

Penetrance of HNPCC-related cancers in a retrolective cohort of 12 large Newfoundland families carrying a MSH2 founder mutation: an evaluation using modified segregation models

Karen A. Kopciuk*¹, Yun - Hee Choi², Elena Parkhomenko³, Patrick Parfrey⁴, John McLaughlin^{5,6}, Jane Green⁷ and Laurent Briollais⁶

¹Division of Population Health Research, Alberta Health Services Board, Calgary, Alberta, Canada

²Department of Epidemiology and Biostatistics, The University of Western Ontario, London, Ontario, Canada

³Research Institute, Hospital for Sick Children, Toronto, Ontario, Canada

⁴Clinical Epidemiology Unit, Memorial University, St. John's, Newfoundland, Canada

⁵Population Studies and Surveillance, Cancer Care Ontario, Toronto, Ontario, Canada

⁶Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada

⁷Discipline of Genetics, Memorial University, St. John's, Newfoundland, Canada

Email: Karen A. Kopciuk* - kakopciu@ucalgary.ca; Yun - Hee Choi - Yun-Hee.Choi@schulich.uwo.ca; Elena Parkhomenko - elena@utstat.toronto.edu; Patrick Parfrey - pparfrey@mun.ca; John McLaughlin - john.mclaughlin@cancerca.on.ca; Jane Green - janeg@mun.ca; Laurent Briollais - laurent@lunenfeld.ca;

*Corresponding author

Abstract

Background: Accurate risk (penetrance) estimates for associated phenotypes in carriers of a major disease gene are important for genetic counselling of at-risk individuals. Population-specific estimates of penetrance are often needed as well. Families ascertained from high-risk disease clinics provide substantial data to estimate penetrance of a disease gene, but these estimates must be adjusted for possible specific sources of bias.

Methods: A cohort of 12 independently ascertained HNPCC families harbouring a founder MSH2 mutation was identified from a cancer genetics clinic in St. John's, Newfoundland, Canada. Carrier status was known for 257 family members but phenotype information on up to 75 additional relatives with unknown carrier status was available; using modified segregation models these additional individuals could be included in the analyses. Three HNPCC-related phenotypes were evaluated as age at diagnosis of: any HNPCC cancer (first cancer), colorectal cancer (CRC), and endometrial cancer (EC) for females.

Results: Lifetime (age 70) risk estimates for male and female carriers were similar for developing any HNPCC cancer (Males = 97.6%, 95% Confidence Interval (CI)=(90.6%, 99.4%); Females = 93.6%, 95% CI=(85.1%, 98.4%)) but female carriers experienced substantially reduced lifetime risk for developing CRC compared to male carriers (Females = 42.7%, 95% CI=(28.3%, 63.4%); Males = 82.6%, 95% CI=(63.2%, 88.3%)). Female non-carriers had very low lifetime risk for these two outcomes while male non-carriers had lifetime risks intermediate to the female carriers and non-carriers. Female carriers had a lifetime risk of developing EC of 63.3% (95% CI=(12.6%, 92.6%)). Relative risks for developing any HNPCC cancer (carriers relative to non-carriers) were substantially greater for females compared to their male counterparts (Females = 15.2, 95% CI=(6.8, 33.5);

Males = 8.2, 95% CI = (3.5, 16.9)). Relative risks for developing CRC were substantially greater for females compared to their male counterparts (Females = 11.9, 95%CI=(5.7, 28.6); Males = 4.3%, 95% CI = (1.6, 5.8)). However, the risk of developing CRC decreased with age among both genders.

Conclusions: The proposed modified segregation-based models used to estimate age-specific risks for HNPCC phenotypes can reduce bias due to ascertainment and missing genotype information as well as provide estimates of absolute and relative risks.

Background

Extensive knowledge about Mendelian disease genes is accumulating as more are discovered, characterized and studied in high risk families. With this accumulating knowledge comes the opportunity for carriers to have improved surveillance and treatment options so disease is detected at an early stage and adverse outcomes are reduced. However, accurate age- and sex-specific penetrance estimates are critical for genetic counseling of at-risk individuals in order to establish and evaluate disease control efforts. For instance, risk models have recently been developed to predict germline mutations linked to Lynch syndrome (hereditary nonpolyposis colorectal cancer or HNPCC) but unlike existing models, these models require a prior specification of the penetrance function to give accurate predictions. [1–3]

HNPCC is an example of an autosomal dominant Mendelian disease which has seen remarkable improvements in associated morbidity and mortality in recent years due to early identification, surveillance and treatment of gene carriers. [4, 5] It is caused by a mutation in one of the mismatch repair genes: MLH1, MSH2, MSH6 or PMS2 and accounts for about 3% of all colorectal cancer (CRC) cases. [6] Carriers of one of these mutations are at increased lifetime risk of developing cancers of the colon, endometrium, ovary, stomach, ureter, upper biliary tract, skin, and brain. [7, 8] Early age at onset is another hallmark of this syndrome.

Studies of heritable genetic diseases are often conducted using families with multiple affected individuals since high risk alleles are often rare in the general population. However, families identified through high risk disease clinics tend to be more likely to carry the disease gene mutation and to have phenotype information, compared to pedigrees iden-

tified through population-based sampling of affected probands. Thus, families identified in clinic-based designs may not be representative of the population and ascertainment may be biased to more severe outcomes. In addition to ascertainment issues, missing genotype information can be problematic in these pedigrees. Genetic testing may not be offered to all family members based on an individual's risk of carrying the deleterious allele and some family members may have already died without testing. Still others may decline testing.

Statistical models based on segregation methods have been developed which can estimate age-specific penetrance in these high risk families and can adjust for ascertainment bias and missing genotype data. Traditionally, segregation-based methods have been used to fit genetic models to phenotype data collected from pedigrees. These methods can easily adjust for complex ascertainment when the rules of sequential sampling of pedigrees are followed, infer the missing genotypes using family structure and evaluate whether another segregating gene might be involved. [9, 10] More precise estimates and confidence intervals (CIs) can be obtained with these methods compared to standard methods, since additional genotyped individuals are included. When the form of the disease risk (e.g., risk increasing or decreasing with age or both) could vary for different age-at-onset phenotypes, a general formulation of the hazard function is necessary.

This paper addresses these key challenges in disease risk estimation by using new statistical methods in a study of 12 HNPCC families, who were ascertained at a Newfoundland (NL) cancer genetics clinic and who share a founder MSH2 mutation. In Newfoundland, founder mutations causing autosomal dominant disease, large family sizes over 8-10

generations and little in or out migration since the founding settlements in the late 18th and early 19th centuries [11] have enabled discoveries of phenotype - genotype correlations in right arrhythmogenic ventricular cardiomyopathy, [12] autosomal dominant polycystic kidney disease, [13] and HNPCC. [14] Risk estimates from other populations may not be appropriate for these NL families who share lifestyle and environmental factors in addition to the founder mutation; each of these risk factors can substantially affect the rates of penetrance. [15–18]

Methods

Family ascertainment and characteristics

Twelve families with hereditary CRC were independently identified from the Medical Genetics Clinic at Memorial University, St. John’s, NL, Canada, and were confirmed as carrying the MSH2 mutation A → T nt942+3. Based on their geographic proximity in an isolated area of the province, as well as having common haplotypes of a subset of at least five microsatellite markers, it is assumed these families have a common ancestor. See Green et al. (2002) for additional details. [7] Data collected from these families have been updated since that initial report, including additional HNPCC outcomes and genotyping of family members.

DNA testing for the MSH2 mutation identified individuals as carriers or non-carriers. For those not tested, family members who presented with an HNPCC cancer and were known to be at least at 50% risk of being a carrier, or who had children known to be carriers, were presumed to be carriers. All other individuals were considered to have unknown mutation status.

Statistical Methods

The clinical outcomes of interest (phenotypes) in this study were age to diagnosis of CRC, age to diagnosis of first HNPCC cancer (included CRC, small bowel, endometrium, ovary, gastric, urinary tract, brain, and bile duct), and among the females, age to diagnosis of endometrial cancer (EC). Lifetime risk

or penetrance at age 70 for these three outcomes is of particular interest. Observations on participants were censored at their age at last followup if no phenotypes were observed; after entering a screening program; or for age to diagnosis of EC for females, after having a hysterectomy/oophorectomy for non-cancer reasons.

The Kaplan-Meier (KM) estimator is often used to estimate age-specific penetrance but has known limitations. [19] Although it is robust to the correlation in outcomes among family members, it does not correct for ascertainment bias associated with these high risk families or infer missing genotype information. We used the KM estimator to assess potential bias and differences in cumulative risks across the different mutation groups – particularly to compare penetrance for those with known versus unknown mutation status.

Modified segregation-based methods can account for the non-random ascertainment of families as well as missing genotype information. [9, 10, 20] We considered a general parametric form for the hazard function, which adopted a Weibull baseline hazard function. The generalized log-Burr hazard form (see Appendix) includes the standard Weibull proportional hazards model or the log-logistic model as special cases. [21] The Weibull model is quite flexible but does have a monotonic functional form of the hazard whereas the log-logistic specification does not. As Jenkins et al. (2006) found, risk can increase and then decrease with age, so assuming monotonicity for risk may not always be appropriate. [22] With three clinical phenotypes of interest, the general formulation permits appropriate functional form choices for each outcome. Likelihood ratio statistics were used to select a simpler model (i.e., either Weibull or log-logistic) that was compatible with the general model. [23] All regression models included the variables for sex and genotype status and their interaction.

Stratified risk estimates were calculated by sex and by mutation status for each phenotype and the corresponding 95% confidence intervals were calculated by simulation. Specifically, 1,000 parameters sets were simulated, assuming a multivariate normal distribution for the estimated parameters. The 95% CI for the cumulative risk to a specific age (range 20-70 years) was estimated as the region between the 2.5th

and 97.5th percentiles of the distribution obtained by substituting the sets of simulated parameters into the penetrance function. [22]

To correct penetrance estimates for families ascertained through an individual with cancer who also satisfied high or intermediate familial risks, we used an ascertainment-corrected retrospective (ACR) likelihood approach. Correction was based on the conditional probability of the observed genotypes in the first degree relatives of the carrier proband, given the observed presence or absence of disease in each person, and corrected for the ascertainment event (i.e. the diagnosis of cancer in the proband) which caused the family to be sampled. [9, 10, 24] We obtained the maximum likelihood estimates of the needed parameters by maximizing this ACR likelihood and then used the parameters to estimate age-specific cumulative and relative risks for each outcome, using the penetrance function in equation (2) of the Appendix. [9, 10]. These modified segregation methods based on an ACR likelihood were implemented in the genetic software program, Mendel. [25]

Results and Discussion

Description of Family Data Set

The number of family members with phenotype information varied from 146 (EC) to 313 (any HNPCC cancer) to 332 (CRC), with 129 males and 127 females having known genotypes (see Table 1). However, family members with unknown mutation status often had a substantial number of outcomes. For instance, nine of the 36 males with unknown carrier status were diagnosed with having any HNPCC cancer. Excluding these family members who were not genotyped and were not presumed carriers will impact the risk estimates.

Table 1 about here

Kaplan-Meier estimator

In Figure 1, the standard Kaplan-Meier (KM) estimator is used to estimate penetrance of the MSH2 gene in the development of any HNPCC cancer and CRC, stratified by genotype status. What is striking about the resulting complementary (1-KM) curves

is the substantial differences among the groups with different genotype status. For development of any HNPCC cancer and CRC, the carriers (denoted by ‘+’) and individuals with unknown genotype status (denoted by ‘?’) have very similar age-specific risks of developing any HNPCC cancer or CRC until age 35. From age 35 onwards, carriers have the highest penetrance while the group with unknown genotype status experience substantially less risk thereafter. Non-carriers (‘-’) have very low penetrance at all ages.

Figure 1 about here

When we further stratify by sex, very similar patterns for developing any HNPCC cancer or CRC are evident (Figure 2). In addition, females generally have lower risks compared to their male counterparts within each genotype category for both outcomes until around age 50. For all outcomes, large jumps in the curves past age 50 are due to sparse data. For estimation of the development of EC, carriers have similar risks to non-carriers and women with unknown genotype status until age 40. Past this age, female carriers exhibit an increasing risk of nearly 40%. Women with unknown carrier status and non-carriers have very low risks over their lifetimes.

These KM plots suggest that missing genotype information will likely bias the age-specific risk estimates to varying degrees for each phenotype of interest. To avoid this problem, we used a segregation-based approach.

Figure 2 about here

Modified Segregation-based Methods

The general formulation for the hazard function (shown in the Appendix and which included the interaction term) yielded these findings:

- Any HNPCC cancer: log-likelihood value was -689.7 and \hat{q} was 29600,
- CRC: log-likelihood value was -544.2 and \hat{q} was 3.1,

- EC: log-likelihood value was -346.8 and \hat{q} was 0.4 .

These results suggest that different models are appropriate for each phenotype. Likelihood ratio test statistics confirmed that the general model formulation was consistent with the Weibull model for any HNPCC cancer (p-value = 0.27), the log-logistic model for CRC (p-value = 0.19) as well as for EC (p-value = 0.39).

Based on these selected models for each phenotype, we estimated the penetrance functions using the corresponding parameter estimates. The estimated penetrance functions, stratified by sex and mutation status, are displayed in Figure 3. The age-related risk of developing any HNPCC cancer for carriers increases quickly after age 25, with females experiencing slightly lower risks than males over their lifetimes. The age-related risks for non-carriers are substantially lower than the risks for the carriers past 25 years of age, with male non-carriers experiencing greater lifetime risk than the female non-carriers. Cumulative risk for CRC was not surprisingly highest for males, compared to their female counterparts. For instance, lifetime CRC risk estimates for male carriers was 82.6% (95% CI = (63.2%, 88.3%)) whereas for female carriers it was 42.7% (95% CI = (28.3%, 63.4%)). The 95% CIs drawn at ages 30, 50 and 70 reveal that within each gender class, carriers have statistically significant higher risk than non-carriers but within each carrier class, males and females have similar ranges of plausible values. Table 2 provides the point and 95% CI estimates of the penetrance for any HNPCC cancer, CRC and EC at ages 30, 50 and 70 (lifetime risk).

Table 2 about here

Figure 3 about here

Endometrial cancer results reveal a relatively low cumulative risk to age 50 (19.2%) for female carriers which rises sharply to 63.3% by age 70. For non-carriers, risk was essentially zero throughout the women’s lifetimes. The low number of women diagnosed with EC resulted in greater estimation uncertainty, so the confidence intervals are somewhat wide and overlapping.

In addition to the absolute risk estimates obtained

with the modified segregation analyses, the relative risks (RR) for developing any HNPCC cancer and CRC were also calculated (Equation 2, Appendix). Table 3 presents the risks for carriers relative to non-carriers, stratified by sex. For age at onset of any HNPCC cancer, the Weibull model was again adopted so the RR are automatically constant with respect to age. The interaction between carrier status and sex was significant in this model. Although males had higher absolute risks compared to females regardless of carrier class, female carriers had double the relative risks (15.2), compared to male carriers (8.2). The wide 95% CI for the female carriers (6.8, 33.5), however, suggests a lack of precision; it also overlaps with the corresponding ones for the male carriers (3.5, 16.9) indicating no statistically significant difference between them. The scant number of events in female non-carriers meant RR estimates for EC were not estimable.

Table 3 about here

For age at onset of CRC, the log-logistic model was adopted so risks could vary with age. Overall, the risk for female carriers relative to female non-carriers was higher than for male carriers relative to male non-carriers over the entire age range considered. At age 30, the RR for male carriers, 18.2 (95% CI=(6.8, 36.7)) was nearly the same as for female carriers, 19.7 (95% CI=(8.5, 45.6)). But by age 70, the relative risk for male carriers (4.3, 95% CI=(1.6, 5.8)) was one third the relative risk for female carriers (11.9, 95% CI=(5.7, 28.6)). Thus, female carriers had greater relative risks of developing CRC compared to their male counterparts through out their adult lifetimes; this risk difference became substantial by age 70.

Discussion

Our paper confirmed the relatively high penetrance associated with MSH2 gene mutations in these NL HNPCC families for various phenotypes, including CRC. Several studies have previously evaluated the risk of developing CRC among identified MMR gene mutation carriers in HNPCC families, where the risk to age 70 ranged from 22% to 100%. [7, 15–18, 26–31] Some studies suggested different risks associated

with MLH1 and MSH2 [16, 28] while others reported males having nearly twice the risk of females [15, 16, 18]; these differences were not consistent across studies. The lower penetrance reported among females suggests they are somehow protected from CRC, perhaps due to environmental/reproductive factors unique to women or to a sex-linked modifier gene.

Our study confirmed a gender-specific effect on the risk of developing CRC with a higher lifetime risk in males than females (82.6% vs. 42.7% by age 70). Our study also underlined the relative importance of other cancers in these large Newfoundland families. Women are susceptible to developing endometrial cancer with a penetrance close to 63.3% by age 70 and when considering all HNPCC-related cancers, the lifetime risks in males and females carriers were similarly very high, 97.6% and 93.6%, respectively. Interestingly, our study also showed that the relative effect of MSH2 on CRC for males and females combined as measured by the hazard ratio, decreases with age (from 17.7 at age 30 to 7.1 at age 70, results not shown). Such an effect was also suggested in Jenkins et al. [22] and in a very recent study from Ontario. [9]

To illustrate the importance of having specific penetrance estimates for the HNPCC families in Newfoundland (NL), we estimated both the probability of being a MSH2 mutation carrier and the probability of developing CRC and EC in cancer-free individuals in Family 1 from our sample. These probabilities can be computed for any family member, i.e. the counselee, and are conditional on the observed phenotypes and genotypes in the family. Gender- and age-specific penetrance are used to derive a posterior probability of being a mutation carrier and then the disease risk estimate is calculated as a weighted average over the mutation status, where the weights are the mutation carrier probabilities. [32] Computations were carried out using the package MMRpro in the R library BayesMendel. [32] We compared the probability of being a mutation carrier and of developing the two outcomes for a selected counselee using either publically available penetrance estimates for MSH2 [1] or the penetrance estimates obtained from our own analyses of the NL family data set.

Figure 4 about here

We chose individual 48 in family 1 as the counselee (see Figure 4) . This is a woman who is cancer free to age 48 and is a mutation carrier. Her mother (the proband in the family) had an endometrial cancer (EC) at age 47 and a colorectal cancer (CRC) at age 67 and was a confirmed mutation carrier. First, we ignore the counselee’s mutation status. Assuming a mutation prevalence varying from 0.1% to 1% in the Newfoundland population, the carrier probability for the counselee ranged from 37% to 41% when using our derived penetrance estimates and was close to 43% when using the published estimates. The probability of developing CRC and EC by age 83 for this person was 18% and 28%, respectively, when using our penetrance estimates and 12% and 19%, respectively, when using the published estimates. Next, we assumed the counselee’s mutation status is known. The probability of developing CRC and EC by age 83 was 42% and 70%, respectively, when using our penetrance estimates but only 22% and 40%, respectively, when using the published penetrance estimates. These risk estimates were not sensitive to the mutation prevalence. Thus, using our calculated penetrance values, we found both the carrier probability and development of both phenotypes to be quite different compared to those values obtained using published data from other populations.

Our paper also demonstrates the strength of the modified segregation-based analysis to estimate the penetrance of HNPCC-related cancers in these large Newfoundland families. First, unlike the classical Kaplan-Meier estimator, this approach can infer missing genotypes by using the Mendelian transmission probabilities and genealogical relationships. As a consequence, the segregation-based analysis was able to use information on up to 75 additional relatives, resulting in more precise and potentially less biased penetrance estimates. Second, the modified segregation-based analysis can correct for ascertainment when families were recruited through several affected individuals. Our recent simulation studies [9] have shown that the inference through the ascertainment-corrected likelihood approach that we proposed was nearly unbiased for various types of family-based designs (population- and clinic-based) and genetic models (including one- and two-gene models). Finally, segregation-based methods allow several parametric hazard functions to be consid-

ered. We proposed a generalized log-Burr formulation of this function, which includes the Weibull or log-logistic forms as particular cases, so a better fit of the cumulative risk function results.

For the aforementioned reasons, our results differ somewhat from those published by Green et al. [7] who used the Kaplan-Meier estimator on an earlier version of the data. For example, their penetrance estimates for CRC were 92% and 64% by age 70 in males and female carriers, respectively, compared to 83% and 43% in our study. The main difference is likely to be due to the treatment of the relatives with unknown genotypes. Omitting this category or combining it with the non-carriers would result in an overestimation of the penetrance in MSH2 carriers, whereas the correct penetrance estimate should be a weighted average over the missing data. It should also be mentioned that Green et al. [7] did not correct for ascertainment and so the estimates are valid only for their clinic-based sample and/or for families with the same pattern of cancer risks.

Our results also differ from those of Quehenberger et al. (2005), even though both studies adjusted for ascertainment of high-risk families using a retrospective likelihood approach and used modified segregation methods to infer missing genotypes. [30] Several important differences might explain the lower risk estimates these authors found in their data. Although they had substantially more data for the three phenotypes they considered (CRC, EC and Minor HNPCC site; $n=1304$ for MSH2 families), they combined families segregating either hMLH1 or MSH2 alleles (number of individuals analysed was 2392). They also had proportionally more missing genotype information (68.3% (67.5%) in MSH2 (combined) families versus 21.8% in our data), and all families likely did not share the same founder mutation.

In addition to the study population differences, methodological differences might also be relevant. Quehenberger et al. assumed a fixed population allele frequency, adopted polynomial functions (constant or quadratic) for the logarithm of the cancer- and sex-specific relative hazard rates and a piecewise constant non-carrier baseline hazard rate with kernel smoother based on the Netherlands Cancer Registry incidence data, and used a competing risk model. In particular, competing risk model estimates are affected by the rates of all phenotypes un-

der consideration, which are also likely not independent for HNPCC cancers. Absolute risk estimates for the time to first HNPCC cancer from our study (Table 2, first two rows) can be compared to their combined (CRC + EC + MC) age dependent cause specific cumulative risks (Table 4, second row, last two columns): our risk estimates at ages 30, 50 and 70 years are consistently higher for both male and female carriers although both approaches had wide 95% confidence intervals. Relative risks for time to first CRC (Table 3 in both studies) suggest decreasing risk over lifetimes, with our study estimating a smaller range of values and different values for males and females.

Conclusions

In summary, the risk estimates we obtained using a modified segregation approach within a general hazard framework are adjusted for ascertainment of the family members and can include those family members with missing genotype information. Thus, this novel and flexible approach reduces several sources of bias in the penetrance estimates for HNPCC-related phenotypes.

Although many sources of bias have been reduced in this study, several limitations may still exist. First, the sample size is relatively small for a risk estimation study and the confidence intervals were often wide, especially for estimating gender-specific penetrances. This problem was only partly overcome by using the modified segregation-based approach. Second, inaccuracy of cancer diagnosis or screening history might introduce error because some cases or dates of entry into screening programs could not be confirmed. Further work on the impact of screening is warranted. Lastly, we did not adjust our analysis for competing risks nor try to model specifically the correlation between the different HNPCC-related cancers or other sources of familial correlation besides the MSH2 mutation. Phenotypic heterogeneity and multiple phenotypes also pose challenges. We are planning to investigate these issues in some future work.

Authors contributions

LB conceived of the study. KAK and Y-HC developed the statistical methods and then performed the

statistical analyses. The manuscript was drafted by KAK. LB provided oversight of the data analyses and critical revision of the manuscript for its intellectual content. JG and PP designed the original study and continue to coordinate follow-up data collection. All authors read and approved the final manuscript.

Acknowledgements

The authors gratefully acknowledge the contributions of the HNPCC families to this research.

This research was supported by a grant from the Institutes of Genetics and Population and Public Health of the Canadian Institutes of Health Research (Grant # 110053), an Interdisciplinary Health Research Team award from the Canadian Institutes of Health Research (Grant # 43821) and a fellowship from the Canadian Breast Cancer Foundation – Ontario Chapter for Y-HC. None of the grant agencies supporting this research had any involvement in the design of this study, the manuscript preparation or decision to submit this paper for publication.

References

1. Chen S, Wang W, Lee S, Nafa K, Lee J, Romans K, Watson P, Gruber S, Euhus D, Kinzler K, Jass J, Gallinger S, Lindor N, Casey G, Ellis N, Giardiello F, Offit K, Parmigiani G, the Colon Cancer Family Registry: **Prediction of Germline Mutations and Cancer Risk in the Lynch Syndrome.** *JAMA* 2006, **296(12)**:1479–1487.
2. Terespolsky D: **Commentary: The MMRpro model accurately predicted the probability of carrying a cancer-susceptibility gene mutation for the Lynch syndrome.** *ACP J Club* 2007, **146(2)**:53.
3. Green RC, Parfrey PS, Woods MO, Younghusband HB: **Prediction of Lynch syndrome in consecutive patients with colorectal cancer.** *J Natl Cancer Inst* 2009, **101(5)**:331–40.
4. Lynch H, Lynch J, Lynch P, Attard T: **Hereditary colorectal cancer syndromes: molecular genetics, genetic counselling, diagnosis and management.** *Fam Cancer* 2008, **7**:27–39.
5. Järvinen H, Mecklin J, Sistonen P: **Screening reduces colorectal cancer rate in families with hereditary nonpolyposis colorectal cancer.** *Gastroenterology* 1995, **108(5)**:1405–11.
6. Vasen H: **Review article: the Lynch syndrome (hereditary nonpolyposis colorectal cancer).** *Aliment Pharmacol Ther* 2007, **26(S2)**:113–126.
7. Green J, O'Driscoll M, Barnes A, Maher ER, Bridge P, Shields K, Parfrey PS: **Impact of gender and parent of origin on the phenotypic expression of hereditary nonpolyposis colorectal cancer in a large Newfoundland kindred with a common MSH2 mutation.** *Dir Colon Rectum* 2002, **45**:1223–1232.
8. Reference GH: **Lynch Syndrome.** *Internet* 2008. [Cited June 12, 2009, <http://ghr.nlm.nih.gov/condition=lynchsyndrome>].
9. Choi Y, Kopciuk K, Briollais L: **Estimating disease risk associated with mutated genes in family-based designs.** *Hum Hered* 2008, **66**:238–251.
10. Kraft P, Thomas DC: **Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods.** *Am J Hum Genet* 2000, **66**:1119–1131.
11. Parfrey P, Davidson W, Green J: **Clinical and genetic epidemiology of inherited renal disease in Newfoundland.** *Kidney Int* 2002, **61**:1925–34.
12. Merner N, Hodgkinson K, Haywood A, Connors S, French V, Drenckhahn J, Kupprion C, Ramadanova K, Thierfelder L, McKenna W, Gallagher B, Morris-Larkin L, Bassett A, Parfrey P, Young T: **Arrhythmogenic right ventricular cardiomyopathy type 5 is a fully penetrant, lethal arrhythmic disorder caused by a missense mutation in the TMEM43 gene.** *Am J Hum Genet* 2008, **82**:809–21.
13. Parfrey P, Bear J, Morgan J, Cramer B, McManamon P, Gault M, Churchill D, Singh M, Hewitt R, Somlo S, et al: **The diagnosis and prognosis of autosomal dominant polycystic kidney disease.** *N Engl J Med* 1990, **323**:1085–90.
14. Stuckless S, Parfrey P, Woods M, Cox J, Fitzgerald G, Green J, Green R: **The phenotypic expression of three MSH2 mutations in large Newfoundland families with Lynch syndrome.** *Fam Cancer* 2007, **6**:1–12.
15. Aarnio M, Sankila R, Pukkala E, Salovaara R, Aaltonen L, de la Chapelle A, Peltomäki P, Mecklin J, Järvinen H: **Cancer risk in mutation carriers of DNA-mismatch-repair genes.** *Int J Cancer* 1999, **81**:214–218.
16. Lin K, Shashidharan M, Thorson A, Ternent C, Blatchford G, Christensen M, Watson P, Lemon S, Franklin B, Karr B, Lynch J, Lynch H: **Cumulative incidence of colorectal and extracolonic cancers in MLH1 and MSH2 mutation carriers of hereditary nonpolyposis colorectal cancer.** *J Gastrointest Surg* 1998, **2**:67–71.
17. Vasen H, Wijnen J, Menko F, Kleibeuker J, Taal C, Grifioen G, Nagengast F, Meijers-Heijboer E, Bertario L, Varesco L, Bisgaard M, Mohr J, Fodde R, Khan P: **Cancer risk in families with hereditary nonpolyposis colorectal cancer diagnosed by mutation analysis.** *Gastroenterology* 1996, **110(4)**:1020–1027.
18. Dunlop M, Farrington S, Carothers A, Wyllie A, Sharp L, Burn J, Liu B, Kinzler K, Vogelstein B: **Cancer risk associated with germline DNA mismatch repair mutations.** *Hum Mol Genet* 1997, **6**:105–110.
19. Kaplan E, Meier P: **Nonparametric estimation from incomplete observations.** *J. Amer. Statist. Assoc* 1958, **53**:457–481.
20. Thomas DC: *Statistical Methods in Genetic Epidemiology.* New York, Oxford University Press 2004.
21. Lawless JF: *Statistical Models and Methods for Lifetime Data, Second Edition.* Hoboken, John Wiley and Sons Inc. 2003.

22. Jenkins M, Baglietto L, Dowty J, Van Vliet C, Smith L, Mead Lea: **Cancer risks for mismatch repair gene mutation carriers: a population-based early onset case-family study.** *Clin Gastroenterol Hepatol* 2006, **4**:489–498.
23. Cox DR, Hinkley DV: *Theoretical Statistics*. New York, Chapman and Hall 1974.
24. Carayol J, Bonaïti-Pellié 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.
25. Lange K, Sinsheimer J, Sobel E: **Association testing with Mendel.** *Genet Epidemiol* 2005, **29**:36–50.
26. Aarnio M, Mecklin J, Aaltonen L, Nystrom-Lahti M, Jarvinen H: **Life-time risk of different cancers in HNPCC syndrome.** *Int J Cancer* 1995, **64**:430–433.
27. Froggart N, Green J, Brassett C, Evans D, Bishop D, Kolodner R, Maher E: **A common MSH2 mutation in English and North American HNPCC families: origin, phenotypic expression, and sex specific differences in colorectal cancer risk.** *J Med Genet* 1999, **36**:97–102.
28. Vasen H, Stormorken A, Menko F, Nagengast F, Kleibeuker J, Griffioen G, et al: **MSH2 Mutation Carriers Are at Higher Risk of Cancer Than MLH1 Mutation Carriers: A Study of Hereditary Non-polyposis Colorectal Cancer Families.** *J Clin Oncol* 2001, **19**:4074–4080.
29. Parc Y, Boisson C, Thomas G, Olschwang S: **Cancer risk in 348 French MSH2 or MLH1 gene carriers.** *J Med Genet* 2003, **40**:208–213.
30. Quehenberger F, Vasen H, van Houwelingen H: **Risk of colorectal and endometrial cancer for carriers of mutations of the hMLH1 and hMSH2 gene: correction for ascertainment.** *J Med Genet* 2005, **42**:491–496.
31. Hampel H, Stephens J, Pukkala E, Sankila R, Aaltonen L, Mecklin J, de la Chapelle A: **Cancer risk in hereditary nonpolyposis colorectal cancer syndrome: later age of onset.** *Gastroenterology* 2005, **129**:415–21.
32. Chen S, Wang W, Browman K, Katk H, Parmigiani G: **BayesMendel: an R environment for Mendelian risk prediction.** *Statistical Applications in Genetics and Molecular Biology* 2004, **3**(1):21.

Figures

Figure 1 - Kaplan-Meier estimates of penetrance for any HNPCC and CRC

Kaplan-Meier estimates of age-specific cumulative risks of any hereditary nonpolyposis colorectal cancer (a) and colorectal cancer (b) stratified by three mutation groups: mutation positive (+), unknown (?) and mutation negative (–).

Figure 2 - Kaplan-Meier estimates of penetrance for any HNPCC, CRC and EC

Kaplan-Meier estimates of age-specific cumulative risks of hereditary nonpolyposis colorectal cancer-related cancers [(a) any hereditary nonpolyposis colorectal cancer, (b) colorectal cancer, (c) endometrial cancer] stratified by sex and three mutation groups: mutation positive (+), unknown (?) and mutation negative (–).

Figure 3 - Segregation model estimates of penetrance for any HNPCC, CRC and EC

Age-specific cumulative risks and 95% confidence intervals of hereditary nonpolyposis colorectal cancer related cancers [(a) any hereditary nonpolyposis colorectal cancer, (b) colorectal cancer, (c) endometrial cancer] among mutation carriers and non-carriers specified by gender, based on the segregation analyses.

Tables

Table 1 - Description of NL Family Data Set

Table 1: Number of each phenotype per subgroup based on sex and mutation status

Descriptive Statistics

Group	Any HNPCC [†] cancer		CRC [‡]		EC [△]	
	Events	n	Events	n	Events	n
Males						
Carriers*	40	78	36	78	-	-
Non-carriers	2	51	1	51	-	-
Unknowns	9	36	5	46	-	-
Females						
Carriers*	40	76	17	76	16	75
Non-carriers	1	51	0	51	0	51
Unknowns	7	21	2	30	1	20
Family Members with Phenotype Information		313		332		146

* includes known carriers and presumed carriers, n = number of relatives

[†] HNPCC = hereditary nonpolyposis colorectal cancer

[‡] CRC = colorectal cancer, [△] EC = endometrial cancer

Table 2 - Penetrance and 95% Confidence Interval Estimates

Table 2: Penetrance estimates (percent) at ages 30, 50 and 70 by carrier status and sex, with corresponding 95% confidence intervals. Estimates based on the specified hazard function for age at diagnosis of any hereditary nonpolyposis colorectal cancer (any HNPCC cancer), colorectal cancer (CRC) and endometrial cancer (EC)

Group	Age 30 (95 % CI)	Age 50 (95 % CI)	Age 70 (95 % CI)
Any HNPCC cancer (Weibull)			
Male Carrier	8.7 (5.1, 13.2)	68.2 (53.3, 77.8)	97.6 (90.6, 99.4)
Female Carrier	6.5 (4.1, 10.7)	57.1 (45.7, 70.1)	93.6 (85.1, 98.4)
Male Non-carrier	1.1 (0.5, 3.1)	12.9 (7.4, 29.4)	36.3 (22.7, 68.3)
Female Non-carrier	0.4 (0.2, 1.1)	5.4 (2.6, 11.1)	16.5 (8.3, 32.1)
CRC (Log-logistic)			
Male Carrier	9.9 (3.3, 15.3)	58.9 (34.4, 67.1)	82.6 (63.2, 88.3)
Female Carrier	1.7 (0.8, 3.9)	18.4 (10.7, 31.8)	42.7 (28.3, 63.4)
Male Non-carrier	0.5 (0.1, 1.2)	6.7 (1.8, 11.9)	19.1 (6.1, 31.9)
Female Non-carrier	0.1 (0.0, 0.3)	1.1 (0.4, 3.0)	3.6 (1.5, 9.7)
EC (Log-logistic)			
Female Carriers	0.3 (0.0, 3.7)	19.2 (1.6, 59.7)	63.3 (12.6, 92.6)
Female Non-carriers	0.0 (0.0, 0.1)	0.3 (0.0, 2.5)	1.9 (0.3, 18.0)

Table 3 - Hazard Ratio and 95% Confidence Interval Estimates

Table 3: Hazard ratio estimates at ages 30, 50 and 70 in mutation carriers compared with mutation non-carriers by sex, with corresponding 95% confidence intervals. Estimates based on the specified hazard function for age at diagnosis of any hereditary nonpolyposis colorectal cancer and colorectal cancer

	Group	Age 30 (95% CI)	Age 50 (95% CI)	Age 70 (95% CI)
<hr/>				
Any HNPCC cancer (Weibull)				
	Male	8.2 (3.5, 16.9)	8.2 (3.5, 16.9)	8.2 (3.5, 16.9)
	Female	15.2 (6.8, 33.5)	15.2 (6.8, 33.5)	15.2 (6.8, 33.5)
<hr/>				
CRC (Log logistic)				
	Male	18.2 (6.8, 36.7)	8.8 (2.6, 12.7)	4.3 (1.6, 5.8)
	Female	19.7 (8.5, 45.6)	16.5 (7.4, 38.1)	11.9 (5.7, 28.6)

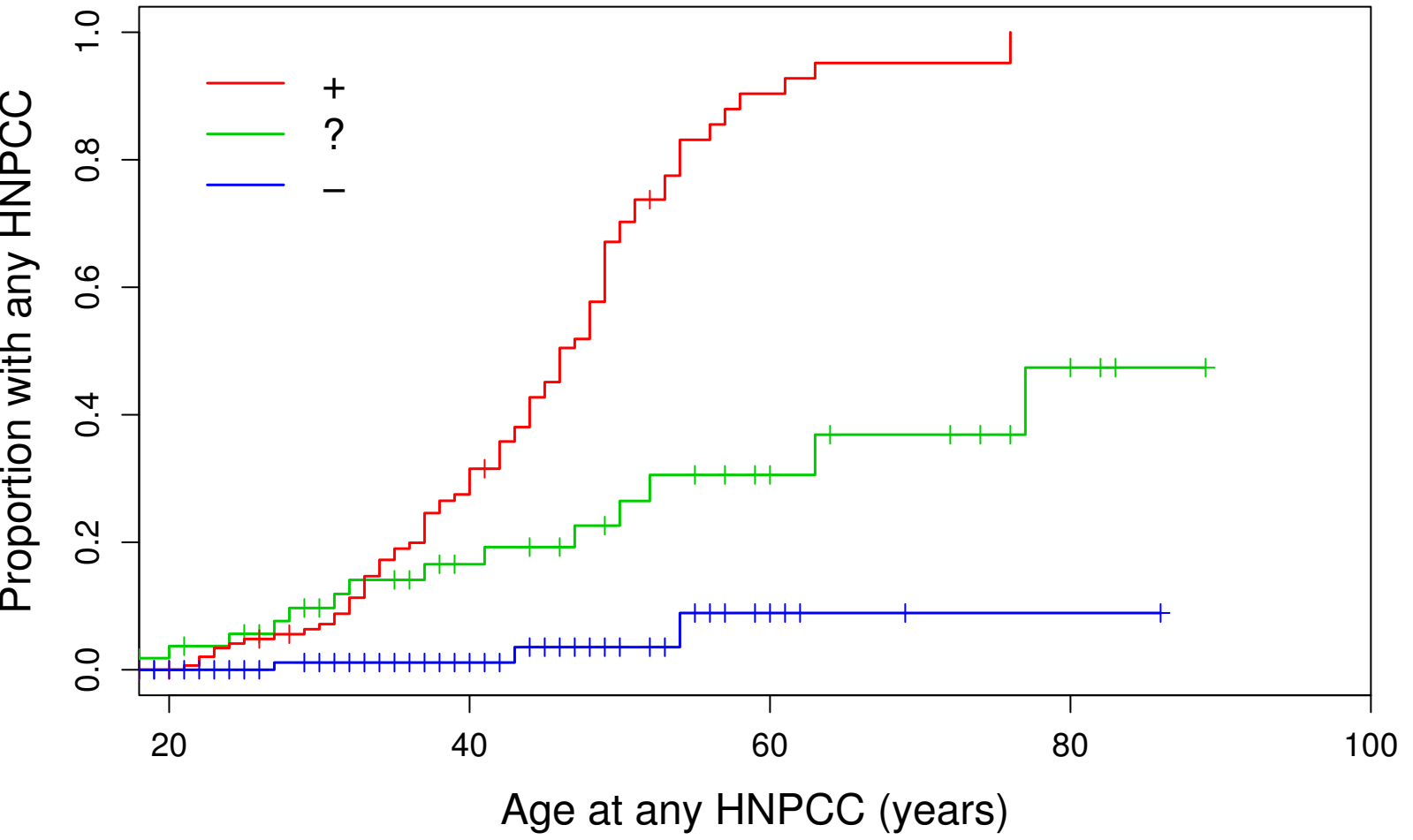
HNPCC = hereditary nonpolyposis colorectal cancer
CRC = colorectal cancer

Additional Files

Appendix — General hazard specification for parametric regression model

The formulation of the hazard function based on a generalized log-Burr specification of the survival function.

(a) Any HNPCC



(b) CRC

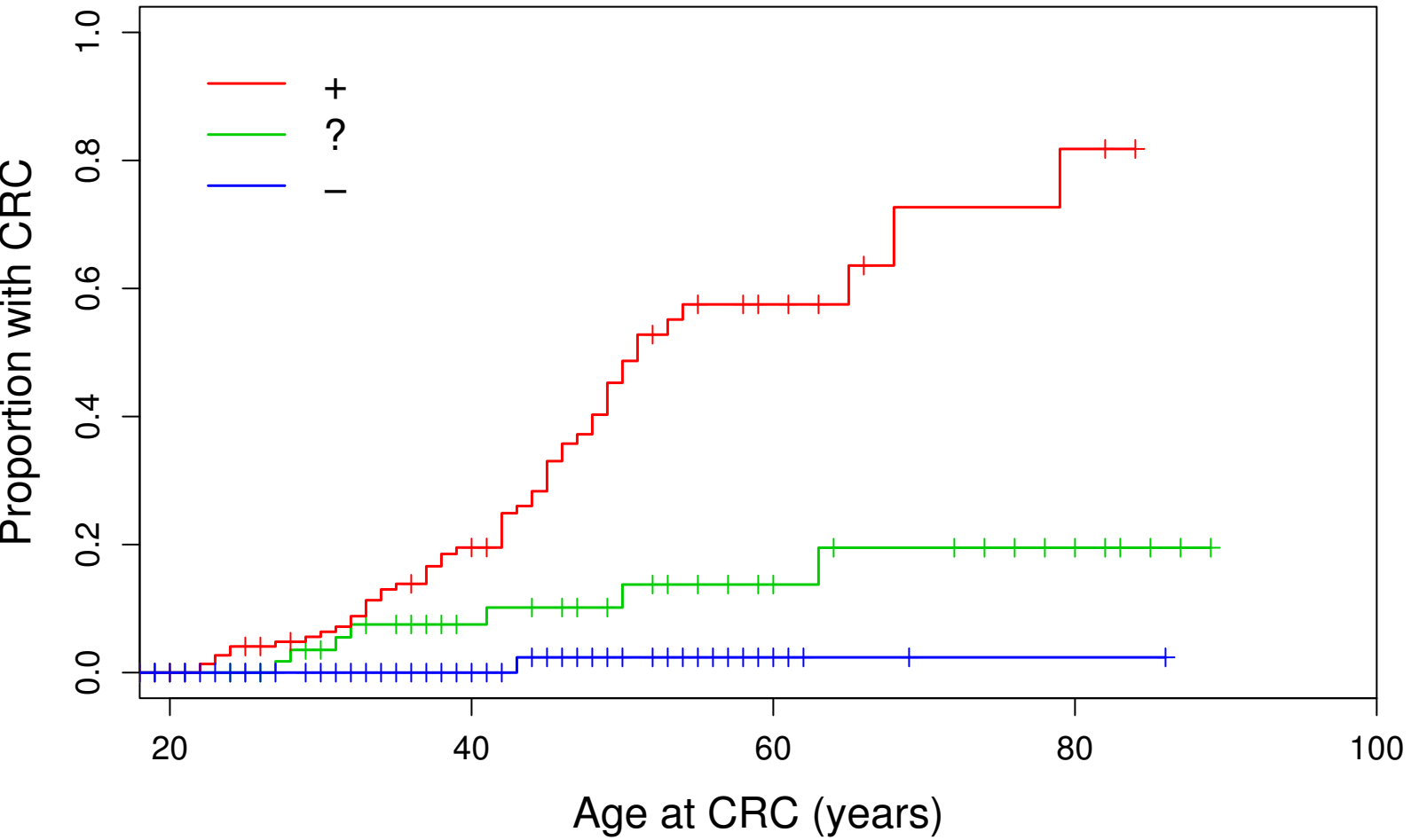


Figure 1

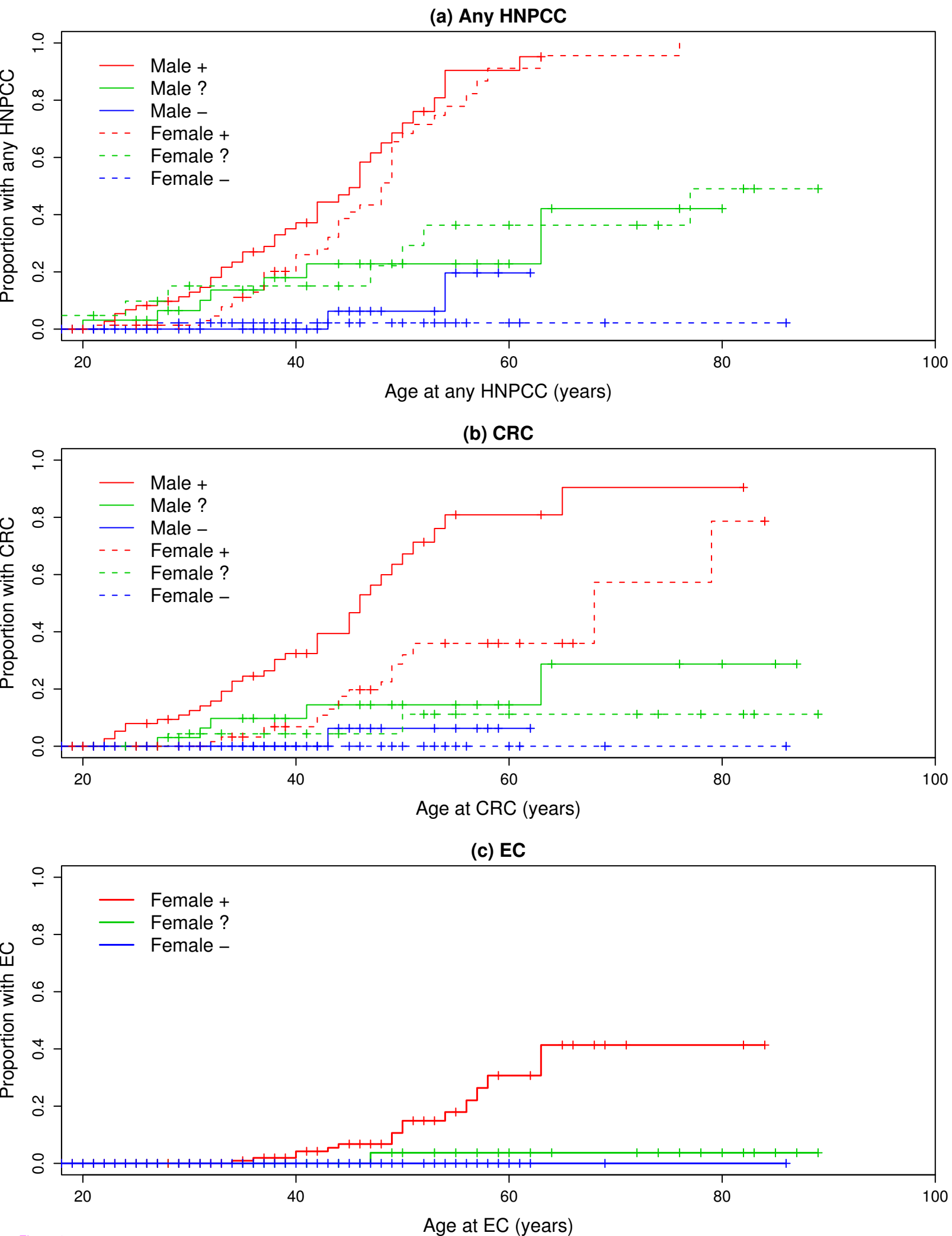
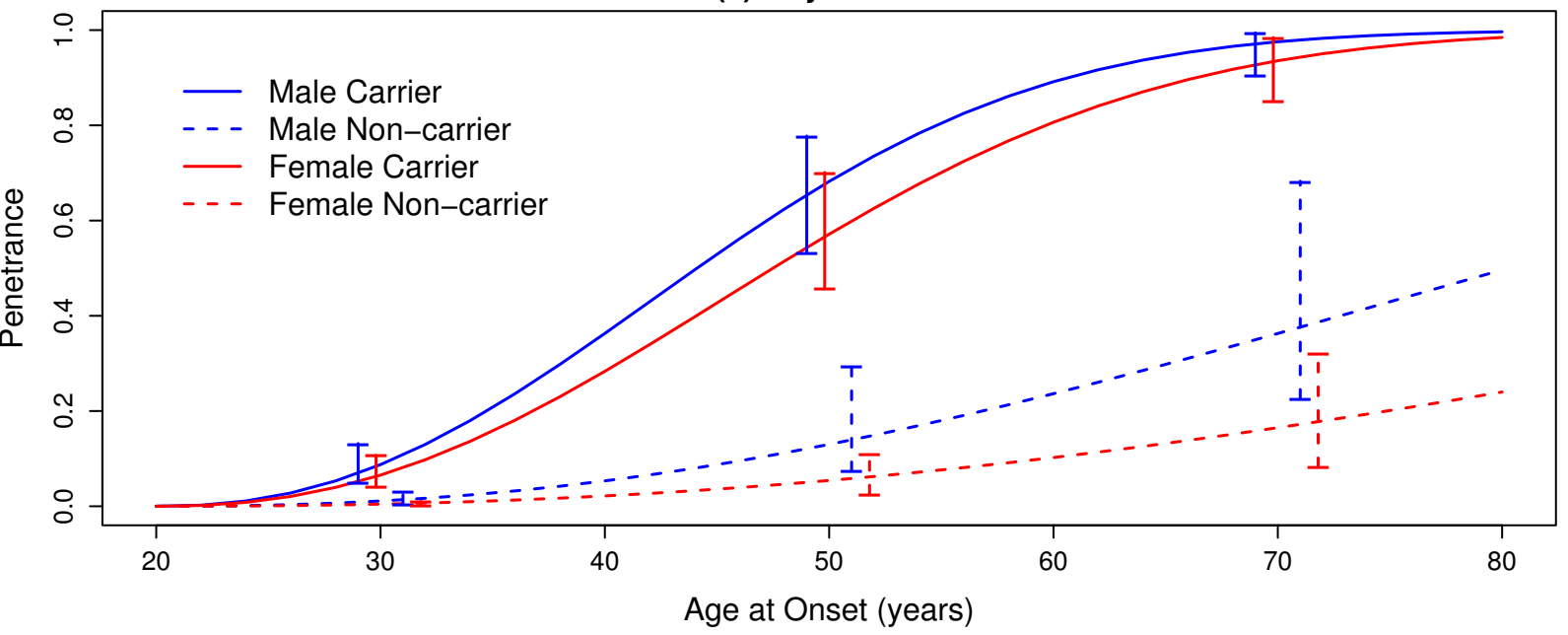
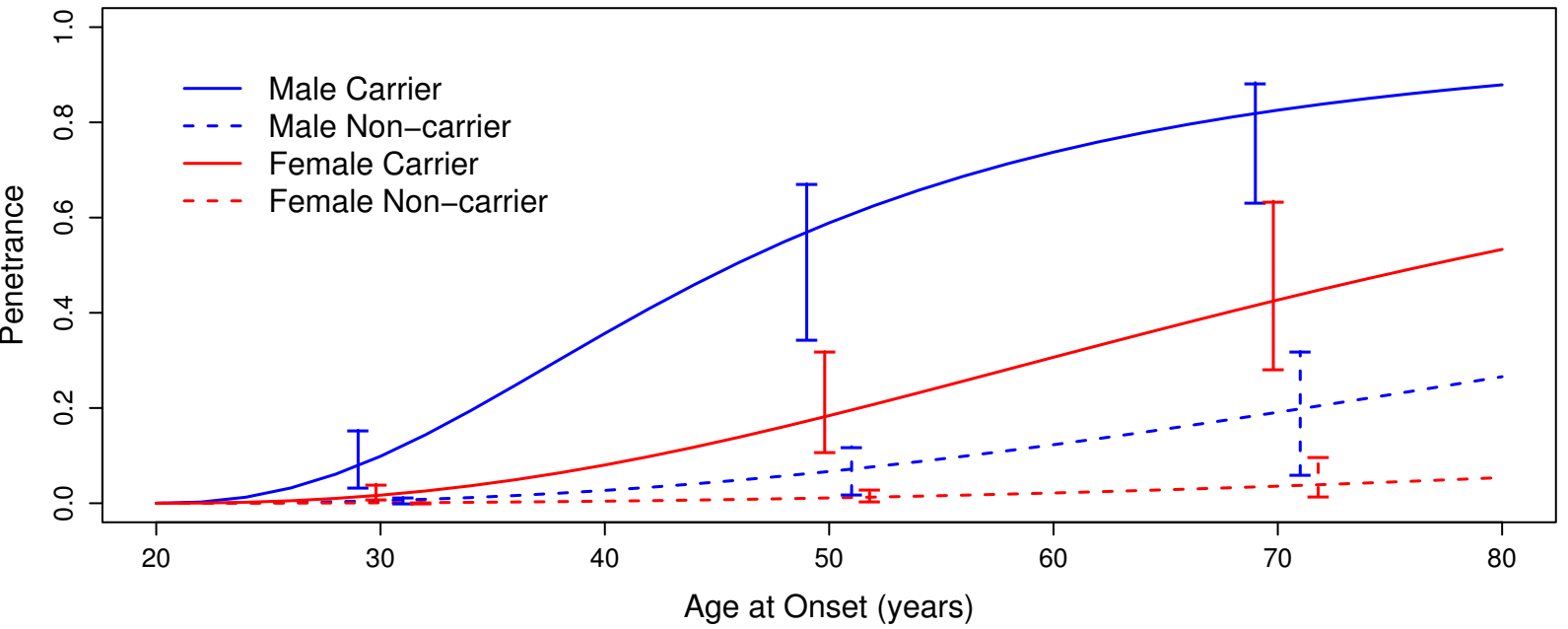


Figure 2

(a) Any HNPPC



(b) CRC



(c) EC

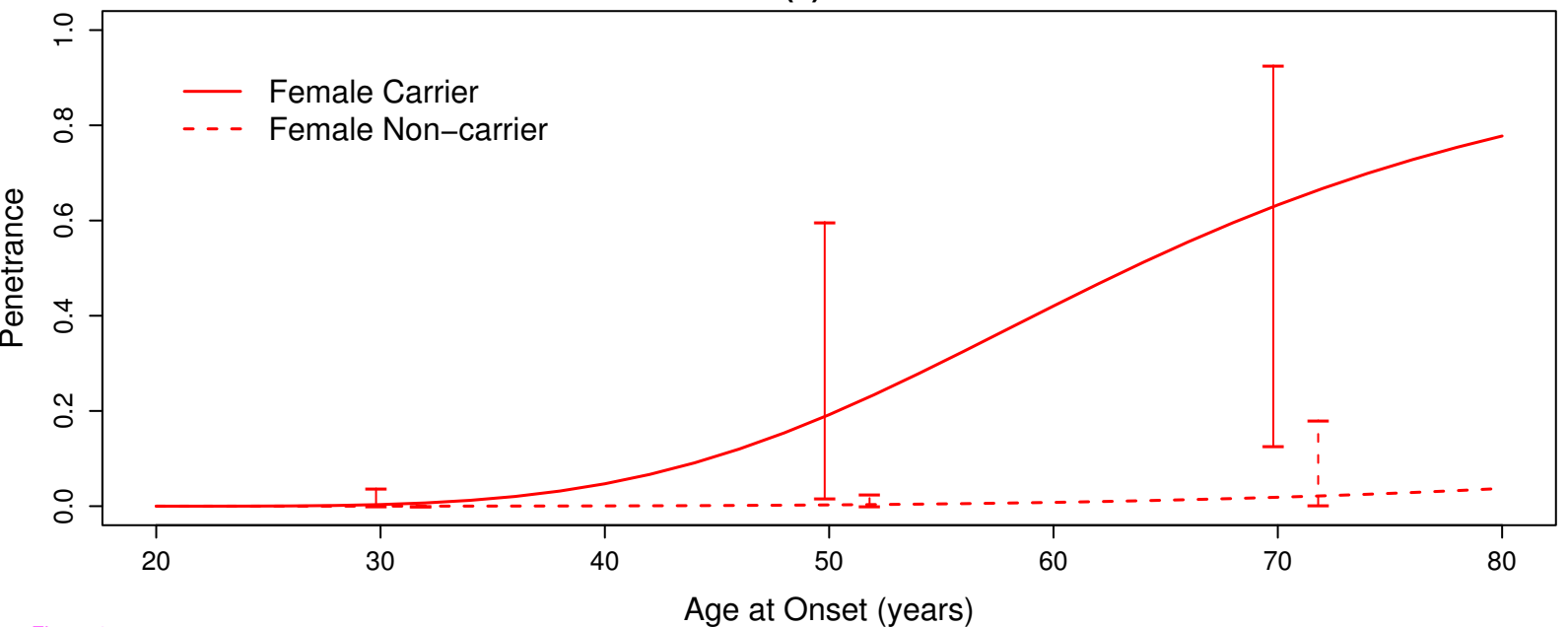


Figure 3

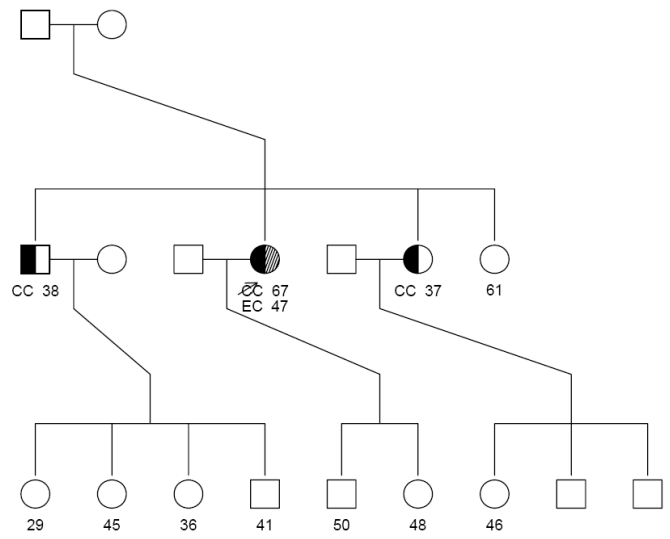


Figure 4

Additional files provided with this submission:

Additional file 1: bmc_kopciuk_2009_appendix.tex, 7K

<http://www.hccjournal.com/imedia/5875090992883343/supp1.tex>