Harvard University

Harvard University Biostatistics Working Paper Series

Year 2014 *Paper* 183

Extending Mendelian Risk Prediction Models to Handle Misreported Family History

Danielle Braun[*] Malka Gorfine[†] Hormuzd A. Katki[‡]

Argyrios Ziogas[**] Hoda Anton-Culver[††] Giovanni Parmigiani[‡‡]

[*] Dana-Farber Cancer Institute and Harvard School of Public Health, dbraun@mail.harvard.edu

[†] Faculty of Industrial Engineering and Management, Technion- Israel Institute of Technology

[‡] Division of Cancer Epidemiology and Genetics, Biostatistics Branch, National Cancer Institute

[**] Department of Epidemiology, University of California Irvine

[††] Department of Epidemiology, University of California Irvine

[‡‡] Dana-Farber Cancer Institute and Harvard School of Public Health

Extending Mendelian Risk Prediction Models to Handle Misreported Family History

Danielle Braun, Malka Gorfine, Hormuzd A. Katki, Argyrios Ziogas, Hoda Anton-Culver, and Giovanni Parmigiani

Abstract

Mendelian risk prediction models calculate the probability of a proband being a mutation carrier based on family history and known mutation prevalence and penetrance. Family history in this setting, is self-reported and is often reported with error. Various studies in the literature have evaluated misreporting of family history. Using a validation data set which includes both error-prone self-reported family history and error-free validated family history, we propose a method to adjust for misreporting of family history. We estimate the measurement error process in a validation data set (from University of California at Irvine (UCI)) using non-parametric smoothed Kaplan-Meier estimators, and use Monte Carlo integration to implement the adjustment. In this paper, we extend BRCAPRO, a Mendelian risk prediction model for breast and ovarian cancers, to adjust for misreporting in family history. We apply the extended model to data from the Cancer Genetics Network (CGN).

Extending Mendelian Risk Prediction Models to Handle Misreported Family History

Danielle Braun[1,2,*], Malka Gorfine[3], Hormuzd Katki[4], Argyrios Ziogas[5], Hoda Anton-Culver[4], ,
and Giovanni Parmigiani[1,2]

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute

[2]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA.

[3]Faculty of Industrial Engineering and Management, Technion- Israel Institute of Technology, Haifa, Israel.

[4]Division of Cancer Epidemiology and Genetics, Biostatistics Branch, National Cancer Institute, Bethesda, MD, USA.

[5]Department of Epidemiology, University of California Irvine, Irvine, CA, USA.

Running title: Adjustment to Handle Misreported Family History.

Contact person:
Danielle Braun
Harvard School of Public Health
677 Huntington Avenue, SPH2, 4th Floor
Boston, MA 02115
United States
(617)-582-7228
dbraun@hsph.harvard.edu

1 Abstract

Mendelian risk prediction models calculate the probability of a proband being a mutation carrier based on family history and known mutation prevalence and penetrance. Family history in this setting, is self-reported and is often reported with error. Various studies in the literature have evaluated misreporting of family history. Using a validation data set which includes both error-prone self-reported family history and error-free validated family history, we propose a method to adjust for misreporting of family history. We estimate the measurement error process in a validation data set (from University of California at Irvine (UCI)) using nonparametric smoothed Kaplan-Meier estimators, and use Monte Carlo integration to implement the adjustment. In this paper, we extend BRCAPRO, a Mendelian risk prediction model for breast and ovarian cancers, to adjust for misreporting in family history. We apply the extended model to data from the Cancer Genetics Network (CGN).

2 Keywords

Mendelian Risk prediction; Measurement error; Non-parametric; Smoothed Kaplan-Meier.

3 Introduction

Cancer is caused by genetic alterations which can either be inherited or can occur during one's lifetime. There is a lot of interest in identifying individuals who are at high risk of cancer due to inherited mutations. Many risk prediction models have been developed to address this problem. These models can be divided into three main types; empirical models, expert-based models, and Mendelian models. Empirical models estimate the probability of a proband (consultand) being a mutation carrier, by summarizing family history and using it as predictors in the models. Expert-based models use algorithms based on clinical judgment to calculate scores summarizing the risk. Mendelian models use Mendelian laws of inheritance of deleterious genetic variants to calculate the probability that the proband is a mutation carrier based on family history and known mutation prevalence and penetrance (the probability of having a disease at a certain age given the mutation status) (Katki et al., 2007; Parmigiani et al., 2007).

Mendelian risk prediction models for various cancers have previously been developed and are available as part of the BayesMendel R pacakge (Chen et al., 2004). These models include BRCAPRO for identifying individuals at high risk for breast or ovarian cancer by calculating the probabilities of germline deleterious mutations in BRCA1 and BRCA2. MMRPro for identifying individuals at high risk of Lynch Syndrome by calculating the probabilities of germline deleterious mutation of the MMR genes: MLH1, MSH2, MSH6. PancPRO for identifying individuals at high risk for pancreatic cancer by calculating the probabilities of germline mutations in CDKN2A, PRSS1, BRCA2, and STK11. The inputs required for these models are the prevalence of deleterious mutations in the general population and their penetrance, which are researched and continually updated. These models have all been validated in the literature (Berry et al., 2002; Chen et al., 2006; Wang et al., 2007).

This paper focuses on BRCAPRO, but the methods developed could be extended to other Mendelian models. Women who carry a mutation in either BRCA1 or BRCA2 have an increased risk of developing breast and ovarian cancer. Mutations in BRCA1 and BRCA2 are rare in the general population (less than 0.2% of women have mutations (Easton et al., 1995)), but are more common in families with high rates of breast or ovarian cancers (Ford et al., 1998; Newman et al., 1997; Weber, 1996). Women with family history of breast or ovarian cancer will seek genetic counseling in order to determine their probabilities of being a BRCA1/BRCA2 carrier (Croyle and Lerman, 1999).

Risk prediction models, and in particular Mendelian risk prediction models, are based on family history which is used to determine which patients are at high risk of cancer. Screening and even treatment strategies have been developed for these high risk individuals (Murff et al., 2004). These models rely on self-reported family history, but inaccurate reporting could lead to inappropriate care. Risk predictions based on underreported (false negatives) family history, can lead to inadequate screening and substandard treatment. On the other hand, risk predictions based on over-reported (false positives) family history, can cause stress, unnecessary procedures and genetic testing (Douglas et al., 1999; Fry et al., 1999; Kerr et al., 1998; Murff et al., 2004; Sweet et al., 2002).

Various studies have evaluated the accuracy of self-reported family history. Murff et al. (2004) provide a review of these studies. Anton-Culver et al. (1996), conducted a population-based study with 359 probands with breast cancer. The probands were interviewed over the phone, and family history was verified using a cancer registry. Kerber and Slattery (1997) conducted a case-control study of colon cancer, with 125 colon cancer cases and 206 controls. They conducted personal interviews, and family history was verified using a cancer registry. Ziogas and Anton-Culver (2003) conducted a study including 1111 families, with both population-based probands and clinic-based probands (with the clinic-based accounting for 6.3% of the families). These probands had either breast, ovarian, or colon cancer. They conducted personal phone interviews followed by self completed reports, and verified using medical records and death certificates. Verkooijen et al. (2004) conducted a population-based study with 219 probands with breast cancer, out of which 110 had at least one relative with breast cancer, 9 had at least one relative with ovarian cancer, and 100 had no relatives with breast or ovarian cancers. Family history was collected using a self-reported survey and was verified using a cancer registry. More recently, Mai et al. (2011) collected data as part of the population-based 2001 Connecticut Family Health Study. Family history for 1,019 individuals was collected on breast, colorectal, prostate, and lung cancers through two phone interviews. The 1,019 participants reported family history for 20,578 first and second degree-relatives, of which a sample of 2,605 were validated. They were validated using state cancer registries, Medicare databases, the National Death Index, death certificates, and health-care facility records.

Sensitivity rates for the accuracy of cancer status reporting of first-degree relatives vary in these studies (Table 1), but are higher in breast cancer compared to ovarian cancer in all studies. For the first four studies sensitivity rates are higher than 80%, and specificity rates are above 90%. Lower sensitivity rates were reported in Mai et al. (2011), probably due to the fact that this was a population based study.

Risk prediction based on reported family history that is inaccurate will also be inaccurate. Suppose you have a family illustrated in Figure 1. The proband in this family underreports cancer in three of the relatives (the mother, a grandmother, and cousin), and misreports the age of diagnosis for an aunt and the grandmother. Based on the reported family history (shown on the left), the probability of being a BRCA carrier is 0.0779, the probability of being a BRCA1 carrier is 0.03407, and the probability of being a BRCA2 carrier 0.04384. Whereas, based on the true family history (shown on the right), the probability of being a BRCA carrier is 0.80314, the probability of being a BRCA1 carrier is 0.59420, and the probability of being a BRCA2 carrier 0.20830. This is an extreme case of misreporting, but in general risk prediction calculations based on the reported family may be very different than based on the true family history, and can lead to inadequate care.

Katki (2006) studies the effect of misreported family history on Mendelian risk prediction models in more detail. Family history is composed of two components; diagnosis (yes/no) and age-at-diagnosis (if diagnosis is yes) or current age or age of death (if diagnosis is no). Katki classifies the types of errors in the reporting of a relative's family history into three categories: (1) Diagnosis is incorrect but age-at-diagnosis is correct, (2) Diagnosis is correct but age-at-diagnosis is incorrect, (3) Both diagnosis and age-at-diagnosis are incorrect. He focuses on the first type of error, but shows the distortion caused by all three types of errors in BRCAPRO.

Katki (2006) considers underreporting of cancer disease status in three types of families: (a) proband alone (b) proband and first-degree relative (c) proband, first-degree relative, and second-degree relative. Katki derives the distortions that different scenarios of underreporting of cancer in these families would cause. Applying this to BRCAPRO, he shows that the worst error would be caused by underreporting of ovarian cancer in probands. This is seen because ovarian cancer yields higher penetrance density ratios than breast cancer. The weakest errors are underreporting of breast cancer in second-degree. In addition, Katki studies the misreporting of age of an affected relative by +/- 5 years and +/- 15 years. Considering both errors in underreporting of disease status and rounding of age, he shows that the errors in underreporting of disease status dominate.

There is extensive literature on measurement error in binary and continuous covariates (Carroll, 2006, among others). Family history, however, is time to event data (since it includes both the disease status and age), which is used as covariates in the Mendelian risk prediction models. We are not aware of any literature directly applicable to this setting in which both the disease status and age of disease can be reported with error. A proposed method to handle this type of error is described in detail elsewhere (Braun et al., 2014). Here, we apply this adjustment method to Mendelian risk prediction models, more specifically to BRCAPRO. We begin by introducing general notation, follow by describing the adjustment method, and illustrate the proposed model extension using real data.

4 Methods

4.1 Mendelian Risk Prediction Models

Mendelian models calculate the probability of being a mutation carrier given family history. Family history has two components; disease status and age. For those who develop disease, their age will be their age of diagnosis, for those who did not develop disease their age will be their current age or age of death. Let $T_i = (T_{i1}, ..., T_{iJ})$, $T_{ij} = \min(T^o_{ij}, C_{ij})$ where T^o_{ij} is the age of disease diagnosis j for individual i, C_{ij} is the current age or age of death for individual i for disease j ,and $\delta_i = (\delta_{i1}, ..., \delta_{iJ})$ is the disease status, where $\delta_{ij} = \mathbf{1}(T^o_{ij} \leq C_{ij})$ for individual i for disease j. Let $H_i = (T_i, \delta_i)$ and $H = (H_0, H_1, ..., H_R)$ for a family of size R. Let $\gamma_i = (\gamma_{i1}, ..., \gamma_{iM})$, where $\gamma_{im} = 1$ indicates carrying the genetic variants that confer disease risk for each individual i at a gene m, and $\gamma_{im} = 0$ otherwise. For example in BRCAPRO, $M = 2$ (BRCA1 and BRCA2).

Our goal is to calculate the proband's carrier probability $P(\gamma_0|H_0, H_1, .., H_R)$. Using Bayes rule, this can be calculated as follows [Blackford and Parmigiani (2010)]:

$$P(\gamma_0|H_0, H_1, .., H_R) = \frac{P(\gamma_0)P(H_0, H_1, ..., H_R|\gamma_0)}{\sum_{\gamma_0} P(\gamma_0)P(H_0, H_1, ..., H_R|\gamma_0)}. \tag{1}$$

$P(\gamma_0)$ is the prevalence of mutation carriers in the general population. $P(H_0, H_1, ..., H_R|\gamma_0)$ is the probability of the phenotypes for the entire family given the genotype for the proband. Using Bayes rule, this probability can be rewritten as:
$P(H_0, H_1, ..., H_R|\gamma_0) = \sum_{\gamma_1,...,\gamma_R} P(H_0, ..., H_R|\gamma_0, ..., \gamma_R)P(\gamma_1, ...\gamma_R|\gamma_0)$.
$P(\gamma_1, ...\gamma_R|\gamma_0)$ are known for all genotype combinations based on Mendelian laws of inheritance. Assuming conditional independence of phenotypes given genotypes implies: $P(H_0, ..., H_R|\gamma_0, ..., \gamma_R) = \prod_{i=1}^{R} P(H_i|\gamma_i)$, where $P(H_i|\gamma_i)$, the probability of phenotype given genotype, is referred to as the penetrance. Under these assumptions we can rewrite the proband's carrier probability as:

$$P(\gamma_0|H_0, H_1, .., H_R) = \frac{P(\gamma_0) \sum_{\gamma_1,...,\gamma_R} \prod_{i=1}^{R} P(H_i|\gamma_i)P(\gamma_1, ...\gamma_R|\gamma_0)}{\sum_{\gamma_0} P(\gamma_0) \sum_{\gamma_1,...,\gamma_R} \prod_{i=1}^{R} P(H_i|\gamma_i)P(\gamma_1, ...\gamma_R|\gamma_0)}. \tag{2}$$

4.2 Adjustment Method

Let H^* indicate the misreported family history; where $H_i^* = (T_i^*, \delta_i^*)$ and $H^* = (H_0, H_1^*, ..., H_R^*)$ for a family of size R. We assume the proband does not misreport his/her own history, therefore $H_0^* = H_0$. Our goal is to estimate $P(\gamma_0|H^*)$. The Mendelian risk prediction models previously developed assume true family history, since penetrance estimates for these models are based on a meta-analysis for which we expect most studies to be based on true family history. Therefore, these models calculate $P(\gamma_0|H)$. Using the total law of probability and Bayes rule, the probability of interest can be rewritten as:

$$P(\gamma_0|H^*) = \sum_H P(\gamma_0, H|H^*) = \sum_H P(\gamma_0|H^*, H)P(H|H^*) = \sum_H P(\gamma_0|H)P(H|H^*). \tag{3}$$

A sum is preformed in Equation (3) rather than an integral, since age in these models is treated as discrete with the following range; 1,...,120. The last equality in Equation (3) follows from the surrogacy assumption, that the carrier probability conditional on both the reported, H^*, and true, H, history is the same as the carrier probability conditional only the true history H. This assumption is plausible since this carrier probability should only be influenced by the true history. $P(\gamma_0|H)$ is then calculated using Equation (2), which is available as part of the BayesMendel R package. The measurement error process, $P(H|H^*)$, is estimated in a validation data set using smoothed Kaplan-Meier estimators. Since summing over all possible H is computationally intensive, we use Monte Carlo integration using $P(H|H^*)$ to select random H's.

This proposed approach assumes that the measurement error model $P(H|H^*)$ is transportable. Ideally, we would like to have a validation as similar as possible to the target population for this assumption to hold. There are scenarios for which $P(H|H^*)$ might not be transportable, but $P(H^*|H)$ is, in which case we could consider a different modeling approach, weighing the penetrance by $P(H^*|H)$. This approach is not the focus of this paper, but is described in detail in the appendix (A.1).

4.3 Measurement Error Model

We model the measurement error process $P(H|H^*)$ using validation data. Treating $H^* = (T^*, \delta^*)$ as covariates, we can model the measurement error process using a survival distribution assuming conditional independence of event and censoring times given T^*, δ^*:

$$P(T,\delta|T^*,\delta^*) = \lambda(T|\delta^*,\delta^*)^{\delta} S(T|T^*,\delta^*) h(T|T^*,\delta^*)^{1-\delta} G(T|T^*,\delta^*). \tag{4}$$

λ and S are the hazard and survival functions of the event time, and h and G are the hazard and survival functions of the censoring time. Depending on the data structure, one could estimate the survival distribution $P(T,\delta|T^*,\delta^*)$ parametrically (for example using a Weibull distribution or accelerated failure time model), semi-parametrically (for example using a Cox model), or non-parametrically (for example using Kaplan-Meier estimators).

We decide to use smoothed Kaplan-Meier estimators, requiring no parametric assumptions on the data. The model is stratified based on δ^*, and information is borrowed from neighborhoods based on the values of our continuous covariate T^* [Braun et al. (2014)].

4.4 Measurement Error Model Involving Multiple Cancers

Family history, H, contains multiple time to event data, corresponding to the multiple diseases in the model. The measurement error process for an individual i can be written as: $P(H_i|H_i^*) = P(T_i, \delta_i|T_i^*.\delta_i^*) = P(T_{i1}, ..., T_{iJ}, \delta_{i1}, ..., \delta_{iJ}|T_{i1}^*, ..., T_{iJ}^*, \delta_{i1}^*, ..., \delta_{iJ}^*)$. We propose different approaches to model this distribution.

The first is a competing risk approach which involves taking the first event for each individual. Define the first event as $T_i^o = min(T_{i1}^o, ..., T_{iJ}^o)$, and $T_i = min(T_i^o, C_i)$, $\delta_i = \mathbf{1}(T_i^o \leq C_i)$, where $\mathbf{1}(J = j)$ indicates a failure of type j occurred. Similarly ,we can define the error-prone failure times as the first event or censoring that is reported; T_i^*, δ_i^*, and $\mathbf{1}(J^* = j^*)$ indicating a failure of type j^* was reported. We propose treating (T^*, δ^*, j^*) as covariates and modeling the measurement error process using a survival distribution assuming conditional independence of event and censoring times given T^*, δ^*, j^*:

$$\begin{aligned} & P(T_i, \delta_i|T_i^*, \delta_i^*) = \\ & \lambda_j(T_i|T_i^*, \delta_i^*, j^*)^{\delta_i} S(T_i|T_i^*, \delta_i^*, j^*) h(T_i|T_i^*, \delta_i^*, j^*)^{1-\delta_i} G(T_i|T_i^*, \delta_i^*, j^*). \end{aligned} \tag{5}$$

λ_j is the cause-specific hazard and S is survival function of the event time, and h and G are the hazard and survival functions of the censoring time. We propose using smoothed Kaplan-Meier estimators, as before. The main limitation of this proposed approach is that it only uses the first event type.

Alternatively, one could model $P(H|H^*)$, as a multivariate distribution. This could be done either by assuming independence of the measurement error across diseases:
$P(T_{i1}, ..., T_{iJ}, \delta_{i1}, ..., \delta_{iJ}|T_{i1}^*, ..., T_{iJ}^*, \delta_{i1}^*, ..., \delta_{iJ}^*) = \prod_{j=1}^{J} P(T_{ij}, \delta_{ij}|T_{ij}^*, \delta_{ij}^*)$.
Or if the independence assumption does not hold, one could model

$P(T_{i1}, ..., T_{iJ}, \delta_{i1}, ..., \delta_{iJ} | T^*_{i1}, ..., T^*_{iJ}, \delta^*_{i1}, ..., \delta^*_{iJ})$ using a multivariate survival distribution (for example using frailty models). Our recommendation is to select an approach based on the context of the problem.

5 Results

5.1 Measurement Error Validation Data Source

We were able to obtain validation data from a family registry of probands with breast, ovarian, and colorectal cancers at the University of California at Irvine (UCI). Detailed information of the cohort characteristics can be found elsewhere [Ziogas et al. (2000)]. Ziogas and Anton-Culver (2003) provide details on the measurement error in this cohort. Briefly, this validation study had 1111 families which had at least one relative verified, of which 670 families had a proband affected with breast cancer, 123 families had a proband affected with ovarian cancer, and 318 families had a proband affected with colorectal cancer. Family history was collected from the proband by an initial phone interview. A verification report and pedigree was produced based on this phone interview. It was then mailed to the proband to complete items that were unknown and verify information. Family history of the relatives was verified by the following methods: (1) obtaining pathology reports, tumor tissue samples, or clinical records (2) obtaining self-reports from the relatives themselves (3) obtaining death certificates on deceased relatives.

Relatives were classified into first-degree: parents, siblings, and children of the proband; second-degree: grandparents, aunts, uncles, half-siblings, nieces, and nephews of the proband; and third-degree: first cousins and grandchildren of the proband. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), for cancer disease status were calculated for various types of cancer by the degree of the relative (only rates for first and second-degree relatives were reported) (Table 2). In general sensitivity and PPV were lower for second-degree relatives compared to first-degree relatives, and even lower for third-degree relatives (with the exception of brain, pancreas, female breast, and leukemia). NPV and specificity were high (greater than 0.95) for all cancer types and degree of relatives. For first-degree relatives PPV and sensitivity was lowest for cancers of the female pelvic organs (pelvic and endometrium) and bladder.

The focus of our analysis is on breast and ovarian cancers, and the data set we obtained for this analysis included 719 families with 1,521 female relatives which were validated. The average family had 2.1 female relatives that were verified. 294 relatives (19.3%) had breast cancer, and 70 relatives (4.6%) had ovarian cancer. Ziogas and Anton-Culver (2003) focus on misclassification of disease status. However, since family history consists of age and disease status, we will consider misreporting of age as well. Figure 2 shows the reported age versus validated age for breast and ovarian cancer. For breast cancer we have 59 relatives (3.9%) for which the reported and validated age are not equal. For ovarian cancer we have 89 relatives (5.9%) for which the reported and validated age are not equal. Both error-free and error-prone family history are available in this data set. We fit the measurement error model on this data set separately for breast and ovarian cancers.

5.2 CGN Model Validation Study

The Cancer Genetics Network (CGN) is a national network funded by the National Cancer Institute. The data obtained for this analysis consists of families with personal or family history of cancer, and includes 2,038 families with 34,310 relatives. The average family in this data set has 16.8 relatives. 3,143 relatives (9.2%) had breast cancer, and 610 relatives (1.8%) had ovarian cancer. Only error-prone, self-reported family history is available in this data set. This data set also contains BRCA1/2 testing results for each proband, in addition to family history. We apply the measurement error adjustment to this data set.

5.3 Applying the Proposed Adjustment Method to the CGN Model Validation Study

We use the UCI data to fit the measurement error model using smoothed Kaplan-Meier estimators (Beran, 1981) using the prodlim R package (Gerds, 2011). We develop a measurement error model for breast and ovarian cancers separately, based on Equation (4). For each cancer type, we stratify the model based on δ^*, and estimate survival and hazards for the failure time as well as censoring time distribution. Thus, four Kaplan-Meier estimators were obtained for each cancer type. For each of these estimators, a bandwidth is required. We select the optimal bandwidth by examining a grid of bandwidths. We examined a four-dimensional grid varying each bandwidth from 0.1 to 0.9 by increments of 0.2, and estimating the measurement error distribution given each quadruplets of bandwidths.

In order to select the optimal bandwidths, 50% of the CGN data was sampled and used to select the optimal bandwidths in terms of overall calibration of being a BRCA, BRCA1, and BRCA2 carrier. This sampling process was repeated ten times. For this data application, three different sets of four-dimensional bandwidths were selected in these ten iterations. In five out of the ten iterations the same four-dimensional set of bandwidths were selected. In the remaining five iterations, one set of four-dimensional bandwidths was selected three times, and one set of four-dimensional bandwidths was selected twice. The final set of four-dimensional bandwidths was the most frequently selected set, which was selected in half of these iterations.

BRCA testing results were available for all probands in this data set. The performance of the measurement error adjustment was evaluated using three different measures. The first, evaluating model calibration by looking at the observed over expected ratios (O/E). The second, evaluating the overall fit by looking at Brier scores. The third, evaluating model discrimination by looking at the area under the receiver operating characteristic curve (ROC-AUC). We also compare our proposed approach to an alternative approach correcting for measurement error assuming measurement error is independent of age. In other words, $P(T, \delta | T^*, \delta^*) = P(\delta|\delta^*)$. We estimate $P(\delta|\delta^*)$ (NPV and PPV) in the UCI validation study, and use it to adjust for measurement error as follows;

$$P(\gamma_0 | T^*, \delta^*) = \sum_T \sum_\delta P(\gamma_0 | T, \delta) P(\delta|\delta^*). \tag{6}$$

We will refer to this as the age independent approach. Alternatively, one could make an even stronger assumption and assume age is reported accurately and only disease status is reported with error. The measurement error adjustment would be as follows;

$$P(\gamma_0|T^* = T, \delta^*) = \sum_{\delta} P(\gamma_0|T, \delta)P(\delta|\delta^*). \tag{7}$$

Results from this approach are not shown in this paper.

Results of these measures based on error-prone family history compared to the full adjustment method and age independent adjustment method are shown in Table 3. The full adjustment method improves the O/E ratio for BRCA1 (1.0113 vs. 1.0734) and BRCA2 (0.9318 vs. 0.9157) although the O/R ratio for BRCA is worse (0.9728 vs. 1.0070). Brier scores are improved for BRCA (0.1393 vs. 0.1409) and BRCA2 (0.0570 vs. 0.0583), but not for BRCA1 (0.1021 vs. 0.1019). ROC-AUC is lower for BRCA, BRCA1, and BRCA2, using the full adjustment compared to based on error-prone family history. For the age independent approach, the O/E ratio is worse for BRCA (1.042 vs. 1.0070), BRCA1 (1.2857 vs. 1.0734), and BRCA2 (0.9136 vs. 0.9157) (Table 3). Brier scores are slightly higher using this approach for BRCA, BRCA1, and BRCA2, and ROC-AUC using this approach are lower for BRCA, BRCA1, and BRCA2.

Overall, the full adjustment preforms better than the age independent adjustment in terms of calibration and Brier score. It is important to note that model calibration for BRCA based on error-prone family history was 1.0070 to begin with. This implies that this data might not have strong rates of misreporting.

Figure 3 shows the log(O/E) stratified by risk deciles for BRCA, BRCA1, and BRCA2. Individuals were stratified into ten strata based on their probabilities of being a carrier given their error-prone family histories. The log of the O/E ratio was calculated within each strata along with 95% confidence intervals. BRCAPRO based on error-prone family history preforms poorly in the low deciles (corresponding to large O/E ratios). Both the full adjustment approach and age independent approach improve calibration especially in these low decile groups, by lowering the O/E ratios. This is of great clinical significance, since individuals in these deciles often face the clinical decision of whether or not to get tested for genetic mutations. Insurance companies often use cutoffs based on risk calculations to determine which individuals would be covered for genetic testing. Individuals in high risk deciles would qualify for testing, however individuals in low risk deciles might not qualify depending on their risk. Thus, it is especially important to have a well calibrated model for these individuals. The fact that O/E ratios in the low risk deciles is greater than one implies that BRCAPRO underestimates the risk for these individuals. This is of great clinical concern, as some of individuals who might be carriers are not tested due to this underestimation. Both adjustments improve model calibration for these individuals. Using the age independent approach, we see that calibration for BRCA1 for higher risk quantiles is worse than using the full adjustment.

5.4 Simulated Families

Extensive simulations have been conducted elsewhere (Braun et al., 2014). Since data sources containing both error-free and error-prone family histories are limited, we decide to illustrate our method on simulated families. These families were generated to have similar characteristics to the families in the CGN data set. We introduce error in disease status using sensitivity=0.649 and specificity=0.990 for breast cancer taken from Mai et al. (2011), and sensitivity=0.833 and specificity=0.989 for ovarian cancer taken from Ziogas and Anton-Culver (2003). We introduce error in age assuming an additive classical model; $T^* = T + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$, and $\sigma = 3$. We use the UCI data to fit the measurement error model for these simulations. The bandwidth was selected to minimize mean squared error in predictions.

Figure 4 compares predictions based on error-free, error-prone, and the full adjustment method in this simulation setting. Mean squared error in prediction was 0.02726 based on error-prone family history and 0.02620 based on the full adjustment. Results for being a BRCA carrier are shown in the top panel in red, BRCA1 carrier in the middle panel in purple, and BRCA2 carrier in the bottom panel in green. The left column shows predictions based on error-free family history compared to error-prone family history. In general there is more underreporting of cancer (individuals whose carrier probabilities are below the 45^o line) due to the low sensitivity in this data. The middle column shows predictions based on error-free family history compared to the full adjustment method. For BRCA1 we have more individuals whose predictions based on the adjustment method are higher than based on the error-free compared to BRCA2. The right column shows predictions based on error-prone family history compared to the full adjustment method. For BRCA and BRCA1 we see that the full adjustment increases the carrier probabilities for low risk individuals and decreases carrier probabilities for high risk individuals, whereas for BRCA2 the adjustment is not as strong. This is likely due to the fact that ovarian cancer, which has high rates of misreporting in the UCI data, affects the BRCA1 carrier status more than the BRCA2 carrier status.

In addition, we compare the full adjustment to an age independent adjustment using PPV and NPV estimates based on the UCI data. Mean squared error in prediction was 0.02726 based on error-prone family history and 0.04803 based on the age independent adjustment. The age independent method does not preform well in this simulation setting.

Overall, the full adjustment method improves predictions by increasing predictions for low risk individuals and decreasing predictions for high risk individuals.

6 Discussion

In this paper we propose a method to adjust for misreporting of family history in Mendelian risk prediction models. Previous literature has focused on evaluating miss-classification of disease status for various cancers by comparing proband reported family history to various gold standards. Katki (2006) studies the effects of misreporting of family history on Mendelian risk prediction models. In this paper, for the first time, we implement a measurement error adjustment in Mendelian risk prediction models.

We compare two methods to adjust for measurement error in the reporting of two cancers (breast and ovarian). The first using non-parametric smoothed Kaplan-Meier estimators, and the second using an age independent approach (using PPV and NPV estimates). Both in the context of the data application and in simulations, the full adjustment outpreforms the age independent adjustment. Using the full adjustment, we see improved calibration in BRCA1 and BRCA2, especially in low risk individuals.

It should be noted that one limitation of the UCI data is that it included only affected probands, meaning that the rates of misreporting might not be generalizable. Also, only a subset of the relatives for a given family were verified. For example, the average number of first-degree relatives for a proband with breast cancer was 6.8 out of which only 1.8 were verified, similarly the average number of second-degree relatives was 18.9 out of which only 2.3 were verified (these numbers are similar for families with probands having ovarian or colorectal cancer). Another limitation is that the data obtained represents only a subset of the families. More specifically these are families for which verification was done before July 2000. This might be a limitation since families who participated in the study earlier might be more likely to have a family history of cancer (Ziogas and Anton-Culver, 2003).

The CGN data on which we illustrate our method also has some limitations. The risk prediction model is well calibrated for being a BRCA carrier based on error-prone family history, implying that there might not be a lot of error in this data set. This data set consists of families with personal or family history of cancer. These families might be more aware of their family history, and therefore have lower rates of misreporting.

Self-reported family history is often reported with error. Inaccurate reporting of family history can lead to inappropriate care. For these reasons, methods to adjust predictions based on self-reported family history are of clinical significance. Insurance companies use fixed cutoffs to determine which patients can receive genetic testing. Without the measurement error adjustment, BRCAPRO is not well calibrated for low risk individuals. Our proposed adjustment improves calibration in this subpopulation. This will lead to better, more accurate clinical decisions for these individuals.

Given a good validation study for measurement error, Mendelian risk prediction models, such as BRCAPRO, can be extended to incorporate the proposed measurement error adjustment. This will allow clinicians and patients to obtain more accurate risk prediction estimates.

References

Anton-Culver H, Kurosaki T, Taylor T, Gildea M, Brunner D, Bringman D. 1996. Validation of family history of breast cancer and identification of the brca1 and other syndromes using a population-based cancer registry. Genetic epidemiology 13:193–205.

Beran R. 1981. Nonparametric regression with randomly censored survival data. Technical Report, Univ California, Berkeley .

Berry D, Iversen Jr E, Gudbjartsson D, Hiller E, Garber J, Peshkin B, Lerman C, Watson P, Lynch H, Hilsenbeck S, et al. 2002. Brcapro validation, sensitivity of genetic testing of brca1/brca2, and prevalence of other breast cancer susceptibility genes. Journal of Clinical Oncology 20:2701–2712.

Blackford A, Parmigiani G. 2010. Biomedical informatics for cancer research. Springer Verlag. chapter Familial Cancer Risk Assessment Using BayesMendel.

Braun D, Gorfine M, Katki H, Ziogas A, Parmigiani G. 2014. Nonparametric adjustment for measurement error in time to event data. Harvard University Biostatistics Working Paper Series .

Carroll R. 2006. Measurement error in nonlinear models: a modern perspective. volume 105. CRC Press.

Chen S, Wang W, Broman K, Katki HA, Parmigiani G. 2004. Bayesmendel: an r environment for mendelian risk prediction .

Chen S, Wang W, Lee S, Nafa K, Lee J, Romans K, Watson P, Gruber S, Euhus D, Kinzler K, et al. 2006. Prediction of germline mutations and cancer risk in the lynch syndrome. JAMA: the journal of the American Medical Association 296:1479.

Croyle R, Lerman C. 1999. Risk communication in genetic testing for cancer susceptibility. JNCI Monographs 1999:59–66.

Douglas F, ODair L, Robinson M, Evans D, Lynch S. 1999. The accuracy of diagnoses as reported in families with cancer: a retrospective study. Journal of medical genetics 36:309.

Easton D, Ford D, Bishop D. 1995. Breast and ovarian cancer incidence in brca1-mutation carriers. breast cancer linkage consortium. American Journal of Human Genetics 56:265.

Ford D, Easton D, Stratton M, Narod S, Goldgar D, Devilee P, Bishop D, Weber B, Lenoir G, Chang-Claude J, et al. 1998. Genetic heterogeneity and penetrance analysis of the brca1 and brca2 genes in breast cancer families. The American Journal of Human Genetics 62:676–689.

Fry A, Campbell H, Gudmundsdottir H, Rush R, Porteous M, Gorman D, Cull A. 1999. Gps' views on their role in cancer genetics services and current practice. Family Practice 16:468–474.

Gerds T. 2011. prodlim: Product limit estimation. R package version 19 .

Katki H. 2006. Effect of misreported family history on mendelian mutation prediction models. Biometrics 62:478–487.

Katki H, Blackford A, Chen S, Parmigiani G. 2007. Multiple diseases in carrier probability estimation: Accounting for surviving all cancers other than breast and ovary in brcapro. Johns Hopkins University, Dept of Biostatistics Working Papers :110.

Kerber R, Slattery M. 1997. Comparison of self-reported and database-linked family history of cancer data in a case-control study. American journal of epidemiology 146:244–248.

Kerr B, Foulkes W, Cade D, Hadfield L, Hopwood P, Serruya C, Hoare E, Narod S, Evans D. 1998. False family history of breast cancer in the family cancer clinic. European journal of surgical oncology 24:275–279.

Mai P, Garceau A, Graubard B, Dunn M, McNeel T, Gonsalves L, Gail M, Greene M, Willis G, Wideroff L. 2011. Confirmation of family cancer history reported in a population-based survey. Journal of the National Cancer Institute 103:788.

Murff H, Spigel D, Syngal S. 2004. Does this patient have a family history of cancer? JAMA: the journal of the American Medical Association 292:1480.

Newman B, Millikan R, King M, et al. 1997. Genetic epidemiology of breast and ovarian cancers. Epidemiologic reviews 19:69.

Parmigiani G, Chen S, Iversen Jr E, Friebel T, Finkelstein D, Anton-Culver H, Ziogas A, Weber B, Eisen A, Malone K, et al. 2007. Validity of models for predicting brca1 and brca2 mutations. Annals of internal medicine 147:441–450.

Sweet K, Bradley T, Westman J. 2002. Identification and referral of families at high risk for cancer susceptibility. Journal of clinical oncology 20:528–537.

Verkooijen H, Fioretta G, Chappuis P, Vlastos G, Sappino A, Benhamou S, Bouchardy C. 2004. Set-up of a population-based familial breast cancer registry in geneva, switzerland: validation of first results. Annals of oncology 15:350.

Wang W, Chen S, Brune K, Hruban R, Parmigiani G, Klein A. 2007. Pancpro: risk assessment for individuals with a family history of pancreatic cancer. Journal of clinical oncology 25:1417–1422.

Weber B. 1996. Genetic testing for breast cancer. Science and Medicine 3:12–21.

Ziogas A, Anton-Culver H. 2003. Validation of family history data in cancer family registries. American journal of preventive medicine 24:190–198.

Ziogas A, Gildea M, Cohen P, Bringman D, Taylor TH, Seminara D, Barker D, Casey G, Haile R, Liao SY, et al. 2000. Cancer risk estimates for family members of a population-based family registry for breast and ovarian cancer. Cancer Epidemiology Biomarkers & Prevention 9:103–111.

7 Illustrations

8 Tables

A Appendix

A.1 Alternative Method to Adjust for Measurement Error in Mendelian Risk Prediction Models

Rather than weighing the predictions by $P(H|H^*)$ as described in section 4.2 , one could consider weighing the penetrances by $P(H^*|H)$. We can write our desired model:

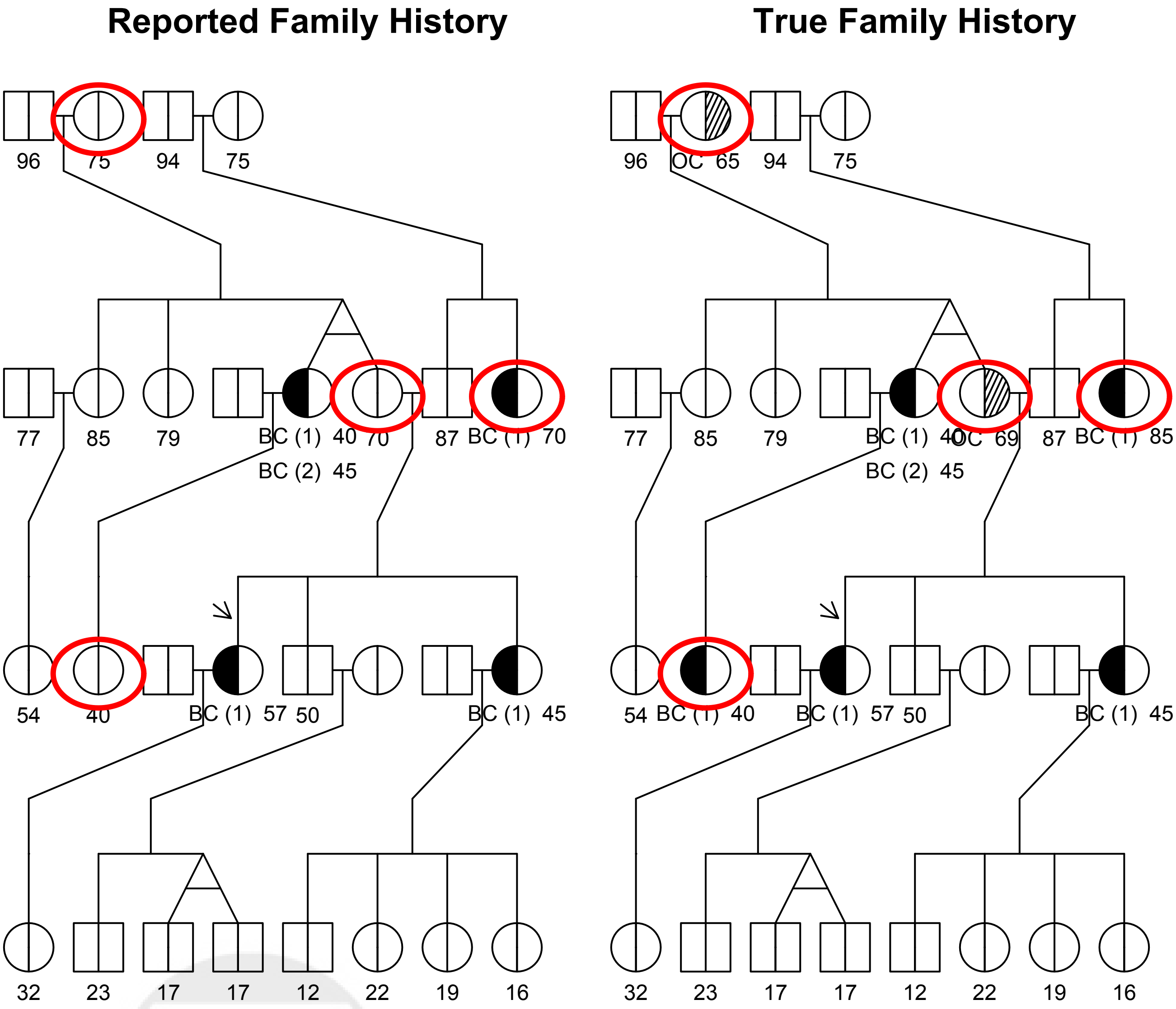
$$\begin{aligned} P(\gamma_0|H^*) \quad &= \frac{P(\gamma_0)P(H^*|\gamma_0)}{\sum_{all\gamma_0' s} P(\gamma_0)P(H^*|\gamma_0)} = \frac{P(\gamma_0)\sum_H P(H^*,H|\gamma_0)}{\sum_{all\gamma_0' s} P(\gamma_0)\sum_H P(H^*,H|\gamma_0)} \\ &= \frac{P(\gamma_0)\sum_H P(H^*|H,\gamma_0)P(H|\gamma_0)}{\sum_{all\gamma_0' s} P(\gamma_0)\sum_H P(H^*|H,\gamma_0)P(H|\gamma_0)} = \frac{P(\gamma_0)\sum_H P(H^*|H)P(H|\gamma_0)}{\sum_{all\gamma_0' s} P(\gamma_0)\sum_H P(H^*|H)P(H|\gamma_0)} \end{aligned}$$

We assume non-differential measurement error, meaning that the measurement error is independent of γ_0. This is plausible since we expect misreporting to be influenced by the true history but not by the carrier status.

Both this approach, and the one in section 4.2 are derived from basic principles. Furthermore, we show that the surrogacy assumption is equivalent to the non-differential misclassification assumption:

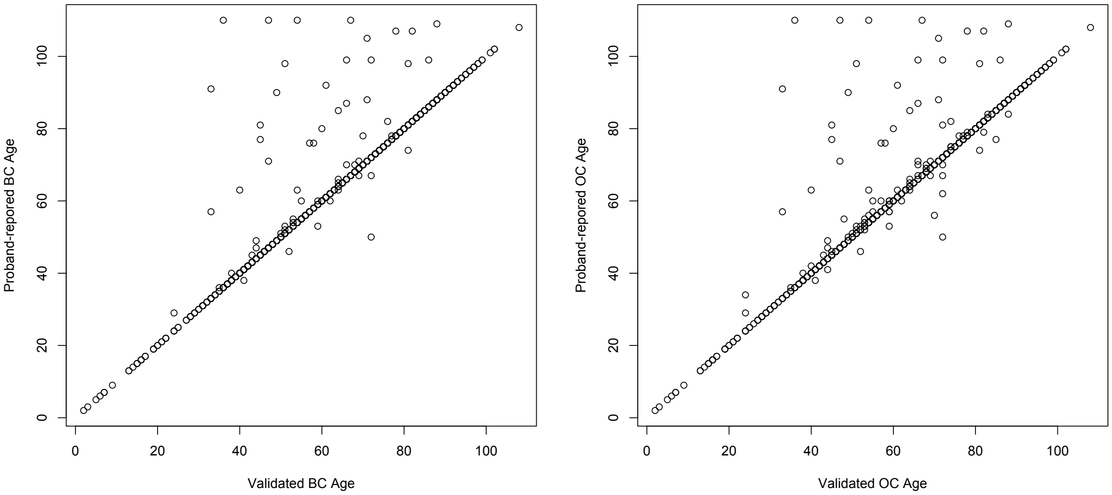
$$\begin{aligned} &P(H^*|H,\gamma_0) = P(H^*|H) \\ &= \frac{P(H^*,H,\gamma_0)}{P(H,\gamma_0)} = \frac{P(H^*,H)}{P(H)} \\ &= \frac{P(H^*,H,\gamma_0)}{P(H^*,H)} = \frac{P(H,\gamma_0)}{P(H)} \\ &= P(\gamma_0|H^*,H) = P(\gamma_0|H) \end{aligned}$$

The advantages of this approach is that it is less computationally demanding, and that the assumption of transportability might hold better for $P(H^*|H)$. However, using this approach $P(H^*|H)$ cannot be modeled using a survival distribution, therefore for the purpose of this paper we use the approach proposed in section 4.2.


Reported Family History

True Family History

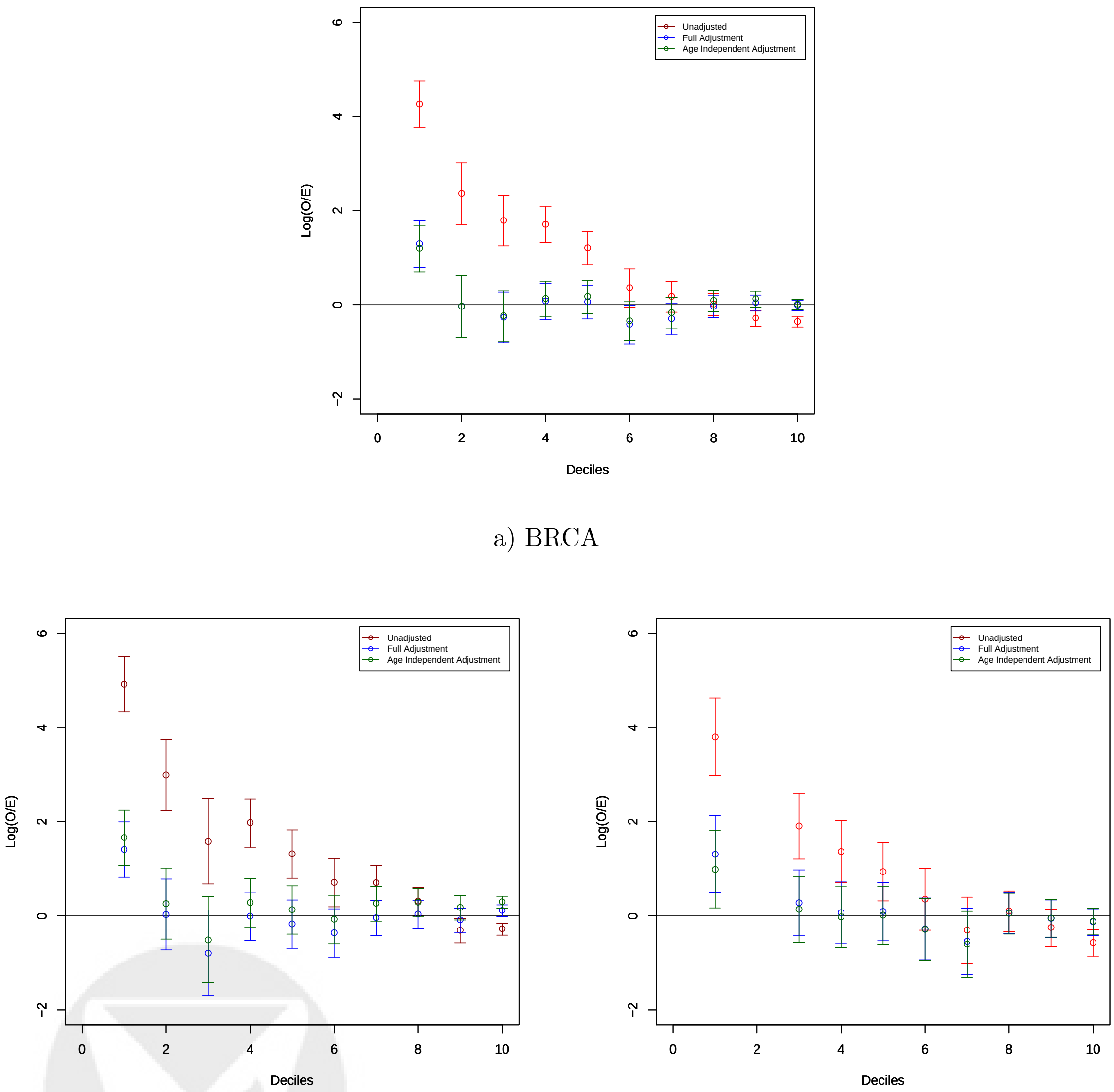
Figure 1: Sample Pedigree. A hypothetical example of a family with reported family history (on the left) and true family history (on the right). The proband is indicated by an arrow.



(a) Breast Cancer

(b) Ovarian Cancer

Figure 2: Reported versus validated ages of breast and ovarian cancers in female Relatives in UCI data.



a) BRCA

b) BRCA1

c) BRCA2

Figure 3: Log(O/E) and 95% confidence intervals for CGN families stratified by risk.

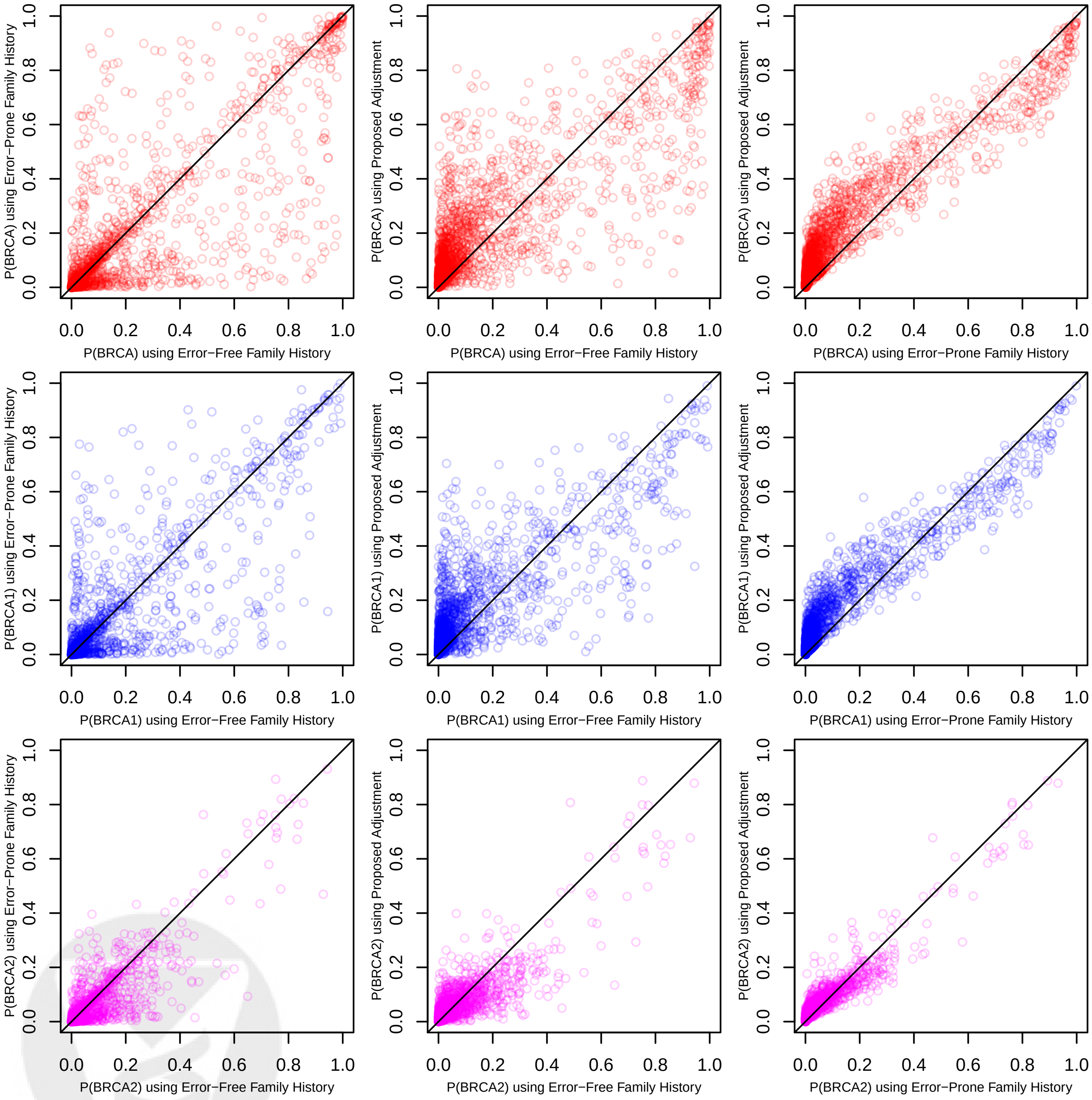


Figure 4: Carrier probabilities based on simulations comparing probabilities based on error-free, error-prone, and the proposed adjustment method. BRCA carrier probabilities (in red), BRCA1 carrier probabilities (in purple), and BRCA2 carrier probabilities (in green).

Table 1: Studies Evaluating Sensitivity and Specificity of Reported Cancer in First-Degree Relatives

Study	Type of Cancer	Proband Affected/Healthy/General Population	Sensitivity	Specificity
Anton-Culver et al. (1996)	Breast	Affected Probands	54/59(92%)	364/370(98%)
Kerber and Slattery (1997)	Breast	Affected Probands	11/13(85%)	107/112(96%)
Kerber and Slattery (1997)	Breast	Healthy Probands	18/22(82%)	167/184(91%)
Ziogas and Anton-Culver (2003)	Breast	Affected Probands	188/197(95%)	850/873(97%)
Verkooijen et al. (2004)	Breast	Affected Probands	60/61(98%)	247/249(99%)
Mai et al. (2011)	Breast	General Population	(65%)	(99%)
Kerber and Slattery (1997)	Ovarian	Affected Probands	2/3(67%)	117/122(96%)
Kerber and Slattery (1997)	Ovarian	Healthy Probands	1/2(50%)	201/204 (99%)
Ziogas and Anton-Culver (2003)	Ovarian	Affected Probands	35/42(83%)	1017/1028(99%)
Verkooijen et al. (2004)	Ovarian	Affected Probands	4/6(67%)	168/170(99%)

[a] Murff et al. (2004), Mai et al. (2011). Mai et al. (2011) report only sensitivity and specificity rates without providing the actual number of individuals

Table 2: Misreporting of Cancer Status in UCI Study, Ziogas and Anton-Culver (2003)

Type of Cancer and Relative	Sensitivity	Specificity	PPV	NPV
Breast, 1st Degree	95.4%	97.4%	89.1%	98.9%
Breast, 2nd Degree	82.4%	97.6%	89.6%	95.8%
Ovarian, 1st Degree	83.3%	98.9%	76.1%	99.3%
Ovarian, 2nd Degree	44.1%	98.5%	62.7%	96.8%

Table 3: Summary of results for CGN families applying proposed methods to adjust for measurement error

		O/E	ROC-AUC	Brier Score
BRCA	Error-Prone	1.0070	0.7769	0.1409
	Proposed Adjustment	0.9728	0.7424	0.1393
	Proposed Naive Adjustment	1.0421	0.7377	0.1401
BRCA1	Error-Prone	1.0734	0.7906	0.1019
	Proposed Adjustment	1.0113	0.7544	0.1021
	Proposed Naive Adjustment	1.2857	0.7544	0.1026
BRCA2	Error-Prone	0.9157	0.7244	0.0583
	Proposed Adjustment	0.9318	0.6907	0.0570
	Proposed Naive Adjustment	0.9136	0.6817	0.0574