PLISKA
STUDIA MATHEMATICA BULGARICA
ПЛИСКА
БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

For further information on
Pliska Studia Mathematica Bulgarica
visit the website of the journal http://www.math.bas.bg/~pliska/
or contact: Editorial Office
Pliska Studia Mathematica Bulgarica
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Telephone: (+359-2)9792818, FAX:(+359-2)971-36-49
e-mail: pliska@math.bas.bg

Pliska Stud. Math. Bulgar. **18** (2007), 111–120

PLISKA
STUDIA MATHEMATICA
BULGARICA

NONPARAMETRIC VERSUS PARAMETRIC STATISTICAL APPROACHES FOR GENETIC ANTICIPATION: THE PANCREATIC CANCER CASE

Gleb R. Haynatzki Vera R. Haynatzka Randall E. Brand
Henry T. Lynch Simon A. Sherman[1]

Genetic anticipation for a particular disease can involve an earlier age of onset, greater severity, and/or a higher number of affected individuals in successive generations within a family. Comparison between nonparametric and semiparametric tests is studied for matched data, and is one of the main focuses of this study. This comparison is investigated for the variable age of diagnosis among different birth cohorts, before and after adjustment for time under observation. The comparison is illustrated on an example of familial pancreatic cancer, which example is the second main focus of this study. The nonparametric test performed on our example better than the two semiparametric tests, and was less sensitive to right censoring. After adjusting for follow up time, all methods detected genetic anticipation.

1. Introduction

Genetic anticipation can involve an earlier age at onset of hereditary/familial disease, greater disease severity, and/or a higher number of affected individuals in successive generations within a family. Different genetic mechanisms (e.g.

[1]This work was supported in part by a grant from the National Cancer Institute (1 R33 CA10595-01A2) to S. A. Sherman. G. R. Haynatzki thanks Mr. Oleg Shats and Mrs. Marsha Ketcham for their help with the PCCR.

2000 *Mathematics Subject Classification:* 62N01, 62N05, 62P10, 92D10, 92D30

Key words: Statistical genetics, genetic epidemiology, anticipation, pancreatic cancer, PCCR

trinucleotide repeat expansion) have been suggested to generate this phenomenon for different diseases (e.g. familial leukaemia, lymphoma, Huntington's disease, myotonic dystrophy) [1]-[16]. Established anticipation provides cues to the nature of the disease and facilitates prediction of age of disease onset. On the other hand, it is an area of genetic epidemiology dealing with practical problems that do not have easy and unquestionable solutions.

Serious dangers to the analysis of anticipation are posed by possible biases from several sources. These are ascertainment bias (e.g. fecundity bias in selection of probands parents with later age of onset while early diagnosed potential parents are excluded or when children with early diagnosis are the probands while later diagnosed children are excluded), difference in length of follow-up "at risk" time between generations, effect of secular trends (e.g. increased smoking rate or changing dietary habits or changing quality of health care). Additionally, possible family-clustered structure of the data, i.e. intra-familial correlation due to shared genotype and/or environment, as well as the information in the censored data representing the unaffected family members have to be taken into account when analyzing the data.

Pancreatic cancer is a deadly disease with average time to death after diagnosis of less than ten months. This, in particular, means that age of diagnosis and age of death are indistinguishable for the purpose of statistical analysis. There are no early diagnostic tests for pancreatic cancer, and therefore established genetic anticipation could improve the chances for early diagnosis of individuals and their survival. Several studies claiming that genetic anticipation for pancreatic cancer exists [17], [18] have been small and hence inadequate, whereas a more recent one and of much larger sample size [19] has been more convincing. The present study goes beyond purely theoretical comparison of advanced statistical methods for anticipation. It focuses on their application to the case of familial pancreatic cancer, and is an important step in establishing the presence of genetic anticipation in this disease on firm scientific basis.

2. Methods and Results

Here we do not consider/detect disease severity but rather age of onset and possible higher number of affected individuals in successive generations. The disease we have used as an example in our analyses is pancreatic cancer whereas the database is the Pancreatic Cancer Collaborative Registry (PCCR) maintained at the University of Nebraska Medical Center.

Several standard statistical methods are usually used to test for genetic anticipation:

- Paired t-test for age at onset of affected parent-offspring pairs;
- Non-parametric ANOVA of age at onset of all affected on the predictor
  - Generation (G1, G2, G3); or
  - Cohort (C1, C2, C3);
- Survival analysis of type
  - Semi-parametric (Cox proportional hazards model); or
  - Non-parametric (log-rank test).

Unfortunately, both the paired t-test and the non-parametric ANOVA use only the affected individuals whereas the information in the unaffected is left unused. Both approaches are less powerful statistically and are prone to the previously mentioned biases. Thus, the paired t-test, while incorporating the family correlated structure to a certain degree, fails to do it in a systematic way yielding biased estimates. For example, if there is an affected parent and three affected children, the three matched affected parent-child pairs are treated as independent whereas they are likely correlated. Furthermore, both the basic Cox proportional hazards approach and the log-rank test are appropriate for independent observations, thus ignoring completely the family correlated structure of the data.

We have investigated the following methods that incorporate the family correlated structure and use the censored ages of unaffected individuals:

1. The non-parametric paired test by Hsu et al. [20], which is a generalization of the log-rank test; and
2. Two semi-parametric methods
   - Cox proportional hazards model with robust sandwich estimate of the covariance matrix [21]; and
   - Gamma frailty model [22].

The results yielded by these methods on a dataset of familial pancreatic cancer are subsequently compared. Whereas both semi-parametric methods above have been in use for some time, and are even implemented in statistical software, Hsu' non-parametric paired test [20] is less well-known. Therefore, we are briefly sketching it next.

2.1. Hsu's Nonparametric Matched Test Statistic

Let (X, δ) be an index age and a disease indicator. That is, X is the age of disease onset if the individual has the disease as indicated by $\delta = 1$, and X is the age at last follow-up or the age of death if the individual does not have the disease as indicated by $\delta = 0$, the latter case being a censored observation. Next, consider K families with n_{1k} individuals in the parental generation G1, and n_{2k} individuals in the children generation G2. For each family k $(k = 1, \ldots, K)$, let (X_{1ki}, δ_{1ki}) $(i = 1, \ldots, n_{1k})$ and (X_{2ki}, δ_{2ki}) $(i = 1, \ldots, n_{2k})$ be the index ages and disease indicators for individuals from the kth family in generations G1 and G2, respectively. It is further assumed that each matched pair has a different baseline hazard function. The following notation has been used [20] to simplify the expression for the test statistic:

$$\begin{aligned} \bar{Y}_{1k}(t) &= \sum_{i=1}^{n_{1k}} I(X_{1ki} \geq t), \\ \bar{Y}_{2k}(t) &= \sum_{i=1}^{n_{2k}} I(X_{2ki} \geq t), \\ \bar{Y}_{k}(t) &= \bar{Y}_{1k}(t) + \bar{Y}_{2k}(t). \end{aligned}$$

The left-hand side of the three equations above designate the number of individuals in the kth family who are at risk for developing the disease at age t in G1, G2, and in the two generations combined, respectively. Then the formula for the matched test statistic will be

$$U_1 = K^{-1/2} \sum_{k=1}^{K} \left[\sum_{i=1}^{n_{1k}} \delta_{1ki} \frac{\bar{Y}_{2k}(X_{1ki})}{\bar{Y}_{k}(X_{1ki})} - \sum_{i=1}^{n_{2k}} \delta_{2ki} \frac{\bar{Y}_{1k}(X_{2ki})}{\bar{Y}_{k}(X_{2ki})} \right] ,$$

which is the standardized sum of the weighted differences of affecteds between generations G1 and G2. In particular, when no censored or tied data are present, this test statistic is simply that for the sign test. The latter test uses the proportion of pairs per family in which pairs the subject in generation G1 is affected earlier than the one in generation G2. The variance of the test statistic U_1 is also shown in [20], and it is also known that U_1 is asymptotically normal.

2.2. An Application to Pancreatic Cancer

The methods from the previous section have been applied to a subset of the Pancreatic Cancer Collaborative Registry (only data of Evanston Northwestern

Healthcare and Creighton University, Dr. Henry T. Lynchs Hereditary Cancer Database) maintained at the University of Nebraska Medical Center in Omaha. Families with at least two affected members and with verified Pancreatic Cancer (PC) syndrome were included in the analyses. Either two consecutive generations (G1 and G2) or just one (G1) per family were selected. For a family to provide two generations, the younger (G2) had to include 1+ (affected or unaffected) member of 65+ years, and the youngest such generation was named G2. Otherwise, if a family had only one generation with members of 65+ years and with 1+ affected, it was obviously the oldest generation affected from that family and was therefore selected as generation G1 in the analyses. Thus, a total of 1,063 individuals from 52 families, with 12.6% affected, were analyzed. The statistical package SAS 9.1 (SAS Inst, Inc, Cary, NC, USA) was used in all calculations.

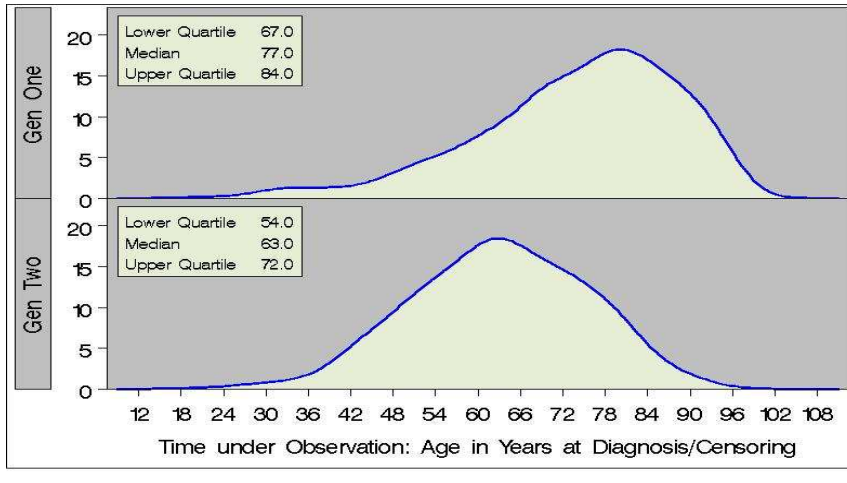

Figure 1: Distibution of time under observation, by generation.

Fig. 1 summarizes the median time under observation per generation, showing that generation G1 (median = 77.0 years) has been observed longer than generation G2 (median = 63.0 years). This indicates that possible bias may be introduced in the data analyzes if no adjustment for time under observation is made.

We next estimated the hazard function for each generation using the life-table approach, as shown on Fig. 2. The older generation, G1, has hazard function smaller than the one for G2 up to and about age 63 years. However, it is obvious that the relation between the two hazard functions is dependent on age, and it changes after age 63 years, when the hazard function for G2 looks better than

that for G1. On the other hand, estimates of the hazard function at larger ages may be less stable and less reliable since more censored observations are present.

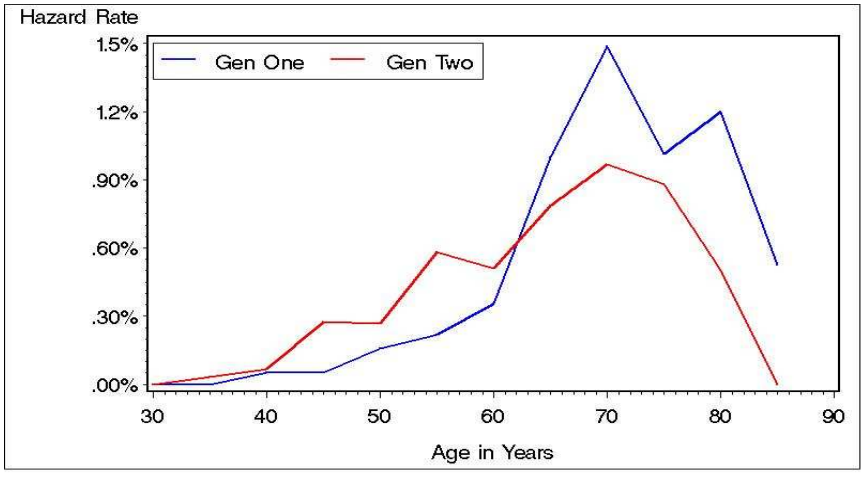

Figure 2: Hazard function estimated by the lifetable method.

We first compared the methods for detection of genetic anticipation by keeping age as is, i.e. unrestricted. The values of the test statistic and the respective p-values are shown below.

2.3. Comparing Methods: Unrestricted Age

The methods performed in this case as follows.

- Hsu's nonparametric paired test:
  - Z-score = 6.36, P-value < 0.0001
- Cox proportional hazards model with robust sandwich estimate of the covariance matrix:
  - Hazard Ratio (SE) = 1.13, P-value = 0.581
- Gamma frailty model:
  - Hazard Ratio (SE) = 1.01, P-value = 0.976

The difference between Hsu's nonparametric matched test and the two semiparametric tests is considerable, the former yielding statistically significant result, whereas the latter two are far from significant at level of significance $\alpha = 0.05$.

2.4. Comparing Methods: Age Restricted at 65 Years

In order to decrease the likely bias generated by difference in length of follow-up "at risk" time between generations, similar to the approach in [19], we have censored age at 65 years for all individuals, which also helped meet the proportional hazards assumption. This resulted in the following values of the respective test statistic and its p-value.

- Hsu's nonparametric paired test:
  - Z-score = 6.70, P-value < 0.0001
- Cox proportional hazards model with robust sandwich estimate of the covariance matrix:
  - Hazard Ratio (SE) = 1.94, P-value = 0.0163
- Gamma frailty model:
  - Hazard Ratio (SE) = 1.83, P-value = 0.0373

Now the results of the three tests are much more coherent, all being statistically significant at level of significance $\alpha = 0.05$. Finally, the results clearly show that Hsu's nonparametric matched test is less sensitive to right censoring than the two semiparametric methods, and is the method of choice at this stage of the analyses.

3. Discussion

Here we do not consider/detect disease severity but rather age of onset and possible higher number/incidence of affected individuals in successive generations. In order to further focus just on age of onset, one has to adjust for disease incidence in the cohort, as has been attempted in [23]. Our study found evidence of anticipation in pancreatic cancer confirming the results from an even larger study [19] while minimizing the effect of artifacts. We used data from the Pancreatic Cancer Collaborative Registry (PCCR) maintained at the University of Nebraska Medical Center. The PCCR is a recently created international

database comprising data from several medical centers based in the United States and Europe.

Our comparison of Hsu's nonparametric matched test [20] with the two semiparametric methods [21], [22] showed that both allow heterogeneity of disease and genetic prevalence across families by stratifying the subjects by families. Also, neither approach makes distributional assumptions about age of onset. However, the semi-parametric methods require the proportional hazards assumption. Inclusion of time-dependent covariates may improve the model fit for the semiparametric methods. On the other hand, it is difficult to adjust for covariates in the nonparametric approach. The results clearly show that Hsu's nonparametric matched test is less sensitive to right censoring than the two semiparametric methods, and is the method of choice at this stage of the analyses. However, further improvements in the semiparametric approaches may change this situation if their sensitivity to censoring is considerably improved and the available information is utilized more efficiently.

Age at onset may appear to decrease at subsequent generations due to cohort effects. These could arise from changes in environmental, treatment and diagnostic factors. Other competing risks or secular (i.e. non genetic anticipation) trends may also be present. Some of these trends may be in reverse direction of anticipation, and decreasing it. Smoking is the only confirmed environmental risk factor for both sporadic and familial pancreatic cancer. However, there are several genetic mutations, including germline mutations in the BRCA2 gene, that have been identified as risk factors and these may occur in up to 20% of familial pancreatic cancer families [24], [25].

Our results are preliminary, and we currently work on further developing both the database and the methods used. In particular, we would like to extend the capabilities of the described nonparametric methods to incorporate covariates. We also work on testing a wider set of approaches for genetic anticipation to an augmented PCCR database.

REFERENCES

[1] Sutherland G. R., Richards R. I. Simple tandem DNA repeats and human genetic disease *Proc. Natl. Acad. Sci. USA* (1995) **92**, 3636–41.

[2] Horwitz M., Goode E. L., Jarvik G. P. Anticipation in familial leukemia *Am. J. Hum. Genet.* (1996) **59**, 990–8.

[3] Goldstein A. M., Clark W. H. Jr, Fraser M. C. et al. Apparent anticipation in familial melanoma, *Melanoma Res.* (1996) **6**, 441–6.

[4] Paterson A. D., Kennedy J. L., Petronis A. Evidence for genetic anticipation in non-Mendelian diseases, *Am. J. Hum. Genet.* (1996) **59**, 264–8.

[5] Fraser F. C. Trinucleotide repeats not the only cause of anticipation, *Lancet* (1997) **350**, 459–60.

[6] La Spada A. R. Trinucleotide repeat instability: genetic features and molecular mechanisms, *Brain Pathol.* (1997) **7**, 943–63.

[7] Petronis A., Kennedy J. L., Paterson A. D. Genetic anticipation: fact or artifact, genetics or epigenetics?, *Lancet* (1997) **350**, 1403–4.

[8] Grandbastien B., Peeters M., Franchimont D. et al. Anticipation in familial Crohn's disease, *Gut* (1998) **42**, 170–4.

[9] Siegel A. M., Andermann F., Badhwar A. et al. Anticipation in familial cavernous angioma: ascertainment bias or genetic cause, *Acta Neurol. Scand.* (1998) **98**, 372–6.

[10] De Lord C., Powles R., Mehta J. et al. Familial acute myeloid leukaemia: four male members of a single family over three consecutive generations exhibiting anticipation, *Br. J. Haematol.* (1998) **100**, 557–60.

[11] Yuille M. R., Houlston R. S., Catovsky D. Anticipation in familial chronic lymphocutic leukaemia, *Luekemia* (1998) **12**, 1696–8.

[12] Kehoe P., Krawczak M., Harper P. S. et al. Age of onset in Huntington disease: sex specific influence of apolipoprotein E genotype and normal CAG repeat length, *L. Med. Genet.* (1999) **36**, 108–11.

[13] Wiernik P. H., Wang S. Q., Hu X. P. et al. Age of onset evidence for anticipation in familial non-Hodgkin's lymphoma, *Br. J. Haematol.* (2000) **108**, 72–9.

[14] Annese V., Andreoli A., Astegiano M. et al. Clinical features in familial cases of Crohn's disease and ulcerative colitis in Italy: a GISC group. Italian Study Group for the Disease of Colon and Rectum, *Am. J. Gastroenterol.* (2001) **96**, 2939–45.

[15] Radstake T. R., Barrera P., Albers M. J. et al. Genetic anticipation in rheumatoid arthritis in Europe. European Consortium on Rheumatoid Arthritis Families, *J. Rheumatol.* (2001) **28**, 962–7.

[16] Shugart Y. Y., Hemminki K., Vaittinen P. et al. Apparent anticipation and heterogeneous transmission patterns in familial Hodgkin's and non-Hodgkin's lymphoma: report from a study based on Swedish cancer database, *Leuk. Lymphoma* (2001) **42**, 407–15.

[17] Rieder H., Sina-Freey M., Ziegler A. et al. German national case collection of familial pancreatic cancer – clinical genetic analysis of the first 21 families, *Onkologie* (2002) **25**, 262–6.

[18] Rulyak S. J., Lowenfels A. B., Maisonneuve P. et al. Risk factors for the development of pancreatic cancer in familial pancreatic cancer kindreds, *Gastroenterology* (2003) **124**, 1292–9.

[19] Greenhalf W., McFaul C., Earl J. et al. Anticipation in familial pancreatic cancer, *Gut* (2006) **55**, 252–8.

[20] Hsu L., Zhao L. P., Malone K. E. et al. Assessing changes in ages at onset over successive generation: an application to breast cancer, *Genet. Epidemiol.* (2000) **18**, 17–32.

[21] Binder D. A. Fitting Cox's proportional hazards models from survey data, *Biometrika* (1992) **79**, 139–47.

[22] Klein J. Semiparametric estimation of random effects using Cox model based on the EM algorithm, *Biometrics* (1992) **48**, 795–806.

[23] Daugherty S. E., Pfeiffer R. M., Mellemkjaer L. et al. No evidence for anticipation in lymphoproliferative tumors in population-based samples, *Cancer Epidemiol. Biomarkers Prev.* (2005) **14**, 1245–50.

[24] Klein A. P., Brune K. A., Petersen G. M. et al. Prospective risk of pancreatic cancer in familial pancreatic cancer kindreds, *Cancer Res.* (2004) **64**, 2634–8.

[25] Hahn S. A., Greenhalf B., Ellis I. et al. NRCA2 germline mutations in familial pancreatic carcinoma, *J. Natl. Cancer Inst.* (2003) **95**, 214–21.

Gleb R. Haynatzki
Creighton University
Omaha, NE, USA
email: glebhaynatzki@creighton.edu