Statistical and Machine Learning Approaches for Family History Data

A DISSERTATION PRESENTED
BY
Theodore Huang
TO
The Department of Biostatistics
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy
IN THE SUBJECT OF
Biostatistics

Harvard University
Cambridge, Massachusetts
May 2019
Statistical and Machine Learning Approaches for Family History Data

Abstract

Germline mutations in many genes have been shown to increase the risk of developing cancer, and numerous statistical models have been developed to predict genetic susceptibility to cancer. Mendelian models predict risk by using family histories with estimated cancer penetrances (age- and sex-specific risk of cancer given the genotype of the mutations) and mutation prevalences. This dissertation is focused on using statistical and machine learning tools to improve Mendelian risk prediction models, as well as exploring assumptions in these models.

Mendelian models assume conditional independence between families members’ cancer ages given the genotype and sex. However, this assumption is often violated due to residual risk heterogeneity even after accounting for the mutations in the model. In chapter 1, we aim to account for this heterogeneity by incorporating a frailty model that contains a family-specific frailty vector, impacting the cancer hazard function. We apply the proposed approach to directly improve breast cancer prediction in BRCAPRO, a Mendelian model that accounts for inherited mutations in the $BRCA1$ and $BRCA2$ genes to predict breast and ovarian cancer. We evaluate the proposed model’s performance in simulations and real data from the Cancer Genetics Network and show improvements in model calibration and discrimination. We also discuss other approaches for incorporating frailties and their strengths and limitations.
In chapter 2, we continue to explore this assumption by determining the extent and sources of the heterogeneity across and within families. We quantify the heterogeneity by evaluating the ratio between the number of observed cancer cases in a family and the number of expected cases under a model where risk is assumed to be the same across families. We perform this analysis for both carriers and non-carriers in each family and visualize the results. We then introduce frailty models as a method to generatively mimic risk heterogeneity, and use synthetic data to explore the impact of various sources of the observed heterogeneity. We apply this approach to data on colorectal cancer in families carrying mutations in Lynch syndrome genes from Creighton University’s Hereditary Cancer Center. We show that colorectal cancer risk in carriers can vary widely across families, and that this variation is not matched by a corresponding variation in the non-carriers from the same families. This suggests that the sources of variation are to be found mostly in variants harbored in the mutated MMR gene considered, or in variants interacting with it.

Compared to training new models from scratch, improving existing widely-adopted prediction models is often a more efficient and robust way towards progress. Existing models may (a) incorporate complex mechanistic knowledge, (b) leverage proprietary information and, (c) have surmounted barriers to adoption. In chapter 3, we propose to combine gradient boosting with any previously developed model to improve existing models while retaining important existing characteristics. To exemplify, we consider the context of Mendelian models, and show via simulations that integration of gradient boosting with an existing Mendelian model can produce an improved model that outperforms both the existing Mendelian model and the model built using gradient boosting alone. We then illustrate the approach on genetic testing data from the USC-Stanford Cancer Genetics Hereditary Cancer Panel Testing study.
# Table of Contents

Title page i
Copyright ii
Abstract iii
Table of Contents v
Acknowledgements vii

1 Practical Implementation of Frailty Models in Mendelian Risk Prediction
   1.1 Introduction ................................................. 1
   1.2 BRCAPRO .................................................. 4
      1.2.1 Notation ........................................... 4
      1.2.2 Mendelian Risk Prediction ......................... 5
   1.3 Frailty Model ............................................. 6
      1.3.1 Model Definition .................................... 7
      1.3.2 Deriving the Baseline Hazard Functions .......... 7
      1.3.3 Risk Prediction ..................................... 9
   1.4 Simulation Study .......................................... 11
      1.4.1 Simulation Design ................................. 11
      1.4.2 Performance Measures ............................. 13
      1.4.3 Results ............................................ 14
   1.5 Application to Cancer Genetics Network Data ....... 15
      1.5.1 Cancer Genetics Network Data .................... 16
      1.5.2 Results ............................................ 16
   1.6 Lessons Learned from Alternative Approaches ....... 20
      1.6.1 Conditional Approach ............................. 22
      1.6.2 Normal Frailty Distribution ...................... 22
   1.7 Discussion ................................................. 25

2 Variation in Cancer Risk among Families with Genetic Susceptibility 30
   2.1 Introduction ............................................. 30
   2.2 Methods .................................................. 33
      2.2.1 Data ............................................... 33
2.2.2 Quantification of Heterogeneity ........................................ 34
2.3 Results ............................................................................. 39
2.4 Impact of Sources of Heterogeneity .................................... 42
  2.4.1 Frailty Models .............................................................. 42
  2.4.2 Synthetic Data .............................................................. 43
2.5 Discussion .......................................................................... 47

3 Extending Models Via Gradient Boosting: An Application to
Mendelian Models ................................................................. 52
  3.1 Introduction ...................................................................... 52
  3.2 Gradient Boosting ............................................................. 55
  3.3 Mendelian Risk Prediction Models ....................................... 57
    3.3.1 Notation .................................................................... 57
    3.3.2 Mendelian Carrier Probability Estimation ...................... 58
    3.3.3 Gradient Boosting with Mendelian Models .................... 59
  3.4 Simulation Study ............................................................... 60
    3.4.1 Gradient Boosting Approach to Incorporating Family History
          Information ................................................................. 60
    3.4.2 Generating Families ................................................... 63
    3.4.3 Simulation Setup ....................................................... 65
    3.4.4 Simulation Results .................................................... 67
  3.5 Data Application ............................................................... 71
    3.5.1 USC-Stanford Data .................................................... 71
    3.5.2 Results .................................................................... 74
  3.6 Discussion .......................................................................... 78

References .............................................................................. 83

Supplementary Materials ......................................................... 89
  S0.1 Code ............................................................................ 89
  S1.1 Deriving Baseline Hazards .............................................. 89
  S1.2 Generating Families ...................................................... 90
  S1.3 Estimating the Data-generating Frailty Distribution .............. 92
  S1.4 Conditional Approach .................................................. 93
  S1.5 Profile Likelihood ........................................................ 95
  S1.6 Twin Approach to Estimating \( \Sigma \) .................................. 97
  S2.1 Derivation of O/E Ratios ................................................ 106
Acknowledgments

First and foremost, I would like to thank God and give all credit and glory to Him for my dissertation. Everything I have is from Him, and without Him none of my accomplishments would be possible.

I would also like to thank my PhD advisor, Giovanni Parmigiani, as well as Danielle Braun, who has essentially been my co-advisor. Both of them have taught me many valuable skills and showed me how to think and approach problems like a statistician. They also taught me the importance of data visualization, reproducibility, clear communication to different audiences, and clinical relevance. They were very supportive and patient during my PhD studies, both in the successful and in the frustrating times.

In addition, I would like to thank the BayesMendel group for their feedback and support. In particular, I am grateful for Zoe Guan, Tom Madsen, and Jane Liang, who all provided valuable code.

I am also grateful for my collaborators—Malka Gorfine, Li Hsu, Gregory Idos, Christine Hong, and Henry Lynch. They provided important ideas and feedback as well as valuable data sets to validate my statistical methods. In addition, I am thankful for my committee members, Lorenzo Trippa and Sebastien Hanusee, for their suggestions and constant support.

None of my research would be possible without my family—my parents Ming-Hsing and Rachel, my sister Tiffany, and my brother Timothy. My parents raised me and have supported me at every step in my life, making countless sacrifices along the way so that I can be where I am today. My siblings have also taken care of me and helped me academically over the years.
Lastly, I want to thank my church, Antioch Baptist Church. I would like to thank Pastor Paul and Becky JDSN for founding the Berkland Baptist Church network, without which I would not be a Christian. I am also grateful for many spiritual shepherds, in particular Pastor David, Pastor Thomas, Pastor Sang, and Ed hyung, who have taught and guided me through Bible studies, sermons, discipleship, and their own life examples. I am extremely grateful for all the people in the church who have loved, supported, and prayed for me for my past five years in Boston.
1

Practical Implementation of Frailty Models in Mendelian Risk Prediction

1.1 Introduction

Inherited susceptibility to various cancers is becoming better understood, as many genetic mutations have been linked to increased risk of cancer. For example, germline mutations of the \textit{BRCA1} and \textit{BRCA2} genes have been shown to confer a markedly increased risk in breast and ovarian cancer (King et al., 2003). Thus, familial cancer risk prediction has become increasingly important in guiding decision-making. Being able to accurately assess the future risk of cancer can significantly reduce the cancer burden by identifying individuals who would benefit from genetic testing, disease screening, and prevention.

Many statistical models have been developed to assess future risk for various cancers based on family histories and known cancer susceptibility genes. Mendelian models predict risk by using cancer-specific penetrances (the age- and sex-specific probability of having the cancer given the mutation status), population mutation prevalences, and family history of cancer. They use Mendelian laws of inheritance to calculate the proband’s genotypic distribution and future risk of cancer, conditional on the family history. These models rely on the assumption of conditional independence of the cancer diagnoses among a family given the genotypes. However, studies have
shown that this assumption is often violated. For example, Claus et al. (1998) showed that among families with a moderate history of breast cancer, family history remains a significant risk factor for breast cancer even after accounting for the BRCA1 and BRCA2 mutations. Antoniou et al. (2001) found evidence that numerous other genes contributed to the residual correlation even after accounting for mutations in the BRCA1 and BRCA2 genes. Mavaddat et al. (2015) developed a polygenic risk score by considering 77 single nucleotide polymorphisms (SNPs) associated with breast cancer, and showed that this risk score could successfully stratify breast cancer risk.

Some Mendelian risk prediction models do account for the heterogeneity of risk across families. For example, the BOADICEA model (Antoniou et al., 2008) aims to capture genetic effects across many genes by using a polygenic model. In addition to information from mutations in BRCA1 and BRCA2, the current version of the model incorporates information from mutations in PALB2, CHEK2, and ATM, three other breast cancer susceptibility genes; a polygenic risk score based on 313 breast cancer-associated SNPs; a polygenic component aimed to capture residual genetic effects; and other non-genetic risk factors (Lee et al., 2019). The polygenic component helps to account for risk heterogeneity by capturing residual genetic effects, but the assumption of conditional independence given the additional genotypes and risk factors may still be violated due to unobserved non-heritable risk factors.

One alternative approach to handling the risk heterogeneity is to use frailty models. These models include a frailty term, which is a family-specific random effect acting multiplicatively on the hazard function that aims to capture the unobserved risk shared by the family members. Gorfine et al. (2013) propose a frailty model for a single disease, introducing a marginalized approach that averages the familial risk over the family’s frailty distribution and a calibrated-conditional approach that estimates
the frailty term and then calibrates the risk predictions using an external data set. They then extend their approach (Gorfine et al., 2014) to handle multiple diseases by using a competing risks framework, which considers the time to the first cancer and defines the frailty model with cause-specific hazard functions. However, while they developed the theoretical foundation for frailty modeling in Mendelian risk prediction, incorporating these models into existing Mendelian risk prediction software requires additional considerations. For example, their model uses continuous hazard functions, while Mendelian risk prediction software typically rely on discrete hazard functions, and family history data are provided in discrete form. In addition, their approach ignores scenarios in which family members could experience multiple events, such as the case where an individual develops both breast and ovarian cancer. Lastly, the methods described in these papers assume a normal frailty distribution with a fixed unknown covariance matrix, which can be challenging to estimate. Since the uptake of genetic testing has been increasing, it is important to build upon prediction models that are already used clinically in genetic counseling.

In this work we propose a frailty model that accounts for risk heterogeneity and directly applies to Mendelian models to improve future cancer risk. We use a discrete frailty distribution, which is compatible with Mendelian models and allows us to efficiently utilize a marginalized approach for risk prediction. We then illustrate the potential gains in model performance as well as the practical challenges in implementing this approach by applying it to BRCAPRO (Parmigiani et al., 1998; Berry et al., 1997), a Mendelian model that, given an individual’s family history, estimates the probabilities of carrying mutations of the \textit{BRCA1} and \textit{BRCA2} genes as well as the future risk of breast and ovarian cancer. Although BRCAPRO can estimate the probability of being a mutation carrier, we focus this work on future risk due to its larger
The rest of the paper is organized as follows. In Section 1.2, we explain BRCAPRO in detail. In Section 1.3, we present our proposed frailty model. Section 1.4 presents a simulation study, generating data from various frailty distributions, to analyze the extent of improvement over BRCAPRO without the frailty adjustment as well as robustness to misspecification of the population-level frailty distribution. In Section 1.5, we apply our approach to data from the Cancer Genetics Network and show an improvement in model calibration and discrimination. Lastly, while the focus of this work is on the discrete frailty distribution and marginalized approach to risk prediction, in Section 1.6 we also provide a discussion of alternative approaches for applying a frailty model.

1.2 BRCAPRO

In this section, we explain how BRCAPRO predicts risk of breast and ovarian cancer based on mutations of BRCA1 and BRCA2. However, the general structure applies to any Mendelian model, irrespective of these cancers and genes.

1.2.1 Notation

Consider a family with $n$ family members. Let $T_{ri}$ be the failure time of the $r$-th cancer for the $i$-th family member, where we let $r = 1$ represent breast cancer and $r = 2$ represent ovarian cancer. Let the subscript $i = 1$ denote the proband, the individual who receives the risk assessment. Let $C_i$ indicate the censoring time, $X_{ri} = \min(T_{ri}, C_i)$ indicate the observed event time for the $r$-th cancer, and $\delta_{ri} = I(T_{ri} \leq C_i)$. Also let $G_i = (G_{i1}, G_{i2})$ indicate the genotype for the $i$-th individual, where
$G_{i1}$ is a binary indicator of the carrier status for $BRCA1$ and $G_{i2}$ indicates the carrier status for $BRCA2$, and let $U_i$ be the indicator of the $i$-th individual being male. Lastly, let $T_r = (T_{r1}, \ldots, T_{rn})$, $T = (T_1, T_2)$, $X_r = (X_{r1}, \ldots, X_{rn})$, $X = (X_1, X_2)$, $G = (G_1, \ldots, G_n)$, and $U = (U_1, \ldots, U_n)$.

1.2.2 Mendelian Risk Prediction

In practice, the family members’ genotypes $G_i$ are often unknown and hence need to be estimated. Mendelian risk prediction models predict the carrier statuses by using the prevalence of genetic mutations and the relationship between genotypes and phenotypes. They rely on Mendelian laws of inheritance to estimate the joint distribution of the family’s genotypes given the family history. This is in contrast to empirical risk prediction models which use data to directly model the distribution of the genotype given the phenotype, and thus do not incorporate knowledge of the genetic mode of inheritance.

BRCAPRO estimates each family member’s probability of carrying the $BRCA1$ and $BRCA2$ mutations; i.e., $P(G_i|H)$, where $H = (H_1, \ldots, H_n)$ and $H_i = (X_{1i}, X_{2i}, \delta_{1i}, \delta_{2i})$ is the history of the $i$-th family member. For example, the proband’s probability of having genotype $G_1$ is

$$P(G_1|H, U) = \frac{P(G_1) \sum_{G_2, \ldots, G_n} \prod_{i=1}^n P(H_i|G_i, U_i) P(G_2, \ldots, G_n|G_1)}{\sum_{G_2} P(G_1) \sum_{G_3, \ldots, G_n} \prod_{i=1}^n P(H_i|G_i, U_i) P(G_2, \ldots, G_n|G_1)}.$$ 

Here, we use Bayes’ rule and the assumption of conditional independence of the family members’ phenotypes given the genotypes. Mendelian laws of inheritance are used to obtain $P(G_2, \ldots, G_n|G_1)$. $P(G_1)$ is derived from the prevalence of mutations in the general population, which is assumed to be known, and $P(H_i|G_i, U_i)$ is derived
from the penetrance, which is the age- (from ages 1 to 94) and sex-specific probability in the general population of developing each cancer given the genotype, which is also assumed to be known. BRCAPRO uses literature-based estimates for the penetrances, including discrete distributions for female and male breast cancer and female ovarian cancer, as well as prevalences; however, the user can also input their own values.

Using the estimated carrier probability, \( P(G_1|H,U) \), we can estimate the future risk of the cancers. If the proband is free of the \( r \)-th cancer at the current age \( X_{r1} \) (i.e., \( T_{r1} > X_{r1} \)), then the \( t \)-year risk of the \( r \)-th cancer is

\[
P(T_{r1} \leq X_{r1} + t|H,U) = \sum_{G_1} P(T_{r1} \leq X_{r1} + t|G_1, U_1, T_{r1} > X_{r1})P(G_1|H,U) \\
= \sum_{G_1} \frac{P(X_{r1} < T_{r1} \leq X_{r1} + t|G_1, U_1)}{P(T_{r1} > X_{r1}|G_1, U_1)}P(G_1|H,U). \tag{1.1}
\]

1.3 Frailty Model

Let \( W = (W_1, W_2) \) be the family-specific frailty vector, where \( W_1 \) represents the latent risk factors associated with breast cancer and \( W_2 \) for ovarian cancer. We assume that \( W \) is independent of the genotypes \( G \) and sex \( U \) so that the frailty represents risk factors unrelated to the \( BRCA1 \) and \( BRCA2 \) mutations and the sex.

This is a standard assumption in frailty models. Furthermore, we assume that conditional on \( W, G, \) and \( U \), the family members’ failure times are independent; i.e., \( T_{ri} \perp T_{i'j'} | W, G, U \) for \( (r,i) \neq (r',i') \). Thus, the frailty vector is assumed to capture the residual correlation between the family members’ failure times after accounting for the genotypes and sex.
1.3.1 Model Definition

We define the proposed frailty model as \( \lambda_{ri}(t|G_i, U_i, W) = \lambda_{0rU_iG_i}(t) \exp(W_r) \) for \( r = 1, 2, i = 1, \ldots, n \). \( \lambda_{ri}(t|G_i, U_i, W) \) is the hazard for the \( i \)-th individual in the family, and \( \lambda_{0rU_iG_i}(t) \) is the sex- and genotype-specific baseline hazard function. In this model we do not consider other covariates, but the model can easily be extended to account for additional covariates that are risk factors for the two cancers. In fact, BR-CAPRO, which will be used for the risk prediction, considers risk factors such as race, whether an individual is Ashkenazi Jewish, and preventative risk-reducing interventions including prophylactic mastectomies and oophorectomies; however, it incorporates these factors by modifying the penetrance and prevalence and not as covariates in a regression model.

BR-CAPRO uses discrete-time penetrances; i.e., the probability of having the cancer at a certain age given the genotype and sex. Thus, we only observe age in grouped one-year interval form (i.e., 25, 26, 27, ...) and will consider the discrete failure times \( T_{ri}^d = \lfloor T_{ri} \rfloor \), where \( \lfloor \cdot \rfloor \) is the floor function and the \( d \) superscript is used to denote discrete time. The discrete hazard functions can then be defined as \( \lambda_{ri}^d(t|G_i, U_i, W) = P(T_{ri}^d = t|T_{ri}^d \geq t, G_i, U_i, W) \). The frailty model is then reformulated in terms of these discrete hazard functions as \( \lambda_{ri}^d(t|G_i, U_i, W) = 1 - [1 - \lambda_{0rU_iG_i}(t)]^{\exp(W_r)} \). This is often called the complementary log-log link model, and the equivalence between the two models is shown in Kalbfleisch & Prentice (2011).

1.3.2 Deriving the Baseline Hazard Functions

The baseline hazard functions are unknown, and are derived from the BR-CAPRO penetrances using the following approach described in Gorfine et al. (2013). BR-
CAPRO provides age- and sex-specific penetrances, defined as
\[ f^d_{ri}(t|G_i, U_i) = P(T^d_{ri} = t|G_i, U_i). \] These are population (not family-specific) quantities, and are considered to be the average over the frailty distribution:
\[ P(T^d_{ri} = t|G_i, U_i) = \sum W P(T^d_{ri} = t|G_i, U_i, W) P(W). \] Here we assume that \( W \) is discrete, but we could replace the summation with integration if \( W \) is continuous. The marginal (with respect to the frailty) survival functions \( S^d_{ri}(t|G_i, U_i) = P(T^d_{ri} > t|G_i, U_i) \) and hazard functions \( \lambda^d_{ri}(t|G_i, U_i) = P(T^d_{ri} = t|G_i, U_i, T_{ri} \geq t) \) are directly obtained from the BRCAPRO penetrances. The baseline hazard functions \( \lambda^d_{0rUG}(t) \) are estimated by equating the marginal survival functions from BRCAPRO to the marginal survival functions obtained by averaging the conditional survival functions based on the frailty model over the frailty distribution. Explicitly, this involves finding the \( h_r \) that minimizes the following objective function:

\[
\left\{ S^d_r(t|G, U) - \sum_{W_r} \left\{ \prod_{s=1}^{t-1} \left[ 1 - \lambda^d_{0rUG}(s) \right] \right\}^{\exp(W_r)} (1 - h_r)^{\exp(W_r)} P(W_r) \right\}^2.
\]

This minimization to obtain \( \lambda^d_{0rUG}(t) \) is performed sequentially for \( t = 1, 2, \ldots, 94 \) (94 is the maximum age in BRCAPRO), and for all \( G \in \{0, 1\}^2 \) and \( r = 1, 2 \) and \( U = 0, 1 \). Here we suppress the subscript \( i \) to emphasize that the baseline hazard functions are not specific to family members but are population quantities. We also sum over the marginal frailties \( W_r \) instead of the entire vector \( W \) since we assume that \( S^d_r(t|G, U, W) = S^d_r(t|G, U, W_r) \); i.e., that given the genotype and sex, only the cancer-specific frailty \( W_r \) impacts the cancer-specific survival \( S^d_r \). Further details including a derivation of the above objective function are provided in Supplementary Section S1.1.
1.3.3 Risk Prediction

In order to explicitly derive the baseline hazard functions, we need to specify a distribution for the frailty vector. In our implementation, we let the support of the distribution of $W_1$ and $W_2$ each be $\{-1.5, -1, \ldots, 1, 1.5\}$. This support is chosen to allow for a wide range for the frailties while maintaining realistic frailty-adjusted penetrances $f^{d}_{ri}(t|G_i, U_i, W)$, which are obtained by

$$f^{d}_{ri}(t|G_i, U_i, W) = P(T^{d}_{ri} = t|G_i, U_i, W) = S^{d}_{ri}(t-1|G_i, U_i, W_r)\lambda^{d}_{ri}(t|G_i, U_i, W_r)$$

$$= \left\{ \prod_{s=1}^{t-1} \left[ 1 - \lambda^{d}_{brU_iG_i}(s) \right]^{\exp(W_s)} \right\} \left\{ 1 - \left[ 1 - \lambda^{d}_{brU_iG_i}(t) \right]^{\exp(W_r)} \right\}.$$

Frailties that are too high lead to unrealistic penetrances, as they cause a dramatic leftward shift of the carrier penetrances that imply a high probability of cancer at young ages and a near-zero probability at older ages (see Supplementary Figure S1.9). In this work, we will assume that the frailty distribution is uniform over this support. The uniform distribution is a natural choice, as it assumes no prior knowledge about the frailty distribution; however, it makes the implicit assumption that $W_1$ and $W_2$ are independent, which may be invalid if families’ residual risk factors tend to impact both breast and ovarian cancers in the same direction. One can estimate the frailty distribution to relax this assumption, but this process can be difficult, as explained in Section 1.7.

We use the marginalized approach as described in Gorfine et al. (2013) for future risk prediction. Given a family’s frailty vector, the proband’s future risk of developing cancer can be predicted by inputting the frailty-adjusted penetrances into BR-CAPRO. Since a family’s frailty vector is unknown, the predicted risk is averaged
over the family’s frailty distribution, conditional on the family history:

\[
P(T_{r1} \leq X_{r1} + t|H, U) = \sum_{W} P(T_{r1} \leq X_{r1} + t|H, U, W)P(W|H, U). \tag{1.2}
\]

\(P(T_{r1} \leq X_{r1} + t|H, U, W)\) is obtained by averaging over the genotypic distribution using equation (1.1). Directly obtaining the family-specific frailty distribution \(P(W|H, U)\) can be computationally challenging for large families, as it requires summing over all \(4^n\) possible genotypic configurations for the family. Instead, in order to obtain the sum efficiently, we use the peeling algorithm (Fernando et al., 1993), which is a recursive method that calculates genotype probabilities for each member in the family. Suppose we are interested in calculating \(P(G_i|H, U)\) for the \(i\)-th individual in the family, and let \(G_{-i} = (G_1, \ldots, G_{i-1}, G_{i+1}, \ldots, G_n)\) be the (potentially unknown) genotypes for everyone else in the family. Then, \(P(G_i|H, U) \propto P(G_i|U)P(H|G_i, U) = P(G_i)\sum_{G_{-i}} P(H|G_i, U)P(G_{-i}|G_i)\). Given the likelihood \(P(H_i|G_i, U_i)\) for each individual, the mutation prevalence \(P(G_i)\), and the pedigree structure, the peeling algorithm efficiently sums over the family members’ genotypes to obtain the \(i\)-th family member’s genotype probabilities for all possible genotype configurations. Similarly, we can use this approach to obtain the family-specific frailty distribution:

\[
P(W|H, U) \propto P(W|U)P(H|W, U) = P(W)\sum_{G_i} P(H|W, G_i, U)P(G_i|W, U) \\
= P(W)\sum_{G_i} P(G_i) \left[ \sum_{G_{-i}} P(H|W, G, U)P(G_{-i}|G_i) \right]. \tag{1.3}
\]

For each possible frailty vector, we obtain a quantity proportional to the family-
specific frailty probability. After obtaining probabilities for all possible frailty vectors, the family-specific frailty distribution is obtained by normalization. This normalized distribution is then used to calculate the predicted future cancer risk for the proband, using equation (1.2). If there are missing ages in the family history, then the missing current ages are imputed using the BRCAPROLYTE-Plus approach in Biswas et al. (2013) and the missing cancer ages are imputed using the non-carrier penetrances through multiple imputation.

1.4 Simulation Study

1.4.1 Simulation Design

We conduct a simulation study to quantify the gains in performance from the proposed frailty model in realistic scenarios, including when the frailty distribution is misspecified. We generate data using three population-level frailty distributions: a bivariate normal with mean $\mathbf{0}$, variances of 0.3 for both frailties, and correlation of 0; a bivariate normal with mean $\mathbf{0}$, variances of 2 for both frailties, and correlation of 0; and a discrete uniform distribution on $\{-1.5, -1, \ldots, 1, 1.5\}^2$. These different frailty distributions are used to generate the data; however, when implementing our frailty model we always assume a discrete uniform distribution. Thus the two normal distributions represent misspecified frailty distributions. The bivariate normal with variances of 0.3 has a density that lies mostly (with probability 0.988) in the support of the discrete uniform distribution used in our model. The bivariate normal with variances of 2 has only probability 0.506 of lying within this support, so it represents a more misspecified distribution. We also generate data using population-level as well as high-risk $BRCA1$ and $BRCA2$ allele frequencies. The population-level
allele frequencies are the ones in BRCAPRO (BRCA1: 0.014 for Ashkenazi Jews, 0.0006 for non-Ashkenazi Jews; BRCA2: 0.012 for Ashkenazi Jews, 0.0007 for non-Ashkenazi Jews). For the high-risk group we used an allele frequency of 0.05 for both BRCA1 and BRCA2 for Ashkenazi Jews and 0.02 for both BRCA1 and BRCA2 for non-Ashkenazi Jews. We generated Ashkenazi Jew status using a probability of 0.019, based on an estimate of the proportion of Ashkenazi Jews in the United States.

We generate families based on these population-level frailty distributions and allele frequencies. Family data is generated at two time points: baseline (to apply the model) and 5-year follow-up (to validate the model). We generate the follow-up data first, then look back 5 years to obtain the baseline data. The data generation process is as follows: for each choice of the population-level frailty distribution and allele frequencies, we generate 100,000 families with varying family sizes (average family size of around 32) and with female probands. For each family, we sample a frailty vector based on the frailty distribution and calculate the family-specific penetrance, \( f_d(t|G, U, W) \), which is obtained by modifying the penetrances in BRCAPRO by the family’s frailty. For the founders in the pedigree, we assign genotypes based on the chosen allele frequencies; genotypes for the rest of the family are assigned following Mendelian laws of inheritance. Using the family-specific penetrance and genotypes, we sample the ages that the family members develop the cancers, which may or may not be observed due to censoring. To account for censoring, we sample the family members’ current ages at the time of the proband’s 5-year follow-up. A family member’s cancer age is only observed if the cancer age is less than or equal to the current age. Finally, we subtract 5 years to obtain the ages and cancer statuses at the time of the proband’s consultation. Further details about the data-generating process are provided in Supplementary Section S1.2.
After generating 100,000 families, we exclude probands who had already developed breast or ovarian cancer at the time of consultation (but not excluding family members who developed these cancers). We exclude proband breast cancer cases, as we are interested in future breast cancer risk prediction, and we exclude proband ovarian cancer cases as the presence of ovarian cancer may impact these probands’ risk of developing breast cancer. We then calculate the frailty-adjusted 5-year breast cancer risk for the remaining probands by averaging over the family-specific frailty distributions using the approach described in Section 1.3.3. All results were obtained using BRCAPRO in the BayesMendel R package, version 2.1-5.

1.4.2 Performance Measures

We assess the performance of our proposed risk prediction model using three measures (Steyerberg et al., 2010): calibration as measured by the ratio of the number of observed events to the number of expected events (O/E), discrimination as measured by the area under the receiver operating characteristic (ROC) curve (ROC-AUC), and accuracy as measured by the root Brier score (rBS), defined as the square root of the average squared difference between the observed event and the predicted risk, mathematically expressed as \(\sqrt{\frac{1}{N} \sum_{k=1}^{N} \{ I(T_{k11} \leq X_{k11} + t) - \hat{p}_k(t) \}^2} \). Here, \(I(T_{k11} \leq X_{k11} + t)\) is the indicator of the proband in the \(k\)-th family developing breast cancer within \(t\) years of the age of consultation \(X_{k11}\) (we only consider probands who are breast cancer-free at baseline, i.e., \(T_{k11} > X_{k11}\)), and \(\hat{p}_k(t)\) is the predicted \(t\)-year risk of breast cancer for this proband. In this work we focus on 5-year risk, or \(t = 5\).
1.4.3 Results

Table 1.1 shows the simulation results—the top section for data generated using population-level allele frequencies, and the bottom section for data generated using high-risk allele frequencies. Overall, we see that the frailty risk prediction results show improvements compared to the BRCAPRO predictions, demonstrating that the frailty approach using the discrete uniform distribution is robust to misspecification of the data-generating frailty distribution.

For the data generated using population-level allele frequencies, we see that under a misspecified setting in which the data is generated from a bivariate normal distribution with variances of 0.3, the frailty model performs similarly to BRCAPRO without the frailty adjustment. However, the frailty adjustment results in improvements in discrimination when the data is generated from the other two distributions (one correctly specified and one misspecified), without much change in the calibration and accuracy. This is because these two data-generating distributions provide a larger frailty effect since they have larger variances, which may allow the frailty model to more easily differentiate between high-risk and low-risk families. We even see this phenomenon when comparing the results using the misspecified bivariate normal distribution with variances of 2 and the correctly specified discrete uniform distribution. Without the frailty adjustment, both distributions had similar AUC values, whereas the frailty model displayed a larger improvement in discrimination for the normal distribution with variances of 2, despite the distribution being misspecified. When we generate data using high-risk allele frequencies, we find similar results. The AUC values were all higher compared to the corresponding values in the data generated using population allele frequencies, showing that regardless of the frailty modification, BRCAPRO
Table 1.1: Performance measures comparing BRCAPRO with and without the frailty model for the families simulated using three different frailty distributions two sets of allele frequencies, with 95% bootstrap confidence intervals.

<table>
<thead>
<tr>
<th>Dist</th>
<th>Var</th>
<th>Frailty</th>
<th>O/E</th>
<th>AUC</th>
<th>rBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>BVN 3</td>
<td>Yes</td>
<td>0.888 (0.793, 0.983)</td>
<td>0.727 (0.705, 0.749)</td>
<td>0.061 (0.058, 0.064)</td>
<td></td>
</tr>
<tr>
<td>BVN 0.3 No</td>
<td>0.886 (0.801, 0.971)</td>
<td>0.720 (0.695, 0.744)</td>
<td>0.061 (0.058, 0.064)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BVN 2 Yes</td>
<td>1.037 (0.944, 1.130)</td>
<td>0.814 (0.793, 0.834)</td>
<td>0.065 (0.062, 0.068)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BVN 2 No</td>
<td>0.995 (0.903, 1.093)</td>
<td>0.733 (0.712, 0.755)</td>
<td>0.065 (0.062, 0.068)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DU Yes</td>
<td>1.005 (0.903, 1.097)</td>
<td>0.751 (0.731, 0.771)</td>
<td>0.065 (0.062, 0.068)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DU No</td>
<td>0.998 (0.905, 1.091)</td>
<td>0.720 (0.698, 0.740)</td>
<td>0.065 (0.062, 0.068)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>High-risk allele frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>BVN 0.3 Yes</td>
</tr>
<tr>
<td>BVN 0.3 No</td>
</tr>
<tr>
<td>BVN 2 Yes</td>
</tr>
<tr>
<td>BVN 2 No</td>
</tr>
<tr>
<td>DU Yes</td>
</tr>
<tr>
<td>DU No</td>
</tr>
</tbody>
</table>

a Data-generating frailty distribution
b Variances of bivariate normal distribution
c Yes if the frailty adjustment was used, No if not
d Bivariate normal
e Discrete uniform

We used the bootstrap to obtain confidence intervals for the performance measures. Ideally, we would generate many data sets to avoid resampling the data; however, we did not do this due to the computational burden. Using this approach could tighten our confidence intervals. Many of the bootstrap confidence intervals we generated overlap, making it difficult to derive strong conclusions from the data.

1.5 Application to Cancer Genetics Network Data

In this section, we aim to validate our proposed approach using data from the Cancer Genetics Network (CGN).
1.5.1 Cancer Genetics Network Data

The CGN (Anton-Culver et al., 2003) is a network established by the National Cancer Institute in 1999 to explore the impact of genetics on cancer susceptibility. Individuals with a family history of cancer were enrolled at clinical centers across the United States and completed questionnaires regarding family history of cancer and other relevant information. They were then followed up annually to update their baseline information. The data contains many useful variables that are incorporated into BRCAPRO such as family histories, histories of mastectomies and oophorectomies, \( BRCA1 \) and \( BRCA2 \) testing results, race, and Ashkenazi Jewish status. For this model validation, we consider a subset of \( N = 8,888 \) (out of 26,465) families with female probands who have not developed breast or ovarian cancer or had a prophylactic mastectomy at baseline. This subset has an average family size of 19.85, with an average baseline proband age of 49.55. Probands were followed up for an average of 7.38 years, which makes this data ideal for validating our risk prediction model using 5-year breast cancer risk. Additional details of this subset of the CGN data are shown in Table 1.2.

1.5.2 Results

We first explore the family-specific frailty distributions \( P(W|H, U) \) for each family in the breast cancer subset. Figure 1.1 shows the distribution of the means of each families’ frailty distributions \( \left( \sum_{w_1} w_1 P(W_{k1} = w_1|H, U), \sum_{w_2} w_2 P(W_{k2} = w_2|H, U) \right) \), where the subscript \( k \) denotes the \( k \)-th family in the data, stratified by having relatives with breast or ovarian cancer. Having relatives with breast or ovarian cancer is strongly associated with the family’s frailty mean. Having relatives with one of
### Family
- Number of families: 8888
- Average family size: 19.85
- Average number of relatives with BC: 1.34

### Proband
- Average age: 49.55
- Average follow-up time: 7.38
- Number with BC within 5 years: 133

### Proband Genetic Testing
*Baseline (Follow-up)*
- Number tested: 14 (547)
- Number positive for BRCA1 only: 4 (81)
- Number positive for BRCA2 only: 0 (47)
- Number positive for BRCA1 and BRCA2: 0 (10)

### Proband Cancer History
*Number Affected (Average age at onset)*
- Colon: 682 (55.48)
- Endometrium: 141 (47.90)
- Melanoma: 1116 (45.73)
- Pancreas: 49 (58.49)

### Proband Ethnicity/Race
*Number (Percentage)*
- Ashkenazi Jew: 2594 (29.19)
- White: 7405 (83.31)
- Hispanic: 816 (9.18)
- Black: 334 (3.76)
- Asian: 113 (1.27)
- Native American: 37 (0.42)
- Unknown Race: 183 (2.06)

---

**Table 1.2:** Summary of the subset of the CGN Data used for validating breast cancer risk prediction
the cancers often results in having a positive mean for that cancer’s frailty variate, and having no relatives with that cancer often results in having a negative mean for the frailty variate. Supplementary Figure S1.4 provides a heatmap of the medians of the family-specific frailty probabilities $P(W_k = w_k|H_k, U_k)$. Overall, we see that the families’ frailty probabilities are higher at the higher frailty values, suggesting an aggregate of breast and ovarian cancers in families. The equivalent heatmap for the means, along with a figure that compares the means and medians, is shown in Supplementary Figure S1.8.

After obtaining the family-specific frailty distributions, we then apply the marginalized approach described in Section 1.3.3 to obtain 5-year risk predictions. Since some probands were not followed for 5 years, we cannot directly obtain the performance measures described in Section 1.4.2. We account for censoring using inverse probabil-
Table 1.3: Performance measures on the CGN data comparing BRCAPRO with and without the frailty model, with 95% bootstrap confidence intervals in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>O/E</th>
<th>AUC</th>
<th>rBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frailty</td>
<td>1.035 (0.858, 1.211)</td>
<td>0.679 (0.634, 0.724)</td>
<td>0.126 (0.115, 0.136)</td>
</tr>
<tr>
<td>No Frailty</td>
<td>1.23 (1.021, 1.445)</td>
<td>0.638 (0.596, 0.681)</td>
<td>0.126 (0.115, 0.136)</td>
</tr>
</tbody>
</table>

ity of censoring weighting (IPCW) (Uno et al., 2007). The censoring distribution is estimated using a Kaplan-Meier estimate, and inverse weights are used based on the estimated censoring distribution. Table 1.3 shows the IPCW performance measures comparing BRCAPRO with and without the proposed frailty model. We see the most noticeable difference in the discrimination, as the AUC increases from 0.638 to 0.679 when using the frailty model. The calibration is also improved, as the frailty risk predictions on average are higher and hence closer to the proportion of 5-year breast cancer cases in the data. The rBS remains unchanged when looking at the first three decimal places. The 95% bootstrap confidence intervals, obtained from 1000 bootstrap replicates, do overlap for all three measures. This may be due to the low number (133 out of 8888) of probands who develop breast cancer within five years.

The calibration for the frailty model is in particular better for families with higher numbers of first-degree relatives with breast cancer at baseline, while the discrimination and accuracy do not vary as substantially (Supplementary Figure S1.5). The improved calibration is perhaps due to the families with a richer history of breast cancer having higher probabilities of larger breast cancer frailty variates and hence higher predicted breast cancer risks. We see evidence of this assertion in Supplementary Figure S1.7. We also see significant improvement in both calibration and discrimination for families who are not Ashkenazi Jewish, as seen in Supplementary Figure S1.6. The BRCA mutation variants in Ashkenazi Jews have been extensively studied, so there
may be less variability in the penetrances and hence less of a frailty effect. Thus families who are not Ashkenazi Jewish may have a greater need for the frailty model. Moreover, it is possible that other genetic mutations that increase susceptibility to breast or ovarian cancer may be less prevalent among Ashkenazi Jews. The frailty model can account for the increase in heterogeneity in the families who are not Ashkenazi Jewish due to these unobserved genetic effects.

Some families had sizable differences in predictions when using the frailty model compared to not using it (Figure 1.2). The predictions using the frailty model tend to be higher than the predictions without the frailty model, illustrating why the frailty model improves the model calibration, as BRCAPRO without the frailty model generally underpredicts. The difference in these increases in risk between probands who develop (blue hexagons) and do not develop (red dots) breast cancer within 5 years is difficult to assess in the plot, but on average those who did develop breast cancer in 5 years had a higher increase in predicted risk with the frailty model compared to those who did not. This also explains the overall improvement in discrimination.

1.6 Lessons Learned from Alternative Approaches

In Section 1.3.3 we studied a discrete frailty distribution and used the marginalized approach to predict breast cancer risk. However, Gorfine et al. (2013) and Gorfine et al. (2014) provide additional approaches, using a normal frailty distribution as well as a conditional approach to risk prediction. In this section we explore these alternative specifications of the frailty model and document their strengths and weaknesses.
Figure 1.2: Differences between the risk predictions comparing BRCAPRO with and without the frailty model. The blue portion is a hexagonal heatmap for probands who did not develop breast cancer in 5 years, and the red dots represent probands who did develop breast cancer in 5 years.

Figure 1.3: Heatmap of the medians of the CGN families’ frailty probabilities.
1.6.1 Conditional Approach

Gorfine et al. (2013) introduced a conditional approach where a frailty vector is estimated for each family, in contrast to the marginalized approach where the average is taken over the family’s frailty distribution. After estimating the frailty vector using a likelihood approach (see Supplementary Section S1.4), we can obtain the frailty-adjusted penetrances and apply them to the Mendelian model.

However, one challenge with this approach is that in data such as the CGN data considered here, the likelihood tends to be flat over a wide range, as different frailty vectors can have similar likelihood values. For example, with a uniform frailty distribution, the estimated breast cancer frailty for a family without any cases of breast cancer will always be the minimum value in the support, even though other frailties may produce similar likelihood values. This can make a significant impact on the risk predictions, as the frailty-adjusted penetrances can vary widely depending on the value of the frailty vector. One solution is to use a non-uniform population-level distribution for the frailty, e.g., the normal frailty distribution; however, the parameters in the normal distribution are difficult to estimate (see Section 6.2). The marginalized approach proposed in this work which averages over the family-specific frailty distribution is more reliable, as it accounts for the uncertainty of the family’s frailty.

1.6.2 Normal Frailty Distribution

In this section we consider in detail the case where the frailty distribution is bivariate normal with mean $\mathbf{0}$ and unknown covariance $\Sigma$, in contrast to the discrete one adopted in our main analysis. This is a more natural choice in many respects. For example, using a continuous distribution allows for more variation in frailties between
families. In addition, using a parametric distribution can be statistically advantageous due to its parsimony.

1.6.2.1 Marginalized Approach

The marginalized approach can be applied in this setting, using equation (1.3) for each frailty vector. However, since the frailty distribution is continuous, we have an infinite number of possible frailty vectors, so obtaining the full family-specific frailty distribution as in equation (1.3) is not straightforward. One solution could be to use an MCMC algorithm to obtain the full family-specific frailty distribution, though this could be computationally demanding and impractical in a clinical setting where a proband would require a quick prediction of their future risk.

1.6.2.2 Conditional Approach

The conditional approach can also be applied in this setting. However, maximizing the log-likelihood is not straightforward when \( W \) is normally distributed compared to when \( W \) follows a discrete, finite distribution. When \( W \) is continuous, an optimization algorithm is required to numerically find the maximum likelihood frailty. This can become computationally intensive, as each iteration in the algorithm requires running BRCAPRO to calculate the carrier probabilities. In addition, optimization algorithms can be sensitive to starting values. In order to help ensure that the resulting maximum likelihood estimate is a global maximum, the algorithm should be run using various starting values, which would add further computation time. The derivative of the log-likelihood cannot be used to obtain an estimating equation as in Gorfine et al. (2013), since the log-likelihood includes frailty-adjusted carrier probabilities which are estimated by BRCAPRO.
1.6.2.3 Unknown Covariance Matrix

In addition to the computational challenges, the unknown frailty covariance matrix $\Sigma$ is another challenge arising when using a bivariate normal distribution. Gorfine et al. (2013) and Gorfine et al. (2014) assume a fixed $\Sigma$; however, the choice of $\Sigma$ can significantly impact the risk predictions. For example, larger variance terms will result in more extreme frailty variates, thereby inducing more correlation in the cancer outcomes between the family members, and larger values of the correlation term will similarly result in increased correlation between the two cancer outcomes. Thus, in order to find the optimal value of $\Sigma$, several approaches are considered.

Likelihood Approach for Estimating $\Sigma$. If we have a data set with multiple families, one approach to estimating $\Sigma$ is to regard it as an additional parameter to be estimated in the likelihood. In this approach, the likelihood can be written as a function of $W_1, \ldots, W_N, \Sigma$, where $W_k$ is the frailty vector for the $k$-th family and $N$ is the number of families in the data set. The optimization is performed over $N + 1$ variables, which could become computationally infeasible with a large sample size $N$. A Bayesian approach can also be used, placing a prior distribution on $\Sigma$ and using MCMC methods to obtain posterior estimates of all $N+1$ parameters. This too would be computationally intensive, as each iteration of the algorithm would require $N$ calls to the BRCAPRO algorithm.

A simpler approach would be to use a profile likelihood, profiling out the nuisance parameters $W_k$. We could then perform a grid search to find the value of $\Sigma$ that maximizes the profile log-likelihood. Although this approach seems promising, it can be shown that the profile likelihood is unbounded (see Supplementary Section S1.5), rendering estimation infeasible.

24
External Data. An alternative approach for estimating $\Sigma$ is to use an external data set along with simulations. We could simulate families under various values of $\Sigma$. For each simulated data set, we could then obtain a summary measure of association or sharedness within the families. We could then choose the value of $\Sigma$ that produces a summary measure that most closely matches the summary measure in the external data set. An example approach using twin data is presented in Supplementary Section S1.6. However, one limitation of this approach is that the data-generating mechanism in the external data set is unknown and may not match the penetrances used for simulations. This could result in nonviable estimates of $\Sigma$, such as ones with negative correlations, which may be unrealistic since unobserved shared risks for breast and ovarian cancer should intuitively be either both positive or negative.

Sensitivity Analysis. Instead of estimating a specific value of $\Sigma$, a sensitivity analysis can be conducted to assess the effect of varying $\Sigma$ on the risk predictions. However, our interest is in selecting one $\Sigma$ value that would allow for implementation of the frailty model. Although one could select the optimal $\Sigma$ based on a sensitivity analysis, using this $\Sigma$ in practice would require a potentially strong generalizability assumption.

1.7 Discussion

There is an increasing need for well-calibrated and discriminating familial risk prediction models. BRCAPRO performs well in estimating the probabilities of carrying the $BRCA1$ and $BRCA2$ mutations (Parmigiani et al., 2007), but does not perform as well when predicting future risk of cancer (as seen in Table 1.3, as the O/E is 1.227 and AUC is 0.631). This may be due to the failure of BRCAPRO to account for un-
observed risk factors beyond the *BRCA1* and *BRCA2* mutations.

The frailty model proposed in this paper provides a natural extension to the BR-CAPRO model that captures the heterogeneity of risk across families due to unobserved genetic and environmental factors. As our approach directly extends BR-CAPRO, which is widely used in genetic counseling, it is more practical and has more clinical relevance than the previous frailty approaches suggested in Gorfine et al. (2013) and Gorfine et al. (2014). Extending existing cancer risk prediction models is important, as these models are being more widely used in practice due to the increased awareness of hereditary cancer syndromes and increased prevalence of genetic panel testing.

The effectiveness of our proposed model was shown through simulation and data application results. The simulations showed an improvement in discrimination as well as a robustness to misspecified population-level frailty distributions, and validation using data from the Cancer Genetics Network showed an improvement in both discrimination and calibration. We also discussed alternative approaches for utilizing a frailty model to improve BRCAPRO risk prediction. We described a bivariate normal frailty distribution as well as a conditional approach which estimates frailty vectors for each family. However, each alternative has limitations such as increased computational time and difficulty in estimating unknown parameters. Thus we believe that the discrete frailty approach proposed in Section 1.3.3 provides the best balance of efficiency and accuracy in predicting breast cancer risk.

The simulations results and data validation are certainly encouraging signs that this model could provide clear advancements for clinical breast and ovarian cancer risk prediction. However, we acknowledge several limitations in our work. The family sizes were larger on average than the families in the CGN data (32 versus 20). De-
spite this, we did not find that the performance measures varied significantly according to the family size, as seen in Supplementary Table S1.1. In addition, for the data validation, there are some limitations in the data that should be considered. The average follow-up time for individuals was 7.33 years; thus, we were limited to studying 5-year risk prediction and could not explore long-term risk prediction. In addition, only 140 out of the 9373 probands developed breast cancer within 5 years. The model performance measures may not be accurate due to this small sample size. Further validation using data sets with longer follow-up times or more breast cancer cases at follow-up could help confirm the effectiveness of the model.

We assumed a uniform population-level frailty distribution. This is a natural and convenient choice but makes strong assumptions, such as the breast and ovarian frailties being independent. We could try to estimate the population-level frailty distribution using external data (or even the entire data set, not just the subset used for validation). One estimation procedure could be to first initialize the population-level frailty distribution with the discrete uniform, then obtain each family’s frailty distribution, and then aggregate the family-specific distributions by taking a summary measure such as the mean or median (see Supplementary Section S1.3). Once we obtain this updated population-level distribution, we could re-run the model and obtain new family-specific frailty distributions. However, results from the simulations showed that this process did not accurately recover the data-generating population-level frailty distribution (see Supplementary Section S1.3 for a discussion), illustrating the challenges of estimating a discrete, nonparametric distribution. Despite the inability to estimate the population-level frailty distribution, the frailty model provided improvements in the AUC without changes to the O/E and rBS (see Table 1.1), giving us confidence that the choice of the discrete uniform distribution is satisfactory for
implementing the model.

The proposed frailty model relies on several assumptions. As described above, we used a discrete uniform frailty distribution, while the true distribution may be some other bivariate distribution, possibly continuous or even asymmetric. Additionally, we consider a fixed range for the uniform distribution, which can be modified. Modifying the range would likely impact the calibration while leaving the discrimination and accuracy mostly unchanged.

We also assumed one frailty variate per cancer for each family, when the true mechanisms may be more complex. For example, if the frailty partially represents environmental risk factors, it may not make sense for distant relatives in a family to share this same trait. One alternative would be to use different frailties for different generations in the family, or have a frailty weight based on the degree of separation from the proband. Another option would be to have one frailty vector for genetic effects and one for environmental effects, where the one for environmental effects is only shared among the proband’s immediate family. The model can be extended to allow for these alternatives; however, the addition of frailty variates may substantially decrease the precision of the frailty estimators. Recently, Metcalfe et al. (2017) showed that among BRCA carriers, there is no significant difference in breast cancer penetrance between those with and without first-degree relatives with breast cancer. Thus having frailty terms only on non-carriers of the BRCA1 and BRCA2 mutations would be an additional model to consider.

The results in this work do not explicitly account for competing risks. We presented results for 5-year net breast cancer risk prediction, which is the probability of developing breast cancer in 5 years in a hypothetical world where there are no competing risks. Crude risk can also be considered, which is the probability of develop-
ing breast cancer before developing any competing risks. BRCAPRO can provide crude risks by accounting for death from other causes; however, it does not account for other possible competing risks.

Overall, this proposed frailty model makes advancements in breast cancer risk prediction. By accounting for unobserved genetic and environmental risk factors in the family, the model showed significant improvements in both calibration and discrimination. We illustrate the proposed extensions on BRCAPRO, but similar improvements may occur when applying this approach to other existing Mendelian risk prediction models, even models considering an arbitrary number of genes and cancers. We hope these improved predictions can make a significant impact in helping individuals with increased family history of cancer better understand their future risk and make important decisions for screening and prevention.
Variation in Cancer Risk among Families with Genetic Susceptibility

2.1 Introduction

Hereditary cancer syndromes represent five to ten percent of all cancers and are caused by inherited germline mutations in cancer susceptibility genes. Determining the impact of these mutations on cancer risk is important, both on a population level and for individual risk assessment. Consequently, many studies have provided estimates of these mutations’ effects. These estimates are often in the form of penetrance functions, which are age-specific probabilities of developing a cancer given a genotype. Estimates of cancer penetrance typically quantify cancer risk for the high-risk population of mutation carriers by assuming that all deleterious mutations in the same gene have the same effect. In order to properly evaluate the cancer risk among mutation carriers, users of these estimated penetrances would ideally consider not only the average penetrance but also the potentially substantial variability in penetrance estimates within a gene. A better understanding of this variability would improve our ability to help high-risk individuals and families accurately assess their future risk of cancer.

Risk heterogeneity among mutation carriers can arise due to a variety of sources. Within the same gene, there can be multiple pathogenic variants that can lead to dif-
ferent cancer risks (Rebbeck et al., 2015). However, at the present time, only a minority of specific deleterious variants have been studied individually. Examples include the founder Ashkenazi mutations in \textit{BRCA1} and \textit{BRCA2} (Levy-Lahad et al., 1997; Struewing et al., 1997). Most mutations are too uncommon for traditional penetrance estimation. Gene-gene interactions (epistasis), the phenomenon where the phenotypic expression of a mutation can depend on the presence of other genetic mutations, are another source of heterogeneity, and have been shown to occur in cancer susceptibility genes (Turnbull et al., 2011). Environmental factors can also produce variation in risks among carriers, even of the same variant. For example, exposure to tobacco smoke has been linked to an increased risk of multiple cancers (of Health et al., 2004). Levels of these environmental risk factors can vary widely among carriers. Similarly, behavioral choices can impact cancer risk. A common example is physical inactivity, which is associated with several cancers (Brown et al., 2012).

Previous studies have examined sources of cancer risk heterogeneity among mutation carriers. Several studies provide evidence of risk modifiers of cancer among mutation carriers. Burn et al. (2011) used results from the CAPP2 randomized trial to show that among patients with Lynch syndrome (Lynch et al., 1966), long-term aspirin use is associated with a decreased risk of cancer incidence. Movahedi et al. (2015) determined that obesity is linked to an increased risk of colorectal cancer among individuals with Lynch syndrome who have a germline mutation in the \textit{MLH1} gene. Dashti et al. (2018) provide results suggesting an inverse relationship between physical activity and colorectal cancer risk in individuals with Lynch syndrome. In addition, studies have suggested the existence of a modifying effect from other low-penetrant genes on cancer risk among mutation carriers (Begg et al., 2008; Antoniou et al., 2002), while Wang et al. (2010) identified specific SNPs that modified breast
cancer risk in carriers of the BRCA1 and BRCA2 genes. Such studies indicate a residual genetic effect among mutation carriers that can cause cancer risk heterogeneity.

In this work, we further explore the risk heterogeneity on a family level. Using family-level data provides us with two design advantages: first, we typically observe more than a single individual carrying a specific pathogenic variant; second, we observe relatives who do not share that variant, but do share a significant proportion of genetic information, environmental exposures, and often behaviors. This can help us formulate hypotheses about the role of shared genetic, environmental, and behavioral factors. In addition, for sufficiently large families, or for collections of families sharing the same variant, it can enable us to more accurately assess risk.

To quantify the risk heterogeneity across families, we need to measure each family’s risk. We propose to do this by comparing the family’s number of observed cancer cases to the number of expected cases, calculated using an estimate of risk in the absence of heterogeneity. We do so separately for carriers and non-carriers. In Section 2.2, we introduce a novel approach to calculate the ratio between observed and expected cases, accounting for censoring and for the fact that the carrier status may not be known for some of the family members. Using this metric allows us to estimate if a family has higher or lower risk than expected. Visualization across a data set provides us with a measure of risk heterogeneity across families. Moreover, by obtaining separate ratios for carriers and non-carriers in the family, we can formulate hypotheses on the relative importance of alternative sources of heterogeneity.

In Section 2.3, we apply this approach to colorectal cancer family data from Creighton University’s Hereditary Cancer Center and provide results displaying the extent of risk heterogeneity among carriers and non-carriers of mutations of the MLH1, MSH2, and MSH6 genes. In Section 2.4, we suggest frailty models as a math-
ematical approach to modeling risk heterogeneity, and use synthetic data generated from several frailty models to explore the impact of various sources of heterogeneity on the O/E ratio. We also use the results from this synthetic data analysis to explore the influence of censoring and carrier probability prediction on our proposed metric.

2.2 Methods

2.2.1 Data

Creighton University’s Hereditary Cancer Center was established in 1984 to study hereditary cancer syndromes. The center has a registry consisting of families at high risk for various hereditary cancers, with rich family history information as well as genetic testing information for multiple family members. One unique feature of the data set is its large pedigrees. The proband initially fills out a questionnaire that asks for information for themselves as well as all first-, second-, and third-degree relatives. These relatives are then contacted for further information on themselves and their relatives, and the pedigree thus expands in this fashion. As a result, families in this data set are unusually large, making the data ideal for studying family-level risk heterogeneity across and within families. Further information on the hereditary cancer center can be found at Lynch et al. (2015).

We will focus on the subset of families with at least one member who carries a mutation of *MLH1* or *MSH2*. These genes are linked with Lynch syndrome, a hereditary cancer syndrome that is associated with various cancers, most commonly colorectal cancer. They are also considered to have similar penetrance when aggregating across all pathogenic variants (Dowty et al., 2013). There are 53 such families in total, with an average family size of 200.66 members. All families have an extensive family his-
Table 2.1: Table consisting of information for the subset of the Creighton University data of families with at least one member with a mutation of $MLH1$ or $MSH2$.

A history of cancers related to Lynch syndrome, in particular colorectal cancer. The families also contain extensive genetic testing information, as an average of 17.68 family members per family received genetic testing, and an average of 6.87 members were carriers of either $MLH1$ or $MSH2$. Additional details of this subset are provided in Table 2.1.

### 2.2.2 Quantification of Heterogeneity

Consider the hypothetical pedigree in Figure 2.1. This pedigree is large, containing second- and third-degree relatives as in the Creighton University data. Individuals in the pedigree whose left sides are shaded have developed colorectal cancer, and individuals whose right sides are shaded have developed endometrial cancer (two of the most common cancers associated with Lynch syndrome). Individuals colored red are $MLH1$ mutation carriers, while individuals colored red are $MLH1$ non-carriers and individuals colored black have unknown $MLH1$ carrier status. The colorectal cancer in this family seems to be hereditary, as it is prevalent and seemingly passed down in an inherited manner. However, it is unclear if the pedigree displays more cancer cases than
Figure 2.1: Hypothetical pedigree for a large family that is similar to the families in the Creighton University data. Individuals with the left sides shaded have developed colorectal cancer, and individuals with the right side shaded have developed endometrial cancer.

we would expect, given the carrier statuses of each family member. Among carriers, do we observe more than what we would expect in a population of carriers? Among non-carriers, do we observe more than what we would expect in a population of non-carriers? In order to answer these questions, we formally define a metric below.

Technically, we need to surmount two obstacles to implement our approach. First, we need to predict family members’ mutation carrier status for untested relatives. Second, we need to account for the fact that life histories of the living unaffected relatives only provided censored information on cancer.

Consider a family with $n$ members. Let $T_i$ be the (annually discretized) cancer age for the $i$-th individual, with possibilities $1, 2, \ldots, T_{\text{max}}, \infty$. Here $T_{\text{max}}$ is the largest
age we consider, and $T_i = \infty$ means that the individual did not develop the cancer in their lifetime. Let $C_i$ be the censoring age, with possibilities $1, \ldots, T_{\text{max}}$. Let $\delta_i = I(T_i \leq C_i)$ be the indicator of observing the cancer of interest. For each family member, we observe $X_i = \min(T_i, C_i)$. Let $Y_i = I(T_i \leq T_{\text{max}})$ be the indicator of developing the cancer in the individual’s lifetime, and let $G_i = (G_{i1}, \ldots, G_{id})$ be the genotype. $G_i$ is a vector, and $G_{ij}$ indicates the carrier status of a pathogenic mutation in the $j$-th gene of interest—1 if the individual carries a mutation in the gene and 0 otherwise. Lastly, let $H_i = (X_i, \delta_i)$ be the observed age and cancer outcome.

For the data application on the Creighton University data set, we focus on colorectal cancer, and $G_i$ is a vector of length 2 indicating the carrier statuses of mutations in $MLH1$ and $MSH2$.

The number of observed cases among carriers is $\sum_{i=1}^{n} Y_i I(G_i \neq 0)$, where $0 = (0, 0)$. However, we may not have genetic testing information for everyone in the family, and hence we may not know $G_i$ for everyone. To predict carrier status, we can use the peeling algorithm (Elston & Stewart, 1971), which efficiently and recursively estimates the marginal (with respect to the other family members) genotypic distribution for each family member, given the family history likelihood and the population mutation prevalences. For the Creighton University data set, we use family history information on colorectal and endometrial cancers and apply the peeling algorithm in MMRpro version 2.1-5, a Mendelian model that predicts risk of carrying mutations of $MLH1$, $MSH2$, and $MSH6$ (Chen et al., 2006). The peeling algorithm in MMRpro accounts for individuals with genetic testing results. To obtain the likelihood, we use the MMRpro age- and sex-specific penetrances, with ages ranging from 1 to $T_{\text{max}} = 94$. We also use the mutation prevalences from MMRpro. A further challenge is that $Y_i$ is unknown if the $i$-th family member is censored. In order to address this
issue, we predict the binary outcome $Y_i$ by estimating the probability of developing the cancer during the unobserved part of the person’s lifetime, given survival up to the censoring age. We can estimate this using a known age-specific penetrance for the cancer of interest. For the data application, again we use the MMRpro penetrances. Putting these together, we obtain our estimate of each family member’s contribution to the number of observed cases among carriers as

$$E_{Y_i,G_i}[Y_iI(G_i \neq 0)|H] = \sum_{g_i \neq 0} P(C_i < T_i \leq T_{max}|G_i = g_i)^{1-\delta_i} \frac{P(G_i = g_i)}{P(T_i > C_i|G_i = g_i)^{1-\delta_i}} P(T_i \leq T_{max}|G_i = g_i) P(G_i = g_i|H) \quad (2.1)$$

The full derivation is provided in the Supplementary Section S2.1. Here we sum over the set of all genotypes corresponding to mutation carriers, that is \( \{g_i \in \{0,1\}^d : g_i \neq 0\}\), where \( d \) is the number of genes. The left side of the summand in (2.1) is obtained using the cancer penetrance, and the right side is obtained using the peeling algorithm. Then to obtain the number of observed cases among carriers in the family, we simply add each family member’s estimated contribution.

We now obtain the number of expected cancer cases among carriers, defined as

$$\sum_{i=1}^{n} E[Y_iI(G_i \neq 0)|G_i].$$

This calculation is more straightforward, since we do not have to worry about censored cancer outcomes. However, we still have the issue of unknown genotypes. Thus we can estimate each family member’s contribution with

$$E_{G_i}[E[Y_iI(G_i \neq 0)|G_i]|H] = \sum_{g_i \neq 0} P(T_i \leq T_{max}|G_i = g_i) P(T_i > C_i|G_i = g_i)^{1-\delta_i} P(G_i = g_i|H). \quad (2.2)$$

The derivation is again provided in Supplementary Section S2.1. Now that we have both the observed and expected number of cancer cases among carriers in the family, we can evaluate their ratio \((O/E)\). In the example pedigree in Figure 2.1, the observed
number among carriers is 18.34 and the expected number is 13.01, giving us a ratio of 1.41. Since the ratio is greater than 1, we conclude that the family has higher risk among carriers than expected based on the aggregate penetrance for carriers. Perhaps the family has a mutation variant that produces higher cancer risk, or alternatively the family may contain other genetic mutations that interact negatively with the mutations of interest.

We are also interested in comparing the observed and expected number of cancer cases among non-carriers of the mutations of interest. These quantities are analogous to the ones for carriers, with $I(G_i \neq 0)$ being replaced with $I(G_i = 0)$. Similarly, when estimating each family member’s contribution, we only consider the case where $g_i = 0$ instead of summing over $\{g_i \neq 0\}$. Among non-carriers, the hypothetical pedigree in Figure 2.1 has 6.08 observed cases and 5.81 expected cases, or a ratio of 1.05. This ratio suggests that the non-carriers in the family develop cancer at a rate consistent with the general population. A possible implication is that the factors that contribute to the excess risk among carriers are not on average shared by non-carriers, pointing to the specific variant as the most likely source.

Before drawing conclusions, however, we need confidence intervals to assess the uncertainty of our estimates. We use a bootstrap procedure where we sample family members with replacement. If individual $i$ is sampled $k$ times, we multiply the individual’s contribution in the observed and expected quantities by $k$. We still use the complete family history information to calculate the carrier probabilities through peeling. In this manner we obtain 95% bootstrap percentile confidence intervals for the ratios for carriers and non-carriers. In the example family, the confidence interval for the ratio for carriers is $(1.13, 1.65)$, giving us confidence that the carriers in the family indeed have higher risk than expected. On the other hand, the confidence interval
for the ratio for non-carriers is (0.88, 1.30), giving us uncertainty about the level of risk among non-carriers compared to expected, but generally supporting the lack of a specific increase.

2.3 Results

We apply the above O/E metric to the data from Creighton University’s Hereditary Cancer Center. The plot of the ratios and their 95% bootstrap confidence intervals is shown in Figure 2.2. There is a wide range of O/E ratios among carriers and non-carriers, indicating a large amount of risk heterogeneity. Many families have an O/E ratio for carriers greater than 1 with an O/E for non-carriers close to 1. This “horizontal” positioning on the plot suggests a genetic component specific to the genes of interest to the heterogeneity, as only the carriers have risks different than expected in those families. Many of the families in this data set may carry mutation variants that confer a higher risk of colorectal cancer than average, or there may be gene-gene interactions that are increasing the risk only in carriers. Other families may carry mutation variants that confer a lower risk than average, or there may be gene-gene interactions that decrease the risk. In Figure 2.2, we visualize MLH1 and MSH2 mutation carriers with different colors and do not identify different patterns depending on the gene.

Some families fall around the “vertical” plane in the plot, having O/E ratios for non-carriers different than 1 but O/E ratios for carriers close to 1. This suggests the presence of unobserved genetic mutations that increase risk of colorectal cancer, and that may come from founders other than those carrying the MLH1 or MSH2 mutations found in this study. We categorize the non-carriers as not having mutations
of $MLH1$ or $MSH2$, but in reality they may be carriers of other cancer susceptibility gene mutations. In particular, the data do not contain genetic testing information for $MSH6$; thus, some non-carriers may actually carry mutations of $MSH6$ which increases colorectal cancer risk. There are also families where the O/E ratios for both carriers and non-carriers are greater than 1. These families on the may share environmental exposures or behavioral choices that increase risk. Note that the sources of heterogeneity may be complex; a combination of the above possible sources can impact a family’s colorectal cancer risk. It is also generally possible that protective and harmful factors can both be present.

It is important to not only focus on the ratios but also consider the confidence intervals. We see that many of the confidence intervals are wide and include 1. Thus even when the O/E ratios are greater than 1, the difference may not be statistically significant. The widths of these confidence intervals is strongly related to the family
Figure 2.3: Plot comparing the widths of the confidence intervals for carriers and non-carriers in the Creighton University data to the family size.

size, as shown in Figure 2.3. There is a negative correlation between family size and confidence interval width, both for carriers and non-carriers. Thus the families with large confidence intervals are the ones with smaller pedigrees. We see that our metric works best for large pedigrees that provide enough sample size to assuredly categorize the ratios as greater or less than 1. We also notice that the widths of the O/E confidence intervals for carriers are generally larger on average (0.53) compared to the O/E confidence intervals for non-carriers (0.36). This difference is also related to sample size, as there are more non-carriers than carriers on average in families with a large number of founders.
2.4 Impact of Sources of Heterogeneity

2.4.1 Frailty Models

In order to better understand the impact of the various sources of heterogeneity, we need a tool to statistically model the heterogeneity. We would like to identify a data-generating mechanism that can explain the results in Figure 2.2. A plausible candidate are frailty models, which are commonly used in survival analysis to account for unobserved heterogeneity (Vaupel et al., 1979). In such models, each individual has their own random effect, called a frailty, that has a multiplicative effect on their disease hazard function. Shared frailty models are an extension where groups of individuals share the same frailty. In the context of family-level cancer risk heterogeneity, our groups will be families, and the frailty allows us to adjust the level of cancer risk for the family. Families with large frailties will have higher risk than average, while families with small frailties will have lower risk than average. We define our model formally as follows.

Let \( W_k \) be the frailty for the \( k \)-th family, realized from a frailty distribution \( f_W(w) \). Let \( T_{ki} \) be the cancer age for the \( i \)-th individual in the \( k \)-th family (defined analogously to Section 2.2.2), and let \( U_{ki} \) be the sex (1 being male and 0 being female). Let \( G_{ki} \) be the genotype (vector of indicators for the mutations of interest). Then our frailty model is

\[
\lambda_{ki}(t|G_{ki}, U_{ki}, W_k) = 1 - \left[ 1 - \lambda_0 U_{ki} G_{ki}(t) \right] \exp(W_k).
\] (2.3)

Here \( \lambda_{ki}(t|G_{ki}, U_{ki}, W_k) \) is the hazard function for the (discrete) cancer age, and is defined as \( P(T_{ki} = t|T_{ki} \geq t, G_{ki}, U_{ki}, W_k) \). \( \lambda_0 U_{ki} G_{ki}(t) \) is a prespecified baseline.
hazard function obtained when the frailty is 0. This formulation of the frailty model is equivalent to the usual formulation using continuous times (Kalbfleisch & Prentice, 2011).

We now apply the frailty model to our setting focusing on colorectal cancer and the \textit{MLH1} and \textit{MSH2} genes. We will consider frailties in \{-2, -1.9, \ldots, 1.9, 2\}. For the baseline hazard function, we use the MMRpro hazard functions (derived directly from the MMRpro penetrances, which are age- and sex-specific). In order to model various sources of heterogeneity, we consider three types of frailty models: one where the frailty only affects carriers in the family, one where it only affects non-carriers, and one where it affects the entire family. The model in (2.3) is the third type. The first type has \(\exp(W_k I(G_{ki} \neq 0))\) instead of \(\exp(W_k)\), and the second type has \(\exp(W_k I(G_{ki} = 0))\). Thus the first type models carrier-only effects such as variant-level differences and gene-gene interactions; the second type models non-carrier-only effects such as unobserved cancer susceptibility gene mutations; and the third type models both carrier and non-carrier effects such as shared environmental and behavioral risk modifiers.

\textbf{2.4.2 Synthetic Data}

We are interested in exploring the impact of a frailty on its family’s O/E ratio. There is no explicit formula, since the O/E ratios are complex and depend on carrier probabilities. Thus we determine the frailty effect numerically by calculating O/E ratios on synthetic data generated from the frailty models described above.

Since \textit{MSH6} is another gene that is linked with Lynch syndrome and MMRpro provides penetrances for this gene, we include it in our simulations. The complete generating process is as follows. Each family has the same pedigree structure: a proband,
two parents, four grandparents, 2 paternal uncles, 2 paternal aunts, 2 maternal uncles, 2 maternal aunts, 2 brothers, 2 sisters, 2 sons, 2 daughters, and each sibling having 2 sons and daughters. The uncles, aunts, and siblings also have spouses. Thus each family has 44 members. For each family, we generate genotypes for the $MLH1$, $MSH2$, and $MSH6$ genes using allele frequencies of 0.1 and Mendelian laws of inheritance. We then generate the current ages using the MMRpro penetrances. Then, within each family, we generate 41 different family history information results corresponding to the 41 frailty values in consideration. For each frailty, we obtain the frailty-adjusted penetrance (derived from the frailty-adjusted hazard function using the frailty model) and generate colorectal and endometrial cancer ages. Then we obtain the O/E ratios for both carriers and non-carriers, using the baseline penetrance, as if we did not know the true frailty for the family. Similarly, we obtain the carrier probabilities and the cancer outcome imputation using the baseline penetrance. For the carrier probabilities in the simulations, we apply a modified version of the peeling algorithm called peeling-paring (Madsen et al., 2018). This method of obtaining the O/E ratios is the same across the different frailties, but the family history information changes.

Since we know the family members’ genotypes in this synthetic data, we can also obtain O/E ratios for the true carriers and non-carriers, along with the previous way using carrier probabilities from peeling-paring. In addition, we can also consider the family history if there had been no censoring (i.e., if everyone had lived to age $T_{max} = 94$). In this scenario, every family member has current age 94, so we can always observe if they developed colorectal cancer before our study horizon limit. Thus, for each family, we have 41 family histories corresponding to the 41 frailties, and for each frailty we have 4 types of O/E ratios (using true genotypes versus carrier probabilities, and having censoring versus not having censoring). We then repeat this process,
generating 1000 such families. For each frailty and O/E ratio, we average over the corresponding ratios for the 1000 families to obtain a 41 by 4 matrix of O/E ratios. Lastly, we repeat the entire process for the three different frailty models.

Plots for the 12 scenarios (4 O/E ratios, 3 frailty models) are shown in Figure 2.4. As expected, for the frailty model on both carriers and non-carriers, increases in the frailty \( W \) result in increases in the O/E ratios for both carriers and non-carriers. Through these results we can see that the relationship is not on the diagonal identity line and is non-linear for larger frailty values. In turn, for the frailty model on carriers only, increases in the frailty mainly cause increases in the O/E for carriers, and for the frailty model on non-carriers only, increases in the frailty mainly cause increases in the O/E for non-carriers. Thus the frailty models seem to be a good approach to modeling the different sources of risk heterogeneity. However, none of these relatively simple options capture in full the Creighton University data in Figure 2.2. We speculate that different subpopulation clusters, each with different models, may be needed to address the complexity of the heterogeneity displayed here. That said, the main horizontal cluster, accounting for a substantial fraction of families, is consistent with the “carriers-only” frailty hypothesis.

We can also use the synthetic data to analyze the impact of carrier probability estimation and censoring on the O/E ratios. Carrier probabilities are calculated using the baseline penetrance, regardless of the frailty in the data-generating mechanism. For high frailties, non-carriers are comparatively more likely to develop the cancer, and hence are more likely to have higher carrier probabilities when their carrier status is unknown. Their cancer outcomes seem unlikely for non-carriers according to the baseline penetrance, and thus they are given a higher probability of being a carrier. We can see this effect clearly when comparing plots I and K in Figure 2.4. In plot K,
Figure 2.4: Plots comparing the O/E for carriers and non-carriers in the frailty model synthetic data for the 12 scenarios. Plots A-D correspond to the frailty model on both carriers and non-carriers; plots E-H correspond to the frailty model on carriers only; and plots I-L correspond to the frailty model on non-carriers only. For each model, the first plot uses carrier probabilities and has censoring; the second plot uses carrier probabilities and has no censoring; the third plot uses true genotypes and has censoring; and the fourth plot uses true genotypes and has no censoring. The x-axis for each plot is the O/E for carriers, and the y-axis for each plot is the O/E for non-carriers.
we use the true genotypes, and thus the increase in frailty only impacts non-carriers as expected. On the other hand, in plot I we see that the carriers are also impacted. This is because we use carrier probabilities to account for unknown genotypes, and some non-carriers are being given large probabilities for genotypes corresponding to carriers. This also causes the impact of the frailty on non-carriers to be diminished, as seen by comparing plots A and C, E and G, and I and K.

We also can inspect the effect of censoring on the ratios. When we have censoring, we notice that the O/E ratios are much smaller compared to when we do not have censoring, as seen in the first and third columns of Figure 2.4. This is because our observed outcome is not truly binary when an individual’s outcome is censored. In this case, we impute the outcome, using the baseline penetrance, with the probability of developing the cancer in the rest of the lifetime, given survival up to the censoring age. Thus, for high frailties, individuals are more likely to have developed the cancer, but they may still be censored, since the current ages are unaffected by the frailty. These individuals who would have developed the cancer in their lifetime now have imputed outcomes less than 1, and thus we observe fewer cancer cases than expected. The imputed outcomes are further decreased since we impute using the baseline penetrance, which will impute fewer outcomes than if we imputed using the true frailty-adjusted penetrance. This effect is also stronger for non-carriers compared to carriers, simply because there are more non-carriers than carriers.

2.5 Discussion

Understanding cancer risk heterogeneity is important in genetic counseling. It could impact the prediction of carrying cancer susceptibility gene mutations, which af-
fects the decision to receive genetic testing, and it could impact decision making after the genetic testing results become available. Clinicians, genetic counselors, and their counselees would benefit from a better understanding of the extent and determinants of the risk heterogeneity. We developed a metric to do this by comparing the observed and expected number of cancer cases in a family, accounting for unknown genotypes and censoring. When applying this metric to colorectal cancer data from Creighton University’s Hereditary Cancer Center, we detected a substantial amount of risk heterogeneity among both carriers and non-carriers of \textit{MLH1} and \textit{MSH2}. Our visualization also strongly suggests that the predominant sources of heterogeneity are to be found in variation affecting the carriers only. Examples include variants with different penetrance and gene-gene interactions that only affect carriers.

Goldgar et al. (1994) proposed a similar approach of comparing the number of observed and expected cancer cases in family data. However, they did not consider individual families but aggregate information across all families. Moreover, they did not compare mutation carriers and non-carriers and only considered first-degree relatives. Thus they did not focus on understanding the risk heterogeneity across and within families, but on examining the overall familiality of cancer and providing a basis for hypotheses on genetic and environmental determinants of cancer. Our proposed approach builds upon the their results by seeking to quantify the familial risk heterogeneity and identify its sources.

To explore potential data-generating mechanisms that could explain the results in the data, we introduced three different frailty models. We then conducted a numerical study using synthetic data generated from these models to statistically describe the heterogeneity as well as investigate the impact of carrier probability estimation and censoring on the O/E ratios. We applied our methods to colorectal cancer and three
Lynch syndrome genes. Most other cancer syndromes are similarly affected by an insufficient understanding of heterogeneity. We describe our method in general terms, and believe it can be directly applied to other cancer syndrome equally well.

Our metric uses lifetime development of the cancer as the outcome of interest \( Y_i \), rather than development of cancer by the current age. Oftentimes family history data for individuals who developed cancer only contain cancer ages, without current ages, i.e., we observe \( X_i = T_i \) for individuals with \( \delta_i = 1 \). This is because information after the cancer age is often not needed in typical analyses. Suppose we used development of cancer by the current age as our outcome of interest, and consider an individual with genotype \( g_i \) who develops cancer at age \( t_i \) and whose unobserved current age is \( s_i \), where \( s_i > t_i \). The individual’s contribution to the expected number of cases would be \( P(T_i \leq s_i|G_i = g_i) \). If we replaced the current age \( s_i \) with the cancer age \( t_i \), we would be decreasing the expected value while maintaining the same observed value, thereby inflating the O/E ratios. On the other hand, using lifetime cancer development as the outcome avoids this issue for individuals who developed cancer, as their expected outcome no longer requires a current age.

Our approach may not be as accurate in the presence of substantial censoring and incomplete genetic testing information. As seen in section 2.4.2, the imputation process shrinks the O/E ratios closer to 1, especially when the family is at a higher risk than average (i.e., for high frailties in a frailty model). The synthetic data analysis also revealed a potential bias for high-risk families when many genotypes are unknown, as the carrier probability estimation, which uses the baseline penetrance, may be inaccurate. These biases could be reduced by using an approach that estimates a frailty for each family, such as the maximum likelihood approach proposed in Gorfine et al. (2013), and then uses frailty-adjusted penetrances for the cancer outcome and
genotype imputations. In addition, the confidence intervals for the O/E ratios in the data point to another limitation of our approach, as many of them are wide and include 1. This reveals a need for large pedigrees to produce confidence intervals that exclude 1. As the pedigrees in the data are already much larger than usually found in cancer family history data, such pedigree sizes may be unrealistic.

One aim of this work is to formulate hypotheses on the relative importance of potential sources of heterogeneity across and within families. However, the observed ratio is a convolution of protective and harmful sources of heterogeneity, which cannot be disentangled without additional information. In addition, our proposed metric only captures aggregate, family-level information, while there may be substantial individual-level differences. Information on variation in risk factors at the individual level could be incorporated in a more general version of our approach.

The results of this work could have useful implications for Mendelian risk prediction models. These models use the peeling algorithm, along with given penetrances and mutation prevalences, to calculate the probability of a proband’s genotype given the family history. Thus they rely on accurate penetrance functions. Our work shows that the assumption that the penetrances are only dependent on the gene harboring the mutation is unlikely to be accurate. If this is the case, Mendelian models may fail to account for important sources of risk heterogeneity. Gorfine et al. (2013) proposed a frailty model to account for this heterogeneity and applied the model to breast cancer prediction. Their model has a single frailty for each family; however, the synthetic data analysis in this work showed that this model may be overly simplistic. Instead, a more complex model having frailties acting on carriers and non-carriers separately may be necessary to acknowledge the various sources of heterogeneity. Using such a model to directly improve existing Mendelian models could provide significant im-
provements in predictive performance.

Overall, this work provides a tool to gauge cancer risk heterogeneity across and within families with genetic susceptibility to cancer. In particular, we saw through the data application that colorectal cancer risk fluctuates significantly across families, especially among carriers of mutations in MMR genes. This suggests a need to better understand the sources of heterogeneity and their impact on cancer risk, as our current bins for risk stratification are too coarse. In particular, our results illustrate the importance of estimating penetrance for specific pathogenic variants of cancer susceptibility gene mutations. An improved grasp of these variant-level differences, along with statistical models that can utilize these risk estimates, can help genetic counselors and counselees with risk comprehension, screening, and prevention.
3

Extending Models Via Gradient Boosting: An Application to Mendelian Models

3.1 Introduction

In a world in which more and more data are being collected, building prediction models has become an increasingly important task, and countless prediction models have been developed. With the expanding wealth of data, much emphasis has been made on training new prediction models; however, improving existing models, both by upgrading the model’s current mechanisms and by incorporating new features, is often more efficient. Existing models already embed the complex mechanistic relationships between variables in the model and some have been extensively trained using large amounts of data. Training a new model from scratch fails to take advantage of the valuable previous work. Furthermore, some existing models are trained using proprietary data. These data can be indispensable, to the point where building a new model that outperforms the existing model may be infeasible. In these settings, using an existing prediction model output as a starting point when training a new model can leverage the existing models. Using existing models can also be advantageous from an implementation perspective. For example, clinicians rely on prediction models to better understand patients’ risks of developing diseases and to accordingly make informed decisions. A new model may have difficulty being adopted for clinical use,
while clinics may quickly and readily adopt a modification to one of its existing models.

However, although directly improving prediction models is crucial, in practice it may not be straightforward. Incorporating new data or new features to existing models often requires the original data used to train the model, and obtaining these data can be impractical. In addition, some models may be complex and depend on prior scientific evidence. Consequently, incorporating new features in the same manner that the current features are utilized in the existing model may be challenging due to limitations in scientific knowledge. Lastly, the modeling mechanisms of some models are proprietary, and hence improving these models by modifying their structure can be infeasible.

Modeling algorithms that take an existing model’s predictions and directly improve them using additional data can prove valuable. Gradient boosting (Friedman et al., 2000), a popular ensemble prediction algorithm for combining weak prediction models to produce a more accurate model, represents one convenient approach to do so. The algorithm identifies the model’s weaknesses and learns how additional data can be leveraged to overcome these weaknesses. It performs this process in an iterative manner, incrementally improving the model’s predictive accuracy. The algorithm is typically used to build prediction models from scratch but not as often applied to improve existing models.

One application where gradient boosting could be effective is in predicting genetic predisposition to cancer. Some cancers are caused by inherited germline mutations, and thus a family history of cancer can inform an individual of their risk of having a mutation in a cancer susceptibility gene. Consequently, genetic counselors play an increasingly important role in helping individuals with a family history of certain can-
cers better understand their risk of having a genetic predisposition to these cancers and formulate a plan for genetic testing, screening and prevention. Statistical models that can accurately and expeditiously predict this risk are thus critical tools for these counselors to advise and guide counselees.

Numerous risk prediction models have been developed to predict genetic predisposition to cancer. Some of these models are empirical and use training data on genetically tested individuals to directly estimate the probability of carrying a mutation given the cancer history, usually through regression models (Couch et al., 1997; Vahteristo et al., 2001; Barnetson et al., 2006). Other models are Mendelian, and use the age-specific probability of developing the cancer given the genotype, called penetrance; the population-level distribution of the genotypes, called prevalence; and Mendelian laws of inheritance to estimate the carrier probability (Murphy & Mutalik, 1969). Examples of Mendelian models include BOADICEA (Antoniou et al., 2008) and MMRpro (Chen et al., 2006).

Mendelian models are complex and embed the pedigree structure to acknowledge the inherited pattern of the cancer susceptibility mutations. As a result, directly improving the model by updating the impact of existing features and accounting for additional features that may help in the prediction can be challenging. Gradient boosting provides a means for empirically incorporating the information to directly improve upon the existing Mendelian model predictions. One can initialize the gradient boosting algorithm with the Mendelian model predictions and then train the new model to learn how the information from the features can correct the shortcomings of the original model. In this work, we explore the potential symbiotic relationship between gradient boosting and Mendelian prediction—how gradient boosting can improve Mendelian prediction, and how Mendelian prediction can improve gradient boosting.
The models are introduced in Sections 3.2 and 3.3. In Section 3.4, we conduct a simulation study to compare the effectiveness of gradient boosting and Mendelian models in incorporating cancer family history information. In Section 3.5, we apply this approach to data from the USC-Stanford Cancer Genetics Hereditary Cancer Panel (HCP) Testing study (Idos et al., 2018). Finally, we conclude with a discussion in Section 3.6.

3.2 Gradient Boosting

Gradient boosting is an iterative ensemble method that combines boosting, which is a machine learning technique that sequentially adds weak learners to create a strong learner, and gradient descent, which is an iterative optimization procedure that takes steps proportional to the negative gradient to find local minima. Consider training data \((z_i, y_i)\) for \(i = 1, \ldots, N\), where \(y_i\) is the outcome and \(z_i\) is a feature or vector of features. Our goal is to obtain a prediction \(P(z_i)\) of \(y_i\). We first initialize the prediction as \(P_0(z_i)\). Then at the \(m\)-th iteration, the algorithm calculates the residuals

\[
r_{im} = -\left[ \frac{\partial L(y_i, P(z_i))}{\partial P(z_i)} \right]_{P(z_i)=P_{m-1}(z_i)}
\]

between the outcome \(y_i\) and the prediction \(P(z_i)\), where \(P_{m-1}(z_i)\) is the prediction at the \((m - 1)\)-th iteration, based on a specified loss function \(L\). We then fit a base learner \(h_m(z_i)\) (often a decision tree) using \((z_i, r_{im})\), \(i = 1, \ldots, N\), learning the residuals through the features. Lastly, the learner is added to the current prediction: \(P_m(z_i) = P_{m-1}(z_i) + \gamma_m h_m(z_i)\), where

\[
\gamma_m = \arg \min_{\gamma} \sum_{i=1}^{N} L(y_i, P_{m-1}(z_i) + \gamma h(z_i)).
\]
This process is repeated, and after $M$ iterations, we obtain the final prediction

$$P_M(z_i) = P_0(z_i) + \sum_{m=1}^{M} \gamma_m h_m(z_i).$$

This prediction model can then be applied to testing data. Overall, gradient boosting identifies the weaknesses of the current prediction model by using the residuals, and uses features in the data to predict these residuals and hence overcome the initial model weaknesses.

The choice of the loss function $L$ depends on the type of outcome variable $y$. For continuous outcomes, common choices of $L$ are squared-error loss, absolute loss, and Huber loss; for binary outcomes, common choices are logistic loss and Adaboost loss. Natekin & Knoll (2013) provide an overview of commonly-used loss functions in gradient boosting.

There are several options for fitting the base learners to the residuals. Linear regression is a natural choice; however, a sum of linear regression models is still linear, and thus the boosting process of sequentially adding linear models would not improve our predictions over one linear regression step. An alternative choice is decision trees, which partition the feature space into rectangles by making linear splits. These are more commonly used in gradient boosting because they are more robust to overfitting, as their depth (i.e., the number of splits) can be controlled. Decision trees more naturally fit the goal of gradient boosting of combining weak learners to create a stronger one. Natekin & Knoll (2013) provide a complete discussion of the various options for the base learners.

In addition to the maximum decision tree depth as well as the number of iterations $M$, gradient boosting can also be customized through other tuning parameters. We can introduce a shrinkage parameter $\nu$ such that each $\gamma_m$ is changed to $\nu \gamma_m$, thereby decreasing the learning rate in each iteration. We can also sample a fraction $f$, called the bag fraction, of the training set at each iteration in the algorithm. Both of these
tuning parameters are regularization methods aimed at reducing overfitting.

Without a prior prediction, gradient boosting commonly initializes with $P_0(z_i) = \arg \min_\gamma \sum_{i=1}^N L(y_i, \gamma)$. In the case of continuous outcomes with squared-error loss, this is the sample mean $\bar{y}$; in the case of binary outcomes with logistic loss, this is the sample log odds $\log(\bar{y}/(1 - \bar{y}))$. Alternatively, gradient boosting can be used to improve existing risk models by initializing with their predictions.

### 3.3 Mendelian Risk Prediction Models

Mendelian risk prediction models are used to predict an individual’s risk of having a mutation in a cancer susceptibility gene. Unlike empirical prediction models, they incorporate family history information by acknowledging the Mendelian inheritance pattern of genetic mutations. In this section, we explain how Mendelian models predict risk and how existing Mendelian models can be used in combination with gradient boosting.

#### 3.3.1 Notation

Consider a proband with $n$ family members. Let $G_i$ denote the genotype for the $i$-th family member, where $i = 1, \ldots, n$, and $i = 1$ denotes the proband, and $G = (G_1, \ldots, G_n)$. $G_i$ can be a vector representing multiple genes, where each component of the vector is a binary indicator of carrying the corresponding mutation. Let $T_{ri}$ denote the time that the $i$-th individual develops the $r$-th cancer, $r = 1, \ldots, R$, where we consider discrete ages $1, 2, \ldots, T_{max}$. Let $C_i$ denote the censoring time (either the time of death or the current age) for the $i$-th individual, and let $X_{ri} = \min(T_{ri}, C_i)$ be the observed time for the $i$-th individual for the $r$-th disease. Let $X_i = (X_{1i}, \ldots, X_{Ri})$.
and $X = (X_1, \ldots, X_n)$. Let $\delta_{ri} = I(T_{ri} \leq C_i)$, $\delta_i = (\delta_{1i}, \ldots, \delta_{Ri})$, and $\delta = (\delta_1, \ldots, \delta_n)$. Let $H_i = (X_i, \delta_i)$ be the observed age and cancer status for the $i$-th individual, and let $H = (H_1, \ldots, H_n)$ be the complete family history.

### 3.3.2 Mendelian Carrier Probability Estimation

Mendelian risk prediction models estimate the proband’s genotype probability conditional on their family history by performing the following Bayes’ rule calculation (Chen et al., 2004):

$$ P(G_1 | H) = \frac{P(H | G_1) P(G_1)}{\sum_{G_1} P(H | G_1) P(G_1)} = \frac{\sum_{G_2, \ldots, G_n} P(H | G_2, \ldots, G_n | G_1) P(G_2, \ldots, G_n | G_1) P(G_1)}{\sum_{G_1} \sum_{G_2, \ldots, G_n} P(H | G_2, \ldots, G_n | G_1) P(G_2, \ldots, G_n | G_1) P(G_1)} = \frac{P(G_1) \sum_{G_2, \ldots, G_n} P(G_2, \ldots, G_n | G_1) \prod_{i=1}^n P(H_i | G_i)}{\sum_{G_1} P(G_1) \sum_{G_2, \ldots, G_n} P(G_2, \ldots, G_n | G_1) \prod_{i=1}^n P(H_i | G_i)}. \tag{3.1} $$

These models are called Mendelian models because they use Mendelian laws of inheritance to estimate the relatives’ genotype distribution conditional on the proband’s genotype: $P(G_2, \ldots, G_n | G_1)$. $P(G_1)$ is the proband’s marginal (with respect to the family members) probability of carrying the genotype $G_1$ and is called prevalence, and $P(H_i | G_i)$ is the age-specific probability of observing the cancer status and observed age given the genotype, which is obtained using the penetrance. Note that we make the fundamental assumption of conditional independence of the family members’ family histories given the genotypes, which may not be true due to risk heterogeneity from genetic, environmental, or behavioral factors. Both the prevalence and penetrance are usually derived from studies in the literature. Since the summation over all genotypic configurations in the family can be computationally intensive,
Mendelian models often use the Elston-Stewart peeling algorithm (Elston & Stewart, 1971) to efficiently estimate the proband’s genotype distribution.

Mendelian models can incorporate information from any cancer and gene, as long as there are reliable corresponding penetrance estimates. In practice, for a given set of cancer susceptibility genes, there may only be one or two cancers that have reliable penetrance estimates ($R = 1$ or 2). However, Mendelian models are flexible, and a user can easily include information from additional cancers by inputting their corresponding penetrances.

### 3.3.3 Gradient Boosting with Mendelian Models

We propose to apply gradient boosting to improve existing Mendelian models and incorporate new features. We define the outcome as the binary indicator of the proband carrying at least one of the genetic mutations, $y = I(G_1 \neq 0)$, and $z_i$ are features that are associated with carrying the mutations (the outcome). These features can include ones already incorporated in the existing Mendelian model as well as new ones. For example, $z_i$ can contain family history information both for cancers already considered in the existing Mendelian model as well as additional cancers. It can summarize these features through a metric such as the proportions of family members who developed each cancer. Besides cancer information, $z_i$ can also include risk modifiers such as medications that impact risk.

We propose to initialize the gradient boosting algorithm with the existing Mendelian model predictions $P(G_1|H)$ from equation 3.1. Here $H$ is the family history information for only the cancers used in the Mendelian model. Note that the initial prediction in gradient boosting is typically a function of the features $z_i$ (as seen in Section 3.2); however, in this context where we use gradient boosting to improve an
existing model, the initial prediction is the Mendelian prediction, and hence a function of the input of the Mendelian model \( H \). For the base learner \( h_m \), we propose to use decision trees with \( z_i \) as the features.

Since our outcome is binary, we propose to use the logistic loss function \( L(y, \theta) = \log(1+\exp(\theta)) - y\theta \), where the prediction \( \theta \) is the log odds of being a mutation carrier. The logistic loss function is the negative of the likelihood for the Bernoulli distribution, and the residual is \( -\frac{\partial}{\partial \theta} L(y, \theta) = y - \frac{1}{1+\exp(-\theta)} \).

### 3.4 Simulation Study

#### 3.4.1 Gradient Boosting Approach to Incorporating Family History Information

In order to assess the performance of our proposed approach, we conduct a simulation study, with a focus on exploring various methods of incorporating cancer family history information to an existing Mendelian prediction model. Mendelian models rely on penetrances for cancers related to the genetic mutations of interest. These penetrances are often estimated from studies in the literature (Chen & Parmigiani, 2007; Marroni et al., 2004), and this process can be challenging due to low mutation prevalences, difficulty in finding data cohorts with genetically tested individuals, bias in cohort ascertainment, and heterogeneity in study populations. Thus Mendelian models may only incorporate information for cancers with accurate and precise penetrance estimates, as using inaccurate penetrances can lead to poor model performance. This restriction also causes existing Mendelian models to omit potentially informative family history data for other cancers that are known to be associated with a genetic mu-
tation but for which there is not enough evidence in the literature to estimate the penetrance. Using inaccurate penetrances to add additional cancers to the model could fail to make improvements in the risk prediction or even impair the predictive performance.

In such settings where the cancer penetrances are inaccurate, gradient boosting can be an effective alternative approach to incorporating family history information. Gradient boosting does not need to rely on penetrance estimates but instead can integrate the information in an empirical manner. It can thus overcome two main limitations of existing Mendelian models: (1) inaccurate penetrances of the cancers incorporated in the model, and (2) difficulty in including information from additional cancers. Using the notation in Section 3.3.3, we let the features $z_i$ represent the family history information from all the cancers of interest. These cancers can include both the ones used in the existing Mendelian model as well as others for which we may have family history information but no accurate penetrance estimates. Suppose the existing Mendelian model uses information from $R$ cancers, and we are also interested in incorporating information from an additional $R'$ cancers (new features). One approach to incorporating the entire cancer family history information is by defining $z_i = (z_{i1}, \ldots, z_{iR}, z_{i(R+1)}, \ldots, z_{i(R+R')})$, where $z_{ir}$ is the proportion of family members who develop the $r$-th cancer (for sex-specific cancers, we only consider family members of that sex), $r = 1, \ldots, R + R'$. Here $z_{i1}, \ldots, z_{iR}$ represents cancers in the existing Mendelian model and $z_{i(R+1)}, \ldots, z_{i(R+R')}$ represent the additional cancers (new features). Note that in addition to these proportions, other approaches to incorporating the family history information into the features $z_i$ can also be used.

We use simulations to compare the gradient boosting and Mendelian approaches to incorporating cancer family history information. We also analyze the combination
of gradient boosting and Mendelian models, where the gradient boosting algorithm is initialized with the Mendelian predictions, and explore settings in which the added gradient boosting component can overcome deficiencies in the existing Mendelian model. In addition, we compare the effectiveness of the gradient boosting approach in adding new cancers with the Mendelian approach of adding these cancers through penetrance estimates.

We consider two scenarios for the penetrances of the cancers used in the Mendelian model. The first uses information from high-penetrant cancers to mimic situations where the cancer information is highly predictive of the mutation carrier status. The second uses low-penetrant cancers in order to understand the added value of information from the additional cancers. The relative infrequency of these cancers results in a low amount of signal for prediction, leaving potential for improved performance using additional cancers. For both scenarios, to further reduce the predictive ability of the existing Mendelian model, we use misspecified (different from data-generating) penetrances for the existing cancers, obtained by taking the square root of the survival functions and converting back to penetrances. This allows us to precisely compare the gradient boosting approach to incorporating family history information to the Mendelian approach when the penetrances are misspecified.

Specifically, we evaluate the proposed approach using MMRpro, an existing Mendelian model which predicts mutation carrier status of the \( MLH1 \), \( MSH2 \), and \( MSH6 \) genes. These mutations are related to DNA mismatch repair (MMR) and have been shown to be linked with Lynch syndrome (Papadopoulos et al., 1994; Fishel et al., 1993; Miyaki et al., 1997), a hereditary cancer syndrome that predisposes individuals to increased risks of many cancers, most commonly colorectal (CRC) and endometrial (EC), but also gastric (GC), ovarian, small intestine, and others (Lynch &
Smyrk, 1996). MMRpro uses family history information on colorectal and endometrial cancers ($R = 2$), as these cancers have been extensively studied and hence have reliable penetrance estimates derived from a meta-analysis (Chen et al., 2006). However, MMRpro ignores family history information on other Lynch syndrome cancers whose penetrance estimates are not as reliable. Gastric cancer, for example, has penetrance estimates (Braun et al., 2018; Dowty et al., 2013; Barrow et al., 2009) but has not been studied as extensively as colorectal and endometrial cancers. In this work, we incorporate family history information on gastric cancer ($R' = 1$) to the existing MMRpro predictions (using version 2.1-5 of the BayesMendel R package), both through the gradient boosting and completely Mendelian approaches.

### 3.4.2 Generating Families

We generate 10,000 families. Each family consists of 3-4 generations, where each proband has at least a mother, father, and maternal and paternal grandparents. We then sample the number of sisters, brothers, maternal and paternal aunts and uncles, daughters, sons, and each sibling’s number of daughters and sons randomly from \{0, 1, 2, 3\}. Thus, family sizes range from 7 to 67 to mimic the variability of family sizes in real data.

We first generate genotypes for the family members. For these simulations, we use inflated allele frequencies of 0.01 for each of the three MMR mutations in order to increase the number of carriers. After having generated genotypes for the founders of the pedigree, we generate the genotypes for the rest of the family using the Mendelian laws of inheritance.

Using these genotypes, we then generate cancer statuses and ages for every individual in the family for colorectal, endometrial, and gastric cancers. For colorectal
and endometrial cancers, we generate data using two scenarios. To simulate high-penetrant cancers, we use the MMRpro penetrances, and to simulate low-penetrant cancers, we use scaled, low-penetrant versions of the MMRpro penetrances, where the lifetime risk for each of the three genotypes corresponding to carrying exactly one MMR mutation is 0.2. For gastric cancer, we use (sex-specific) penetrance estimates from the clinical decision support tool ASK2ME (Braun et al., 2018; Dowty et al., 2013; Barrow et al., 2009) for carriers. For non-carriers, we use SEER penetrances. We scale the gastric penetrances so that the lifetime risk of non-carriers is 0.05 and the lifetime risk for each of the three genotypes corresponding to carrying exactly one MMR mutation is 0.5. We use this scaled version of the penetrance in order to increase the association between carrying an MMR mutation and developing gastric cancer. This helps us clearly evaluate the ability of gradient boosting to leverage information from the additional gastric cancer. After obtaining the penetrances for non-carriers and carriers of exactly one of the MMR mutations, we obtain penetrances for carriers of multiple mutations by multiplying the survival functions of the individual mutations, thus calculating penetrances for all $3^3 = 27$ genotypes.

Using these penetrances, which are defined from ages 1 to 94, we generate cancer ages for the family members (95 if they do not develop the cancer). We also generate current ages for all individuals. The proband’s grandmothers (who are founders in our pedigrees) have their current ages generated from a normal distribution with mean 100 and variance 4, and current ages of their spouses (as well as ages of other spouses in the pedigree) are generated from a normal distribution with their current age as the mean and variance 4. All children are generated iteratively using a normal distribution with the mother’s current age minus 30 as the mean and variance 25 (or simply assigned to be the mother’s current age minus 15, if this is less than the gener-
ated current age).

### 3.4.3 Simulation Setup

We generate family data using the above approach and then randomly select half of the data to be used for training and the other half to be used for testing. We do this for two data-generating scenarios—one with high-penetrant colorectal and endometrial cancers, and one with low-penetrant colorectal and endometrial cancers.

**Model Training.** We train the following models using the training data: (1) gradient boosting initialized with the Mendelian predictions (without gastric cancer information), using information from the three cancers as features, with 50 iterations; (2) gradient boosting without the Mendelian predictions (initialized with the sample log odds), using information from the three cancers as features, with 50 iterations; (3) gradient boosting initialized with the Mendelian predictions, using information from colorectal and endometrial cancers as features, with 50 iterations; (4) gradient boosting without the Mendelian predictions (initialized with the sample log odds), using information from colorectal and endometrial cancers as features, with 50 iterations; (5-6) gradient boosting initialized with Mendelian predictions, using information from the three cancers as features, with both 25 and 100 iterations; and (7-8) gradient boosting without Mendelian predictions (initialized with the sample log odds), using information from the three cancers as features, with both 25 and 100 iterations.

**Model Testing.** On each of the families in the testing set we estimate predictions based on models 1-6 trained above. In addition, we estimate predictions based on the following models: (9) the Mendelian model without gastric cancer information; (10) the Mendelian model with gastric cancer information incorporated in a Mendelian manner through the data-generating gastric cancer penetrance; (11-14) MMRpro with
gastric cancer information incorporated in a Mendelian manner through misspecified gastric cancer penetrances, obtained from the gastric cancer survival function being raised to the powers of 0.25, 0.5, 2, and 4; (15) the “oracle” Mendelian model, incorporating colorectal and endometrial cancer information through the data-generating penetrances, without gastric cancer; and (16) the “oracle” Mendelian model, incorporating colorectal, endometrial, and gastric cancer information through the data-generating penetrances. In models (9-14), the Mendelian model is run using misspecified colorectal and endometrial cancer penetrances, obtained by taking the square root of the survival functions. Predictions from model (9) are used to initialize the gradient boosting algorithms in models (1), (3), (5), and (6). In all cases, the Mendelian model is run using the data-generating allele frequencies. The 16 models allow us to evaluate the effectiveness of combining gradient boosting with Mendelian models, while exploring the impact of changing the number of iterations as well as determining the settings in which gradient boosting provides a modeling advantage over a completely Mendelian approach.

In order to obtain misspecified gastric cancer penetrances, we raise the gastric cancer survival functions (which are age- and sex-specific) to various powers. This has the effect of increasing (for powers greater than 1) and decreasing (for powers less than 1) the penetrance (a plot comparing the misspecified penetrances is provided in Supplementary Figure S3.1). To run gradient boosting, we use XGBoost (Chen & Guestrin, 2016), a scalable and popular implementation of gradient boosting, through the xgboost R package. For the base learners, we use decision trees with three features: the proportion of family members with colorectal cancer, the proportion of female family members with endometrial cancer, and the proportion of family members with gastric cancer. We limit each tree to a maximum depth of 2, use a shrinkage
parameter of $\nu = 0.1$, and a bagging fraction of $f = 0.5$. These choices of tuning parameters are commonly-used and based on suggestions in Friedman (2001), Friedman (2002), and Friedman et al. (2009). In order to account for the variability in the selection of the training set, we use 100 bootstrap replicates for the entire algorithm, where in each replicate we randomly select half of the data to be the training data. Lastly, in order to assess the uncertainty of our performance measures, we generate 100 data sets to obtain percentile confidence intervals for each performance measure.

### 3.4.4 Simulation Results

In order to evaluate our model performance (on the testing data), we consider three metrics (Steyerberg et al., 2010): calibration as measured by the ratio of the number of observed to expected events (O/E), discrimination as measured by the area under the ROC curve (AUC), and accuracy as measured by the root Brier score (rBS), defined as $\sqrt{\sum_i(y_i - P_M(z_i))^2}$.

Results for the simulation for 8 of the models (under two data-generating scenarios for the colorectal and endometrial cancers) are shown in Table 3.1, where we provide average performance measures across the 100 simulated data sets, along with 95% percentile confidence intervals. Since we used misspecified colorectal and endometrial cancer penetrances, we see that the resulting Mendelian model (without gastric cancer information) has relatively poor performance. Gradient boosting, without the Mendelian prediction and without gastric cancer information and using 50 iterations, has a better O/E but worse AUC and similar rBS. The combination of gradient boosting and the Mendelian model, using only colorectal and endometrial cancers as features, combines the strengths of the two approaches by having the best O/E while matching the AUC and rBS of the Mendelian model.
<table>
<thead>
<tr>
<th></th>
<th>O/E</th>
<th>AUC</th>
<th>rBS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Low-penetrant data-generating CRC and EC penetrances</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No gastric cancer information</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mendelian(^a)</td>
<td>0.856 (0.797, 0.911)</td>
<td>0.784 (0.764, 0.802)</td>
<td>0.218 (0.210, 0.224)</td>
</tr>
<tr>
<td>GB(^b)</td>
<td>0.961 (0.943, 0.976)</td>
<td>0.730 (0.709, 0.749)</td>
<td>0.228 (0.220, 0.234)</td>
</tr>
<tr>
<td>GB with Mendelian(^c)</td>
<td>1.003 (0.982, 1.017)</td>
<td>0.780 (0.759, 0.798)</td>
<td>0.219 (0.211, 0.224)</td>
</tr>
<tr>
<td><strong>Incorporating gastric cancer information</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mendelian(^d)</td>
<td>0.872 (0.813, 0.927)</td>
<td>0.845 (0.825, 0.859)</td>
<td>0.212 (0.206, 0.218)</td>
</tr>
<tr>
<td>GB(^e)</td>
<td>0.962 (0.945, 0.977)</td>
<td>0.794 (0.771, 0.812)</td>
<td>0.224 (0.217, 0.230)</td>
</tr>
<tr>
<td>GB with Mendelian(^f)</td>
<td>1.002 (0.986, 1.017)</td>
<td>0.833 (0.813, 0.848)</td>
<td>0.216 (0.208, 0.221)</td>
</tr>
<tr>
<td><strong>Oracle Mendelian models</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mendelian(^g)</td>
<td>1.002 (0.937, 1.058)</td>
<td>0.785 (0.763, 0.804)</td>
<td>0.217 (0.209, 0.224)</td>
</tr>
<tr>
<td>Mendelian+GC(^h)</td>
<td>1.001 (0.935, 1.061)</td>
<td>0.848 (0.827, 0.862)</td>
<td>0.211 (0.205, 0.217)</td>
</tr>
<tr>
<td><strong>High-penetrant data-generating CRC and EC penetrances</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No gastric cancer information</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mendelian(^a)</td>
<td>0.735 (0.671, 0.787)</td>
<td>0.914 (0.900, 0.924)</td>
<td>0.196 (0.189, 0.203)</td>
</tr>
<tr>
<td>GB(^b)</td>
<td>0.958 (0.943, 0.971)</td>
<td>0.867 (0.851, 0.878)</td>
<td>0.215 (0.207, 0.222)</td>
</tr>
<tr>
<td>GB with Mendelian(^c)</td>
<td>1.000 (0.987, 1.013)</td>
<td>0.912 (0.897, 0.922)</td>
<td>0.196 (0.189, 0.203)</td>
</tr>
<tr>
<td><strong>Incorporating gastric cancer information</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mendelian(^d)</td>
<td>0.765 (0.702, 0.817)</td>
<td>0.928 (0.917, 0.938)</td>
<td>0.193 (0.187, 0.198)</td>
</tr>
<tr>
<td>GB(^e)</td>
<td>0.958 (0.943, 0.971)</td>
<td>0.880 (0.868, 0.888)</td>
<td>0.214 (0.206, 0.220)</td>
</tr>
<tr>
<td>GB with Mendelian(^f)</td>
<td>1.001 (0.987, 1.015)</td>
<td>0.920 (0.908, 0.929)</td>
<td>0.195 (0.189, 0.202)</td>
</tr>
<tr>
<td><strong>Oracle Mendelian models</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mendelian(^g)</td>
<td>0.995 (0.906, 1.065)</td>
<td>0.915 (0.900, 0.925)</td>
<td>0.194 (0.187, 0.201)</td>
</tr>
<tr>
<td>Mendelian+GC(^h)</td>
<td>0.996 (0.911, 1.063)</td>
<td>0.929 (0.919, 0.939)</td>
<td>0.191 (0.184, 0.196)</td>
</tr>
</tbody>
</table>

\(^a\) Mendelian model with misspecified CRC and EC penetrances, without GC information  
\(^b\) GB initialized without Mendelian predictions, using CRC and EC information  
\(^c\) GB initialized with Mendelian predictions (without GC), using CRC and EC information  
\(^d\) Mendelian model with misspecified CRC and EC penetrances, and incorporating GC information through the data-generating penetrance  
\(^e\) GB without Mendelian predictions, using all 3 cancers  
\(^f\) GB initialized with Mendelian predictions (without GC), using all 3 cancers  
\(^g\) Oracle Mendelian model with data-generating CRC and EC penetrances, without GC information  
\(^h\) Oracle Mendelian model with data-generating CRC and EC penetrances, and incorporating GC information through the data-generating penetrance

**Table 3.1:** Performance measures for the families in the simulated data. The entries are the mean performance measures across the 100 simulated data sets (each data set has 10,000 families), with 95% percentile confidence intervals in the parentheses. The Mendelian model used is MMRpro, with misspecified CRC and EC penetrances. The top represents data generated from low-penetrant (scaled MMRpro) CRC and EC penetrances, and the bottom represents high-penetrant (unscaled MMRpro) CRC and EC penetrances. The oracle Mendelian model uses the data-generating CRC and EC penetrances. All gradient boosting models are run with 50 iterations.
We can also assess the ability of the two approaches to incorporating gastric cancer information. The completely Mendelian method of using a gastric cancer penetrance (where we use the true data-generating gastric cancer penetrance) shows improvement in AUC over the Mendelian model without gastric cancer information (significant for the low-penetrance scenario, insignificant for the high-penetrant scenario), without much change in O/E and rBS. Again, the gradient boosting approach, without the Mendelian prediction and using 50 iterations, had a better O/E but worse AUC and similar rBS. The gradient boosting and Mendelian model combination again produced the best of the three models with gastric cancer information, with an improved O/E and similar AUC and rBS values to the completely Mendelian model that incorporates gastric cancer information.

These results show that penetrances which are misspecified in a similar manner to our method (taking the square root of the survival function, then converting back to penetrance) can affect the calibration of Mendelian models. However, the discrimination of the models is fairly robust to this type of misspecification. Gradient boosting is strong in calibration but is comparatively lacking in discrimination. Thus the combination of gradient boosting and Mendelian models provides the best of both worlds, re-calibrating the Mendelian model with misspecified penetrances while utilizing the discriminatory power of these Mendelian predictions.

We can compare our models to the “oracle” Mendelian models which use the true data-generating penetrances. The oracle Mendelian models (one which uses all three cancers and one which only uses colorectal and endometrial cancers) are both well-calibrated, while the model that uses gastric cancer information naturally has better discrimination. These oracle models do not actually provide improvements in AUC and rBS compared to their corresponding models with misspecified colorectal and
endometrial cancer penetrances. Instead, the main improvements are seen in the calibration. Thus we note that our proposed gradient boosting approach, initializing with the Mendelian predictions, performs similarly to the oracle Mendelian models in all three performance measures.

We can also compare the results from the two data-generating scenarios. Overall, the Mendelian model performs worse in calibration for the high-penetrant scenario, as the difference between the correctly and misspecified penetrances is larger in this situation. However, the increase in penetrance allows the Mendelian model to have better discrimination and accuracy. Similarly, gradient boosting performs better in discrimination and accuracy in the high-penetrant scenario, without much change in calibration (as gradient boosting is unaffected by the misspecified penetrances). Thus, as expected, high-penetrant cancers provide more information for predicting the mutation carrier status, though using misspecified penetrances can affect the model calibration. Overall, the comparisons between the Mendelian and gradient boosting approaches are similar for both scenarios.

In Supplementary Table S3.1, we also explore the impact of the number of iterations on the gradient boosting performance. For the gradient boosting and Mendelian model combination, the three choices of the number of iterations (25, 50, and 100) led to similar performance measures, demonstrating a general robustness to the number of iterations. However, gradient boosting without the Mendelian predictions had differences in the O/E ratios, with the O/E increasing with an increase in the number of iterations. Without a base prediction with which to initialize, gradient boosting struggled to provide well-calibrated results without enough iterations. The AUC and rBS values did not change based on the number of iterations. For all three choices of the number of iterations, initializing with the Mendelian predictions provided better
discrimination, exhibiting the importance of the Mendelian component.

When incorporating gastric cancer through the completely Mendelian approach, we used the data-generating gastric cancer penetrance. However, in practice the gastric cancer penetrance used in the Mendelian model may be misspecified. We explore the impact of this misspecification in Supplementary Table S3.2. The results show that the O/E ratios tend to decrease as the estimated penetrance is decreased and increase as the estimated penetrance is increased (see also Supplementary Figure S3.1), while the AUC and rBS levels remain similar. This again highlights the dangers of using misspecified penetrances in Mendelian models. Incorporating gastric cancer information through a misspecified penetrance can actually impair the calibration, compared to the Mendelian model without gastric cancer information. Gradient boosting, on the other hand, does not need these penetrance estimates and can provide well-calibrated predictions.

3.5 Data Application

3.5.1 USC-Stanford Data

We apply our methods to data collected from the University of Southern California (USC) and Norris Comprehensive Cancer Center and Stanford Cancer Institute Cancer Genetics Hereditary Cancer Panel (HCP) Testing study (Idos et al., 2018). Investigators collected these data to study the impact of several genetic mutations on cancer outcomes as well as the role of genetic testing in clinical settings. Participants were meant to be representative of a high-risk population; thus, individuals with greater than a 2.5% risk of carrying a cancer susceptibility gene mutation, as measured by various risk prediction models, were eligible to be included in the study. The
enrollment process lasted between 2014 and 2016, and in total 2000 participants were enrolled. Each participant received the Myriad Genetics myRisk Hereditary Cancer test, which identifies mutations for 25 genes which have been shown to increase cancer susceptibility, including the three MMR genes included in MMRpro. The panel also includes other genes which have been shown to be associated with colorectal, endometrial, and gastric cancers, such as *EPCAM* and *PMS2*. Each proband completed baseline questionnaires providing information on risk factors and family history.

Of the 2000 probands, 102 had variants of uncertain significance (VUS) for at least one of the MMR genes. By definition, it is unknown whether or not VUS carriers have pathogenic or benign variants; thus, we therefore exclude individuals with VUS’s from our analysis. Also, 211 probands tested positive for genes outside of the three MMR genes we are considering. As carriers of a mutation that leads to increased cancer susceptibility, these probands could have a richer cancer history, which would inflate the predictions of their risk of carrying the MMR mutations of interest. In order to avoid this issue, we excluded these individuals as well. Lastly, we excluded 5 families resulting in errors when running MMRpro (these families have relationships or typos that led to errors, such as siblings being married). After removing these individuals, we are left with a final subset of 1687 probands, with 25 probands who tested positive for an MMR mutation. A summary of this subset is provided in Table 3.2.

We ran 14 prediction models on these data. These include the Mendelian model with and without gastric cancer information; gradient boosting with and without the Mendelian predictions, with gastric cancer information, with 25, 50, and 100 iterations; gradient boosting with and without the Mendelian predictions, without gastric cancer information, with 50 iterations; and the Mendelian model with altered gastric cancer penetrances (gastric cancer survival functions raised to the power of 0.25,
<table>
<thead>
<tr>
<th>General</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of families</td>
<td>1687</td>
</tr>
<tr>
<td>Average family size</td>
<td>33.67</td>
</tr>
<tr>
<td>Number of male probands</td>
<td>323</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Proband Cancer History</th>
<th></th>
</tr>
</thead>
</table>
| Number of probands with:
  Colorectal cancer (average age) | 260 (50.18) |
  Endometrial cancer (average age)  | 65 (50.62)  |
  Gastric cancer (average age)     | 40 (50.03)  |

<table>
<thead>
<tr>
<th>Family Member Cancer History</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of families with at least one non-proband with:</td>
</tr>
<tr>
<td>Colorectal cancer</td>
</tr>
<tr>
<td>Endometrial cancer</td>
</tr>
<tr>
<td>Gastric cancer</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Proband MMR Mutations</th>
<th></th>
</tr>
</thead>
</table>
| Number of probands with mutations of:
  MLH1                        | 7        |
  MSH2                        | 11       |
  MSH6                        | 7        |

<table>
<thead>
<tr>
<th>Proband Race</th>
<th></th>
</tr>
</thead>
</table>
| Number of probands whose race is:
  Asian                       | 184      |
  Black                       | 68       |
  Hispanic                    | 659      |
  Native American             | 6        |
  White                       | 691      |
  Unknown                     | 79       |

Table 3.2: Summary of the USC-Stanford data
0.5, 2, and 4). Unlike the simulations, we used the actual (unscaled) penetrance estimates for gastric cancer, with lifetime risks for non-carriers, carriers of MLH1 mutations, carriers of MSH2 mutations, and carriers of MSH6 mutations of 0.006, 0.163, 0.006, and 0.035, respectively, for males; and 0.006, 0.163, 0.201, and 0.035, respectively, for females. Since the true “data-generating” gastric cancer penetrance for real data is unknown, we do not know if the estimated gastric cancer penetrance or altered penetrances are misspecified. We also used the actual MMRpro penetrance estimates for colorectal and endometrial cancers and the MMRpro allele frequencies of 0.0004, 0.0005, and 0.0002 for MLH1, MSH2, and MSH6, respectively. MMRpro can account for information on race, microsatellite instability (MSI) testing, and immunohistochemistry (IHC) testing; since the data included such information, we use this information to obtain our MMRpro predictions. We used the same gradient boosting tuning parameters and decision trees as in the simulations. Again we split the data in half to training and testing sets and then used a bootstrap procedure to account for uncertainty in the splitting process.

3.5.2 Results

A summary of the results is provided in Table 3.3. The Mendelian model without gastric cancer information had an O/E of 0.803 (0.509, 1.140), AUC of 0.860 (0.810, 0.911), and rBS of 0.128 (0.104, 0.148). Gradient boosting, without the Mendelian predictions, without gastric cancer information, and using 50 iterations, showed slightly improved calibration (0.82 (0.369, 1.549)) but worse discrimination (0.718 (0.604, 0.825)) and similar rBS (0.121 (0.098, 0.141)). The combination of gradient boosting and the Mendelian model (without gastric cancer) had the best calibration (1.028 (0.427, 2.191)) but worse discrimination compared to the Mendelian model.
Thus, like the results in the simulations, the proposed gradient boosting approach provided well-calibrated predictions; however, here the discrimination suffered. This may be due to the fact that we used misspecified colorectal and endometrial cancer penetrances in the simulations, while here our estimated colorectal and endometrial cancer penetrances may be more accurate. Thus, the penetrances effectively provided strong discrimination, and the gradient boosting approach to incorporating the information from the same cancers actually reduced the discriminatory power.

<table>
<thead>
<tr>
<th></th>
<th>O/E</th>
<th>AUC</th>
<th>rBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>No gastric cancer info.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mendelian</td>
<td>0.803 (0.509, 1.140)</td>
<td>0.860 (0.810, 0.911)</td>
<td>0.128 (0.104, 0.148)</td>
</tr>
<tr>
<td>GB</td>
<td>0.82 (0.369, 1.549)</td>
<td>0.718 (0.604, 0.825)</td>
<td>0.121 (0.098, 0.141)</td>
</tr>
<tr>
<td>GB with Mendelian</td>
<td>1.028 (0.427, 2.191)</td>
<td>0.809 (0.677, 0.896)</td>
<td>0.124 (0.105, 0.142)</td>
</tr>
<tr>
<td>Incorporating gastric cancer info</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mendelian</td>
<td>0.703 (0.446, 0.991)</td>
<td>0.843 (0.787, 0.901)</td>
<td>0.134 (0.112, 0.153)</td>
</tr>
<tr>
<td>GB</td>
<td>0.801 (0.356, 1.478)</td>
<td>0.718 (0.608, 0.827)</td>
<td>0.120 (0.098, 0.141)</td>
</tr>
<tr>
<td>GB with Mendelian</td>
<td>0.981 (0.426, 2.005)</td>
<td>0.816 (0.695, 0.894)</td>
<td>0.123 (0.102, 0.142)</td>
</tr>
</tbody>
</table>

| a Mendelian model, without gastric cancer information |
| b Gradient boosting initialized without Mendelian predictions, using colorectal and endometrial cancer information |
| c Gradient boosting initialized with Mendelian predictions (without gastric cancer), using colorectal and endometrial cancer information |
| d Mendelian model, incorporating gastric cancer information through the penetrance |
| e Gradient boosting without Mendelian predictions, using all 3 cancers |
| f Gradient boosting initialized with Mendelian predictions (without gastric cancer), using all 3 cancers |

**Table 3.3:** Performance measures for the USC-Stanford HSCP study data. The numbers represent the mean performance measures across 1000 bootstrap replicates, with 95% bootstrap confidence intervals in the parentheses. All gradient boosting models are run with 50 iterations.

We also explored the various approaches to incorporating gastric cancer informa-
tion. The completely Mendelian approach of using the gastric cancer penetrance actually provided worse calibration (0.703 (0.446, 0.991)) and discrimination (0.843 (0.787, 0.901)) compared to the Mendelian model without gastric cancer information. This suggests that the estimated gastric cancer penetrance used in the Mendelian model is misspecified and not representative of the true data-generating mechanism in the data. Thus, we see a clear example of a situation where penetrance estimates may be misspecified and hence harmful to Mendelian prediction. The gradient boosting approach to incorporating information from the three cancers, without the Mendelian prediction, had similar performance measures as the equivalent gradient boosting approach without using gastric cancer as a feature (O/E of 0.801 (0.356, 1.478), AUC of 0.718 (0.608, 0.827), rBS of 0.120 (0.098, 0.141)). Thus we see that the gastric cancer information did not add information on top of the colorectal and endometrial cancer information for predicting the carrier status of a MMR mutation. The combination of gradient boosting with the Mendelian prediction, using all three cancers as features, provided the best calibration but had worse discrimination than the Mendelian model with gastric cancer information, with similar rBS values.

The impact of the number of iterations was similar to that seen in simulations. The calibration for gradient boosting initialized with the Mendelian predictions was fairly stable, although the O/E ratios did increase as the number of iterations increased. The AUC values also decreased as the number of iterations increased, while the rBS values remained similar. The O/E ratios for gradient boosting without the Mendelian predictions increased dramatically as the number of iterations increased, while the AUC and rBS values changed slightly.

As in the simulations, the Mendelian model calibration was sensitive to alterations of the gastric cancer penetrance. The O/E tended to increase as the power used to
raise the gastric cancer survival function increased. All levels of alterations resulted in overprediction on average, as the O/E ratio was 0.839 when using the gastric cancer survival function raised to the power of 4. This suggests that the gastric cancer penetrance from ASK2ME corresponds to drastically higher gastric cancer risks than the "true penetrance" in the data.

Overall, although the mean performance measures seem to provide evidence of these assertions, all the confidence intervals were much wider than the ones from the simulations, constraining our ability to arrive at definitive conclusions. Compared to the simulated data, we have a smaller data set with fewer mutation carriers. In addition, the confidence intervals for the data application are obtained through a bootstrap resampling procedure, while the confidence intervals in the simulations are obtained by generating multiple data sets.

Figure 3.1 compares the mean gradient boosting predictions, initializing with the Mendelian model, using all three cancers as features, and with 50 iterations, with a completely Mendelian model using gastric cancer information. The gradient boosting predictions are the average predictions over the 1000 bootstrap replicates. For each proband, we only average over the bootstrap replicates where the proband was selected to be in the testing set. The plot shows the improvement in calibration for gradient boosting, as the risk predictions are decreased on average compared to the completely Mendelian model with gastric cancer. In particular, we notice that several non-carriers had high predicted risks using the completely Mendelian model with gastric cancer, while the gradient boosting predictions were much lower. Out of these non-carriers, many of the ones with the largest differences in risk predictions were also in families with at least one member with gastric cancer, showing how gradient boosting may handle gastric cancer information better than the Mendelian model.
with the current gastric cancer penetrance estimate. We can see this more clearly in Figure 3.2, which compares the performance measures of the two models when subsetting to families with at least one member with gastric cancer versus families without. Based on these results, gradient boosting has much better calibration for families with a history of gastric cancer, with improvements in discrimination and accuracy as well.

**Figure 3.1:** The mean (over the bootstrap replicates) gradient boosting predictions initialized by MMRpro compared to the completely Mendelian MMRpro predictions with gastric cancer for the USC-Stanford HCP study data. The gradient boosting predictions are initialized with the MMRpro predictions without gastric cancer information and obtained using 50 iterations, and the completely Mendelian MMRpro predictions are with gastric cancer information. There are a total of 26 carriers (triangles) and 1670 non-carriers (circles). There are also 307 families with at least one member with gastric cancer (blue), while 1389 families do not have anyone with gastric cancer (red).

### 3.6 Discussion

Many existing prediction models have the potential to be improved by upgrading the model mechanisms and by incorporating additional features known to be associated with the outcome; however, improving the models in practice can be difficult due to model complexity and proprietariness. Gradient boosting provides an opportunity to
empirically incorporate features, both ones already used in the existing model as well as new ones, while taking advantage of the strengths of the existing model. In this work, we illustrated the gradient boosting approach by applying it to Mendelian models, which are complex and require accurate cancer penetrance estimates. Obtaining accurate penetrances is not always feasible, and results from the simulations showed that when the penetrances are misspecified, Mendelian models do not provide well-calibrated predictions. In addition, while MMRpro is open-source, provided through the BayesMendel R package, other Mendelian models such as BOADICEA are proprietary and hence difficult to enhance by users. Results from both simulated data as well as data from the USC-Stanford Genetics HCP Testing study provide evidence that our proposed gradient boosting approach for integrating cancer information improves the Mendelian model calibration compared to utilizing this information using a completely Mendelian approach as well as compared to using gradient boosting with-

Figure 3.2: Comparison of the O/E, AUC, and rBS values for gradient boosting initialized with MMRpro with 50 iterations and a completely Mendelian MMRpro model with gastric cancer information in the USC-Stanford data. Comparison of families who have at least one member with gastric cancer to families without members with gastric cancer.
out the Mendelian model predictions.

Although the data application results are promising, there are some limitations in the data. Since there are only 25 MMR mutation carriers, the confidence intervals for the performance measures in Table 3.3 are wide. In particular, we see that even though the O/E ratios for gradient boosting initialized with the Mendelian predictions are much closer to 1 on average compared to the Mendelian models, the confidence intervals overlap considerably. The low number of carriers is especially concerning since the algorithm was trained by splitting the data into training and testing sets, so validation on the testing set only considers 13 carriers on average. Some splits may have a disproportionate amount of carriers in either the training or testing sets, leading to unstable risk predictions. The comparatively increased effectiveness of gradient boosting in the simulated data is due to several factors. In the simulations we generated carrier statuses using allele frequencies of 0.01, which provides a significant amount of carriers. In addition, we scaled the gastric cancer penetrances to artificially create a wider gap in gastric cancer risk between mutation carriers and non-carriers; consequently, we improved the predictive power of gastric cancer outcomes, which helped the gradient boosting algorithm improve the predictions. We also used missetpecified colorectal and endometrial cancer penetrances when running the Mendelian models, which reduced their predictive performance. Lastly, we used a larger sample size of 10,000 families per simulated data set and generated 100 different data sets, allowing us to obtain tight confidence intervals.

All models applied to the USC-Stanford Genetics HCP Testing study data tended to overpredict the risk. This may be because the participants represent individuals who are likely to have a genetic predisposition to cancer. The Myriad Genetics panel considers 25 genes, ignoring countless other cancer susceptibility genes. It is possible
that some of the individuals in the data who tested negative for all 25 genes were carriers of a cancer susceptibility gene that was not included in the panel. We attempted to mitigate the issue of being a carrier of a non-MMR mutation in the panel by excluding such individuals; however, we could not do this for genes not included in the panel. Validating our gradient boosting approach on data that has 1) genetic testing results for more genetic mutations and 2) a larger sample size, would help determine our model's effectiveness.

We used a simplistic gradient boosting model, as we only included three features in each base learner corresponding to the three cancers of interest. This was to directly compare the gradient boosting and completely Mendelian approaches of incorporating cancer information. However, including additional features could improve gradient boosting's ability to predict the carrier status. For example, we could include information from other cancers or separate the cancer information based on relationship type. A family member with cancer who is not a blood relative of the proband does not carry as much significance as a first-degree relative with cancer. This distinction is incorporated in Mendelian models, but gradient boosting may be able to detect residual non-Mendelian impacts of the cancers. We could also incorporate other risk factors besides cancer history. The USC-Stanford data includes information about aspirin use, which has been shown to be associated with an decreased risk of colorectal cancer (Flossmann et al., 2007). Thus, since MMR mutation carriers have an increased risk of developing colorectal cancer, aspirin use may be different on average between carriers and non-carriers, so including this information in the base learner could improve the model's performance. We also fixed the gradient boosting tuning parameters (the maximum tree depth, bagging fraction, and shrinkage parameter) based on recommended values from the literature, but could explore other values to
see their impact on the results. Lastly, other binary loss functions to apply to the gradient boosting algorithm can be considered.

Gradient boosting is an example of an ensemble method, which is a technique where prediction models are combined to increase predictive power. Thus, in addition to gradient boosting, other ensemble methods could be considered for improving existing models. For example, bagging (Breiman, 1996), or bootstrap aggregating, averages the predictions from models fit on bootstrap samples of the data. Bagging works best with unstable models by reducing overfitting through its resampling procedure; however, it may not reduce model bias as well as gradient boosting, since gradient boosting reduces the bias by predicting the model residuals. Multiple ensemble techniques can also be combined to further improve predictive performance. Stacking (Wolpert, 1992) combines predictions from various models through a combiner algorithm such as logistic regression. It thus learns how to best take advantage of the strengths of each model, thereby creating a new model that performs better than any of the individual models.

Extending existing risk models via machine learning techniques is an important and underutilized approach compared to training new models from scratch. Through our illustration on Mendelian models, we see that gradient boosting can be applied as a convenient and powerful tool to extend models, leveraging the advantages of the existing model while providing opportunities for improvements through new data and features. We hope that this work will encourage others to further apply and develop methods under this framework that have the potential to lead to significant upgrades for numerous complex prediction models.
References


Supplementary Materials

S0.1 Code

Code for the simulations in chapter 1 can be found at https://github.com/theohuang/Frailty-Project. Code for the simulations in chapter 2 can be found at https://github.com/theohuang/Gradient-boosting. Code for the simulations in chapter 3 can be found at https://github.com/theohuang/Variation-Familial-Cancer-Risk.

S1.1 Deriving Baseline Hazards

We show the derivation of the procedure to derive the baseline hazard functions.

\[ S_d^d(t|G, U) = P(T_d^d > t|G, U) \]
\[ = P(T_r^c \geq t + 1|G, U) \]
\[ = \sum_{W_r} P(T_r^c \geq t + 1, W_r|G, U) \]
\[ = \sum_{W_r} P(T_r^c \geq t + 1|G, U, W_r)P(W_r|G, U) \]
\[ = \sum_{W_r} S_r^c(t + 1|G, U, W_r)P(W_r) \]
\[ = \sum_{W_r} \exp\{-\Lambda_r G(t + 1|W_r, U)\}P(W_r) \]
\[ = \sum_{W_r} \exp\{-\lambda_0 U G(t + 1) \exp(W_r)\}P(W_r) \]
\[ = \sum_{W_r} \exp\{-\lambda_0 U G(t + 1)\} \exp(W_r)P(W_r) \]
The left hand side is the marginal survival function from BRCAPRO, while the right hand side gives the expression in terms of the baseline hazard functions. Thus, the baseline hazards at each time \( t \) can be estimated iteratively by minimizing the squared difference between the two.

**S1.2 Generating Families**

Each family consists of 5 generations, where the proband is in the third generation. Each family consists of a female proband, father and mother, paternal and maternal grandparents, paternal and maternal aunts and uncles, sons and daughters, and nephews and nieces. For each family, the number of maternal aunts, maternal uncles, paternal aunts, paternal uncles, brothers, sisters, sons, daughters, sons for each sibling, and daughters for each sibling are each generated randomly and independently from \( \{0, 1, 2, 3\} \). Thus families can be as small as a proband with a mother, father, and paternal and maternal grandparents (size 7); and as large as a proband with a mother, father, maternal and paternal grandparents, 3 paternal uncles, 3 paternal
aunts, 3 maternal uncles, 3 maternal aunts, 3 brothers, 3 sisters, 3 sons, 3 daughters, and 3 sons and 3 daughters for each sibling (size 67).

Each family is randomly selected to be Ashkenazi Jewish or not with a probability of $p = 5.5/285$, based on an estimate of the proportion of Ashkenazi Jews in the United States. For Ashkenazi Jewish families, founders of the pedigree have genotypes ($BRCA1$ and $BRCA2$ mutation carrier statuses) generated using the BRCAPRO Ashkenazi Jewish allele frequencies, while founders of families who are not Ashkenazi Jewish have genotypes generated using the BRCAPRO non-Ashkenazi Jewish allele frequencies. The genotypes for the rest of the family are generated following Mendelian laws of inheritance.

Current ages at the 5-year follow-up are generated as follows. The grandmothers’ ages are generated from a normal distribution with mean 100 and standard deviation 2, and the grandfathers’ ages are generated from a normal distribution with their spouse’s age as the mean and standard deviation 2. Then the parents and aunts and uncles of the proband have their current ages generated from a normal distribution with their mother’s age minus 30 as the mean and standard deviation 5. If their mother’s age minus 15 is less than this generated age, then the mother’s age minus 15 will instead be used as the current age. The current ages of the proband and siblings are similarly generated based on the mother’s age. The current ages of the spouses of the proband and her siblings are generated from a normal distribution with their spouses’ ages as the mean and standard deviation 2. The current ages of the children of the proband and her siblings are then generated in the same way as the ages of the proband and her siblings, based on their mothers’ ages. Lastly, current ages less than 7 are assigned to be 7, and current ages greater than 94 are assigned to be 94.

The cancer ages of the family members are based on the family-specific penetrance.
For each family, a frailty vector is generated based on the specified data-generating frailty distribution. Then, based on this frailty vector and the BRCAPRO penetrances (which are defined from ages 1 to 94), the family-specific penetrances are obtained, using the method described in Section 1.3.3. Then each family member’s cancer age is generated from this family-specific penetrance and their genotype. These cancer ages can go up to age 95, meaning that the individual would not develop cancer in their lifetime. If an individual developed cancer, we only observe it if their cancer age is less than or equal to the current age.

This family history represents the known information at the 5-year follow-up, used for validating our model. We then go back in time 5 years to obtain the information at the time of the proband’s consultation.

**S1.3 Estimating the Data-generating Frailty Distribution**

For the simulations, we can explore our model’s ability to recover the data-generating frailty distribution. For each family, we can think of $P(W)$ as the prior frailty distribution and $P(W|H,U)$ as the posterior frailty distribution. Thus, in order to detect the overall frailty distribution in the data, we can aggregate all the posterior frailty distributions $P(W_k|H_k)$, where the subscript $k$ denotes the $k$-th family in the data set, $k = 1, \ldots, N$. A natural way to do this is to treat this aggregated frailty distribution as a mixture of the family posterior frailty distributions, using equal mixing weights, to obtain an aggregated posterior distribution $P_{agg}(W = w|H_1, \ldots, H_N) = \frac{1}{N} \sum_{k=1}^{N} P(W_k = w_k|H_k)$. An alternative approach is to take the medians of the family-specific frailty probabilities instead of the means. Heatmaps of the means and medians of the family-specific frailty probabilities are shown in Supplementary Fig-
ures S1.1 and S1.2, respectively.

$P_{agg}(\mathbf{W}|\mathbf{H}_1, \ldots, \mathbf{H}_N)$ can also be seen as an estimate of the data-generating distribution, although it is important to note that obtaining $P_{agg}$ depends on the choice of prior. In the situation where we generate data from the discrete uniform distribution, which is the prior distribution $P(\mathbf{W})$, we might expect the aggregated posterior distribution to be similar to the prior. However, as seen in Supplementary Figures S1.1 and S1.2, even in the case where the data-generating distribution is correctly specified, we are unable to recover the correct discrete uniform distribution. This shows how estimating a nonparametric distribution, which in this case has 48 parameters, is challenging. We can explore this further by comparing the families’ mean frailties (mean of the family-specific frailty distribution) to their true frailty in the case where the data-generating distribution is the discrete uniform. We can visualize the means of these family-specific frailty means, subsetted by the true frailties, in Supplementary Figure S1.3. Although families with higher true frailties have higher frailty means, the means themselves are not close to the true frailties. Despite this limitation, we observe improvements in the performance measures, suggesting that our approach is robust to misspecification in the data-generating frailty distribution.

**S1.4 Conditional Approach**

The likelihood construction for the conditional approach as described in Section 1.6.1 is as follows:

$$L(\mathbf{W}|\mathbf{G}, \mathbf{U}) = P(\mathbf{X}^d, \mathbf{W}|\mathbf{G}, \mathbf{U}) = P(\mathbf{X}^d_1, \mathbf{X}^d_2|\mathbf{G}, \mathbf{U}, \mathbf{W})P(\mathbf{W}|\mathbf{G}, \mathbf{U})$$
\[ = \prod_{r=1}^{2} \prod_{i=1}^{n} \left\{ f_{r_i} (X_{r_i}^d|G_i, U_i, W_r) \right\} \delta_{r_i} \left\{ S_{r_i} (X_{r_i}^d|G_i, U_i, W_r) \right\}^{1-\delta_{r_i}} f_W(W) \]
\[ = \prod_{r=1}^{2} \prod_{i=1}^{n} \left\{ \lambda_{r_i}^d (X_{r_i}^d|G_i, U_i, W_r) \right\} \delta_{r_i} \left\{ S_{r_i} (X_{r_i}^d - 1|G_i, U_i, W_r) \right\}^{1-\delta_{r_i}} f_W(W) \]

The log-likelihood can be written as:

\[ \ell(W|G, U) = \sum_{r=1}^{2} \sum_{i=1}^{n} \left[ \delta_{r_i} \log \left\{ 1 - \lambda_{0rU_iG_i}^d (X_{r_i}^d)^{\text{exp}(W_r)} \right\} \right] + \delta_{r_i} \text{exp}(W_r) \sum_{s=1}^{X_{r_i}^d - 1} \log \left\{ 1 - \lambda_{0rU_iG_i}^d (s) \right\} \]

\[ + (1 - \delta_{r_i}) \text{exp}(W_r) \sum_{s=1}^{X_{r_i}} \log \left\{ 1 - \lambda_{0rU_iG_i}^d (s) \right\} \]

The construction of the likelihood is based on the genotypes of the family members; however, in most cases these are unknown. Thus for family members whose genotypes are unknown, we replace the term in the log-likelihood that depends on genotypes with its conditional expectation (over the genotypic distribution) given the family history and the frailty, based on the BRCAPRO carrier probabilities. If we let \( a_{r_i} \)
denote the term inside the square brackets, then the log-likelihood can be written as:

$$\ell(W|G, U) = \sum_{r=1}^{2} \sum_{i=1}^{n} a_{ri} + \log f_W(W).$$

The log-likelihood can then be rewritten by summing over all possible unobserved genotypes \(G_i\) and weighting each by the probability of observing each genotype, \(P(G_i|H, U, W)\):

$$\sum_{r=1}^{2} \sum_{i=1}^{n} \sum_{G_i \in \{0, 1\}^2} a_{ri} P(G_i|H, U, W) + \log f_W(W).$$

The carrier probabilities \(P(G_i|H, U, W)\) can be calculated using BRCAPRO. The estimated frailty is then the value of \(W\) that maximizes this log-likelihood.

**S1.5 Profile Likelihood**

Let \(\ell_k\) be the log-likelihood for the \(k\)-th family. Then the profile log-likelihood can be defined as \(\ell^P(\Sigma) = \sum_{k=1}^{N} \ell_k(\hat{W}_k|G_k, U_k; \Sigma)\). Here \(G_k\) is the vector of genotypes for the \(k\)-th family, \(\hat{W}_k\) is the estimated frailty for the \(k\)-th family when using \(\Sigma\) as the frailty covariance matrix, and \(\ell_k(\cdot|G_k; \Sigma)\) is the log-likelihood when using \(\Sigma\) as the frailty covariance matrix. For each value of \(\Sigma\), the maximum likelihood estimates of each family’s frailty vectors is estimated and incorporated into the overall log-likelihood.

Now let \(\Sigma_{11}\) and \(\Sigma_{22}\) be fixed. We consider the impact of \(\rho\) on the profile
log-likelihood. Each family’s contribution to the profile log-likelihood is

$$\sum_{r=1}^{2} \sum_{i=0}^{n} \left[ \delta_{ri} \log \left\{ 1 - [1 - \lambda_{0rU_iG_i}^{d}(X^{d}_{ri})]^\text{exp}(W_{r}) \right\} + \delta_{ri} \exp(W_{r}) \sum_{s=1}^{X_{ri}^{d}-1} \log \left\{ 1 - \lambda_{0rU_iG_i}^{d}(s) \right\} \right]$$

$$+ (1 - \delta_{ri}) \exp(W_{r}) \sum_{s=1}^{X_{ri}^{d}} \log \left\{ 1 - \lambda_{0rU_iG_i}^{d}(s) \right\} + \log \phi(W; \Sigma),$$

where $\phi(\cdot; \Sigma)$ is the density of the bivariate normal distribution with mean $(0, 0)$ and variance $\Sigma$. Here $\rho$ only impacts each family’s contribution to the likelihood through the second component $\log \phi(W; \Sigma)$. $\Sigma_{11}$ and $\Sigma_{22}$ impact the first component by affecting the baseline hazard functions as well as the penetrance functions used when calling BRCA2 to calculate the carrier probabilities.

The density $\phi(W; \Sigma)$ is simply the density of the bivariate normal distribution:

$$\phi(W; \Sigma) = \frac{1}{2\pi\Sigma_{11}\Sigma_{22}\sqrt{1 - \rho^2}} \exp \left\{ -\frac{1}{2(1 - \rho^2)} \left( \frac{W_{1}^2}{\Sigma_{11}} - \frac{2\rho W_{1} W_{2}}{\Sigma_{11} \Sigma_{22}} + \frac{W_{2}^2}{\Sigma_{22}} \right) \right\}.$$

Since $\hat{W}_{k}$ maximizes the log-likelihood for the $k$-th family,

$$\lim_{|\rho| \to 1} \ell_{k}(\hat{W}_{k} | G_{k}, U_{k}; \Sigma) \geq \lim_{|\rho| \to 1} \ell_{k}(W = (0, 0) | G_{k}, U_{k}; \Sigma) = \infty.$$

The last equality holds since

$$\lim_{|\rho| \to 1} \phi(W = (0, 0); \Sigma) = \lim_{|\rho| \to 1} \frac{1}{2\pi\Sigma_{11}\Sigma_{22}\sqrt{1 - \rho^2}} = \infty.$$
S1.6 Twin Approach to Estimating $\Sigma$

This section proposes an approach to estimating the unknown covariance matrix $\Sigma$ in the frailty distribution using external twin data. Suppose we have data consisting of twin pairs with breast and ovarian cancer outcomes. For the summary measure, we could use concordance for three outcomes: breast cancer, ovarian cancer, and breast/ovarian cancer. For either cancer, a twin pair could be considered concordant if both twins have that cancer. For the breast/ovarian cancer outcome, a twin pair could be concordant if both breast and ovarian cancer are present between the two twins (one twin may have breast cancer and the other ovarian cancer, or one twin may have both). For each of the three outcomes, we could then calculate the proportion of twins who are concordant in simulated data under various values of $\Sigma$. We could then select the value of $\Sigma$ that most closely matches these observed measures in an external data set. For example, we could choose the value of $\Sigma$ that produces a vector of length 3 of concordance proportions in the simulated data for the three outcomes that has the least $L_2$ distance from the corresponding vector of concordance proportions in the twin data.

However, this kind of approach using external data can be problematic, even resulting in estimates of $\Sigma$ with negative correlations. For example, when using BRCAPRO penetrances for the simulations, if the breast/ovarian concordance proportion in the data set is much lower than the ones generated using the BRCAPRO penetrances, then very negative correlation terms would be needed to generate data that match the external data. Negative values of $\rho$ would cause each family’s breast and ovarian cancer frailty variates to be further apart, resulting in less breast/ovarian cancer concordance. However, negative values of $\rho$ may not be suitable for our model, as the un-
observed shared risks for breast and ovarian cancer should intuitively be either both positive or negative.

**Figure S1.1:** Heatmap of the means of the family-specific frailty probabilities $P(W_k = w|H_k)$ (where $k$ denotes the $k$-th family) in the simulated data sets. There are three data-generating frailty distributions (bivariate normal with variances of 0.3, bivariate normal with variances of 2, and discrete uniform), and two levels of allele frequencies (population-level and high-risk).
Figure S1.2: Heatmap of the medians of the family-specific frailty probabilities $P(W_k = w|H_k)$ (where $k$ denotes the $k$-th family) in the simulated data sets. There are three data-generating frailty distributions (bivariate normal with variances of 0.3, bivariate normal with variances of 2, and discrete uniform), and two levels of allele frequencies (population-level and high-risk).
Figure S1.3: Heatmap showing the means of the families’ mean breast cancer and mean ovarian cancer frailties in the simulated data generated from the discrete uniform distribution. Each family has a mean breast cancer frailty and mean ovarian cancer frailty, derived from the family-specific frailty distribution. Each tile represents the families who had the corresponding true frailty vector, and the numbers in the tiles represent the means of the family-specific frailty means (first breast cancer, then ovarian cancer). The color of each tile represents the mean of the two numbers in the tile.
<table>
<thead>
<tr>
<th>Family size Distribution</th>
<th>Variance</th>
<th>O/E</th>
<th>AUC</th>
<th>rBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Population-level allele frequency</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes BVN</td>
<td>0.3</td>
<td>0.863</td>
<td>0.739</td>
<td>0.060</td>
</tr>
<tr>
<td>No BVN</td>
<td>0.3</td>
<td>0.909</td>
<td>0.718</td>
<td>0.062</td>
</tr>
<tr>
<td>Yes BVN</td>
<td>0.2</td>
<td>1.049</td>
<td>0.809</td>
<td>0.065</td>
</tr>
<tr>
<td>No BVN</td>
<td>0.2</td>
<td>1.027</td>
<td>0.819</td>
<td>0.065</td>
</tr>
<tr>
<td>Yes DU</td>
<td></td>
<td>1.045</td>
<td>0.746</td>
<td>0.066</td>
</tr>
<tr>
<td>No DU</td>
<td></td>
<td>0.974</td>
<td>0.755</td>
<td>0.064</td>
</tr>
<tr>
<td><strong>High-risk allele frequency</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes BVN</td>
<td>0.3</td>
<td>1.138</td>
<td>0.745</td>
<td>0.089</td>
</tr>
<tr>
<td>No BVN</td>
<td>0.3</td>
<td>1.150</td>
<td>0.758</td>
<td>0.089</td>
</tr>
<tr>
<td>Yes BVN</td>
<td>0.2</td>
<td>1.119</td>
<td>0.822</td>
<td>0.087</td>
</tr>
<tr>
<td>No BVN</td>
<td>0.2</td>
<td>1.169</td>
<td>0.844</td>
<td>0.088</td>
</tr>
<tr>
<td>Yes DU</td>
<td></td>
<td>1.087</td>
<td>0.780</td>
<td>0.087</td>
</tr>
<tr>
<td>No DU</td>
<td></td>
<td>1.135</td>
<td>0.794</td>
<td>0.088</td>
</tr>
</tbody>
</table>

a Data-generating frailty distribution
b Variances of the data-generating frailty distribution (if bivariate normal)
c Bivariate normal distribution with mean (0, 0)
d Discrete uniform distribution

Table S1.1: Sensitivity analysis of family size on performance measures for the frailty model on the simulated data.
Figure S1.4: Heatmap of the medians of the frailty probabilities for the families in the CGN breast cancer subset.

Figure S1.5: Comparison of BRCAPRO with and without the frailty model for different numbers of first degree relatives with breast cancer in the CGN data.
Figure S1.6: Comparison of BRCAPRO with or without the frailty model for those who are and are not Ashkenazi Jews in the CGN data.

Figure S1.7: Kernel density estimates of the breast cancer frailty means for the CGN families, subsetted by the number of first degree relatives with breast cancer.
**Figure S1.8:** Top: Heatmap of the means of the values of the family-specific frailty distributions in the CGN breast cancer subset. Each of the 49 possible frailty vectors have a probability for each family, and the mean of the probabilities is taken across all the families. Bottom: Histograms for the family-specific frailty probabilities for different breast cancer frailty variates. Ovarian cancer frailty variate is fixed to be 0. The red lines indicate the median of each distribution, and the blue lines indicate the mean.
Figure S1.9: Plots showing the frailty-adjusted breast cancer penetrance functions for different values of breast cancer frailty variates. a) Penetrances for non-carriers of BRCA1 and BRCA2. b) Penetrances for carriers of BRCA1 only.
S2.1 Derivation of O/E Ratios

The number of observed cases among carriers is derived as follows:

\[
E_{Y_i,G_i}[Y_i I(G_i \neq 0)|H] = \sum_{g_i} E_{Y_i}[Y_i I(G_i \neq 0)|G_i = g_i, H] P(G_i = g_i|H)
\]

\[
= \sum_{g_i \neq 0} E_{Y_i}[Y_i|G_i = g_i, H] P(G_i = g_i|H)
\]

\[
= \sum_{g_i \neq 0} P(T_i \leq T_{max}|G_i = g_i, H) P(G_i = g_i|H)
\]

\[
= \sum_{g_i \neq 0} P(T_i \leq T_{max}|G_i = g_i, H) P(G_i = g_i|H)
\]

\[
= \sum_{g_i \neq 0} \frac{P(C_i < T_i \leq T_{max}|G_i = g_i)I_{1-\delta_i}}{P(T_i > C_i|G_i = g_i)I_{1-\delta_i}} P(G_i = g_i|H)
\]

The number of expected cases among carriers is as follows:

\[
E_{G_i}[E[Y_i I(G_i \neq 0)|G_i]|H] = \sum_{g_i} E[Y_i I(G_i \neq 0)|G_i = g_i] P(G_i = g_i|H)
\]

\[
= \sum_{g_i \neq 0} E[Y_i|G_i = g_i] P(G_i = g_i|H)
\]

\[
= \sum_{g_i \neq 0} P(T_i \leq 94|G_i = g_i) P(G_i = g_i|H)
\]

The observed and expected for non-carriers is analogous, with \(G_i \neq 0\) becoming \(G_i = 0\).
<table>
<thead>
<tr>
<th>With Mendelian</th>
<th>Iterations</th>
<th>O/E</th>
<th>AUC</th>
<th>rBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-penetrant data-generating CRC and EC penetrances</td>
<td>Yes</td>
<td>25</td>
<td>0.990 (0.973, 1.003)</td>
<td>0.835 (0.816, 0.849)</td>
</tr>
<tr>
<td>Yes</td>
<td>50</td>
<td>1.002 (0.986, 1.017)</td>
<td>0.833 (0.813, 0.848)</td>
<td>0.216 (0.208, 0.221)</td>
</tr>
<tr>
<td>Yes</td>
<td>100</td>
<td>1.003 (0.987, 1.017)</td>
<td>0.831 (0.811, 0.846)</td>
<td>0.216 (0.209, 0.222)</td>
</tr>
<tr>
<td>No</td>
<td>25</td>
<td>0.630 (0.609, 0.649)</td>
<td>0.791 (0.767, 0.806)</td>
<td>0.227 (0.22, 0.234)</td>
</tr>
<tr>
<td>No</td>
<td>50</td>
<td>0.962 (0.945, 0.977)</td>
<td>0.794 (0.771, 0.812)</td>
<td>0.224 (0.217, 0.23)</td>
</tr>
<tr>
<td>No</td>
<td>100</td>
<td>1.004 (0.988, 1.021)</td>
<td>0.793 (0.771, 0.81)</td>
<td>0.225 (0.218, 0.231)</td>
</tr>
<tr>
<td>High-penetrant data-generating CRC and EC penetrances</td>
<td>Yes</td>
<td>25</td>
<td>0.975 (0.957, 0.99)</td>
<td>0.921 (0.909, 0.93)</td>
</tr>
<tr>
<td>Yes</td>
<td>50</td>
<td>1.001 (0.987, 1.015)</td>
<td>0.92 (0.908, 0.929)</td>
<td>0.195 (0.189, 0.202)</td>
</tr>
<tr>
<td>Yes</td>
<td>100</td>
<td>1.006 (0.992, 1.018)</td>
<td>0.919 (0.907, 0.928)</td>
<td>0.197 (0.19, 0.204)</td>
</tr>
<tr>
<td>No</td>
<td>25</td>
<td>0.626 (0.598, 0.647)</td>
<td>0.875 (0.863, 0.885)</td>
<td>0.217 (0.21, 0.224)</td>
</tr>
<tr>
<td>No</td>
<td>50</td>
<td>0.958 (0.943, 0.971)</td>
<td>0.88 (0.868, 0.888)</td>
<td>0.214 (0.206, 0.22)</td>
</tr>
<tr>
<td>No</td>
<td>100</td>
<td>1.007 (0.992, 1.025)</td>
<td>0.88 (0.869, 0.889)</td>
<td>0.214 (0.207, 0.221)</td>
</tr>
</tbody>
</table>

\(a\) Initializing boosting with Mendelian predictions (without gastric cancer)

\(b\) Number of iterations in gradient boosting (\(M\) from Section 3.2)

Table S3.1: Sensitivity analysis of the number of iterations in gradient boosting for the simulated data. The top represents data generated using low-penetrant (scaled MMRpro) colorectal and endometrial cancer penetrances, and the bottom represents data generated using high-penetrant (unscaled MMRpro) colorectal and endometrial cancer penetrances.
<table>
<thead>
<tr>
<th>Exponent&lt;sup&gt;a&lt;/sup&gt;</th>
<th>O/E</th>
<th>AUC</th>
<th>rBS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Low-penetrant data-generating CRC and EC penetrances</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0.799 (0.744, 0.849)</td>
<td>0.839 (0.82, 0.854)</td>
<td>0.214 (0.207, 0.219)</td>
</tr>
<tr>
<td>0.5</td>
<td>0.798 (0.743, 0.849)</td>
<td>0.846 (0.826, 0.86)</td>
<td>0.212 (0.206, 0.218)</td>
</tr>
<tr>
<td>2</td>
<td>1.105 (1.025, 1.175)</td>
<td>0.838 (0.819, 0.853)</td>
<td>0.213 (0.206, 0.22)</td>
</tr>
<tr>
<td>4</td>
<td>1.691 (1.581, 1.807)</td>
<td>0.824 (0.802, 0.84)</td>
<td>0.217 (0.21, 0.224)</td>
</tr>
<tr>
<td><strong>High-penetrant data-generating CRC and EC penetrances</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0.708 (0.649, 0.758)</td>
<td>0.925 (0.913, 0.935)</td>
<td>0.194 (0.188, 0.2)</td>
</tr>
<tr>
<td>0.5</td>
<td>0.716 (0.657, 0.766)</td>
<td>0.928 (0.916, 0.938)</td>
<td>0.193 (0.187, 0.199)</td>
</tr>
<tr>
<td>2</td>
<td>0.896 (0.824, 0.956)</td>
<td>0.924 (0.914, 0.935)</td>
<td>0.193 (0.187, 0.198)</td>
</tr>
<tr>
<td>4</td>
<td>1.19 (1.099, 1.272)</td>
<td>0.915 (0.905, 0.926)</td>
<td>0.195 (0.189, 0.202)</td>
</tr>
</tbody>
</table>

<sup>a</sup> Power used to raise the gastric cancer survival function

**Table S3.2:** Analysis of Mendelian model performance on the simulated data when incorporating gastric cancer using misspecified penetrances. The exponent is the power used to raise the gastric cancer survival function. The top represents data generated using low-penetrant (scaled MMRpro) colorectal and endometrial cancer penetrances, and the bottom represents data generated using high-penetrant (unscaled MMRpro) colorectal and endometrial cancer penetrances.

<table>
<thead>
<tr>
<th>With MMRpro&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Iterations&lt;sup&gt;b&lt;/sup&gt;</th>
<th>O/E</th>
<th>AUC</th>
<th>rBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>25</td>
<td>0.957 (0.440, 1.879)</td>
<td>0.836 (0.745, 0.901)</td>
<td>0.124 (0.104, 0.143)</td>
</tr>
<tr>
<td>Yes</td>
<td>50</td>
<td>0.981 (0.426, 2.005)</td>
<td>0.816 (0.695, 0.894)</td>
<td>0.123 (0.102, 0.142)</td>
</tr>
<tr>
<td>Yes</td>
<td>100</td>
<td>1.032 (0.431, 2.114)</td>
<td>0.783 (0.63, 0.883)</td>
<td>0.123 (0.102, 0.141)</td>
</tr>
<tr>
<td>No</td>
<td>25</td>
<td>0.267 (0.155, 0.409)</td>
<td>0.686 (0.542, 0.803)</td>
<td>0.127 (0.110, 0.146)</td>
</tr>
<tr>
<td>No</td>
<td>50</td>
<td>0.801 (0.356, 1.478)</td>
<td>0.718 (0.608, 0.827)</td>
<td>0.120 (0.098, 0.141)</td>
</tr>
<tr>
<td>No</td>
<td>100</td>
<td>1.091 (0.442, 2.242)</td>
<td>0.707 (0.571, 0.820)</td>
<td>0.121 (0.098, 0.141)</td>
</tr>
</tbody>
</table>

<sup>a</sup> Initializing boosting with the Mendelian predictions (without gastric cancer)

<sup>b</sup> Number of iterations in gradient boosting ($M$ from Section 3.2)

**Table S3.3:** Sensitivity analysis of the number of iterations in gradient boosting for the USC-Stanford data.
Exponent | O/E | AUC | rBS |
--- | --- | --- | --- |
0.25 | 0.668 (0.426, 0.943) | 0.844 (0.788, 0.902) | 0.135 (0.114, 0.154) |
0.5 | 0.680 (0.433, 0.959) | 0.844 (0.787, 0.902) | 0.135 (0.113, 0.154) |
2 | 0.749 (0.474, 1.060) | 0.843 (0.785, 0.902) | 0.132 (0.110, 0.151) |
4 | 0.839 (0.529, 1.204) | 0.842 (0.784, 0.901) | 0.129 (0.107, 0.148) |

\(^a\) Power used to raise the gastric cancer survival function

**Table S3.4:** Analysis of Mendelian model performance on the USC-Stanford data when incorporating gastric cancer using altered penetrances. The exponent is the power used to raise the gastric cancer survival function.

![Non-carriers of the MMR mutations](image1)

![Carriers of MLH1 Only](image2)

**Figure S3.1:** Plots of the male gastric cancer penetrances where the survival functions are raised to different powers.