World Applied Sciences Journal 6 (1): 77-82, 2009
ISSN 1818-4952

Kin-cohort Estimate of Penetrance with Piecewise Weibull Model

[1]*Amir Hossein Hashemian,* [1]*Ebrahin Hajizadeh,*
[1]*Anoushirvan Kazemnezad,* [2]*Mohammad Reza Meshkani and* [3]*Parvin Mehdipour*

[1]Department of Biostatistics, Tarbiat Modares University, Tehran, Iran
[2]Department of Statistics, Shahid Beheshti University, Tehran, Iran
[3]Department of Medical Genetics, Faculty of Medicine, Tehran University of Medical Sciences, Tehran, Iran

Abstract: The estimation of cancer risk in individuals is of great interest for research teams. The kin-cohort method is a method of choice for estimating hazards and penetrance of cancer, because of its capability to illustrate the correlation between genotype and phenotype. In the estimation of cancer penetrance, many models omit to take into account the fact that hazards are likely to increase, decrease, or remain constant with respect to age variations. In order to estimate cancer penetrance by rectifying the said drawback of the commonly used models, we utilized a modified piecewise exponential model using weibull distribution, i.e. a piecewise Weibull model. We considered mutations in BRCA1 and BRCA2 genes, which are related to ovarian and breast cancer. A set of data similar to true values was generated. We analyzed the data set using both piecewise exponential and piecewise Weibull models. Our results showed that the Weibull model was closer than the exponential model to the true values in terms of estimating the hazards and penetrance and a significant difference between two models was recognized. For the persons who are at risk of developing cancer, Due to the importance of the estimation of incidence probability, methods that can generate most accurate estimations are preferable. Therefore, we recommend the piecewise Weibull model as a proper model for the estimation of hazard and penetrance.

Key words: Cancer . Kin-cohort . Penetrance . Piecewise exponential model . Piecewise weibull model

INTRODUCTION

Cancer is fundamentally a genetic disease and is partially due to gene mutation. Many types of cancer have a higher incidence in the relatives of patients than in the general population and some of them exhibit the Mendelian inheritance [1]. For many diseases like cancer, age is one of the primary risk factors. The risk of developing cancer increases with age; nonetheless, not everyone experiences the disease in his or her lifetime [2]. The risk of cancer increases up to threefold if one first-degree relative and up to tenfold if more than one first-degree relative is affected [1]. These familial risks tend to increase even further if the onset of disease in the affected first-degree relative is at age 40 or younger [1].

The two genes of BRCA1 and BRCA2 are the most important predisposing genes in the causation of breast or ovarian cancer [1, 3-8]. Individuals who carry the mutation in BRCA1 are at increased risk of developing breast or ovarian cancer [3]. Detection of cancer cases and/or estimation of incidence probabilities in persons at risk in different age groups are of great importance for the medical community [1].

There are several methods such as cohort designs for estimating the risk of cancer in those at risk. However, penetrance estimation in a cancer gene, like BRCA1 or BRCA2, via such common methods as cohort and case-control designs is hardly feasible. Studying families with ovarian or breast cancer in order to estimate the risk of cancer yields unreal results with respect to penetrance.

The Kin-cohort approach, a method for estimating penetrance, uses probands and their relatives to study the effects of mutations in such genes as BRCA1 and BRCA2 on cancer risk. In the kin-cohort method, the probands are genotyped in the first step. The set of all the carrier's relatives is called carrier kin and the set of all the non-carrier's relatives is called non-carrier kin. The phenotypes of all these relatives are, thereafter, determined and data on disease history among relatives are collected so that cancer penetrance can be estimated [2, 9-15].

The ability to study multiple phenotypes simultaneously is an advantage of this design [12]. This design was employed by Wacholder *et al*. [9] to estimate the cumulative probability of developing breast or ovarian cancer, as a function of age, for

Corresponding Author: Dr. Ebrahim Hajizadeh, Department of biostatistics, Tarbiat Modares University, Tehran, Iran

carriers of mutations of BRCA1 or BRCA2 in Ashkenazi Jews from the region surrounding Washington, DC. In this study, penetrance for mutations in BRCA1/2 genes for the first occurrence of breast or ovary cancer in 70-year-old women was estimated at 63%.

A common assumption for all the existing methods for the kin-cohort estimation is that the censoring mechanism does not depend on the mutation under study [2, 9-11]. When the events have the censoring mode, the competing risk model provides the better method for analyzing data. In this model, the estimation of cancer penetrance is in terms of 'cause-specific hazard' functions. A likelihood-based estimation with piecewise exponential modeling of cause-specific hazard functions has been proposed to estimate these hazards [15].

The Weibull distribution is the generalized form of exponential distribution and is commonly used as a lifetime distribution (most useful model for lifetime distribution). Given that hazards are likely to increase, decrease, or remain constant with respect to age variations, the piecewise Exponential model in this situation can not be accepted as a proper solution. In this study, we used a modified piecewise exponential model, i.e. a piecewise Weibull model.

MATERIALS AND METHODS

Competing risk: A common assumption for all of the existing methods for the kin-cohort estimation is supposed that the censoring is independent of genotype [9-11, 16]. However, the onset of any of the known BRCA1/2-related cancers can be treated as a censoring event. Therefore in the estimation of the risk of ovarian cancer, for example, subjects may be censored because of a death that came from breast cancer it could occur before diagnosing of ovarian cancer.

Let y denote the indicator of whether the subject is a carrier. The cumulative risk (penetrance) of a disease up to age t associated with genotype Y = y can be showed by F_y (t). The estimates of F_0 (t) and F_1 (t) are needed [2, 9-12, 15]. In the presence of two competing events (for example), E_1 and E_2, T_1 and T_2 are the time to these events. The follow-up for the second event ends at the onset of the first event. It is showed that the cause-specific hazard function for the i^{th} event at time t for an individual with genotype Y = y, the instantaneous probability that an individual with genotype Y = y will experience the event E_i at time t, given that s/he has been "at risk", can be computed as [15]

$$h_{iy}(t) = \lim_{\delta t \to 0} \frac{P_r\{T_i \in [t, t+\delta t) \mid T_1 \geq t, T_2 \geq t, Y = y\}}{\delta t}$$

Both h_{i0} (t) and h_{i1} (t) are hazards for non-carriers and carriers respectively that are defined in defined intervals by:

$$0 = t_0^{(i)} < t_1^{(i)} < t_2^{(i)} < \cdots < t_{k_i}^{(i)} < t_{k_i+1}^{(i)} < \infty$$

It is showed that piecewise exponential model can be used for the estimation of cause specific hazard values. In this method composite likelihood with regard to the likelihood contribution of family history data of the relatives of a volunteer is used as the product of the probabilities of the phenotype history of the individual relatives, given the genotype of the volunteers as below.

Composite likelihood: Let *m* be the number of probands and Y_{0i} the genotype of the i^{th} proband. If the i^{th} volunteer reports the family history of a phenotype γ for n_i relatives and γ_{ij} the value of γ for the j^{th} relative of the i^{th} proband therefore the composite-likelihood of the family history data of the relatives can be showed by

$$\prod_{i=1}^{m}\prod_{j=1}^{n_i} P_r\left(\gamma_{ij} \mid Y_{0i}\right) = \prod_{i=1}^{m}\prod_{j=1}^{n_i}\sum_{g=0}^{1} P_r\left(\gamma_{ij} \mid Y_{ij} = y\right) \times P_r\left(Y_{ij} = y \mid Y_{0i}\right) \quad (1)$$

Which in this equation P_r ($\gamma_{ij}|Y_{0i}$) defined the marginal probability density of the phenotype history of the j^{th} relative of the i^{th} proband, given the genotype of the i^{th} proband and on the right-hand side, this probability is computed as the weighted sum of the probability density of the phenotype history of the relative if the relative was a noncarrier or a carrier, with weights defined by the corresponding probabilities of the relative being a noncarrier and a carrier given the genotype of the volunteer.

With regard to "competing risks" definition, we need to define the triplet observations by $\gamma = (t, \delta_1, \delta_2)$, which T denote the time to the first of the two events, or censoring if neither of the events occurred during follow-up. δ_1, δ_2 are the indicator variables of occurring of two competing events, E_1 and E_2. It is obvious that both of them cannot occur simultaneously.

Given that hazards are likely to increase, decrease, or remain constant with respect to age variations, the piecewise exponential model in this situation may not be regarded as a proper solution. Therefore as Weibull distribution is the generalized form of exponential distribution and is commonly used as a lifetime distribution, with respect to Weibull distribution the composite-likelihood of the event history data of relatives can be computed by replacing P_r ($\gamma_{ij}|Y_{ij} = y$) in

Equation (1) with the corresponding likelihood for competing risk data, given by

$$\left[\beta_{1y}\lambda_{1y}{}^{\beta_{1y}}T_{ij}^{\beta_{1y}-1}\right]^{\delta_{1ij}}.\left[\beta_{2y}\lambda_{2y}{}^{\beta_{2y}}T_{ij}^{\beta_{2y}-1}\right]^{\delta_{2ij}}.\exp\left[-(\lambda_{1y}T_{ij})^{\beta_{1y}}\right].\exp\left[-(\lambda_{2y}T_{ij})^{\beta_{2y}}\right]$$

Maximization of the composite-likelihood: With using of EM (Expectation-Maximization) algorithm for maximization of the composite-likelihood with respect to the hazard parameters we have the following steps for each iteration. In the E-step of the algorithm we compute the conditional probability of each relative being a carrier and a noncarrier, given their individual event history and the genotype of the index proband. In this state W_{0ij} and W_{1ij} denoted the corresponding probabilities of being a noncarrier and a carrier, respectively, for the j^{th} relative of the i^{th} proband. We improved the weighted values W_{0ij} and W_{1ij} by participating the weibull model to computing theme's and hazards,

$$W_{yij}=\frac{\left[\beta_{1y}\lambda_{1y}{}^{\beta_{1y}}T_{ij}^{\beta_{1y}-1}\right]^{\delta_{1ij}}.\left[\beta_{2y}\lambda_{2y}{}^{\beta_{2y}}T_{ij}^{\beta_{2y}-1}\right]^{\delta_{2ij}}.\exp\left\{-\left[(\lambda_{1y}T_{ij})^{\beta_{1y}}+(\lambda_{2y}T_{ij})^{\beta_{2y}}\right]\right\}.P_r\left(Y_{ij}=y\mid Y_{0i}\right)}{\sum_{y'=0}^{1}\left[\beta_{1y'}\lambda_{1y'}{}^{\beta_{1y'}}T_{ij}^{\beta_{1y'}-1}\right]^{\delta_{1ij}}.\left[\beta_{2y'}\lambda_{2y'}{}^{\beta_{2y'}}T_{ij}^{\beta_{2y'}-1}\right]^{\delta_{2ij}}.\exp\left\{-\left[(\lambda_{1y'}T_{ij})^{\beta_{1y'}}+(\lambda_{2y'}T_{ij})^{\beta_{2y'}}\right]\right\}.P_r\left(Y_{ij}=y'\mid Y_{0i}\right)}$$

In the M-step of the algorithm, hazard values can be obtained by

$$\hat{h}_{yk}^{(l)}=\frac{\sum_{i,j}N_{ijk}^{(l)}W_{yij}}{\sum_{i,j}P\gamma_{ijk}^{(l)}W_{yij}},\ y=0,1;k=1,2,\ldots,k_l,\ l=1,2$$

$P\gamma_{ijk}^{(l)}$ computes the number of person years the i^{th} relative of the j^{th} proband contributes to the age interval $\left[t_k^{(l)},t_{k+1}^{(l)}\right)$; and $N_{ijk}^{(l)}$ denotes the indicator of whether or not the relative has an event of type *l* in that interval.

Iterating takes place between the E-step and the M-step of the algorithm until convergence yields the final estimates of hazard values. With regard to weibull hazard function, h (t) = $\beta\lambda^{\beta}t^{\beta-1}$ and its logarithm, ln h(t) = ln ($\beta\lambda^{\beta}$) + (β-1) ln t, it is obvious that depend on a regression relation, we can estimate β and λ parameters.

Steps of EM algorithm: We carry out these processes for the piecewise Weibull model in each step of EM algorithm:

- Finding the primary estimates of h (t) and S (t) (Survival)
- Estimating the Weibull distribution parameters with bootstrap sampling of primary simulated data
- Revising h (t) and S (t) values
- Revising weighted values W_{0ij} and W_{1ij}
- Computing h (t) and S (t) values
- Iterating between steps 3 and 5 until convergence to good solution

Simulation: Because of mutations in BRCA1/2 which are related to Ovarian or Breast cancer, we used the simulation experiments to evaluate the performance of the proposed piecewise Weibull model for the estimation of the default hazard and the penetrance functions of ovarian cancer (i.e.: in the absence of breast cancer). Thus, the onset of breast cancer treated as censoring events for ovarian cancer. We generated data in a setting similar to the true values (following setting). We used an allele frequency of 0.0112 to generate the mutation status for 10,000 probands. For describing ovarian cancer risk, we chose the corresponding shape and the scale parameters to be 0.0051 and 4.0051, respectively, for the non-carriers and 0.0081 and 2.9837, respectively, for the carriers. We assumed that relatives can be censored either at their death from other causes, or at the time of the interview of the proband. For the relatives of carriers and noncarriers, we generated their age at mortality from a normal distribution that had a mean age of 81 and standard deviation of 10. We repeated these steps 50 times [15].

We analyzed each data set, using both the piecewise exponential and Weibull models. Graphs of the hazard and the cumulative incidence functions were plotted to compare the visual differences of the two models.

AIC (Akaike's Information Criterion) was used for a comparison of the two methods. In this step, by computing the AIC in both piecewise exponential and Weibull models and differences of these AIC, we have

$$\Delta = AIC_{Exponential} - AIC_{Weibull}$$

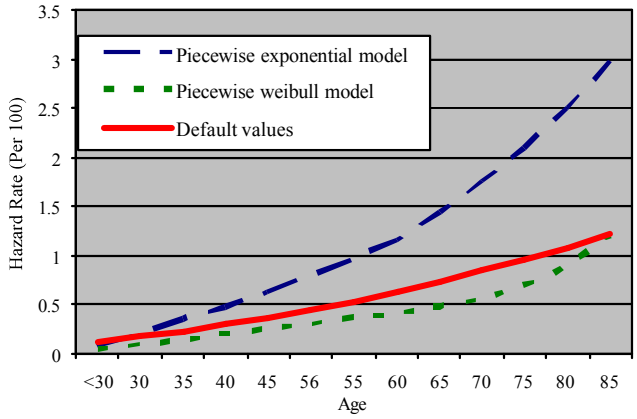

Fig. 1: Estimates and default values of age-specific hazard function of ovarian cancer in the absence of breast cancer in carriers

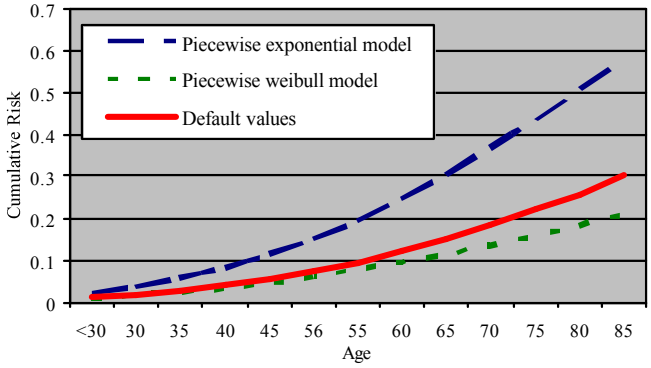

Fig. 2: Estimates and default values of cumulative risk (penetrance) of ovarian cancer in the absence of breast cancer in carriers

Table 1: Confidence intervals of age-specific hazard functions of ovarian cancer in the absence of breast cancer in BRCA1/BRCA2 gene carriers in both piecewise models on the basis of different age groups

Age group	95% Confidence Intervals	
	Piecewise exponential model	Piecewise Weibull model
< 30	0.00069-0.00075	0.00025-0.00040
30-35	0.00279-0.00313	0.00092-0.00157
35-40	0.00382-0.00423	0.00117-0.00202
40-45	0.00512-0.00570	0.00149-0.00266
45-50	0.00684-0.00748	0.00183-0.00338
50-55	0.00836-0.00912	0.00220-0.00437
55-60	0.01006-0.01107	0.00244-0.00514
60-65	0.01202-0.01331	0.00261-0.00572
65-70	0.01536-0.01696	0.00309-0.00704
70-75	0.01812-0.01975	0.00341-0.00827
75-80	0.02208-0.02431	0.00445-0.01145
80-85	0.02559-0.02821	0.00503-0.01390
85-90	0.03101-0.03479	0.00744-0.02195

The Δ is easy to interpret and allow a quick strength of evidence comparison and ranking of candidate models. If $\Delta \leq 2$, there is strong evidence of the equality of the two models. $4 \leq \Delta \leq 7$ has a considerably less support and models having $\Delta > 10$ have essentially no support. The model with less *AIC* is suitable [17].

Because of discontinuity of hazard functions in these models, Plots for hazard estimates are obtained after smoothing original estimates, using a moving average method of length 10-year

RESULTS

Results from simulated experiments show the bias in estimation of age-specific hazard (Fig. 1 and 3) and penetrance (Fig. 2 and 4) of ovarian cancer (in the absence of breast cancer). Solid curves show hazard/penetrance functions corresponding to default underlying Weibull distribution. Dashed lines show mean of estimates over 50 simulated data via piecewise exponential model and Dotted lines show

Table 2: AIC results in piecewise exponential and Weibull models

	AIC		
	Piecewise exponential model	Piecewise Weibull model	$\Delta = AIC_{Exponential} - AIC_{Weibull}$
Non-carriers	13075	12764	311
Carriers	16880	14127	2753

corresponding mean estimates of ovarian cancer, using method developed in this study.

Our results showed that the estimate of age-specific hazard values and cumulative risk (penetrance) of ovarian cancer in the absence of breast cancer in the carriers in all age groups was closer to the default values in the piecewise Weibull model by comparison with the piecewise exponential model (Fig. 1 and 2).

For penetrance estimation shown in Fig. 2, the bias seems to be more important, not as much as piecewise exponential model, for older ages than younger in carrier group by proposed method.

Computing the 95% confidence intervals showed that in almost all age groups, there was no overlap of age-specific hazards between the two models (Table 1).

As was mentioned previously in Methods, given $\Delta>10$ and computed differences of AIC in both groups of carriers and non-carriers in the two surveyed models, there was evidence of difference in the estimates yielded by the two models (Table 2).

CONCLUSSION AND SUGGESTIONS

Detection of persons susceptible to cancer and estimation of incidence probabilities for different age groups are of utmost importance [1]. Several methods for the estimation of incidence probabilities have been developed over the years. The cohort design is one such estimation method with the ability to study the etiology of multiple diseases. In the kin-cohort analysis, estimation and interpretation of parameters while studying the effect of a gene depend on proper accounting for any other competing events that may be strongly influenced by the same gene [15].

Given that hazards are likely to increase, decrease, or remain constant with respect to age variations and that age is one of the primary causes of cancer incidence, it is not advisable to assume that cause-specific hazards at any given time interval remain constant. Therefore, a hazard model based on the Weibull distribution, which covers these variations, can be of great interest.

In this article, we analyzed the piecewise Weibull model against the piecewise exponential model. We used simulated data for the purpose. In our study, the results from simulated data showed that we can expect better results when the hazard model is based on the Weibull distribution. Our results demonstrated that the estimated cause-specific hazard functions and cumulative risks (penetrance) of ovarian cancer in the absence of breast cancer in the two groups of carriers and non-carriers via the piecewise Weibull model were closer than piecewise exponential model to the default theoretical distribution.

For comparing the recommended model (piecewise weibull) with piecewise exponential model,We have used the AIC. The marked differences of AICs (311 for Non-carriers and 2753 for Carriers) showed significant differences between two methods and bear out the superiority of the piecewise Weibull model

Because of importance of estimation of the cancer risk in individuals, the methods which can generate most accurate estimations are preferable. Therefore, we recommend the piecewise Weibull model as a proper model for the estimation of hazard and penetrance in case of competing event.

REFERENCES

1. Nussbaum, R.L., R.R. McInnes and H.F. Willard, 2001. Thompson and Thompson genetics in medicine. W.B. Saunders, Philadelphia
2. Gail, M.H., D. Pee and R. Carroll, 1999. Kin-cohort designs for gene characterization. J. Nat. Cancer Inst. Monogr., 26: 55-60.
3. Teimori, H., P. Mehdipour, M. Atri and M.R. Mirzai, 2002. Mutation detection in exons 3, 10, 12 of BRCA1 gene in 30 patients affected with familial breast cancer. J. Gorgan University Med. Sci., 3 (8): 19-24.
4. Yasaee, W.R., A. Dalton and D.P. Hornby, 2006. The New Genetically Mutations in the Breast Cancer's main Genes (BRCA1, BRCA2) in Iranian women affected by Unripe Cancer. Research in Medicine. R. J. Med. Sci., 28 (2): 101-108.
5. Gayther, S.A., P. Russell, P. Harrington, A.C. Antoniou, D.F. Easton and B.A.J. Ponder, 1999. The contribution of germline BRCA1 and BRCA2 mutations to familial ovarian cancer: No evidence for other ovarian cancer-susceptibility genes. Am. J. Hum. Genet., 65: 1021-1029.
6. Streuwing, J.P., P. Hartge, S. Wacholder, S.M. Baker, M. Berlin, M. McAdams, M.M. Timmerman, L.C. Brody and M.A. Tucker, 1997. The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. N. Engl. J. Med., 336 (20): 1401-1408.

7. Easton, D.F., D. Ford and D.T. Bishop, 1995. Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. Am. J. Hum. Genet., 56 (1): 265-271.
8. Hartge, P., J.P. Struewing, S. Wacholder, L.C. Brody and M.A. Tucker, 1999. The prevalence of common BRCA1 and BRCA2 mutations among ashkenazi jews. Am. J. Hum. Genet., 64: 963-70.
9. Wacholder, S., P. Hartge, J.P. Streuwing, D. Pee, M. McAdams, L. Brody and M. Tucker, 1998. The kin-cohort study for estimating penetrance. Am. J. Epidemiol., 148: 623-630.
10. Moore, D.F., N. Chatterjee, D. Pee and M.H. Gail, 2001. Pseudo-likelihood estimates of the cumulative risk of an autosomal dominant disease from a kin-cohort study. Genetic Epidemiology, 20: 210-227.
11. Chatterjee, N. and S. Wacholder, 2001. A marginal likelihood approach for estimating penetrance from kin-cohort designs. Biometrics, 57: 245-252.
12. Elston, R.C., J.M. Olson and L. Palmer, 2002. Biostatistical Genetics and Genetic Epidemiology. John Wiley & Sons Ltd., pp: 419-424.
13. Saunders, C. and C. Begg, 2003. Kin-cohort evaluation of relative risks of genetic variants. Genetic Epidemiology, 24: 220-229.
14. Sigurdson, A.J., M. Hauptmann, N. Chatterjee and B.H. Alexander, 2004. Kin-cohort estimates for familial breast cancer risk in relation to variants in DNA base excision repair. BRCA1 interacting and growth factor Genes. BMC Cancer, 12: 4-9.
15. Chatterjee, N., P. Hartge and S. Wacholder, 2003. Adjustment for competing risk in kin-cohort estimation. Genetic Epidemiology, 25: 303-313.
16. Gail, M.H., D. Pee, J. Benichou and R. Carroll, 1999. Designing studies to estimate the penetrance of an identified autosomal dominant mutation: Cohort, case-control and genotyped-proband designs. Genetic Epidemiology, 16: 15-39.
17. Raftery, A.E., 1996. Approximate bayes factors and accounting for model uncertainty in generalized linear regression models. Biometrika, 83: 251-266.