values are estimated for categorical and continuous variables missing values, respectively. **Figure 4** depicts an example of a breast cancer risk with continuous variables variability and a risk confidence interval.

	Complete	**No missing values with variability**	**Missing values with no variability**	**Missing values with variability**
High-risk (HR)	579	462	531	427
Low-risk (LR)	1271	1167	1287	1205
Uncertain (U)	-	221	32	218
HR → HR	-	462	531	427
LR → LR	-	1167	1253	1174
HR → LR	-	0	34	31
LR → HR	-	0	0	0
HR → U	-	117	14	121
LR → U	-	104	18	97

Table 3: Summary of interpretation changes by introducing the C.I. associated with the risk prediction. The uncertain category highlights the cases where the original risk estimate was either above or below 1.67%, but when a range of possible input values is considered, the decision threshold falls within the C.I. of the risk estimate.

Evaluation

Using the generated predictions of risk, we performed two types of analyses. In the first type of analysis, we explored the impact of intentionally varying continuous variables such as current age and age at menarche when complete information was available (i.e., no imputed values) to determine the effect of rounding on predicted risk. Within the testing set, variability in known values of current age and age at menarche (collectively referred to as the continuous age variables) was introduced by calculating the risk based on 0.1 increments between -1 and 1 years from the inputted value, resulting in 20 risk estimates. In the second type of analysis, we examined the effect of imputing missing values using the MCI approach and the effect of intentionally varying continuous age variables. We fitted the continuous variables of all the similar cases on a normal distribution around the median and defined the range as the 50% C.I. around the median. This range was also split into 10 linear steps. Overall, we present four analyses: the original complete dataset without introducing variability on the continuous age variables ("complete"), the original complete dataset with variability introduced on the continuous age variables ("No missing values with variability"), the imputed dataset without variability introduced on the continuous age variables ("Missing values with no variability"), and the imputed dataset with variability introduced on the continuous age variables, if ranges were not already imputed due to the value being missing ("Missing values with variability"). We evaluated how often the risk model resulted in a predicted absolute cancer risk that would change the management of a patient (e.g., the risk range predicted for each test case crossed the 1.67% threshold). We also evaluated which combinations of feature values would change the categorization of a given patient (e.g. if the patient moves from low-risk to high-risk).

Results

Implication of Risk Predictions under Uncertainty

The MCI method was used to impute the test set of 1,850 cases. **Table 3** summarizes changes in management interpretation when C.I.s surrounding a risk prediction is provided. In the "complete" column, there are 579 and 1,271 high and low-risk predictions, respectively. When variability was introduced in the continuous age variables, as summarized in the "no missing values with variability" column, the risk category of 221 cases changed from high- or low-risk to "uncertain" given that the CI overlaps with the decision threshold of 1.67%. Out of those 221 cases, 117 were originally high-risk, and 104 were low-risk individuals. No cases changed status form high to low-risk or vice versa. When we randomly introduced missing values into the test data, represented by the "missing values with no variability" column, the total number of uncertain individuals were 32 of which 14 were originally high-risk, and 18 were low-risk. 34 cases changed category from high-risk to low-risk. No cases changed from low-risk to high-risk. In the "missing values with variability" column, a higher number of high-risk cases changed to uncertain cases compared with other columns. Of the 218 uncertain cases, 121 were previously high-risk, and 97 were low-risk. Additionally, 31 cases that were originally high-risk changed to low-risk. No low-risk cases changed to high-risk.

Analysis of Risk Predictions under Uncertainty

In the "no missing values with variability" analysis, when age and age at menarche were used independently (varied one at a time), the number of uncertain cases was 96 and 124 for current age and age at menarche, respectively. Age at menarche had a stronger impact on predicted risk than current age. **Table 4** summarizes the average value for each variable, stratified by risk group (HR, LR, U). In addition, several trends that reinforce prior findings were noted: 1) older women were associated with a higher risk of breast cancer; 2) women who started menarche at an older age were associated with a lower cancer risk; 3) the number of biopsies was proportionate with risk; and 4) women with more 1st-degree relatives with a history of breast cancer had higher risk themselves. Uncertain cases had average values for variables in-between average values found in high and low-risk groups. Moreover, when missing values were introduced in the analysis, the number of uncertain cases increased when variability was introduced in the continuous age variables. We estimated the percentage of missing values per variable in the high to low-risk, low to high-risk, and high or low to uncertain risk groups. The variables with the highest percentage of missing values in the group that changed from high- to low-risk were and the number of 1st-degree relatives with breast cancer, the age at first live birth, and the number of prior biopsies, in descending order. The main variables with the highest percentage of missing values in the groups that changed from high- or low-risk to uncertain were the age at menarche and the number of prior biopsies.

		B	W	A	H	Age	AM	AFLB	Biopsies	FDR_BC
Complete	HR	0.07	0.78	0.09	0.08	60.35	12.63	27.36	0.50	0.56
	LR	0.07	0.65	0.17	0.17	46.07	12.91	21.97	0.07	0.09
No missing values with variability	HR	0.06	0.78	0.10	0.09	60.89	12.61	27.51	0.56	0.65
	LR	0.08	0.65	0.17	0.17	45.24	12.91	21.57	0.06	0.08
	U	0.07	0.75	0.11	0.10	56.85	12.80	26.61	0.22	0.24
	HR → HR	0.05	0.84	0.06	0.06	60.93	12.65	27.01	0.54	0.64
	LR → LR	0.08	0.64	0.17	0.18	44.68	12.88	21.61	0.05	0.08
	HR → LR	0.00	0.73	0.22	0.05	62.59	13.76	20.51	0.24	0.11
	LR → HR	0.11	0.06	0.53	0.39	57.28	12.31	32.44	0.19	0.42
	HR → U	0.04	0.85	0.09	0.06	59.04	12.79	24.60	0.22	0.22
	LR → U	0.14	0.58	0.16	0.19	53.07	12.81	30.07	0.21	0.26
Missing values with no variability	HR	0.06	0.82	0.10	0.10	60.06	12.79	25.43	0.55	0.58
	LR	0.06	0.67	0.16	0.18	45.46	12.96	19.91	0.07	0.07
	U	0.03	0.88	0.06	0.06	57.91	13.00	24.69	0.09	0.25
	HR → HR	0.05	0.87	0.05	0.07	60.38	12.82	25.22	0.55	0.57
	LR → LR	0.06	0.66	0.16	0.18	44.60	12.96	19.92	0.06	0.07
	HR → LR	0.02	0.77	0.19	0.05	58.18	13.05	19.74	0.31	0.07
	LR → HR	0.13	0.31	0.57	0.43	55.25	12.60	27.78	0.19	0.50
	HR → U	0.00	0.95	0.10	0.00	58.76	13.00	27.38	0.05	0.29
	LR → U	0.09	0.73	0.00	0.18	56.27	13.00	19.55	0.18	0.18
Missing values with variability	HR	0.05	0.79	0.09	0.09	61.36	12.64	24.88	0.54	0.60
	LR	0.07	0.68	0.15	0.16	45.49	12.94	19.20	0.05	0.07
	U	0.08	0.75	0.10	0.10	58.05	12.75	24.37	0.22	0.19
	HR → HR	0.05	0.85	0.05	0.07	61.42	12.68	24.34	0.51	0.60
	LR → LR	0.07	0.67	0.15	0.17	44.69	12.92	19.26	0.04	0.08
	HR → LR	0.03	0.76	0.18	0.08	58.71	13.22	18.17	0.18	0.06
	LR → HR	0.10	0.13	0.55	0.29	57.84	12.36	31.02	0.19	0.36
	HR → U	0.03	0.86	0.07	0.06	60.03	12.80	22.15	0.21	0.19
	LR → U	0.17	0.54	0.17	0.17	54.36	12.67	28.49	0.23	0.18

Table 4: Summary of high-risk (HR), low-risk (LR), and uncertain (U) cases average feature values. Left: Binary variables' mean frequency in each risk group. Right: Continuous/ordinal variables' mean value in each risk group. B: Black, W: White, A: Asian, H: Hispanic, AM: Age at menarche, AFLB: Age at first live birth, Biopsies: Number of biopsies, and FDR_BC: 1st degree relatives with breast cancer.

Availability

We have made our analysis available in the form of Jupyter notebooks[1].

Discussion

In this study, we examine the effect of uncertainty on the input values of the Gail model when estimating risk. In addition, we evaluate an approach for imputing a range of missing values for a patient to generate an individualized breast cancer risk C.I., using previously observed cases. While many of the variables collected as part of the Gail model are straightforward to provide, risk models are becoming increasingly complex, and the impact of uncertainty or invalid data should be explored. For example, breast cancer risk models such as Tyrer-Cuzick [23] and BRCAPRO

[1] https://github.com/panas89/multipleClusteringImputation

[24] ask for a detailed family history of cancer from first-, second-, and even third-degree relatives, which may be difficult to report precisely. A better understanding of how uncertain or unreliable inputs into these risk models is needed to better inform subsequent management decisions on whether a patient is considered "high-risk" or not. From our analysis, we conclude that uncertainty in input and missing values can potentially change the risk category of an individual when using the Gail model. Interestingly, throughout the four analyses shown in **Table 3**, low-risk cases never changed to high-risk, implying that the Gail model is more robust to low-risk uncertainty than it is to high-risk (high-risk cases being downgraded to low-risk). In **Table 4**, we demonstrated that a significant number of uncertain cases was classified primarily due to uncertainty in variables such as current age and age at menarche. Additionally, the majority of cases classified as uncertain were primarily missing values such as the age of first live birth and number of biopsies. Cases that were classified as "low-risk" but were actually "high-risk" upon further analysis had missing values for age at menarche, the number of biopsies, and breast cancer history among 1st-degree relatives.

Our study has several limitations. The imputation approach had difficulty providing reasonable estimates for current age; therefore, we chose not to introduce missing values to that variable. We believe the information provided by the other variables was not sufficient to provide meaningful estimates of the age variables. We also assumed that the distribution of continuous variables was normal. For example, the variable age at first live birth had zero values for women without a first birth. A normal distribution was not suitable for this variable; hence it was instead modeled as an ordinal/categorical variable. Future work may consider additional clinical risk factors that could serve as surrogate measures. In addition, while we employed and examined the effect of varying model parameters such as the range of the continuous variables' values and tolerance values, a full search was not performed, hence the performance of the algorithm may not be optimal. We introduced missing values at random into the dataset, but values were frequently missing not at random in real-world scenarios. Bias could be introduced into the missing value generation by adding weights to specific variables that are more frequently missing in practice. We also weighted variables equally when using the MCI method; future work can examine how these weights can be customized for individual variables.

While variables such as current age should be readily accessible, this analysis underscores the need to ensure that all of these variables are accurately recorded, given their impact on the final risk estimate. Cases that were unchanged generally had low percentages of missing values for all variables and any variability introduced on the continuous age variables had no effect on risk as their values were either very high or low. Our work highlights the utility of conducting sensitivity analyses as part of validating risk prediction models. Furthermore, we believe that reporting of C.I.s may be more informative than simply interpreting a point estimate of risk. Several studies have shown the utility and potential challenges of representing uncertainty associated with risk predictions to decision makers, including clinicians and patients [25]–[27]. By conveying the risk as a distribution, clinicians can understand the uncertainty associated with a risk estimate and better determine whether the patient's situation is clearly "high-risk" and should be given risk-reducing interventions or "uncertain" and should undergo further testing. Narrower C.I.s imply less variability (more confidence) in risk estimate and vice-versa.

Acknowledgments

This work is supported by the National Science Foundation (#1722516), an Impact Award from the UCLA Jonsson Comprehensive Cancer Center, and the UCLA Department of Radiology under the Data-Driven Diagnostic Decision Support (D4S) initiative. Data utilized in this study were collected as part of the Athena Breast Health Network.

References

[1] B. Fisher, J. P. Costantino, D. L. Wickerham, C. K. Redmond, M. Kavanah, W. M. Cronin, V. Vogel, A. Robidoux, N. Dimitrov, J. Atkins, M. Daly, S. Wieand, E. Tan-Chiu, L. Ford, and N. Wolmark, "Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study.," *J. Natl. Cancer Inst.*, vol. 90, no. 18, pp. 1371–88, Sep. 1998.

[2] "Patient education: Medications for the prevention of breast cancer (Beyond the Basics) - UpToDate." [Online]. Available: https://www.uptodate.com/contents/medications-for-the-prevention-of-breast-cancer-beyond-the-basics. [Accessed: 04-Mar-2018].

[3] M. H. Gail, "Personalized estimates of breast cancer risk in clinical practice and public health," *Stat. Med.*, vol. 30, no. 10, pp. 1090–1104, 2011.

[4] L. C. Walter and K. E. Covinsky, "Cancer Screening in Elderly Patients," *JAMA*, vol. 285, no. 21, p. 2750, Jun. 2001.

[5] E. Peters, J. Hibbard, P. Slovic, and N. Dieckmann, "Numeracy skill and the communication, comprehension, and use of risk-benefit information," *Health Affairs*, vol. 26, no. 3. pp. 741–748, 01-May-2007.

[6] R. Al-Abri and A. Al-Balushi, "Patient satisfaction survey as a tool towards quality improvement," *Oman*

Medical Journal, vol. 29, no. 1. Oman Medical Specialty Board, pp. 3–7, Jan-2014.
[7] N. Timmins, "NHS goes to the PROMS," *BMJ*, vol. 336, no. 7659, pp. 1464–1465, Jun. 2008.
[8] J. A. C. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter, "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls.," *BMJ*, vol. 338, p. b2393, Jun. 2009.
[9] J. H. Flory, J. Roy, J. J. Gagne, K. Haynes, L. Herrinton, C. Lu, E. Patorno, A. Shoaibi, and M. A. Raebel, "Missing laboratory results data in electronic health databases: implications for monitoring diabetes risk," *J. Comp. Eff. Res.*, vol. 6, no. 1, pp. 25–32, Jan. 2017.
[10] C. Mosleh, A.; Smidts, C. ; Siu, N.; Lui, "Model uncertainty: Its characterization and quantification," in Proceedings of workshop I in advanced topics in risk and reliability analysis. Model uncertainty: Its characterization and quantification.
[11] L. Esserman, Y. Shieh, and I. Thompson, "Rethinking Screening for Breast Cancer and Prostate Cancer," *JAMA*, vol. 302, no. 15, p. 1685, Oct. 2009.
[12] A. N. Freedman, D. Seminara, M. H. Gail, P. Hartge, G. A. Colditz, R. Ballard-Barbash, and R. M. Pfeiffer, "Cancer risk prediction models: A workshop on development, evaluation, and application," *J. Natl. Cancer Inst.*, vol. 97, no. 10, pp. 715–723, 2005.
[13] M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, and J. J. Mulvihill, "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually," *J. Natl. Cancer Inst.*, vol. 81, no. 24, pp. 1879–1886, Dec. 1989.
[14] Y. Li, L. Chen, X. Wan, and A. Chiang, "Implementation of Breast Cancer Risk Assessment Tool using SAS ®," *PharmaSUG 2013*, 2013.
[15] J. P. Costantino, M. H. Gail, D. Pee, S. Anderson, C. K. Redmond, J. Benichou, and H. S. Wieand, "Validation Studies for Models Projecting the Risk of Invasive and Total Breast Cancer Incidence," *JNCI J. Natl. Cancer Inst.*, vol. 91, no. 18, pp. 1541–1548, 1999.
[16] W. H. Nova F. Smedley, N Y Elizabeth Chau, Antonia Petruse, Alex A. T. Bui, Arash Naeim, "A Platform for Generating and Validating Breast Risk Models from Clinical Data: Towards Patient-Centered Risk Stratified Screening - Semantic Scholar," in *AMIA*, 2015.
[17] P. Loukopoulos, S. Sampath, P. Pilidis, G. Zolkiewski, I. Bennett, F. Duan, and D. Mba, "Dealing with missing data for prognostic purposes," in *2016 Prognostics and System Health Management Conference (PHM-Chengdu)*, 2016, pp. 1–5.
[18] J. Schafer, *Analysis of Incomplete Multivariate Data*, vol. 72. Chapman & Hall, 1997.
[19] S. Zhang, J. Zhang, X. Zhu, Y. Qin, and C. Zhang, "Missing Value Imputation Based on Data Clustering," in *Transactions on Computational Science I*, Berlin, Heidelberg: Springer, 2008, pp. 128–138.
[20] G. Toshniwal , Durga, Satish, "Missing Value Imputation Method Based on Clustering and Nearest Neighbours," *Int. J. Futur. Comput. Commun.*, vol. 1, 2012.
[21] T. Liu, A. W. Moore, and A. Gray, "New Algorithms for Efficient High-Dimensional Nonparametric Classification," *J. Mach. Learn. Res.*, vol. 7, no. Jun, pp. 1135–1158, 2006.
[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. Oct, pp. 2825–2830, 2011.
[23] J. Tyrer, S. W. Duffy, and J. Cuzick, "A breast cancer prediction model incorporating familial and personal risk factors," *Stat. Med.*, vol. 23, no. 7, pp. 1111–1130, Apr. 2004.
[24] D. M. Euhus, K. C. Smith, L. Robinson, A. Stucky, O. I. Olopade, S. Cummings, J. E. Garber, A. Chittenden, G. B. Mills, P. Rieger, L. Esserman, B. Crawford, K. S. Hughes, C. A. Roche, P. A. Ganz, J. Seldon, C. J. Fabian, J. Klemp, and G. Tomlinson, "Pretest prediction of BRCA1 or BRCA2 mutation by risk counselors and the computer model BRCAPRO," *J. Natl. Cancer Inst.*, vol. 94, no. 11, pp. 844–851, Jun. 2002.
[25] W. Jorritsma, F. Cnossen, and P. M. A. van Ooijen, "Improving the radiologist–CAD interaction: designing for appropriate trust," *Clin. Radiol.*, vol. 70, no. 2, pp. 115–122, Feb. 2015.
[26] J. M. McGuirl and N. B. Sarter, "Supporting Trust Calibration and the Effective Use of Decision Aids by Presenting Dynamic System Confidence Information," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 48, no. 4, pp. 656–665, Dec. 2006.
[27] P. K. J. Han, W. M. P. Klein, T. Lehman, B. Killam, H. Massett, and A. N. Freedman, "Communication of Uncertainty Regarding Individualized Cancer Risk Estimates," *Med. Decis. Mak.*, vol. 31, no. 2, pp. 354–366, Mar. 2011.