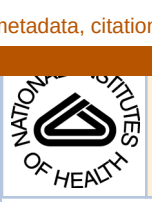

Published in final edited form as:

J Stat Plan Inference. 2013 February ; 143(2): 368–377. doi:10.1016/j.jspi.2012.08.006.

Semiparametric inference on the penetrances of rare genetic mutations based on a case-family design

Hong Zhang[1,2], **Donglin Zeng**[3], **Sylviane Olschwang**[4,5], and **Kai Yu**[2,3]

[1]Institute of Biostatistics, School of Life Science, Fudan University, P.R.C

[2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, U.S.A

[3]Department of Biostatistics, University of North Carolina, U.S.A

[4]Institut National de la Sante et de la Recherche Médicale (INSERM), France

[5]Department of Oncogenetics, Institut Paoli-Calmettes, France

Abstract

A formal semiparametric statistical inference framework is proposed for the evaluation of the age-dependent penetrance of a rare genetic mutation, using family data generated under a case-family design, where phenotype and genotype information are collected from first-degree relatives of case probands carrying the targeted mutation. The proposed approach allows for unobserved risk factors that are correlated among family members. Some rigorous large sample properties are established, which show that the proposed estimators were asymptotically semi-parametric efficient. A simulation study is conducted to evaluate the performance of the new approach, which shows the robustness of the proposed semiparamteric approach and its advantage over the corresponding parametric approach. As an illustration, the proposed approach is applied to estimating the age-dependent cancer risk among carriers of the *MSH2* or *MLH1* mutation.

Keywords

Case-family design; kin-cohort design; penetrance; proportional hazards model

1 Introduction

A precise estimation of the age-dependent risk for people carrying disease-causing mutations is critical for defining prevention/intervention strategies and understanding the underlying mechanisms of disease progression. A large number of disease-associated genetic mutations have been found to be rare. To estimate the penetrance function for a rare mutation, the cohort design is not cost-efficient for estimating penetrance function, since a large number of subjects are needed in order to have enough cases to ensure a sufficiently precise estimate. The case-control design is a more cost-efficient design but it needs additional knowledge such as composite incidence estimates from cohort data to estimate penetrance function (Gail et al., 1989). Compared with the cohort or case-control design, the

*Correspondence to: Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Boulevard, Executive Plaza South, Room 5064, Bethesda, MD 20892 U.S.A., 301-594-7206 phone, 301-402-0081 fax, yuka@mail.nih.gov.

kin-cohort design (Wacholder et al., 1998; Gail et al. 1999b; Chatterjee and Wacholder, 2001; Chatterjee et al., 2006; Wang et al., 2007; among others) has several practical advantages, including comparatively rapid execution and a modest reduction in required sample size. However, the kin-cohort design can lead to a biased estimate of penetrance if sampling is dependent on the phenotypes of relatives (Gail et al., 1999a). If the information from the relatives' genotypes is available in the kin-cohort design, it is possible to have a more robust penetrance estimate that is less vulnerable to sampling bias.

In this paper, we focus on a so called case-family design. Under this design, those probands who are affected with disease and are mutation carriers are ascertained; at least one first-degree relative of each case probands is genotyped at the targeted mutation locus; for those relatives, their current ages and the ages at disease onset if affected and other relevant covariate information are recorded. In the case-family design, the ascertainment procedure is related to both probands' mutation status and all individuals' phenotypes including their disease status. In literature, few approaches have been proposed to deal with such kind of data, the exceptions include two parametric approaches by Carayol and Bonaïti-Pellieé (2004) and Zhang et al. (2010) and one nonparametric approach by Wang et al. (2006). Carayol and Bonaïti-Pellieé (2004) proposed a conditional-likelihood based approach by accounting for ascertainment procedure, their approach can reduce penetrance estimation bias considerably when the cumulative penetrance is not small, but it needs to specify the prevalence of the mutation and a parametric form of the penetrance function. In the nonparametric approach of Wang et al. (2006) method, a rare disease assumption is needed and only unaffected relatives are used, and violation of the rare disease assumption can lead to substantial bias in the penetrance estimate. Zhang et al. (2010) recently developed a general parametric statistical approach, which can estimate age-dependent penetrance function with some covariates being adjusted for if needed. The validity of the two parametric approaches rely on the parametric specification of the hazard function.

In this paper, by relaxing the parametric assumption of the baseline hazard function in Zhang et al. (2010), we propose a more robust semiparametric approach for the inference of the age-dependent penetrance of a rare mutation. Large-sample properties of the approach are established using empirical process theory, showing that the proposed estimators of the regression parameters for the effects of rare mutation and covariates are asymptotically semiparametric efficient.

The estimation and inference procedure is described in the next section. A simulation study is performed in Section 3. A real data application can be found in Section 4. Some conclusions and discussions are given in Section 5.

2 Methods

2.1 Notation and Likelihood function

An individual is said to be a carrier of mutation s/he carries at least one copy of mutation allele. We assume that I unrelated case probands carrying mutation are ascertained, and n_i first-degree relatives (sibs, offspring, and parents) of the ith proband are recruited. Denote by $n=\sum_{i=1}^{I} n_i$ the total number of recruited first-degree relatives. The observation information include the current ages for the relatives, the ages at disease onset for those affected relatives, and the genotypes of the relatives at the mutation locus. The genotype at the target mutation locus is coded by 1 for a carrier and 0 for a non-carrier. Let the genotype and the affection status of the jth relative (the 0th relative is the case proband) of the ith case proband be coded by G_{ij} and D_{ij}, respectively; that is, $G_{ij} = 1$ if the jth relative is a carrier and 0 otherwise, and $D_{ij} = 1$ if the jth relative is affected with disease and 0 otherwise. Let

C_{ij} and T_{ij} denote the current age and the age at onset of the jth relative, respectively. Let $Y_{ij} = \min\{T_{ij}, C_{ij}\}$. We assume that a p-vector of covariates X_{ij} is observed for each relative, and allow for an unobserved risk factor vector R_{ij} that can be correlated among family members. We assume a Cox proportional hazards model with survival function for T_{ij} (Cox, 1972):

$$S(t|G_{ij}, X_{ij}, R_{ij}) = \exp\{-\Lambda(t)e^{\beta G_{ij} + \gamma^{\mathrm{T}} X_{ij} + \xi^{\mathrm{T}} R_{ij}}\}, \quad (1)$$

where $S(t \mid G_{ij}, X_{ij}, R_{ij})$ is the survival function of T_{ij} and $\Lambda(t)$ is the baseline cumulative hazard function.

To reduce potential penetrance estimation bias caused by the phenotype-dependent sampling in the case-family design, we formulate a conditional likelihood for the ith family's data as

$$P(\mathbf{G}_i|\mathbf{D}_i, \boldsymbol{Y}_i, \mathbf{X}_i, G_{i0}=1, D_{i0}=1, Y_{i0}, X_{i0}),$$

where $\mathbf{G}_i = (G_{i1}, \ldots, G_{in_i})$, $\mathbf{D}_i = (D_{i1}, \ldots, D_{in_i})$, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$, and $\mathbf{X}_i = (X_{i1}, \ldots, X_{in_i})$. To derive the likelihood function, we make the following assumptions:

- **A1** the targeted mutation is rare;
- **A2** in the general population, Hardy-Weinberg equilibrium holds for the targeted mutation, mating is random, and Mendel's Law of Heredity holds;
- **A3** the unobserved risk factors are independent of the targeted mutation and the covariates in the general population;
- **A4** the disease is rare.

Remark 1—The rare mutation assumption A1 is a key assumption, and A1 and A2 imply that the first-degree relatives' genotypes of a proband carrying mutation are independent and have the probability 0.5 to be a mutation carrier (refer to the proof of Theorem 1). This joint distribution is essential for the derivation of a simple likelihood function presented in Theorem 1. The assumption A3 is used to reduce the impact of the unobserved risk factors on the penetrance estimation. The rare disease assumption A4 is a technical one, and our simulation study demonstrates that this assumption has minor impact on the penetrance estimation, even if there are risk factors that are correlated among family members.

We have the following result:

Theorem 1—Under assumptions A1–A4, the overall likelihood $\prod_{i=1}^{I} P(\mathbf{G}_i|\mathbf{D}_i, \boldsymbol{Y}_i, \mathbf{X}_i, G_{i0}=1, D_{i0}=1, Y_{i0}, X_{i0})$ can be approximated by

$$L_{1n}(\beta, \gamma, \Lambda) = \prod_{i=1}^{I} \prod_{j=1}^{n_i} \frac{\exp\{D_{ij}\beta G_{ij} - \Lambda(Y_{ij})e^{\beta G_{ij} + \gamma^{\mathrm{T}} X_{ij}}\}}{\sum_{G=0}^{1} \exp\{D_{ij}\beta G - \Lambda(Y_{ij})e^{\beta G + \gamma^{\mathrm{T}} X_{ij}}\}}. \quad (2)$$

where $f(\cdot \mid G, X)$ and $S(\cdot \mid G, X)$ are the density function and the survival function, respectively, of the age at onset of an individual with genotype G and covariate vector X.

Refer to Supplemental A for a proof of Theorem 1.

Remark 2—The conditional independence of the relatives' genotypes of a proband given the proband is a carrier is not an assumption, but a derived conclusion under the mild assumptions A1–A4. We allow for unobserved risk factors that are correlated among family members, as often is the case in real applications.

Remark 3—In Supplemental B, we show that the parameters β, γ, and Λ are identifiable under the mild assumptions A1–A4. Therefore, the absolute penetrance functions are estimable. In the conventional case-control design, the absolute penetrances are not estimable when the exposure distribution is unknown, and in practice a conditional likelihood is used to estimate the odds ratio parameters. If the exposure distribution is known, then the absolute penetrances are estimable under the case-control design. In the current case-family design, the probands are assumed to be mutation carriers so that the joint distribution of relatives' genotypes (exposures) are known, this makes sure the absolute penetrances are estimable as shown rigorously in Supplemental B.

If the genotypes are not available for affected relatives, β is not identifiable and the likelihood function (2) is not applicable. For rare diseases, most of the relatives would be unaffected and it is appropriate to assume that the survival function for non-carriers, $S(t \mid 0, X, R)$, is 1. In this situation, we only need to model the survival function for carriers, that is,

$$S(t|1, X, R)=\exp\{-\Lambda(t)e^{\gamma^{\mathrm{T}}X+\xi^{\mathrm{T}}R}\} \quad (3)$$

under the proportional hazards model. Under the model (3), the likelihood function (2) becomes

$$L_{2n}(\gamma, \Lambda)=\prod_{i=1}^{I}\prod_{j=1}^{n_i}\frac{\exp\{-G_{ij}\Lambda(Y_{ij})e^{\gamma^{\mathrm{T}}X_{ij}}\}}{1+\exp\{-\Lambda(Y_{ij})e^{\gamma^{\mathrm{T}}X_{ij}}\}}. \quad (4)$$

Remark 4—It is straightforward to verity that $L_{1n}(\log 2, \gamma, \Lambda) = L_{2n}(\gamma, \Lambda)$. This indicates that the likelihood function (2) is a generalization of the likelihood function (4) that requires the assumption $S(t \mid 0, X, R) = 1$.

Remark 5—In the situation where the genotype is available from only one relative of each proband and no covariate is involved, the likelihood function (4) reduces to

$$\prod_{i=1}^{I}\frac{S(Y_{i1})^{G_{i1}}}{1+S(Y_{i1})},$$

which is exactly the same as that proposed by Wang et al. (2006), where $S(\cdot)$ is the marginal survival function of mutation carriers. In the situation where some probands have more than one relatives, Wang et al. (2006) proposed to weight the likelihoods for relatives with the weights depending on pedigree structure, while our simple likelihood with multiple relatives is derived rigorously under some mild assumptions and can account for covariates in a natural manner.

Remark 6—Notice that the likelihood functions (2) and (4) are independent of ξ, and they can be derived from the survival functions

$$S_1(t|G,X;\beta,\gamma,\Lambda)=\exp\{-\Lambda(t)e^{\beta G+\gamma^{\mathrm{T}}X}\} \quad (5)$$

and

$$S_2(t|X;\gamma,\Lambda)=\exp\{-\Lambda(t)e^{\gamma^{\mathrm{T}}X}\}, \quad (6)$$

respectively, as if the unobserved risk factors R do not exist.

2.2 Maximum likelihood estimation of unknown parameters

The parameters of most interest are the penetrance functions of both carriers and non-carriers, with adjustment of some covariates if needed. Notice that the survival functions, or equivalently the penetrance functions, depend on unknown parameters β, γ, and Λ under model (5) or (6), we can estimate the unknown parameters using a likelihood principle and consequently obtain penetrance estimates using a plug-in rule.

When the baseline cumulative hazard function Λ is known up to a finite number of unknown parameters η, a standard optimization algorithm can be used to obtain the maximum likelihood estimates of β, γ, and η (Zhang et al., 2010). However, such a parametric approach could produce considerable bias in the penetrance function estimate when the parametric form of Λ is not properly specified, as shown by our simulation study in the next section. Therefore, it would be advantageous to estimate the penetrance function without specifying a parametric form for Λ.

We consider as the maximum likelihood estimator $\hat{\Lambda}_n$ of Λ a right-continuous increasing step function with jumps only at Y_{ij}. That is, we define a maximum likelihood estimator of Λ by

$$\widehat{\Lambda}_n(t)=\begin{cases} 0, & 0 \le t<t_1; \\ \widehat{\Lambda}_n(t_k), & t_k \le t<t_{k+1} \text{ for } k=1,\ldots,K-1; \\ \widehat{\Lambda}_n(t_K), & t \ge t_K. \end{cases}$$

Here $0 = t_0 < t_1 < \cdots < t_K$, and $t_1, \ldots, t_K$ are distinct values of Y_{ij}, $j = 1, \ldots, n_i$; $i = 1, \ldots, n$. Let the jump sizes be $\delta_k = \Lambda(t_k) - \Lambda(t_{k-1})$, $k = 1, \ldots, K$; then $\Lambda(Y_{ij}) = \Sigma_{k:Y_{ij}\ge t_k}\,\delta_k$. The parameters to be estimated are $\{\beta, \gamma, \boldsymbol{\delta}\}$ with $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)$, and the number of all unknown parameters is $m = 1 + p + K$.

Remark 7—In the standard prospective cohort design with right-censored data, only uncensored times have positive jump sizes. In the current case-family design, it is possible that censored times ($D_{ij} = 0$) have positive jump sizes. This is particularly true under model (6) where all survival times are censored. Similar phenomenon happens in the situation where all survival times are interval censored in the standard cohort design.

For $G_{ij} = 0$ and $G_{ij} = 1$, it can be immediately verified that

$$\log\left[\frac{\exp\{D_{ij}\beta G_{ij}-\Lambda(Y_{ij})e^{\beta G_{ij}+\gamma^{\mathrm{T}}X_{ij}}\}}{\sum_{G=0}^{1}\exp\{D_{ij}\beta G-\Lambda(Y_{ij})e^{\beta G+\gamma^{\mathrm{T}}X_{ij}}\}}\right]=1+\exp\left\{D_{ij}\beta(1-2G_{ij})-\Lambda(Y_{ij})e^{\gamma^{\mathrm{T}}X_{ij}}(e^{\beta(1-G_{ij})}-e^{\beta G_{ij}})\right\}.$$

Therefore, the logarithm of the likelihood function (2) can be written as

$$l_n(\beta, \gamma, \Lambda) = -\sum_{i=1}^{I}\sum_{j=1}^{n_i} \log\left[1+\exp\left\{D_{ij}\beta(1-2G_{ij}) - \Lambda(Y_{ij})e^{\gamma^{\mathrm{T}}X_{ij}}(e^{\beta(1-G_{ij})} - e^{\beta G_{ij}})\right\}\right]. \quad (7)$$

It is straightforward to verify that $l_n(\beta, \gamma, \Lambda)$ is concave in γ and δ. Notice that the log-likelihood functions (2) and (4) are twice differentiable with respect to $\{\beta, \gamma, \delta\}$, we can use the interior-reflective Newton method that is implemented in the function "fmincon" in the Optimization Toolbox of Matlab (Coleman and Li, 1994; 1996).

Notice that the number of unknown parameters m is an increasing function in the sample size, which has a great impact on the convergence speed of the optimization algorithm. In practice, with little information loss we can round Y_{ij} to integers so that K is no more than the maximum age, say 120, no matter how large the sample size is. When p is small or moderate, the number of all unknown parameters m should also be at most moderate. Our numerical experiment shows that the MLEs of β, γ, and δ can be obtained in a few to no more than 100 seconds using a desktop PC with a 2.33 GHz processor when the sample size is around several hundred. The desired performance of the proposed algorithm might be partly due to the concavity of the likelihood function (7) in γ and δ.

When the number of relatives n is small, K would be close to or even equal to n, and the total number of parameters can be greater than n. The phenomenon that the number of parameters (including nuisance parameters and regression parameters) is greater than the sample size is commonly seen in the framework of semiparametric setting such as Cox's proportional hazards model (Cox, 1972). Although the convergence speed of the nuisance parameter estimates is usually slower than that of the regression parameter estimates (of fixed dimension), the estimation of regression parameters is asymptotically efficient in general (Muphy and van der Vaart, 2000). Our preliminary simulation study shows that all unknown parameters in (2) and (4) can be reliably estimated when the number of unknown parameters is greater than the number of relatives.

The survival functions and the corresponding cumulative penetrance functions can be estimated using a plug-in rule.

2.3 Variance estimation and large sample properties

Denote by $(\hat{\beta}_n, \hat{\gamma}_n, \hat{\Lambda}_n)$ the resulting MLEs of (β, γ, Λ). In Supplemental B, we show that the MLE $\hat{\theta}_n = (\hat{\beta}_n, \hat{\gamma}_n)$ is consistent, asymptotically normally distributed, and asymptotically efficient. Therefore, we can estimate the variances of $\hat{\theta}_n$ by virtue of the asymptotic normality. Here we focus on the likelihood function (2), since (4) is its special form. We can estimate the variance-covariance matrix of $\hat{\theta}_n = (\hat{\beta}_n, \hat{\gamma}_n)$ as follows. Denote the profile log-likelihood function of $\theta = (\beta, \gamma)$ by

$$\mathrm{pl}_n(\theta) = \max_{\Lambda} \log L_{1n}(\beta, \gamma, \Lambda).$$

The (s, t)th element of the inverse of the estimated variance-covariance matrix of $\hat{\theta}_n$ takes the form

$$-(2h_n)^{-2}\{\mathrm{pl}(\widehat{\theta}_n + h_n e_t + h_n e_s) - \mathrm{pl}(\widehat{\theta}_n + h_n e_s - h_n e_t) - \mathrm{pl}(\widehat{\theta}_n - h_n e_s + h_n e_t) + \mathrm{pl}(\widehat{\theta}_n)\}, \quad (8)$$

where h_n is a constant of order $n^{-1/2}$ and e_s and e_t are the sth and tth canonical vectors, respectively (Murphy and van der Vaart, 2000).

Because the survival function $S_1(t \mid G, X; \beta, \gamma, \Lambda)$ defined in (5) depends on the high-dimensional unknown parameter Λ, it is difficult to obtain an explicit expression of the variance-covariance matrix of the resulting survival function estimate. Instead, we propose to use a nonparametric bootstrap strategy (Efron and Tibshirani, 1993). First, we generate a large number of estimates of (β, γ, Λ), $(\widehat{\beta}_j^*, \widehat{\gamma}_j^*, \widehat{\Lambda}_j^*)_{j=1,\ldots,N}$, using bootstrap samples; then, we use $S_1(t|G, X; \widehat{\beta}_j^*, \widehat{\gamma}_j^*, \widehat{\Lambda}_j^*)$ to estimate the variance of $S_1(t|G, X; \hat{\beta}, \hat{\gamma}, \hat{\Lambda})$, and construct the confidence interval of $S_1(t|G, X; \beta, \gamma, \Lambda)$ based on the asymptotic normality of $S_1(t|G, X; \hat{\beta}, \hat{\gamma}, \hat{\Lambda})$.

3 A simulation study

We conducted a simulation study to examine the performance of two proposed semiparametric estimators. For the purpose of comparison, we also examined two parametric estimators proposed by Zhang et al. (2010) with the baseline cumulative hazard function being specified to be of Weibull form. The first parametric estimator uses both affected and unaffected relatives (PARA1) and the second parametric estimator only unaffected relatives (PARA2). The first semiparametric estimator is based on the likelihood function (2) which uses both affected and unaffected relatives (SEMI1) and second semiparametric estimator is based on the likelihood function (4) which uses only unaffected relatives (SEMI2).

We assumed that there were two single-nucleotide polymorphisms (SNPs) responsible for the disease of interest, with one SNP's genotype information available and the other one not available. The observed SNP had minor allele frequency (MAF) 0.001, and the unobserved SNP had MAF 0.2. We considered the Cox model with hazard function

$$\lambda(t|g_O, g_U) = \lambda_0(t) e^{2g_O + \log(\mathrm{OR}) g_U}, \quad (9)$$

where $\lambda_0(t)$ was the baseline hazard function, g_O (g_U) was 1 if the genotype of the observed SNP (unobserved SNP) had at least one copy of minor allele and 0 otherwise (i.e., the mode of inheritance is dominant), and OR was the odds ratio parameter for the unobserved SNP, which took value 1 (the unobserved risk factor is absent) or 1.5 (the unobserved risk factor is present). The baseline hazard function $\lambda_0(t)$ was assumed to be piecewise constant such that the cumulative penetrances at ages 20, 60, and 80 were 0.05, 0.1, and 0.2, respectively. The current age of the proband a was uniformly distributed in the interval (10, 90). The ages of the first-degree relatives (sibs and parents) of a proband were assumed to be independently distributed, with the ages of sibs being uniformly distributed in the interval $(a - 5, a + 5)$ and the ages of the parents being uniformly distributed in the interval $(a + 25, a + 35)$. In each family, the genotypes of the proband's parents at two loci were independently generated under the Hardy-Weinberg equilibrium and random mating, and the genotypes of the proband and sibs were generated under Mendel's Law of Heredity. Conditionally on the genotypes, the ages at onset of disease were generated based on model (9). Those first-degree relatives (sibs and parents of the probands) aged between 10 and 90 were used for analysis, so that the sampling is phenotype dependent. We generated 400 such families and applied SEMI1, SEMI2, PARA1, and PARA2 to these families. The baseline cumulative hazard function and the regression parameter were estimated by ignoring the unobserved SNP, and the cumulative penetrance functions of carriers and non-carriers were then

estimated using a plug-in rule. Based on 2000 replicates, we calculated the mean estimated cumulative penetrance functions of non-carriers and carriers as displayed in Figure 1.

It is seen that SEMI1 have only minor bias no matter the unobserved SNP has effect or not. On the other hand, PARA1 produces noticeable bias. Furthermore, SEMI2 and PARA2 underestimate the penetrance functions systematically, this is due to the fact that these two estimators treat the penetrance of non-carriers to be zero. In the simulation situation, the penetrance function of non-carriers is actually not that small, for example, the cumulative penetrance at age 60 is 0.1. This shows that SEMI1 is quite robust to the rare disease assumption.

We also considered a larger number of families, i.e., 800. The results are given in Figure 2. We can see that the bias of SEMI1 gets even smaller. We conducted additional simulations to study the situation where the penetrance function is smaller. The results (not shown) demonstrate that the bias of SEMI2 gets smaller as expected. When the odds ratio of unobserved SNP is very large, the bias of SEMI1 can be considerable (results not shown).

We further considered a generalized gamma baseline hazard function with shape parameter $\lambda = 50$ ($\lambda \neq 1$ corresponds to a non-Weibull baseline hazard function), and the scale parameter and another scale parameter were chosen such that the baseline cumulative penetrances at ages 20 and 80 were 0.05 and 0.2, respectively. The other settings were the same as those for Figure 1 and 2. The results are given in Figures 3 and 4 for family sizes 400 and 800, respectively. Again, SEMI1 has much smaller biases than PARA1, PARA2, and SEMI2 have.

To check the validity of the bootstrap method, we focused on the settings for upper panel of Figure 2 where the unobserved risk factor was absent. Based on 200 bootstrap samples, we estimated the standard error of the estimated survival function and hence the 95% confidence interval of the true survival function, with each family being treated as a single unit in the bootstrapping. The mean estimated standard error (SEE) and coverage probability (CP) of the confidence interval were then estimated based on 500 simulations. Table 1 displays the relative bias of the estimates (Rbias), standard errors (SEs), SEEs, and CPs, for ages 30, 50, and 70. Overall, the SEEs of all estimators are close to the SEs. Among the four estimators, SEMI1 has CPs closest to the nominal level 95%, and PARA1, PARA2, SEMI2 can have very poor coverage probabilities due to their biased estimates.

4 Application to a study of Lynch syndrome

We applied three considered approaches (the semiparametric approach of Wang et al., 2006, the parametric approach of Zhang et al., 2010, and the proposed semiparametric approach) to a study of Lynch syndrome (Olschwang et al., 2009). In this study, the data were obtained via a retrospective questionnaire sent to eight genetic units in France and Switzerland that offered germline analysis of *MSH2* and *MLH1* genes. Phenotypes and genotypes from 856 asymptomatic first-degree relatives of *MSH2* or *MLH1* carriers were collected from the eight centers. For each relative, the mutation status at genes *MSH2* and *MLH1* were obtained. The phenotypes of the relatives include but are not limited to ages (18 to 89 years) and genders (402 males and 454 females). There is little difference between the disease penetrances of *MSH2* and *MLH1* genes according to Olschwang et al. (2009), so we will not distinct the two genes in our analysis. Since the ages of the relatives were available, we could estimate the age-dependent penetrance function.

For the *MSH2* and *MLH1* genes, it has been estimated that the prevalence of the corresponding mutations in the general population of European origin is between 0.002 and 0.001 (Salovaara et al., 2000). Therefore, the mutations are rare enough and our approach is

applicable. Furthermore, Lynch syndrome only accounts for 3% to 5% of all colorectal cancers (Bonadona et al., 2011), so the prevalence of Lynch syndrome is also very small. Small prevalences of both mutation and Lynch syndrome imply that the penetrance of non-carriers is close to zero by the law of total probability. Therefore, we can use the information from unaffected relatives only to estimate the cumulative penetrance function of carriers. In the parametric approach (PARA2), we assumed a Weilbull baselline hazard function. Gender was adjusted for as a covariate in both PARA2 and SEMI2. The resulting cumulative penetrance functions are plotted in Figure 5 for both males and females. Overall, the difference between PARA2 and SEMI2 is small. By assuming each proband has only one relative, we also applied the nonparametric approach (WANG) of Wang et al. (2006) to male and female relatives, separately. All of the three estimators suggest that male carriers have higher cumulative risk than female carriers at most ages, except that WANG gave reverse conclusion for old carriers. It is believed that there exists a threshold value of age under which an individual will not be affected by Lnych syndrome. The two semiparametric approaches gave a threshold at the age of 26, which is close to that estimated by Olschwang et al. (2009), 28 years.

5 Discussion

In this paper, we propose a robust semiparametric approach for the inference of age-dependent penetrance of a rare mutation. The estimation of unknown parameters, including the baseline cumulative hazard function and regression parameters, can be reliably obtained. Large-sample properties are established via empirical process theory, which shows that the estimator is semiparametric efficient. Through simulations, we demonstrated the robustness of the proposed semiparamteric approach and its advantage over the corresponding parametric approach.

Our penetrance estimate is derived from the likelihood model conditioning on the phenotypes of all subjects and genotypes of the probands. We adopt this conditional approach in order to minimize the bias that could potentially arise when the ascertainment procedure of families is influenced by the proband's genotype and the family members' phenotypes. In the situation where the ascertainment procedure is known, it is more appropriate to use a joint likelihood approach that simultaneously models genotypes and phenotypes of family members. Some methods have been proposed for estimating age-dependent penetrance function by properly modeling residual familial aggregation (unobserved risk factors) in the case-control family design (Li et al., 1998; Shih and Chatterjee, 2002; Hsu et al., 2004; Hsu and Gorfine, 2006), where the ascertainment procedure for genotyped relatives is assumed to be independent of their phenotypes. When the ascertainment procedure depends on both probands' genotypes and family members' phenotypes, it deserves further investigations to develop an efficient approach by fully utilizing the information from genotypes and phenotypes and accounting for residual familial aggregation. In real applications, some relatives' phenotype information is available but their genotype information is not available, making the statistical analysis of such data very complicated.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by the State Key Development Program for Basic Research of China (Grant No. 2012CB316505) (HZ) and the Intramural Program of the National Institutes of Health (HZ and KY). We would like to thank Dr. B. J. Stone for editorial help.

References

Bonadona V, Bonaïti B, Olschwang S, Grandjouan S, Huiart L, Longy M, Guimbaud R, Buecher B, Bignon YJ, Caron O, Colas C, Nogues C, Lejeune-Dumoulin S, Olivier-Faivre L, Polycarpe-Osaer F, Nguyen TD, Desseigne F, Saurin JC, Berthet P, Leroux D, Duffour J, Manouvrier S, Frebourg T, Sobol H, Lasset C, Bonaiti-Pellie C. Cancer risks associated with germline mutations in *MLH1*, *MSH2*, and *MSH6* genes in Lynch syndrome. J Am Med Assoc. 2011; 305:2304–2310.

Carayol J, Bonaïti-Pellieé C. Estimating penetrance from family data using a retrospective likelihood when ascertainment depends on genotype and age of onset. Genet Epidemiol. 2004; 27:109–117. [PubMed: 15305327]

Chatterjee N, Kalaylioglu Z, Shih JH, Gail MH. Case-control and case-only designs with genotype and family history data: estimating relative risk, residual familial aggregation, and cumulative risk. Biometrics. 2006; 62:36–48. [PubMed: 16542227]

Chatterjee N, Wacholder S. A marginal likelihood approach for estimating penetrance from kin-cohort designs. Biometrics. 2001; 57:245–252. [PubMed: 11252606]

Coleman TF, Li Y. On the convergence of reflective newton methods for large-scale nonlinear minimization subject to bounds. Math Program. 1994; 67:189–224.

Coleman TF, Li Y. An interior, trust region approach for nonlinear minimization subject to bounds. SIAM J Optim. 1996; 6:418–445.

Cox DR. Regression models and life tables (with Discussion). J R Statist Soc B. 1972; 34:187–220.

Efron, B.; Tibshirani, RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993.

Gail MH, Pee D, Benichou J, Carroll R. Designing studies to estimate the penetrance of an identified autosomal dominant mutation: cohort, case-control, and genotyped-proband designs. Genet Epidemiol. 1999a; 16:15–39. [PubMed: 9915565]

Gail MH, Pee D, Carroll R. Kin-cohort designs for gene characterization. J Natl Cancer Inst Monog. 1999b; 26:55–60.

Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting indiidualized probabilities of developing breast cancer for white femailes who are being examined annually. J Natl Cancer Inst. 1989; 81:1879–1886. [PubMed: 2593165]

Hsu L, Chen L, Gorfine M, Malone K. Semiparametric estimation of marginal hazard function from the case-control family studies. Biometrics. 2004; 60:936–944. [PubMed: 15606414]

Hsu L, Gorfine M. Multivariate survival analysis for case-control family data. Biostatistics. 2006; 7:387–398. [PubMed: 16368774]

Li H, Yang P, Schwartz AG. Analysis of age at onset data from case-control family studies. Biometrics. 1998; 54:1030–1039. [PubMed: 9750249]

Olschwang S, Yu K, Lasset C, Baert-Desurmont S, Buisine MP, Wang Q, Hutter P, Rouleau E, Caron O, Bourdon V, Thomas G. Age-dependent cancer risk is not different in between msh2 and mlh1 mutation carriers. J Cancer Epidemiol. 200910.1155/2009/791754

Salovaara R, Loukola A, Kristo P, Kääriäinen H, Ahtola H, Eskelinen M, Härkönen N, Julkunen R, Kangas E, Ojala S, Tulikoura J, Valkamo E, Järvinen H, Mecklin JP, Aaltonen LA, de la Chapelle A. Population-based molecular detection of hereditary nonpolyposis colorectal cancer. J Clin Oncol. 2000; 18:2193–2200. [PubMed: 10829038]

Shih JH, Chatterjee N. Analysis of survival data from case-control family studies. Biometrics. 2002; 58:502–509. [PubMed: 12229984]

Murphy SA, van der Vaart AW. On profile likelihood. J Am Statist Asso. 2000; 95:449–465.

Wacholder SPH, Struewing JP, Pee D, McAdams M, Brody L, Tucker M. The kin-cohort study for estimating penetrance. Am J Epidemiol. 1998; 148:623–630. [PubMed: 9778168]

Wang Y, Clark LN, Marder K, Rabinowitz D. Nonparametric estimation of age-at-onset distributions from censored kin-cohort data. Biometrika. 2007; 94:403–414.

Wang Y, Ottman R, Rabinowitz D. A method for estimating penetrance from families sampled for linkage analysis. Biometrics. 2006; 62:1081–1088. [PubMed: 17156282]

van der Vaart, AW.; Wellner, JA. Weak convergence and empirical processes. New York: Springer-Verlag; 1996.

Zhang H, Olschwang S, Yu K. Statistical inference on the penetrances of rare genetic mutations based on a case-proband design. Biostatistics. 2010; 11:519–532. [PubMed: 20179148]

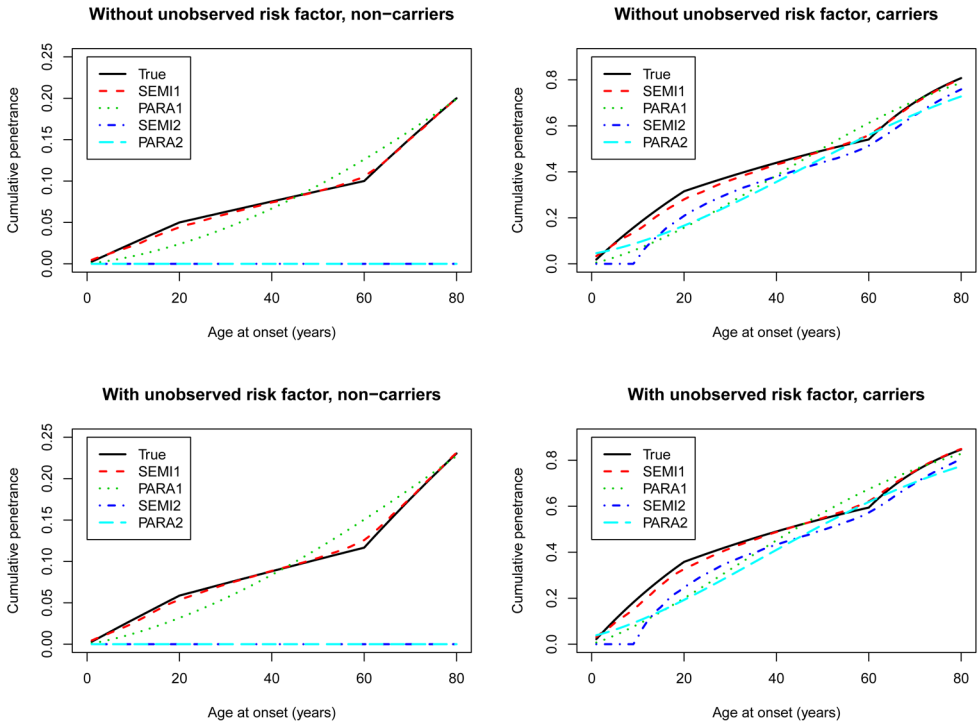

Figure 1.
Mean estimated cumulative penetrances (piecewise constant hazard, number of families = 400).

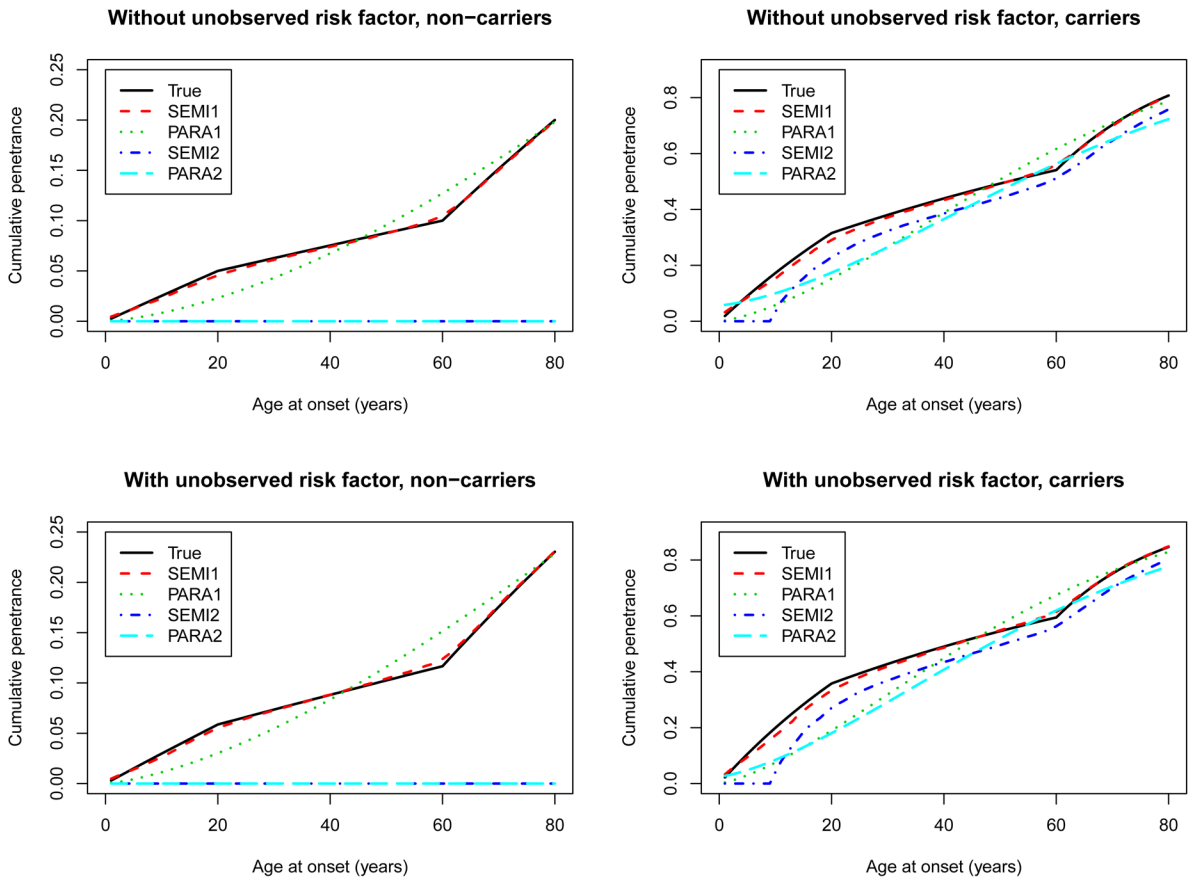

Figure 2.
True and mean estimated cumulative penetrances (piecewise constant hazard, number of families = 800).

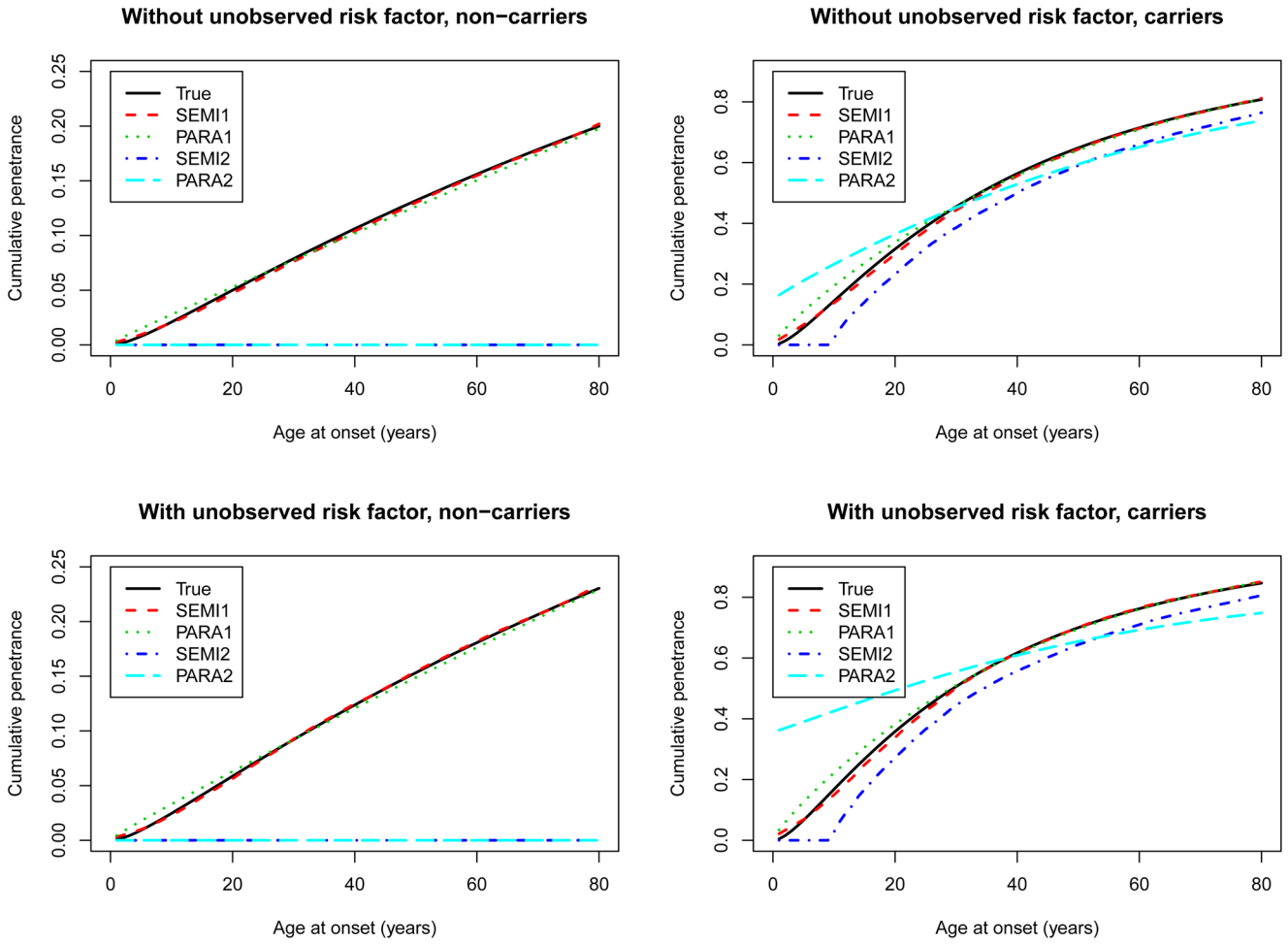

Figure 3.
True and mean estimated cumulative penetrances (generalized gamma hazard, number of families = 400).

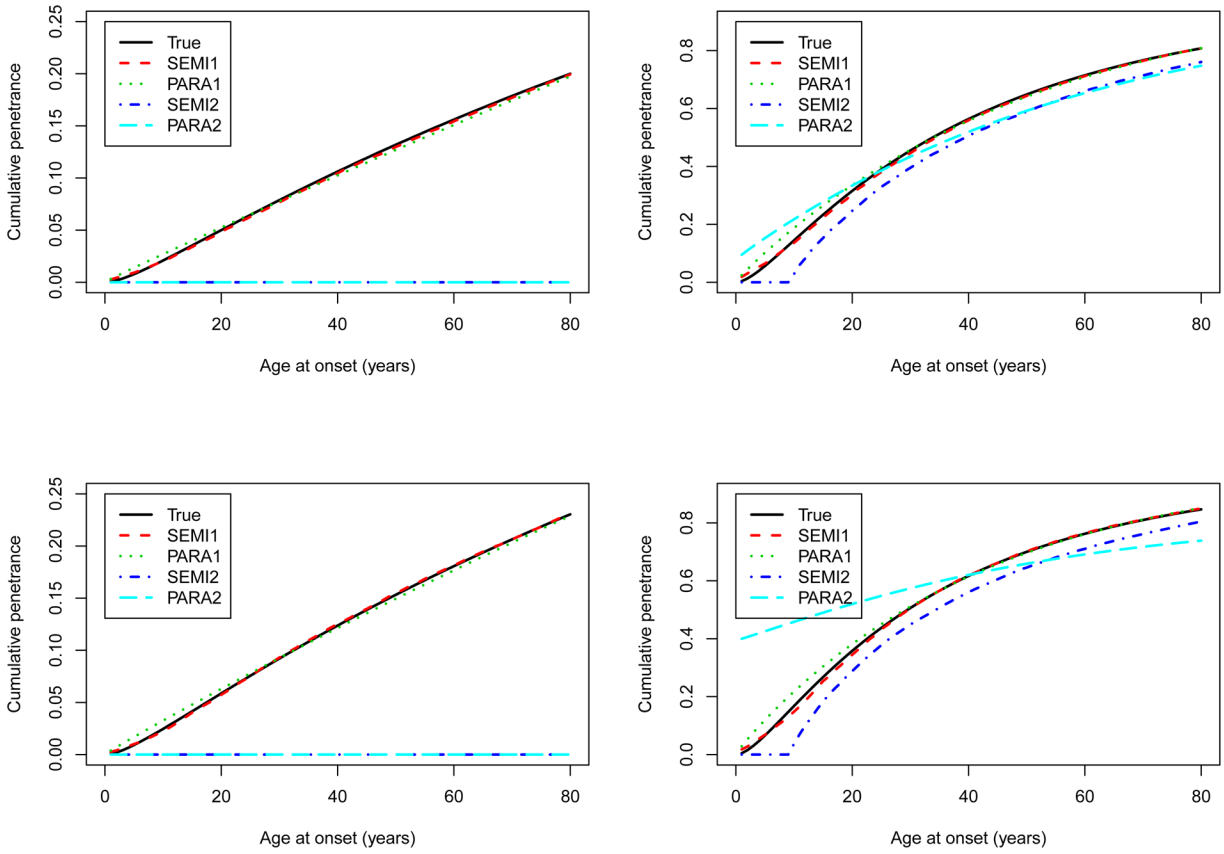

Figure 4.
True and mean estimated cumulative penetrances (generalized gamma hazard, number of families = 800).

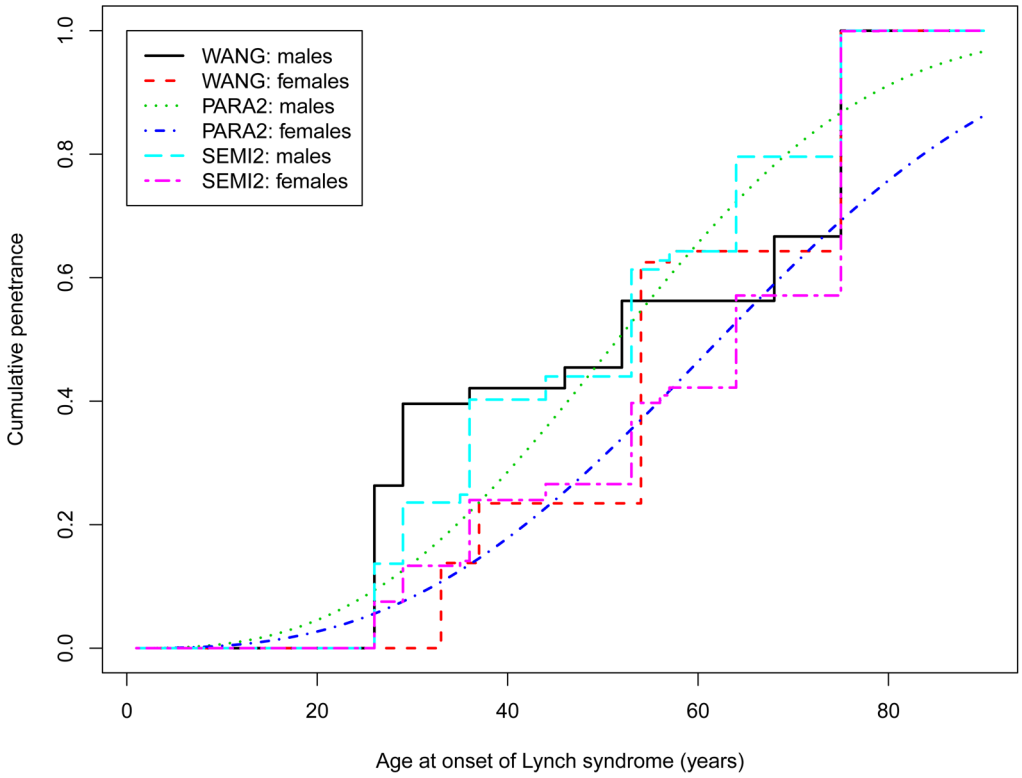

Figure 5.
Estimated cumulative penetrance functions for *MSH2* or *MLH1* carriers in the Lynch syndrome study.

Table 1

Summary of cumulative penetrance estimates

		Age at onset											
		30				50				70			
Mutation	**Estimator**	**Rbias**	**CP**	**SE**	**SEE**	**Rbias**	**CP**	**SE**	**SEE**	**Rbias**	**CP**	**SE**	**SEE**
Non-carriers	SEMI1	0.002	0.94	0.017	0.016	0.000	0.95	0.016	0.016	0.001	0.92	0.024	0.023
Carriers	SEMI1	0.020	0.93	0.086	0.081	0.005	0.94	0.057	0.058	0.017	0.93	0.051	0.048
Carriers	SEMI2	−0.273	0.22	0.088	0.088	−0.388	0.00	0.058	0.060	−0.585	0.00	0.056	0.054
Non-carriers	PARA1	0.022	0.60	0.012	0.012	−0.008	0.94	0.012	0.013	−0.012	0.94	0.017	0.017
Carriers	PARA1	0.193	0.62	0.070	0.072	−0.021	0.94	0.049	0.052	−0.024	0.96	0.019	0.022
Carriers	PARA2	−0.206	0.69	0.122	0.113	−0.407	0.02	0.071	0.075	−0.589	0.00	0.043	0.042

SEMI1, semiparametric estimator using both affected and unaffected relatives; SEMI1, semiparametric estimator using both affected and unaffected relatives; SEMI2, semiparametric estimator using only unaffected relatives; PARA1, parametric estimator using both affected and unaffected relatives; PARA2, parametric estimator using only unaffected relatives; Rbias, relative bias of mean estimated cumulative penetrance; SE, standard error of estimated cumulative penetrances; SEE, mean estimated standard error of estimated cumulative penetrances; CP, 95% coverage probability of nonparametric bootstrap confidence interval.