Title
Nonparametric Adjustment for Measurement Error in Time-to-Event Data: Application to Risk Prediction Models

Journal
Journal of the American Statistical Association, 113(521)

ISSN
0162-1459

Authors
Braun, Danielle
Gorfine, Malka
Katki, Hormuzd A
et al.

Publication Date
2018-01-02

DOI
10.1080/01621459.2017.1311261

Peer reviewed

HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2019 January 01.

Published in final edited form as:

J Am Stat Assoc. 2018 ; 113(521): 14–25. doi:10.1080/01621459.2017.1311261.

Nonparametric Adjustment for Measurement Error in Time-to-Event Data: Application to Risk Prediction Models

Danielle Braun,
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, MA 02115

Malka Gorfine,
Department of Statistics, Tel Aviv University, Tel Aviv, Israel 6997801

Hormuzd A. Katki,
Division of Cancer Epidemiology and Genetics, Biostatistics Branch, National Cancer Institute, NCI Shady Grove, Room 7E606, 9609 Medical Center Drive, Rockville, MD 20850

Argyrios Ziogas, and
Department of Epidemiology, University of California Irvine, Irvine, CA 92697

Giovanni Parmigiani
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, MA 02115

Abstract

Mismeasured time to event data used as a predictor in risk prediction models will lead to inaccurate predictions. This arises in the context of self-reported family history, a time to event predictor often measured with error, used in Mendelian risk prediction models. Using validation data, we propose a method to adjust for this type of error. We estimate the measurement error process using a nonparametric smoothed Kaplan-Meier estimator, and use Monte Carlo integration to implement the adjustment. We apply our method to simulated data in the context of both Mendelian and multivariate survival prediction models. Simulations are evaluated using measures of mean squared error of prediction (MSEP), area under the response operating characteristics curve (ROC-AUC), and the ratio of observed to expected number of events. These results show

Correspondence to: Danielle Braun.

Research Fellow, Department of Biostatistics, Harvard School of Public Health and Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute (dbraun@hsph.harvard.edu). Malka Gorfine is an Associate Professor, Faculty of Industrial Engineering and Management, Technion- Israel Institute of Technology (gorfinm@ie.technion.ac.il). Hormuzd Katki is a Senior Investigator, Division of Cancer Epidemiology and Genetics, Biostatistics Branch, National Cancer Institute (katkih@mail.nih.gov). Argyrios Ziogas is an Associate Professor, Department of Epidemiology, University of California Irvine (aziogas@uci.edu). Giovanni Parmigiani is a Professor of Biostatistics, Department of Biostatistics, Harvard School of Public Health and Chair and Professor of Biostatistics, Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute (gp@jimmy.harvard.edu). Mailing address: Department of Biostatistics, Building 2, 4th Floor, 655 Huntington Avenue, Boston, MA 02115.

SUPPLEMENTARY MATERIALS

Web Appendix: A supplementary appendix with additional simulation results referenced in sections 5.1, 6, 7.
Code for simulations is available online; https://github.com/daniellebraun/NonParamAdjME.

that our method mitigates the effects of measurement error mainly by improving calibration and total accuracy. We illustrate our method in the context of Mendelian risk prediction models focusing on misreporting of breast cancer, fitting the measurement error model on data from the University of California at Irvine, and applying our method to counselees from the Cancer Genetics Network. We show that our method improves overall calibration, especially in low risk deciles.

Keywords

Survival Analysis; Mismeasured Covariates; Smoothed Kaplan-Meier Estimator; Carrier Status Prediction; Family History

1 Introduction

Measurement error in binary and continuous covariates has been studied extensively in the literature (Carroll et al., 2006, among others). The focus of this paper is on measurement error in time to event data which are used as predictors in an established risk prediction model. This work is motivated by Mendelian risk prediction models, which use Mendelian laws of inheritance to calculate the probability that an individual carries a cancer causing inherited mutation based on his/her family history. These models incorporate population parameters such as mutation prevalence, and penetrance (the probability of having a disease at a certain age given the person's genotype) (Murphy and Mutalik, 1969). Several of these models are in wide clinical use. All these models currently assume that family history is error-free. However, in practice they often rely on self-reported family history, which is not always accurate. This trend is increasingly relevant as models are being moved into primary care setting and web-based patient-oriented tools. The accuracy of self-reported family history has been evaluated in several studies which show that sensitivity and specificity estimates for reported disease status vary by degree of relative and type of cancer. For example, for breast cancer in first-degree relatives, sensitivity estimates vary from 65% to 95% while specificity estimates are usually around 98% − 99% (Mai et al., 2011; Ziogas and Anton-Culver, 2003).

More specifically, our interest is estimating risk of being a BRCA1/2 carrier for counselees (individuals seeking genetic counseling) in the Cancer Genetics Network (CGN). We estimate BRCA1/2 carrier risk using BRCAPRO (Berry et al., 1997; Parmigiani et al., 1998), a Mendelian risk prediction model identifying individuals at high risk of breast and ovarian cancer. CGN consists of families with personal or family history of cancer, and includes 2,038 families with 34,310 relatives. However, only error-prone self-reported family history is available. In addition to misreported family history, this data set also contains BRCA1/2 testing results for each counselee, which allows us to evaluate model performance. We also have data from a validation study conducted at the University of California at Irvine (UCI) (Ziogas and Anton-Culver, 2003) evaluating misreporting of family history. This study includes 719 cancer affected individuals with either breast, ovarian, or colon cancer for whom both error-free and error-prone family history are available. Self completed (error-prone) reports were collected from patients, and verified

using medical records and death certificates. Using this validation study, we propose an approach to adjust for misreporting of family history, and apply our proposed adjustment to counselees in CGN.

Time to event data are coded by two variables; T indicating either time to event or censoring whichever occurs first, and δ indicating whether the event occurred. We focus on scenarios in which both T and δ are measured with error. Because of the relationship between T and δ, standard techniques adjusting for measurement error in binary or continuous covariates cannot be applied directly. Previous work in this setting has focused on measurement error in survival outcomes (Meier et al., 2003). Meier et al. consider a discrete setting in which subjects are tested at predetermined time points until the time of first observed failure. Using the sensitivity and specificity rates of failing, they develop a model for the measurement error process based on a validation data set, and incorporate this into an adjusted proportional hazards model. Their method cannot be extended to our setting for two reasons. The first, that our time to event data is not obtained by repeated testing, instead we look at scenarios for which the time to event data is measured with error at one time point. The second, that our interest is in time to event data used as predictors in an existing risk prediction model and not as outcomes. We are not aware of any literature directly applicable to our setting.

The effects of misreported family history on Mendelian risk prediction models have been examined by Katki (2006). Both errors in underreporting of disease status and rounding of age were considered, and it was shown that misreporting of family history, especially in disease status, leads to distortions in predictions. A model based on these inaccurately assessed predictors will not be well calibrated. Katki (2006) studies the effects of misreporting but does not propose a method to adjust for the distortions in predictions, which is the focus of this work. More recently, Daniels et al. (2014) observed an underestimation of BRCA1/2 carrier probabilities in women with high-grade serous ovarian cancer. We hypothesize that some of the underestimation could be due to misreporting of family history (Braun et al., 2014b), motivating the clinical need for this work.

Although our work is motivated by the setting of Mendelian risk prediction models, where the interest is in the prediction of a binary variable (mutation carrier status) based on time-to-event data (family history), it is applicable to other scenarios, particularly in the context of survival prediction models where the interest is in the prediction of the probability of having an event by a given time based on time-to-event data. For example, suppose one is interested in predicting survival. In some disease settings, such as cancer, one possible predictor for survival is time to disease progression (TTP, the length of time until the disease starts to get worse or spread to other parts of the body), a time to event predictor. Suppose one has developed a model predicting survival based on error-free TTP. In practice, TTP in oncology clinical trials is often error-prone due to two main reasons; assessment of tumor size based on imaging varies by the observer and to a lesser extent the equipment, generating variation; and scans are taken at regularly scheduled intervals, generating rounding errors (Korn et al., 2010). Gray et al. (2009) evaluated measurement error in a similar endpoint, progression free survival (PFS, the length of time a patient lives with the disease but the disease does not get worse), by comparing PFS assessment by an independent review facility

(IRF) (which would be considered the error-free predictor) to an investigator-based local assessment (which would be considered the error-prone predictor). They conducted an independent review of trial E2100, an open-label multi-center, randomized, phase III trial conducted by the Eastern Cooperative Oncology Group (ECOG). They saw that for 6% of the patients a PFS event was only identified by IRF, and for 18.1% of the patients a PFS event was only identified by local review. 43.5% of patients had PFS events identified by both IRF and local review, and for those the date of PFS was the same for 54.5% of the patients and within 6 weeks for 70.4% of the patients.

In this context, suppose one has developed a model to predict survival based on IRF determined time to progression. In practice, one might want to use the prediction model to predict survival using as a covariate the time to progression determined by local review, and not by IRF, since local review might be the only feasible option. Another related example arises in the context of using time to event data on short-term outcomes as predictors for survival or for other long-term outcomes (Parast et al., 2012). If one has developed a model using error-free short term survival outcomes as predictors, but in practice these predictors are measured with error, our proposed method is again applicable.

It is important to note that our setting differs from typical measurement error settings involving error-prone covariates. In the typical setting, usually, only the error-prone covariate is observed in the main study, and the goal is to estimate the relationship between the outcome and the true covariate. In our setting, however, we have an already well-established risk prediction model, for which the relationship between the outcome and the true covariate is known. The goal is to use this model to estimate the risk based on an error-prone covariate. However, since the existing risk prediction model uses the relationship between the outcome and the true covariate, calculating the risk predictions based on the error-prone covariate will lead to biased results. In this work, we propose an approach to adjust for this bias.

Using a validation data set, we propose a nonparametric method to adjust for this type of measurement error, where the measurement error model is estimated by a conditional version of the kernel smoothed Kaplan-Meier estimator (Beran, 1981). This is presented in section 2. The proposed adjustment relies on two common assumptions that should be justified; the first that the measurement error model is transportable from the validation study to the main study, and the second of non-differential measurement error (surrogacy). In our motivating example, in the context of family history, both of these assumptions will likely hold. It is reasonable to assume that the mechanism driving the misreporting of family history in the validation study population would be the same as in the main study population. It is also reasonable to assume that the misreporting of family history is independent of the outcome (carrying a cancer causing inherited mutation). These assumptions are discussed in more detail in section 7.

We apply our proposed approach to Mendelian risk prediction models in section 3, and to other multivariate survival prediction models in section 4. Simulation results are presented in section 5. We illustrate our method using a data application in the context of Mendelian risk prediction models in section 6. The illustration focuses on misreporting of breast cancer,

fitting the measurement error model on data from UCI, and applying our method to counselees from the CGN. Finally, we summarize the main conclusions in section 7.

2 Proposed Method for Measurement Error Adjustment

Consider an outcome Y, and assume that time to event data are used as predictors in a model for Y. Specifically, let T^o be the true failure time, let C be the true right-censoring time, $T = min(T^o, C)$ and $\delta = 1(T^o \leq C)$. We denote the error-free predictor as $H = (T, \delta)$. We assume a model $P(Y|H)$ has been previously developed using this error-free predictor. For example, $Y = 1$ if the counselee is a carrier of genetic variants that confer disease risk, and 0 otherwise, and T^o is their mother's age at onset.

Now, suppose when implementing this model, the time to event data used as predictor has error. We denote this error-prone predictor as $H^* = (T^*, \delta^*)$. In our context, it could be that the counselee doesn't know that her/his relative had the disease, or she/he doesn't know the correct age-at-onset. We also assume we have a validation study which includes both the error-free time to event data, H, and the error-prone time to event data, H^*. We do not assume the validation data includes Y.

The available risk prediction model, $P(Y|H)$, uses the error-free predictor, however, in practice, H^* is available instead of H and our goal is to estimate $P(Y|H^*)$ for each new counselee. One could naively plug in H^* into the established risk prediction model, however this will lead to biased results. Instead, we propose to rewrite $P(Y|H^*)$ by applying the law of total probability and Bayes rule, as follows:

$$P(Y|H^*) = \int_H P(Y, H|H^*)dH = \int_H P(Y|H, H^*)P(H|H^*)dH = \int_H P(Y|H)P(H|H^*)dH \quad (1)$$

The last equality in Equation (1) follows from the non-differential measurement error (which is equivalent to surrogacy) assumption: we assume that H^* is a surrogate for H; that is, H^* contains no information on predicting Y in addition to the information already contained in H. This is plausible in many applications, including in Mendelian risk prediction models, when the probability of the outcome conditional on both the error-free and error-prone predictor is only influenced by the error-free predictor (i.e., carrier probability conditional on both the error-prone and true family history is only influenced by the true family history).

The risk prediction model, $P(Y|H)$, is already developed, and we provide an estimator of the measurement error distribution, $P(H|H^*)$, by the validation data. This approach assumes that the measurement error model $P(H|H^*)$ is transportable, meaning that the measurement error distribution estimated in the validation study can be applied to the population of interest. For this to be true, the validation and the target population should be as similar as possible. One should give thought to the choice of an appropriate validation study when applying our proposed method, as will be further discussed in section 7.

We propose to model $P(H|H^*)$ using hazard functions and assuming conditional independence of event and censoring times given T^*, δ^*:

$$P(T, \delta | T^*, \delta^*) = \lambda(T | T^*, \delta^*)^{\delta} S(T | T^*, \delta^*) h(T | T^*, \delta^*)^{1-\delta} G(T | T^*, \delta^*). \quad (2)$$

Here λ and S are the conditional hazard and survival functions of the event time T^o given H^*, and h and G are the conditional hazard and survival functions of the censoring time C given H^*. These hazard and survival functions are estimated by the validation data. Thus, when implementing the adjustment based on Equation (1) for a given counselee, $P(H|H^*)$ is estimated by the validation data, a large study population that does not involve the counselee.

In the validation study, one could estimate the survival distribution $P(T, \delta | T^*, \delta^*)$ parametrically (e.g. using a Weibull distribution), semi-parametrically (e.g. using a Cox model), or non-parametrically (e.g. using Kaplan-Meier estimators). In our implementations we have used the kernel smoothed Kaplan-Meier estimators with nearest neighborhoods (Beran, 1981), requiring no parametric assumptions on the hazard and survival functions involved. This approach is simple, flexible, and is applicable to any measurement error model.

More specifically, assume a validation study with n individuals, and for each individual i we observe $H_i = (T_i, \delta_i)$, where $T_i = min(T_i^o, C_i)$ and $\delta_i = 1(T_i^o \le C_i)$, and $H_i^* = (T_i^*, \delta_i^*)$. Let $N_i(t, l) = I(T_i \le t, \delta_i = 1, \delta_i^* - l)$, $l = 0, 1$, and the at-risk process $X_i(t, l) = I(T_i \ge t, \delta_i^* = l)$, $l = 0, 1$. We estimate the conditional survival function of the event time T^o given $H^* = (t^*, l)$, i.e. $S(\cdot | t^*, l)$, for any $0 < t^* \le \tau$, $\tau > 0$ a pre-specified constant, and $l = 0, 1$, by the following Beran (1981) estimator. Let $n_l = \sum_{i=1}^{n} I(\delta_i^* = l)$, $l = 0, 1$. Define K as a known kernel function and $\{b_{nl}\}$ as sequences of positive constants tending to 0 as $n_l \to \infty$, $l = 0, 1$, i.e. bandwidth sequences. Then,

$$\hat{S}(t | t^*, l) = \prod_{s \le t} [1 - d\hat{\Lambda}(s | t^*, l)] \quad (3)$$

where

$$\hat{\Lambda}(t | t^*, l) = \int_0^t \frac{\sum_{i=1}^{n} W_i(t^*; b_{n_l}, l) dN_i(s, l)}{\sum_{i=1}^{n} W_i(t^*; b_{n_l}, l) X_i(s, l)} \quad (4)$$

and

$$W_i(t; b_{n_l}, l) = I(\delta_i^* = l)\frac{K\left(\frac{t - T_i^*}{b_{n_l}}\right)}{\sum_{j=1}^{n} I(\delta_j^* = l)K\left(\frac{t - T_j^*}{b_{n_l}}\right)}. \quad (5)$$

The estimators of the conditional hazard functions of the event time T^o given $H^* = (t^*, l)$, i.e. $\hat{\lambda}(\cdot|t^*, l)$, are defined as the consecutive differences of $\hat{\Lambda}(\cdot|t^*, l)$. The conditional hazard and survival functions of the censoring distribution C given H^*, i.e. h and G, are obtained in a similar manner, treating the censoring times as events and the event times as censoring. Thus, our proposed estimator of $P(H|H^*)$ is defined by $\hat{P}(H|H^*) = \hat{\lambda}(T|T^*, \delta^*)^{\delta}\hat{S}(T|T^*, \delta^*)\hat{h}(T|T^*, \delta^*)^{1-\delta}\hat{G}(T|T^*, \delta^*)$. The proposed estimator will be consistent even when the n individuals consist of clustered data (as in the case of Mendelian risk prediction models). If one were to use a parametric approach to model $P(H|H^*)$, the dependence would need to be incorporated into the model (for example, using a copula or frailty model) which would require stronger assumptions.

While there is a relationship between T and δ, it is important to realize that when T contains measurement error, δ will not automatically be error-contaminated. The measurement error model, Equation (2), is stratified by δ^* and estimate it separately for $P(T, \delta|T^*, \delta^* = 0)$ and $P(T, \delta|T^*, \delta^* = 1)$. By estimating the measurement error model conditional on T^*, δ^* we are able to account for the relationship between T, δ and T^*, δ^*.

After obtaining the estimator $\hat{P}(H|H^*)$ defined above, we proceed to implementing the adjustment in Equation (1), which requires integration over all possible values of H. Depending on the data, integrating over all possible values of H might be computationally challenging. In these cases, we propose using Monte Carlo integration, and generating the Monte Carlo samples from $\hat{P}(H|H^*)$. More specifically, given H^* which is known, we sample B replicates of H: $H^{(1)}, \ldots, H^{(B)}$, from $\hat{P}(H|H^*)$; calculate $\hat{P}(Y|H^{(b)})$, $b = 1, \ldots, B$; and the final proposed estimator of $P(Y|H^*)$ is given by: $\hat{P}(Y|H^*) = B^{-1}\sum_{b=1}^{B}\hat{P}(Y|H^{(b)})$.

3 Mendelian Risk Prediction Models

Mendelian risk prediction models estimate the probability that a counselee carries an inherited susceptibility to a disease, based on information about his or her family history and are widely used in genetic counseling. Statistical software for evaluating these models is available as part of the BayesMendel R package (Chen et al., 2004), which includes BRCAPRO (Berry et al., 1997; Parmigiani et al., 1998), MMRPro a model identifying individuals at high risk of Lynch Syndrome, PancPRO a model identifying individuals at high risk of pancreatic cancer, and MelaPro a model identifying individuals at high risk of melanoma.

For the purpose of our discussion we consider a counselee who provides information on R members of his/her family, and focus on a single disease. Then, predictors are $H = (H_0, H_1,$

…, H_R). For family member i, $H_i = (T_i, \delta_i)$ and $T_i = min(T_i^o, C_i)$ where T_i^o is the age of disease diagnosis, C_i is the current age or age of death, and $\delta_i = \mathbf{1}(T_i^o \leq C_i)$, where $i = 0$ indicates the counselee. Mendelian models aim to estimate the counselee's carrier probability $P(\gamma_0|H)$, and $\gamma_i = (\gamma_{i1}, \ldots, \gamma_{iM})$, where $\gamma_{im} = 1$ indicates carrying the genetic variants that confer disease risk for each individual $i = 0, \ldots, R$ at a gene $m = 1, \ldots, M$, and $\gamma_{im} = 0$ otherwise. These models are based on known mutation prevalence, $P(\gamma)$ and penetrance $P(H|\gamma)$, where the penetrance is estimated from the literature and is a vector of discrete probabilities of disease conditional on carrier status, for ages=1, …, 110. Using Bayes rule and assuming conditional independence of family members' phenotype given their genotypes, we can write the counselee's carrier probability given his/her current disease status and his/her family history as follows:

$$P(\gamma_0|H_0, H_1, \ldots, H_R) = \frac{P(\gamma_0)\sum_{\gamma_1,\ldots,\gamma_R}\prod_{i=1}^{R}P(H_i|\gamma_i)P(\gamma_1,\ldots,\gamma_R|\gamma_0)}{\sum_{\gamma_0}P(\gamma_0)\sum_{\gamma_1,\ldots,\gamma_R}\prod_{i=1}^{R}P(H_i|\gamma_i)P(\gamma_1,\ldots,\gamma_R|\gamma_0)}. \quad (6)$$

Note that unaffected family relatives are informative for the risk calculation, as the likelihood of being a carrier will decrease when unaffected members are present. These models are typically trained using validated family history H. For example, the current version of the BRCAPRO model assessed $P(H_i|\gamma_i)$ via a meta-analysis of studies, the majority of which use family history information verified using medical records.

In practice, when these models are used clinically, the counselee provides his or her own recollection of the medical history of the family members. We refer to this as the reported history H^*. While validation of this history is sometimes possible, the majority of clinical implementations need to provide a carrier probability using H^* only. Normally, H^* is simply plugged in $\hat{P}(\gamma_0|H)$ instead of H. Our goal is to assess $P(\gamma_0|H^*)$, addressing the measurement error present in H^* for the counselee at hand, and at the same time leveraging the models that have been previously trained on validated data. Using Equation (1), this probability can be rewritten as follows:

$$P(\gamma_0|H^*) = \int_H P(\gamma_0|H)P(H|H^*)dH. \quad (7)$$

$P(\gamma_0|H)$ is estimated based on Equation (6), and the details can be found in (Berry et al., 1997; Parmigiani et al., 1998). The measurement error process, $P(H|H^*)$, is estimated by using kernel smoothed Kaplan-Meier estimators based on a validation study. More specifically, for a family of size R, $\hat{P}(H_1, \ldots, H_R|H_1^*, \ldots, H_R^*) = \prod_{i=1}^{R}\hat{P}(H_i|H_i^*)$, and $\hat{P}(H_i|H_i^*) = \hat{\lambda}(T_i|T_i^*, \delta_i^*)^{\delta_i}\hat{S}(T_i|T_i^*, \delta_i^*)\hat{h}(T_i|T_i^*, \delta_i^*)^{1-\delta_i}\hat{G}(T_i|T_i^*, \delta_i^*)$. $\hat{\lambda}$, $\hat{S}$, $\hat{h}$, $\hat{G}$ would be estimated based on a validation study (not including the counselee data). The validation data consists of independent families, each family consists of multiple relatives. Finally, the

integration is implemented using a Monte Carlo integration, as described earlier, and the H_i are sampled independently for each relative.

4 Survival Prediction Models

The problem of measurement error in time to event data also arises in survival prediction models. We continue our discussion of a hypothetical example in the context of predicting survival using time to progression (TTP). Assuming a prediction model for survival using error-free TTP has been developed, one might be interested in applying it in an environment where only error-prone TTP is available. In this scenario, we let $T_s = min(T_s^o, C_s)$ and $\delta_s = 1(T_s^o \leq C_s)$, where T_s^o is the death time, and C_s is the censoring time for death. We let $T_{ttp} = min(T_{ttp}^o, C_{ttp})$ and $\delta_{tpp} = 1(T_{ttp}^o \leq C_{ttp})$, where T_{ttp}^o is the progression time, and C_{ttp} is the censoring time for progression. We let $H = (T_{ttp}, \delta_{ttp})$ be error-free TTP, and $H^* = (T_{ttp}^*, \delta_{ttp}^*)$ be error-prone TTP. The existing prediction model based on error-free TTP is $P(T_s^o > t|H)$, while our application requires $P(T_s^o > t|H^*)$. Using Equation (1) we can rewrite this probability as:

$$P(T_s^o > t|H^*) = \int_H P(T_s^o > t|H)P(H|H^*)dH\,. \quad (8)$$

$P(H|H^*)$ would be estimated using validation data containing paired error-free TTP and error-prone TTP and will be based on Equation (2). Studies such as the one conducted by Gray et al. (2009), would be a good source of validation data, since they compared progression assessment conducted by IRF review (error-free TTP) to progression assessment conducted by local review (error-prone TTP).

5 Simulations

5.1 Mendelian Risk Prediction Models

We begin with simulations whose goal is to quantify the impact of adjusting for measurement error in the context of Mendelian risk prediction models. For each simulation scenario, we generated two data sets; the first (the measurement-error-estimation study) is used to model the measurement error distribution; the second (the model-evaluation study) is used to estimate the carrier probability of each counselee given their family history, and evaluate our method.

For our measurement-error-estimation study, we simulated 100,000 families with 5 members (mother, father, and three daughters). These simulations focus on one gene, BRCA1, and only on breast cancer. For the marginal carrier probability, $P(\gamma = 1)$, we assumed the value 0.006098, which is the estimated allele frequency of BRCA1 in the Ashkenazi Jewish population. We simulated error-free breast cancer failure times for each member based on fixed penetrance functions for $P(H|\gamma)$, $\gamma = 0, 1$, the same used by BRCAPRO version 2.08. We simulated error-free censoring times from a truncated normal distribution with mean 55

and standard deviation 10 (truncated at 0 so that no ages are below 0). We used a large sample because the allele frequency for BRCA1 is low.

We considered two settings for the measurement error in disease status. The first using sensitivity= 0.954 and specificity=0.974, taken from Ziogas and Anton-Culver (2003), the second using sensitivity=0.649 and specificity=0.990, taken from Mai et al. (2011). We considered four settings for the measurement error in age. For the first three settings, we assume an additive classical model; $T^* = T + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$, and either $\sigma = 5$, 3 or 1. For the fourth setting, we assume a multiplicative measurement error model, $T^* = TU$, where $U \sim \exp(1)$. We estimate $P(H|H^*)$ using smoothed Kaplan-Meier with the nearest neighborhoods kernel using the *prodlim* R package (Gerds, 2015). We calculated the optimal bandwidths using the direct plug in approach proposed by Sheather and Jones (1991).

For our model-evaluation study, we generated 50,000 counselees, whose family history were generated in a similar manner. For each of the 50,000 families, we calculated carrier probability for the counselee based on family history using BRCAPRO for three estimators: a) based on the simulated error-free history $\hat{P}(\gamma_0|H)$; b) based on the simulated error-prone history by naively replacing H by H^*, denoted by $\tilde{P}(\gamma_0|H^*)$, and c) using our proposed adjusted estimator $\hat{P}(\gamma_0|H^*)$ (Table 1). For our adjusted approach, we applied Monte Carlo integration by sampling 100 configurations of H from $\hat{P}(H|H^*)$.

We used three different performance measures, which are standard in the risk prediction literature (Steyerberg et al., 2010), to evaluate the estimators introduced above. Calibration was evaluated by the ratio of observed to expected events (O/E) (a well calibrated model will have O/E close to 1), accuracy of prediction was evaluated by mean squared error of prediction (MSEP), and discrimination was evaluated by the area under the receiver operating characteristics curve (ROC-AUC). More specifically, O/E, can be written as: $\sum_{i=1}^{n} \mathbf{1}(\gamma_i = 1) / \sum_{i=1}^{n} \hat{P}_i$, where $\hat{P}_i$ is replaced by each of the three predictive probabilities above, in turn. We define MSEP, as the mean of the squared differences between the $\hat{P}_i$ and the error-free predictions, or $(1/n)\sum_{i=1}^{n} \{\hat{P}_i - \hat{P}_i(\gamma_0|H)\}^2$. Therefore, MSEP based on the error-free data is always 0. We used the *verification* R package (Gilleland, 2009) which calculates ROC-AUC following the process outlined by Mason and Graham (2002).

Table 1 provides the results. In summary, the ratios of observed to expected events (O/E) based on the error-free family history are reasonably close to 1 in all the simulation settings. The O/E ratios based on the error-prone family history are lower than one, when using sensitivity=0.954 and specificity=0.974, and higher than one when using sensitivity=0.649 and specificity=0.990, with the exception of the scenarios involving multiplicative error in age, for which the O/E is always less than one. A specificity of 0.974 implies that 2.6% of truly unaffected relatives are reported to be affected. The majority of relatives in these simulations are unaffected, therefore this specificity drives a large number of predictions to be higher than they should be, and the O/E to be less than 1. For lower sensitivity, the underreporting of disease drives the O/E to be greater than one, since the expected probabilities are lower. In all the settings, the proposed adjusted estimator improves the O/E substantially, and shifts it closer to 1. For example, in the first simulation scenario, the O/E

is 0.9773 based on error-free family history, 0.8190 based on error-prone family history, and 0.9712 based on the adjustment. Thus, we are able to eliminate almost all the bias induced by errors in reported histories.

MSEP based on adjusting the error-prone data is lower than MSEP based on the error-prone data alone for all simulations, by an amount that varies but can be substantial. For example, in the first simulation scenario the square root of the MSEP multiplied by 1000, is 19.1351 based on the error-prone family history, and 16.7405 based on the adjustment. ROC-AUC in all simulations are higher based on the error-free data compared to the error-prone data. In simulations involving additive error in age, ROC-AUC values are only slightly improved using our adjustment compared to the error-prone data alone. In some cases no improvement was observed. For example, in the first simulation scenario ROC-AUC was 0.8160 based on error-free family history, 0.8090 based on error-prone family history, and 0.8086 based on the adjustment. However, it is not in general the case that the adjustment improves calibration without affecting discrimination. To illustrate, we consider a scenario with multiplicative error in age, where the error in age is likely stronger than what should be expected in real genetic counseling applications. In this case, predictions are substantially reordered by our adjustment, and ROC-AUC is higher compared to the error-prone data alone. For example, in the fourth simulation scenario in Table 1, ROC-AUC was 0.8145 based on error-free family history, 0.7185 based on error-prone family history, and 0.8020 based on the adjustment.

Figure 1 compares the three predictions corresponding to the first row in Table 1, representing a simulation setting with sensitivity 0.954 and specificity 0.974. The first column on the left, shows predictions based on error-free family history compared to error-prone family history. The majority of the families have carrier probabilities less than 0.2, therefore we present a close up of these families in the second row. There is more over-reporting (in the plot these are individuals who are above the 45° line), due to the lower specificity. We have many individuals close to the 45° line, corresponding to simulated families for which error-prone and error-free family histories were very similar. The second column in the middle, shows predictions based on error-free family history compared to our adjustment. In families with high carrier probabilities, we see more individuals below the 45° line, implying our adjustment method slightly over adjusts by shifting probabilities down. Even though this is the case, overall, model calibration is improved (Table 1). In the bottom row, for families with lower carrier probabilities, we see individuals both above and below the 45°. Thus, although it appears that our adjustment method over adjusts for some individuals, the overall O/E shows improvement in calibration. The third column on the right, shows predictions based on error-prone family history compared to our adjustment. We can see that especially in this simulation scenario (which has more over-reporting than underreporting of cancer), our adjustment shifts individuals' carrier probabilities down. Additional analysis of a second simulation scenario, corresponding to the fifth row in Table 1, representing a simulation setting with sensitivity 0.649 and specificity 0.990, can be found in Web Appendix A.

Overall, the proposed adjustment method improves MSEP and calibration in all scenarios, while ROC-AUC either remains the same or improves depending on the scenario.

5.2 Survival Prediction Models

We also preformed simulations in the context of predicting survival based on TTP. For these simulations, we consider a hypothetical scenario in which one has developed a prediction model for survival using error-free TTP as a predictor. In reality, however, TTP is measured with error, as shown in the E2100 review conducted by Gray et al. (2009). Based on the results of this review, Korn et al. (2010), performed simulations to assess the potential bias of measurement error on the conclusions of a proportional hazards analysis in randomized trials. We followed a similar approach in generating the data for our simulations. Using the notation introduced in section 4, we let (T_s, δ_s) indicate the survival data, and (T_{ttp}, δ_{ttp}) indicate the TTP data. To mimic a plausible counseling scenario, the goal is to predict survival of a patient; $P(T_s^o > t | T_{ttp}^*, \delta_{ttp}^*)$, where t is fixed and equal for all patients and T_s^o is the time to death from either a progression event or the patient's last visit to the clinician's office.

Similarly to the Mendelian risk prediction model simulations of Section 5.1, we generated two types of study; the measurement-error-estimation study used to model the correct TTP given the error-prone TTP, and the model-evaluation study, used to estimate the survival probability for each individual given their TTP, and evaluate our method. In addition to these, we generated a third study type to train the prediction model for survival based on error-free TTP. In the Mendelian risk prediction model simulations, we used an existing risk prediction model (BRCAPRO). In this section, an existing model was not available and so we generated data to fit one. In each simulation run we generated all three study types anew.

We begin by describing the training data set and the fitting of the prediction model. We assume a scenario in which patients were followed for progression for a period of 25 months. Some of these patients will have a progression event during this time interval, whereas others will be censored. We assume that censoring represents the patient's last visit to the clinician's office.

We generated survival data (T_s, δ_s) as well as error-free TTP data (T_{ttp}, δ_{ttp}) for 1,000 patients. We began by generating progression event times (T_{ttp}) based on a Weibull distribution with shape parameter=1.456 and scale parameter=11.063 (based on E2100). We then generated censoring times (C_{ttp}) assuming a censoring distribution between 0 and 25 months with density $f(t) = 2(25 - t) / 625$, resulting in approximately 55% censoring. Individuals who had a progression time smaller than their censoring time, were assumed to have a progression event at their progression time. The remainder were assumed to not have a progression event.

Next we generated survival. Survival, T_s, was measured from the time an individual either had a progression event or the individual's last visit to the clinician's office. To simplify our choice of scenarios, we first simulated δ_s from a Bernoulli distribution with the following probabilities; for those who had a progression event, we assumed that $P(\delta_s = 1 | \delta_{ttp} = 1) = 0.5$, for those who did not have a progression event, we assumed that $P(\delta_s = 1 | \delta_{ttp} = 0) = 0.1$. We generated T_s given the value of δ_s. To simplify model evaluation, we assumed that following a progression event or the last encounter, all individuals are followed for survival

until either death or the end of the follow-up period at time $c > t$. So for those who survive ($\delta_s = 0$), we assigned them an observed time $T_s = c$ months. For those who died ($\delta_s = 1$) and had a progression event ($\delta_{ttp} = 1$), we assigned them a survival time, T_s, by taking the absolute value of a random sample from a $N(12, 5^2)$. For those who died and did not have a progression event ($\delta_{tpp} = 0$), we assigned them a survival time, T_s, by taking the absolute value of a random sample from a $N(48, 5^2)$. This generates a dependence between survival and TTP. We fit a prediction model $P(T_s^o > t | T_{ttp}, \delta_{ttp})$ for $t = 60$ using the kernel smoothed Kaplan-Meier estimators of the *prodlim* R package (Gerds, 2015) for $\delta_{ttp} = 0$ and 1, separately. For these simulations 1,000 patients provide a large enough sample, since events are generated at a high rate (around 45%).

Moving to the measurement-error-estimation study, we generated error-free TTP, (T_{ttp}, δ_{ttp}), as well as error-prone TTP, $(T^*_{ttp}, \delta^*_{ttp})$, for 1,000 individuals. We simulated error-free TTP as we did in the training data set. We then generated error-prone TTP conditional on the error-free TTP: we introduced error in TTP event indicators using various sensitivities and specificities (E2100 had 88% sensitivity and 64% specificity). We generated T^*_{ttp} depending on δ_{ttp} and δ^*_{ttp} as follows. If $\delta_{ttp} = 1$, $\delta^*_{ttp} = 1$; 55% of the time we assumed complete agreement in the time of the event (based on E2100); $T^*_{ttp} = T_{ttp}$. For others, 35% were within 6 weeks of each other. Therefore, we used multiplicative measurement error with log-normal distribution and standard deviation log(1.5) (Korn et al., 2010) to generate T^*_{ttp}. If $\delta_{ttp} = 0$, $\delta^*_{ttp} = 0$, we assumed agreement in the time; $T^*_{ttp} = T_{ttp}$. If $\delta_{ttp} = 1$, $\delta^*_{ttp} = 0$, we assigned T^*_{ttp} to be the simulated censoring time. If $\delta_{ttp} = 0$, $\delta^*_{ttp} = 1$, we assigned T^*_{ttp} to be the minimum of the simulated progression failure time and end of study (25 months). We obtained error-free and error-prone TTP, and fit the measurement error model, $P(T_{ttp}, \delta_{ttp} | T^*_{ttp}, \delta^*_{ttp})$ using kernel smoothed Kaplan-Meier estimators based on each simulated data set.

Finally, for the model-evaluation study, we generated error-prone and error-free TTP as well as survival data, for 1,000 individuals as we did for the first two data sets. We preformed survival prediction calculations for each subject in this data set. We compared three different prediction calculations; the first using the error-free TTP as a covariate in the prediction model $\hat{P}(T_s^o > t | T_{ttp}, \delta_{ttp})$, the second by naively replacing the error-free TTP by the error-prone TTP as a covariate and calculating $\tilde{P}(T_s^o > t | T^*_{ttp}, \delta^*_{ttp})$, and the third using our measurement error adjustment $\hat{P}(T_s^o > t | T^*_{ttp}, \delta^*_{ttp})$. All the three predictions above were done for $t = 60$. Although we have a time to event outcome, predictions were calculated for a fixed t, and for model evaluation, ROC-AUC and calibration, we consider a dichotomized outcome, $Y = \mathbb{1}(T_s^o > t)$.

In addition, we compared our proposed method to an alternative approach of modeling the measurement error process, which considers event indicators but not time, that is

$P(T_{tt\,p}, \delta_{tt\,p} | T^*_{tt\,p}, \delta^*_{tt\,p}) = P(\delta_{tt\,p} | \delta^*_{tt\,p})$. We estimated $P(\delta_{tt\,p} | \delta^*_{tt\,p}), \delta_{tt\,p} | \delta^*_{tt\,p} = 0, 1$ in the measurement-error-estimation study, and used it to adjust for measurement error as follows; $P(T^o_s > t | T^*_{tt\,p}, \delta^*_{tt\,p}) = \sum_{\delta_{tt\,p}=0,1} P(T^o_s > t | T^*_{tt\,p}, \delta_{tt\,p}) P(\delta_{tt\,p} | \delta^*_{tt\,p})$. We refer to this as the time independent adjustment.

We conducted simulations for various values of sensitivity and specificity, varying in increments of 0.1 from 0.1 to 1. Figure 2 summarizes the O/E for the four methods considered, across all choices of sensitivity and specificity. In summary, both the full adjustment and the time independent adjustment preform very well; the O/E ratios are very close to one and vary only slightly across sensitivity and specificity. As expected, the O/E ratios based on error-free covariates are even closer to one. The O/E ratios based on the error-prone covariates decrease as sensitivity increases, and increase as specificity increases. A low sensitivity corresponds to underreporting of events, corresponding to higher O/E ratios, whereas a low specificity corresponds to over-reporting of events, corresponding to lower O/E ratios. Thus, as sensitivity increases, O/E decreases; as specificity increases, O/E decreases.

The full adjustment preformed best in terms of MSEP compared to both no adjustment and the time independent adjustment (Figure 3), with the exception of a few scenarios with very high sensitivity and specificity. MSEP based on error-prone covariates decreases as sensitivity and specificity increase, that is, as we introduce less error MSE improves. For both adjustment methods, MSEP increases and then decreases as sensitivity increases, for lower specificities, while for higher specificities MSEP decreases as sensitivity increases.

In general, ROC-AUC was highest using the full adjustment method compared to both no adjustment and the time independent adjustment (Figure 4). ROC-AUC was largest based on the error-free covariates. In theory ROC-AUC based on the error-free covariates should be constant. Small variability in the ROC-AUC based on the error-free covariates was observed (green line in Figure 4) due to the variability in the data generation for each simulation scenario. ROC-AUC based on error-prone covariates increases as sensitivity and specificity increase, that is, as we introduce more error ROC-AUC decreases. For both adjustment methods, ROC-AUC decrease and then increase as sensitivity increases, for lower specificities, while for higher specificities ROC-AUC increases as sensitivity increases.

Overall, the full adjustment method preform best in terms of calibration, overall model accuracy, and discrimination.

6 Data Application

We illustrate our proposed method using a data application in the context of Mendelian risk prediction models focusing on misreporting of breast cancer. For the counselees, we used 2,038 families from the CGN Model Evaluation Study. For these families only self-reported (misreported) family history is available therefore we are unable to estimate the extent of measurement error in the CGN data. 9.2% of the 34,310 relatives were reported to have breast-cancer. Our interest is in estimating the risk of being a BRCA1/2 carrier based on the

available misreported family history using our proposed adjustment, which we can compare to the genetic testing results for BRCA1/2 mutations which are known for all counselees.

We use data from UCI in order to estimate the measurement error process. Briefly, this study includes 719 participants who report family history on 1,521 female relatives, 19.3% of whom are reported to have breast-cancer. Cancer diagnosis and age of onset was verified using medical records and death certificates. Therefore, for these individuals we know both the error-prone and error-free family histories. Sensitivity and specificity for misreporting of breast cancer in this study are 80% and 97% respectively. Although specificity is high, even 3% of unaffected relatives being reported as affected will have a larger impact, as the majority of the relatives are unaffected. Misreporting of age of diagnosis is less frequent, age of breast cancer was misreported for 3.1% of the relatives, with an average 4.5 years difference between the true and misreported ages, and age of ovarian cancer was misreported for 4.2% of the relatives, with an average 4.2 years difference between the true and misreported ages (more details presented in Web Appendix B).

We estimated the measurement error model in the UCI data as described earlier. The average family size in the UCI data is 2.11 relatives per family, therefore we fit a measurement error process on this entire population rather than stratifying by the degree of the relative (the proposed smoothed Kaplan Meier estimator will still be consistent under this setting). We then applied our proposed adjustment for BRCA1/2 carrier prediction, for each counselee in the CGN data. Additionally, we use BRCAPRO to estimate the probability of each counselee being a carrier for a mutation given her error-prone family history. Using the true BRCA1/2 carrier status, we calculated O/E ratios, and ROC-AUC based on the error-prone family history as well as our proposed adjustment. MSEP as defined previously cannot be calculated since H is not observed, however, we can calculate the Brier score, which is defined as the mean of the squared differences between the $\hat{P}_i$ and the event indicator $\mathbb{1}(\gamma_{0i} = 1)$ as $(1/n)\sum_{i=1}^{n}\{\hat{P}_i - \mathbb{1}(\gamma_{0i} = 1)\}^2$.

The UCI data contains a relatively small number of events, and therefore the kernel smoothed Kaplan-Meier was sensitive to bandwidth selection. Rather than estimating the optimal bandwidth using the direct plug in approach proposed by Sheather and Jones (1991), the bandwidths were selected so that calibration of the probability of being a BRCA, BRCA1, and BRCA2 carrier in the CGN dataset were closest to 1. Therefore, this example does not provide a completely independent validation of the calibration performance.

The respective O/E ratios of being a BRCA, BRCA1, and BRCA2 carrier are 1.007, 1.073, and 0.916 based on error-prone family history; and 0.976, 1.037, and 0.892 based on the adjustment. The respective Brier scores for being a BRCA, BRCA1, and BRCA2 carrier are 0.141, 0.102, 0.058 based on error-prone family history; and 0.139, 0.102, and 0.057 based on the adjustment. The respective ROC-AUC for BRCA, BRCA1, and BRCA2 carriers are 0.777, 0.791, and 0.725 based on the error-free family history; and 0.776, 0.787, and 0.722 based on the adjustment.

Overall, the adjustment results in a slight improvement in Brier score but a slightly worse ROC-AUC. Even with the error-prone family history, the BRCAPRO model is well

calibrated overall for being a BRCA carrier, but not as well calibrated for a BRCA1 carrier or BRCA2 carrier separately. The adjustment improves BRCA1 calibration, while the calibration of BRCA2 is slightly worse.

In clinical applications, it is important for risk prediction models to preform well in low risk deciles, since insurance companies will often approve genetic testing only for individuals whose estimated carrier probability is above a cutoff which is relatively low, such as 5% or 10%. Therefore, we further examined model calibration by looking at the O/E ratios of being a BRCA carrier, stratified by risk (Figure 5). Individuals were ordered by their probabilities of being a BRCA carrier based on the error-prone family history, and stratified into deciles. O/E ratios as well as 95% confidence intervals were calculated for each stratum. Using error-prone family history, the model is not well calibrated in the low risk deciles. The O/E ratio is greater than one in these deciles, implying that the model underestimates the risk for these individuals. Our proposed adjustment improves calibration in the low risk deciles by an extent which we expect will lead to better clinical decisions.

7 Discussion

In this paper we explore a method to adjust for measurement error in time to event data. Previous literature has focused on measurement error in survival outcomes, but not on measurement error in a time to event predictor. Our proposed method is applicable to both Mendelian risk prediction models and survival prediction models. Simulations studies in both of these settings show that, when implementing these models using error-prone time to event data, the models are miscalibrated. Our proposed adjustment improves model calibration and total accuracy across all simulation scenarios. Model discrimination either remains the same or is improved, depending on the amount of error introduced in the simulation setting.

In practice, different populations will have different amounts of measurement error. For example Mai et al. (2011) focused on general population and showed higher rates of misreporting compared to Ziogas and Anton-Culver (2003) whose study included only affected probands. It is important that the measurement error model be estimated in a population that is reflective of the population to which the final algorithm will ultimately be applied. In these settings we would recommend deploying our methods, as even a relatively small correction has the potential of providing clinically meaningful improvement, as seen in the CGN data illustration. Since insurance companies use a low cutoff to approve genetic testing, the improvement in calibration in the low risk deciles can have a direct impact, increasing the number of individuals being referred to genetic testing.

Existing software for Mendelian risk prediction models does not report the uncertainty in the estimated risk for an individual i. One approach to estimate the uncertainty for a new counselee i are the resampling procedures proposed by Parmigiani et al. (1998); Gorfine et al. (2013). These address the uncertainty about the model parameters (prevalence and penetrance) by drawing these parameters from a distribution, and for each draw calculating the counselee's carrier probability. This approach can be extended to incorporate the

additional uncertainty due to the variability of the measurement error model estimates, by drawing this parameter from a distribution as well.

Our proposed method relies on two main assumptions. The first is non-differential measurement error, which is equivalent to assuming that H^* is a surrogate for H. While this may be a reasonable assumption in the context of family history, there could be scenarios in which this assumption would be violated. The non-differential measurement error allows us to use an existing and reliable model for $P(Y|H)$, which is a key component of our proposed methodology. Therefore the methods proposed in this paper would not be helpful if this assumption was violated. The second assumption is that the measurement error distribution, $P(H|H^*)$ is transportable from the validation study to the main study. This assumption should be given careful thought, as there may be scenarios for which it is violated. Ideally, we would have multiple validation studies and would be able to test this assumption. Unfortunately, validation studies assessing the accuracy of family history are costly and not widely conducted. Furthermore, data from these studies is not often publicly accessible.

In the context of the UCI validation study, we assume that the mechanisms driving the measurement error (recall error, being unaware of family history, etc), would be similar to the CGN population where we apply the correction. CGN contains probands with cancer and/or a family history of cancer. While the CGN data does contain unaffected probands (and the UCI contains only affected probands), both CGN and UCI are high risk populations, and it is reasonable to assume that the measurement error would be similar across the two populations. When applying this correction to the general population, using the UCI data to fit the measurement error model could be a limitation, as it may not represent error rates realistically. Studies such as the one conducted by Mai et al. (2011) can be used to estimate the measurement error model in the general population. We hope this work will motivate both additional research as well as data sharing and future collaborations, so that we can apply this correction to the general population

There may be scenarios for which $P(H|H^*)$ is not transportable, but it may be plausible to assume that $P(\delta|\delta^*)$ (PPV, NPV) is (that is, a measurement error process which does not depend on time). In such settings we recommend performing the adjustment by estimating the measurement error process $P(\delta|\delta^*)$. We illustrate this approach, referred to as time independent, in simulations, and show that even in simulations where the measurement error depends on time it performs quite well.

There may also be settings for which $P(H|H^*)$ is not transportable, yet $P(H^*|H)$ is. One key component of our adjustment is the estimation of $P(H|H^*)$ using a survival distribution. On the other hand, $P(H^*|H)$ cannot be estimated using standard survival analysis methodology. While T^* and δ^* are dependent, they are not dependent through the traditional censoring mechanism. Since both T^* and δ^* are error-prone, the familiar definitions for the survival time and censoring no longer hold. Therefore the likelihood for $P(H^*|H)$ is not available in standard form. Extending our proposed method to the setting in which $P(H|H^*)$ is not transportable, yet $P(H^*|H)$ is, would be more intensive in terms of calculations, it is an ongoing project and will be addressed in a separate communication.

If the covariate distribution differs across the two population, then there could be settings in which one wouldn't expect $P(H|H^*)$ to be transportable, however it may be plausible to assume that the model would be transportable if we condition on additional covariates (such as counselee's age, race, education, income, etc). In such settings, we would recommend extending the measurement error process to incorporate these additional covariates. If several continuous covariates are being added, a semi-parametric modeling approach could be considered instead of the kernel smoothed Kaplan-Meier approach.

The conditional independence assumption used in Equation (2) means that for all t, within any subgroup of individuals defined by H^*, the subjects who are censored at time t are representative, with respect to their survival experience, of all the subjects in that subgroup who are at risk at time t. In case one suspects that additional covariates (e.g. status of other diseases) should be included (beside H^*) for this conditional independence assumption to hold, such covariates should be included in Equation (2) as well.

The simulations performed in this paper in the context of survival prediction, model survival using time to progression as a covariate. More generally, the method proposed in this paper can be used in any survival prediction model in which the time to event covariate is measured with error. For example, it can be used in the setting of estimating the probability of the counselee remaining free of disease until time t, $P(T^o > t|H^*)$. Mendelian models for $P(T^o > t|H)$ were developed similarly to what is described above, and are also available as part of the BayesMendel R package. Using Equation (1), our proposed method can be implemented in this context: $P(T^o > t|H^*) = \int_H P(T^o > t|H)P(H|H^*)\, dH$.

The main advantage of the nonparametric kernel smoothed Kaplan-Meier estimators is not to require parametric assumptions on the form of the measurement error model. In simulations, with large number of events in the validation data, the estimator was robust to bandwidth selection in terms of calibration (Web Appendix C), and therefore optimal bandwidths were selected using the direct plug in approach proposed by Sheather and Jones (1991). However, when there are relatively few observed events in the validation data, as in the UCI data, the kernel smoothed Kaplan-Meier estimators might be sensitive to bandwidth selection. In this setting, the optimal bandwidth was selected based on the criterion that calibration of the probability of being a BRCA (BRCA1 or BRCA2) carrier in the CGN dataset be closest to 1. Alternatively, if the data allows, one could use a semi-parametric or parametric approach to model the measurement error process. This could be advantageous in terms of model performance, however, in our settings, since T and T^* are highly correlated, a Cox proportional hazards model did not converge due to a monotone likelihood. If one considers a parametric model, one should be aware that observations in the setting of family data are dependent, and this dependence would need to be incorporated into the model.

The CGN Model Evaluation Study through which we illustrate our method has some limitations. First, it may not be representative of the measurement error patterns seen in unselected populations, as it consists mostly of families who self-select for genetic counseling. Second, the Mendelian risk prediction model we used for our analysis considers only one disease, whereas in real applications, Mendelian risk predication models include multiple diseases. An extension of this work to multiple diseases is presented elsewhere

(Braun et al., 2014a). Even with these limitations, we see that our proposed method improved calibration in the low risk deciles.

In addition, the UCI validation data used to estimate the measurement error model is relatively small. However, in simulations our proposed approach was robust to the validation study sample size, and performed well for sample size as low as $N = 100$ as long as the validation study includes both events and non-events (Web Appendix C). The total number of carriers in the CGN Model Evaluation Study is also relatively small. Efforts should be made to obtain larger model evaluation study, which could lead to substantial improvements to the current BRCAPRO carrier predictions using our proposed adjustment technique.

While in theory one could develop new risk prediction models based on error-prone family history, this is not optimal in the context of Mendelian risk prediction models. Mendelian risk prediction models are based on the estimation of two parameters; penetrance function and prevalence. The penetrance function is the probability that an individual is diagnosed with the disease at age t, $t = 0, 1, \ldots, 110$, given the carrier status. Mendelian risk prediction models are not developed by fitting an outcome model conditional on family history, but rather by using Mendelian laws of inheritance, Bayes' rule, and known mutation prevalence and penetrance function. Therefore, when we say the original model was developed based on true family history, we mean that the penetrance function estimates in the original model are based on true disease status and age. Studies estimating these penetrance function are based on individuals who undergo genetic testing and whose disease status and age is generally known with great accuracy (for example, based on medical records). They are not based on family histories, but rather based on individuals who are tested. Therefore, penetrance-function estimate can be considered as being based on true disease status and age. Thus by using these estimates as parameters in the original models, we generate a model that applies to error-free family history.

Mendelian risk predictions models are based on published penetrance estimates and their meta-analysis, which leverages the extensive information available. Although error-prone penetrances could theoretically be estimated by estimating the penetrance based on family history; asking the proband to report their family history, and extrapolating the proband's own genetic testing results to other family members (without testing the other family members), this is not done in practice, and is even less likely to be done in the future as genetic testing becomes more widely available. Estimating an error-prone penetrance would move modeling in the direction of using data of much lower quality. In addition, our proposed approach has the advantage of allowing the adjustment of misreporting of family history in some settings and not others. It also allows for population specific adjustments by estimating different measurement error processes targeting different populations, for example; elderly population, high risk, primary care clinic.

A recent study by Daniels et al. (2014) used BRCAPRO to calculate the risk for being a BRCA1/2 carrier in women with high-grade serous ovarian cancer, and showed that the risk is underestimated. In response to this study, we hypothesized that some of the underestimation could be due to misreporting of family history, and used a simplified version of the methods proposed in this paper to illustrate how adjusting for the misreporting

will lead to better calibrated predictions (Braun et al., 2014b). The underestimation of risk observed by Daniels et al. (2014) illustrates the clinical impact in using methods to adjust for misreporting in this setting.

The method proposed in this paper can be used to generalize the algorithms currently incorporated into the BayesMendel R package, and will be of direct clinical use. Self-reported family history is often affected by inaccuracies which could lead to inappropriate care (Murff et al., 2004). Underreporting (false negatives) of cancer in the family gives rise to an underestimation of cancer risk, which can result in inadequate screening and substandard treatment (Murff et al., 2004). On the other hand, over-reporting (false positives) of cancer, gives rise to an overestimation of cancer risk, which can cause stress (Douglas et al., 1999), unnecessary procedures and unnecessary genetic testing (Kerr et al., 1998; Sweet et al., 2002; Fry et al., 1999). For these reasons, we hope that the methods proposed in this paper will be of clinical significance.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We gratefully acknowledge support from the National Cancer Institute at the National Institutes of Health [5T32CA009337-32 to Parmigiani]. Work of Malka Gorfine is supported by the Israel Science Foundation (ISF) grant 2012898.

References

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

Berry D, Parmigiani G, Sanchez J, Schildkraut J, Winer E. Probability of carrying a mutation of breast-ovarian cancer gene brca1 based on family history. Journal of the National Cancer Institute. 1997; 89(3):227–237. [PubMed: 9017003]

Braun D, Gorfine M, Katki H, Ziogas A, Anton-Culver H, Parmigiani G. Extending mendelian risk prediction models to handle misreported family history. Harvard University Biostatistics Working Paper Series. 2014a

Braun D, Gorfine M, Parmigiani G. Misreported family histories and underestimation of risk. Journal of Clinical Oncology. 2014b; 32(32):3682–3683. [PubMed: 25199751]

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement error in nonlinear models: a modern perspective. CRC press; 2006.

Chen S, Wang W, Broman KW, Katki HA, Parmigiani G. Bayesmendel: an r environment for mendelian risk prediction. Statistical Applications in Genetics and Molecular Biology. 2004; 3(1): 1–19.

Daniels MS, Babb SA, King RH, Urbauer DL, Batte BA, Brandt AC, Amos CI, Buchanan AH, Mutch DG, Lu KH. Underestimation of risk of a BRCA1 or BRCA2 mutation in women with high-grade serous ovarian cancer by brcapro: A multi-institution study. Journal of Clinical Oncology. 2014; 32(12):1249–1255. [PubMed: 24638001]

Douglas FS, O'Dair LC, Robinson M, Evans DGR, Lynch SA. The accuracy of diagnoses as reported in families with cancer: a retrospective study. Journal of Medical Genetics. 1999; 36(4):309–312. [PubMed: 10227399]

Fry A, Campbell H, Gudmundsdottir H, Rush R, Porteous M, Gorman D, Cull A. GPs' views on their role in cancer genetics services and current practice. Family Practice. 1999; 16(5):468–474. [PubMed: 10533942]

Gerds T. prodlim: Product-limit estimation for censored event history analysis. R package version. 2015; 1(5)

Gilleland E. Verification. R package version. 2009; 1(41)

Gorfine M, Hsu L, Parmigiani G. Frailty models for familial risk with application to breast cancer. Journal of the American Statistical Association. 2013; 108(504):1205–1215. [PubMed: 24678132]

Gray R, Bhattacharya S, Bowden C, Miller K, Comis RL. Independent review of E2100: a phase III trial of bevacizumab plus paclitaxel versus paclitaxel in women with metastatic breast cancer. Journal of Clinical Oncology. 2009; 27(30):4966–4972. [PubMed: 19720913]

Katki HA. Effect of misreported family history on mendelian mutation prediction models. Biometrics. 2006; 62(2):478–487. [PubMed: 16918912]

Kerr B, Foulkes WD, Cade D, Hadfield L, Hopwood P, Serruya C, Hoare E, Narod SA, Evans DG. False family history of breast cancer in the family cancer clinic. European Journal of Surgical Oncology. 1998; 24(4):275–279. [PubMed: 9724992]

Korn EL, Dodd LE, Freidlin B. Measurement error in the timing of events: effect on survival analyses in randomized clinical trials. Clinical Trials. 2010; 7(6):626–633. [PubMed: 20819840]

Mai PL, Garceau AO, Graubard BI, Dunn M, McNeel TS, Gonsalves L, Gail MH, Greene MH, Willis GB, Wideroff L. Confirmation of family cancer history reported in a population-based survey. Journal of the National Cancer Institute. 2011; 103(10):788–797. [PubMed: 21562245]

Mason SJ, Graham NE. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Quarterly Journal of the Royal Meteorological Society. 2002; 128(584):2145–2166.

Meier AS, Richardson BA, Hughes JP. Discrete proportional hazards models for mismeasured outcomes. Biometrics. 2003; 59(4):947–954. [PubMed: 14969473]

Murff HJ, Spigel DR, Syngal S. Does this patient have a family history of cancer?: an evidence-based analysis of the accuracy of family cancer history. Journal of American Medical Association. 2004; 292(12):1480–1489.

Murphy EA, Mutalik GS. The application of bayesian methods in genetic counselling. Human Heredity. 1969; 19(2):126–151.

Parast L, Cheng S-C, Cai T. Landmark prediction of long-term survival incorporating short-term event time information. Journal of the American Statistical Association. 2012; 107(500):1492–1501. [PubMed: 23293405]

Parmigiani G, Berry DA, Aguilar O. Determining carrier probabilities for breast cancer–susceptibility genes BRCA1 and BRCA2. The American Journal of Human Genetics. 1998; 62(1):145–158. [PubMed: 9443863]

Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society. 1991:683–690.

Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology. 2010; 21(1):128. [PubMed: 20010215]

Sweet KM, Bradley TL, Westman JA. Identification and referral of families at high risk for cancer susceptibility. Journal of Clinical Oncology. 2002; 20(2):528–537. [PubMed: 11786583]

Ziogas A, Anton-Culver H. Validation of family history data in cancer family registries. American Journal of Preventive Medicine. 2003; 24(2):190–198. [PubMed: 12568826]

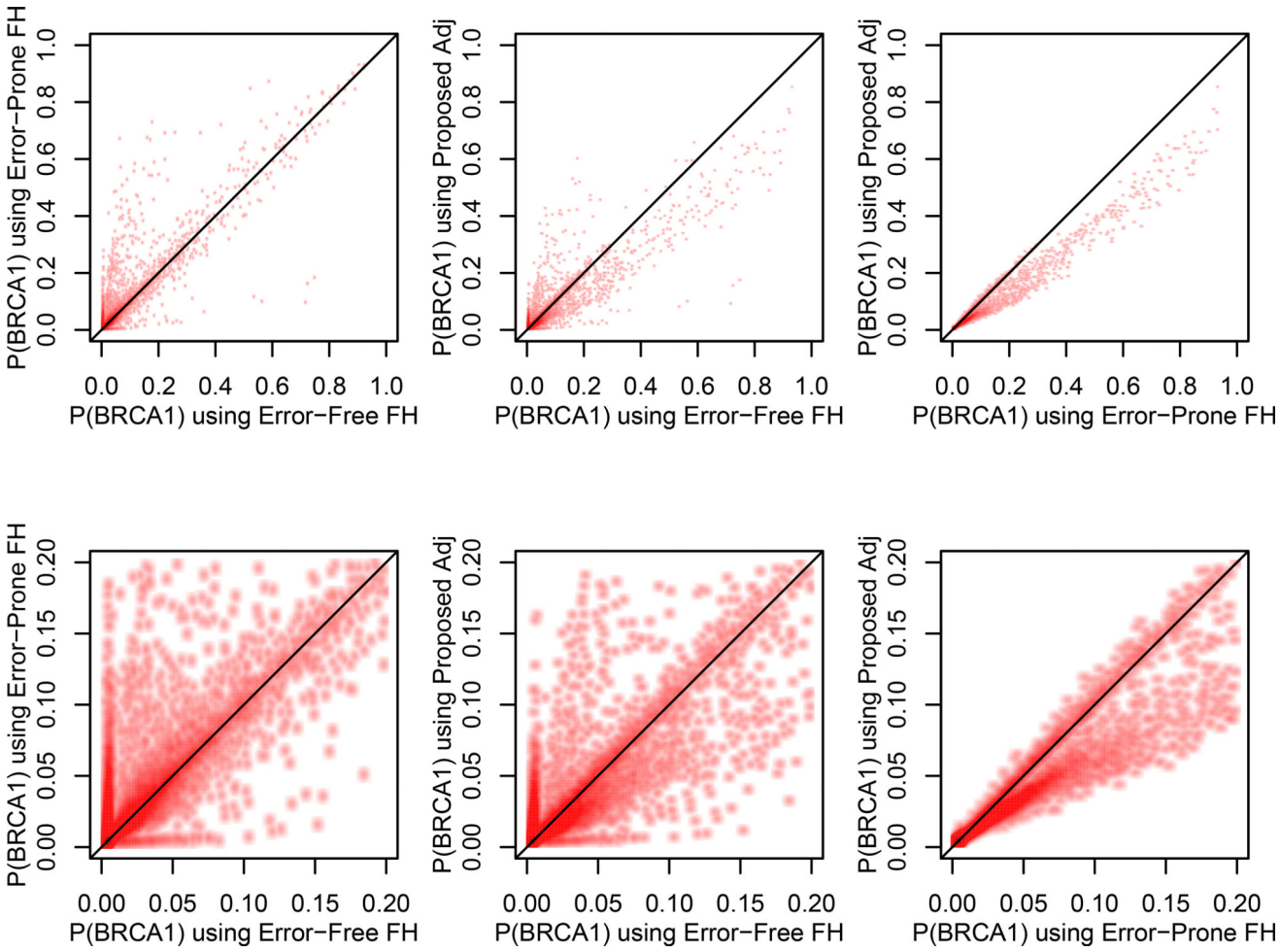

Figure 1.

P(BRCA1) for simulated families based on error-free family history, error-prone family history, and the proposed adjustment. In red, a simulation setting with sensitivity 0.954 and specificity 0.974, classical additive model for error in age $T^* = T + \varepsilon$ where $\varepsilon \sim N(0, 5^2)$, and the counselee is the mother. The bottom row is a close up version of the top row, focusing on counselees with carrier probabilities less than 0.2.

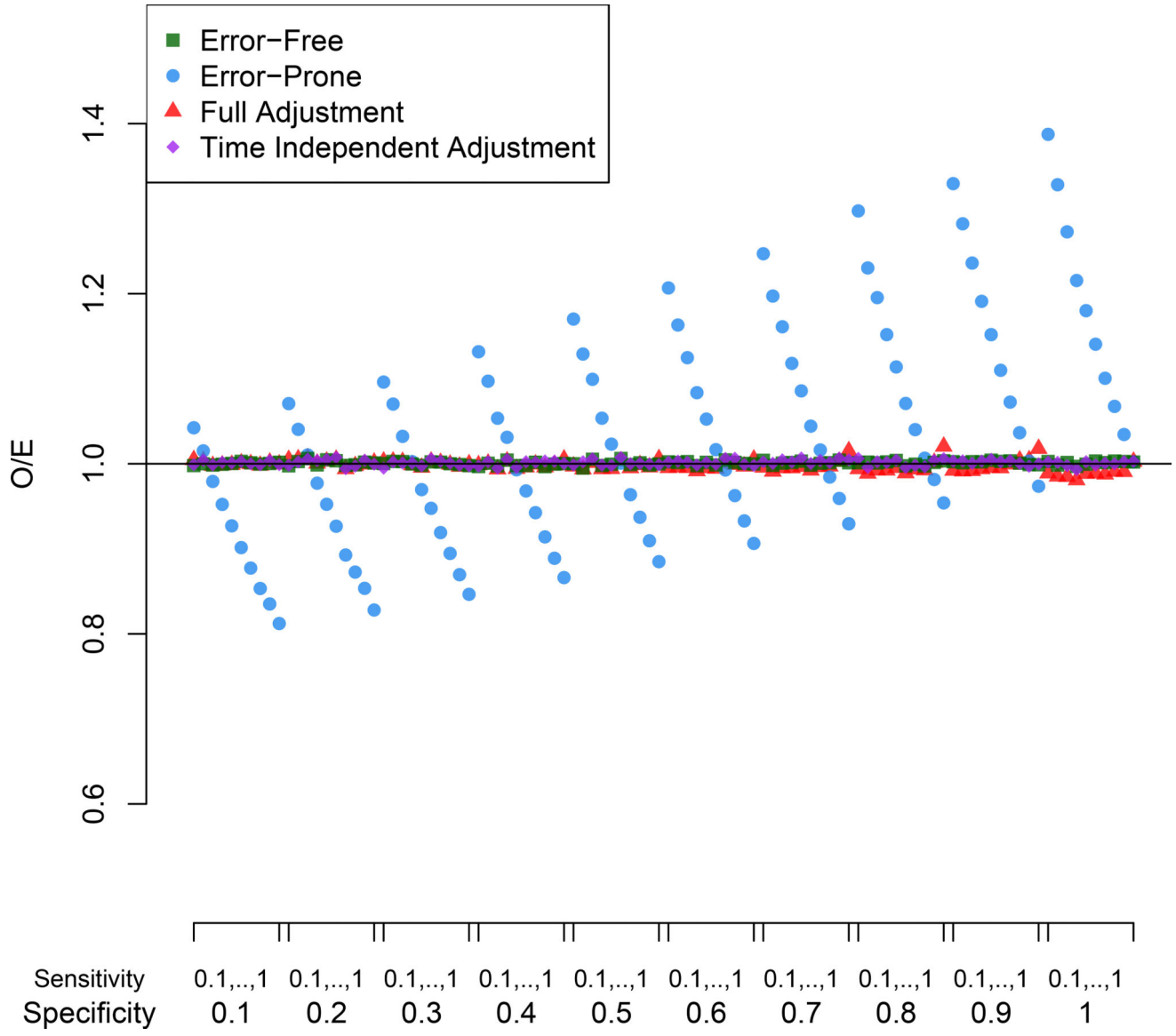

Figure 2.
O/E ratios for survival simulations under varying sensitivity and specificity rates.

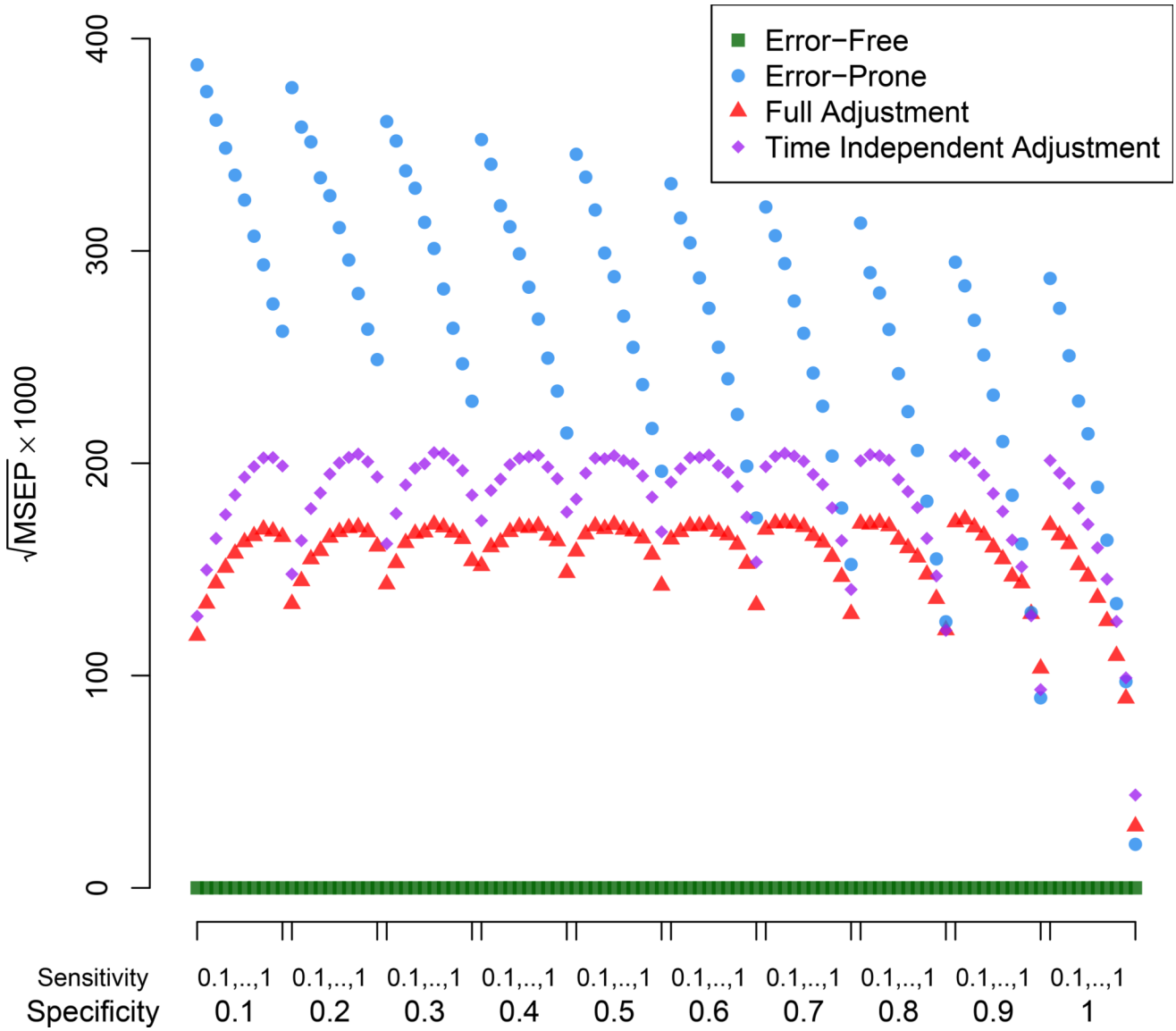

Figure 3.

$\sqrt{MSEP} * 1000$ for survival simulations under varying sensitivity and specificity rates.

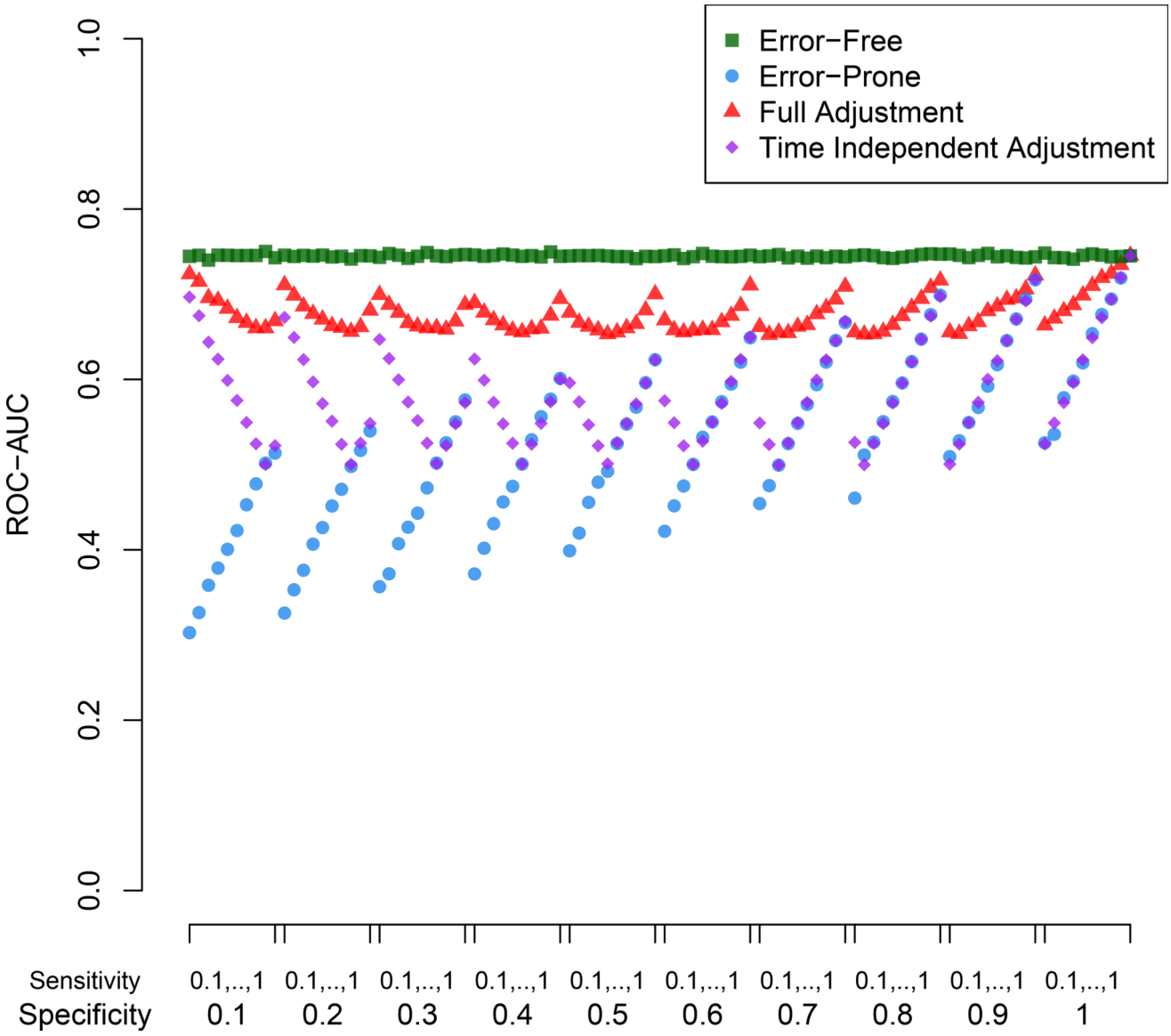

Figure 4.
ROC-AUC for survival simulations under varying sensitivity and specificity rates.

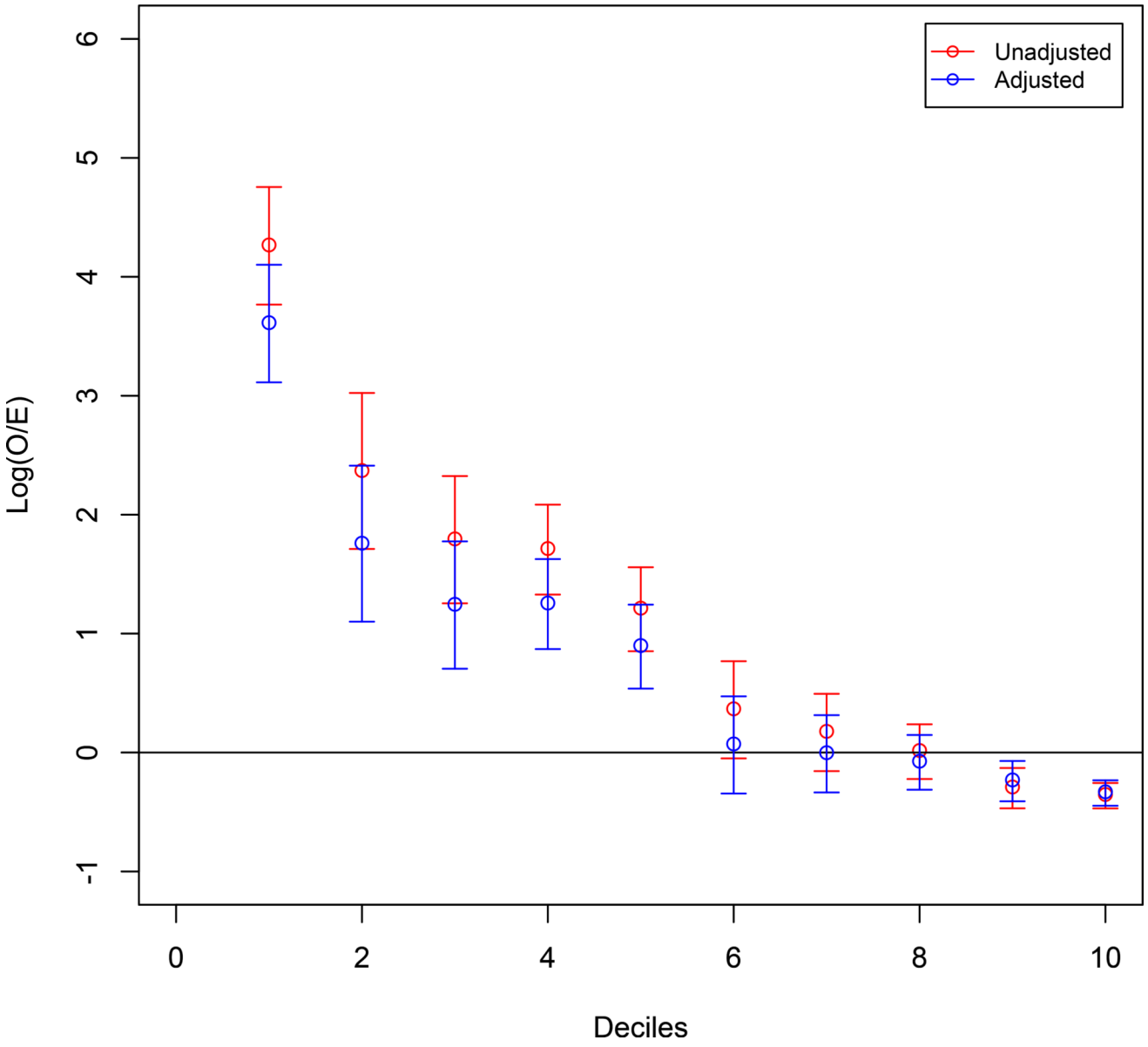

Figure 5.
Log of the observed over expected ratios and 95% bootstrap confidence intervals for being a BRCA carrier for counselees in CGN data set, stratified by risk deciles.

Table 1

Mendelian Risk Prediction Simulation Results. MSEP and O/E improve using the adjusted proposed method, ROC-AUC either improves or remains the same depending on the setting.

Counselee	Sens/Spec	Error in Age	$\sqrt{MSEP^{\dagger}} * 1000$			O/E			ROC-AUC		
			Error-Free	Error-Prone	Adjusted	Error-Free	Error-Prone	Adjusted	Error-Free	Error-Prone	Adjusted
Mother	0.954, 0.974	a: $N(0, 5^2)$	0.0000	19.1351	16.7405	0.9773	0.8190	0.9712	0.8160	0.8090	0.8086
		a: $N(0, 3^2)$	0.0000	18.1006	15.9490	0.9773	0.8280	0.9746	0.8160	0.8098	0.8078
		a: $N(0, 1^2)$	0.0000	17.5526	15.6430	0.9773	0.8327	0.9746	0.8160	0.8102	0.8115
		m: *exp*(1)	0.0000	43.2855	21.3037	0.9833	0.6063	0.9484	0.8145	0.7185	0.8020
	0.649, 0.990	a: $N(0, 5^2)$	0.0000	21.3122	20.8859	0.9773	1.0466	0.9783	0.8160	0.7814	0.7803
		a: $N(0, 3^2)$	0.0000	20.7947	20.5099	0.9773	1.0556	0.9792	0.8160	0.7821	0.7815
		a: $N(0, 1^2)$	0.0000	20.5213	20.1459	0.9773	1.0604	0.9737	0.8160	0.7826	0.7818
		m: *exp*(1)	0.0000	36.0314	23.8184	0.9817	0.8026	0.9565	0.8155	0.7140	0.7752
Daughter	0.954, 0.974	a: $N(0, 5^2)$	0.0000	18.8437	16.5614	0.9719	0.8166	0.9680	0.8171	0.8070	0.8082
		a: $N(0, 3^2)$	0.0000	17.7033	15.6918	0.9719	0.8256	0.9659	0.8171	0.8083	0.8086
		a: $N(0, 1^2)$	0.0000	17.1421	15.1775	0.9719	0.8301	0.9680	0.8171	0.8093	0.8083
		m: *exp*(1)	0.0000	43.4717	21.0365	0.9785	0.6028	0.9445	0.8162	0.7146	0.7976
	0.649, 0.990	a: $N(0, 5^2)$	0.0000	20.1613	20.0763	0.9719	1.0573	0.9872	0.8171	0.7895	0.7862
		a: $N(0, 3^2)$	0.0000	19.6759	19.5498	0.9719	1.0661	0.9834	0.8171	0.7913	0.7904
		a: $N(0, 1^2)$	0.0000	19.4507	19.1871	0.9719	1.0708	0.9803	0.8171	0.7928	0.7911
		m: *exp*(1)	0.0000	35.2381	23.0851	0.9760	0.8087	0.9535	0.8165	0.7229	0.7745

†MSEP: difference between (adjusted) error-prone and error-free predictions.

a: indicates a classical additive model; $T^* = T + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$, m: indicates a multiplicative measurement error model, $T^* = TU$, $U \sim exp(\lambda)$.