Electronic appendix to "Risk of colorectal and endometrial cancer for carriers of mutations of the hMLH1 and hMSH2 gene: Correction for ascertainment"

Franz Quehenberger, Hans Vasen, Hans C. van Houwelingen

Sensitivity analysis

Table 5. Sensitivity analysis of the of the genotype restricted likelihood of a competing risks model for colorectal cancer (CRC), endometrial cancer (EC) and minor HNPCC cancer sites (MC). Results obtained from maximizing the genotype restricted likelihood (L_{GR}) when mutation frequency and non-carrier hazard rates, respectively, are multiplied by a factor or when all phenotype and genotype information earlier than 1990 is discarded.

Freq*	CRC†	ED†	MC†	L_{GR}	a_0‡	a_1‡	a_2‡	b_0§	c_0¶
1	1	1	1	-472.91	2.74	-2.06	-1.24	3.44	1.74
.5	1	1	1	-471.06	2.67	-2.02	-1.24	3.46	1.75
2	1	1	1	-474.68	2.85	-2.14	-1.25	3.41	1.69
1	1.5	1	1	-473.67	2.59	-1.91	-1.15	3.31	1.69
1	1	1.5	1	-473.13	2.74	-2.02	-1.23	3.18	1.71
1	1	1	1.5	-473.41	2.71	-2.02	-1.21	3.39	1.54
1	1	1	1	-389.56	2.84	-2.06	-1.41	3.57	1.85

$*$) Proportionality factor of the gene frequency of the disease mutation.
†) Proportionality factor of the non-carrier hazard with respect to diseases.
‡) Parameter estimates of the CRC log relative hazard function.
§) Parameter estimate of the EC log relative hazard.
¶) Parameter estimate of the MC log relative hazard.
$**$) Phenotypic and genotypic information earlier than 1990 discarded.

Penetrance estimates from hMLH1 families

Table 2a (hMLH1 families). See Table 2 for legend.

Cancer site	Polynomial	Coefficient	Estimate	±	Standard error
CRC	$a_0 + a_1 t + a_2 t^2$				
		a_0	2.6	±	0.71
		a_1	−1.7	±	0.57
		a_2	−1.2	±	0.57
EC	b_0 (constant)				
		b_0	4.4	±	0.46
MC	c_0 (constant)				
		c_0	1.9	±	0.77

Table 3a (hMLH1 families). See Table 3 for legend.

	Age	Relative risk		Absolute risk (%)	
CRC (males)	30	16.9	(9.8 – 29.3)	0.044	(0.026 – 0.076)
	40	24.1	(14.0 – 41.3)	0.22	(0.13 – 0.37)
	50	18.7	(9.5 – 36.8)	0.82	(0.41 – 1.6)
	60	7.9	(3.9 – 16.1)	1.1	(0.54 – 2.3)
	70	1.8	(0.83 – 4.1)	0.56	(0.26 – 1.2)
	80	0.23	(0.068 – 0.81)	0.10	(0.030 – 0.36)
CRC (females)	30	16.9	(9.8 – 29.3)	0.044	(0.026 – 0.076)
	40	24.1	(14.0 – 41.3)	0.22	(0.13 – 0.37)
	50	18.7	(9.5 – 36.8)	0.82	(0.41 – 1.6)
	60	7.9	(3.9 – 16.1)	1.1	(0.54 – 2.3)
	70	1.8	(0.83 – 4.1)	0.56	(0.26 – 1.2)
	80	0.23	(0.068 – 0.81)	0.10	(0.030 – 0.36)
EC	30	90.4	(58.5 – 140)	0.025	(0.016 – 0.039)
	40	90.4	(58.5 – 140)	0.27	(0.18 – 0.42)
	50	90.4	(58.5 – 140)	2.0	(1.3 – 3.2)
	60	90.4	(58.5 – 140)	5.2	(3.4 – 8.1)
	70	90.4	(58.5 – 140)	5.7	(3.7 – 8.8)
	80	90.4	(58.5 – 140)	6.9	(4.4 – 10.6)
MC (males)	30	7.04	(3.3 – 15.0)	0.013	(0.0062 – 0.028)
	40	7.04	(3.3 – 15.0)	0.062	(0.029 – 0.13)
	50	7.04	(3.3 – 15.0)	0.28	(0.13 – 0.60)
	60	7.04	(3.3 – 15.0)	0.82	(0.38 – 1.7)
	70	7.04	(3.3 – 15.0)	2.0	(0.93 – 4.2)
	80	7.04	(3.3 – 15.0)	3.0	(1.3 – 6.3)
MC (females)	30	7.04	(3.3 – 15.0)	0.041	(0.019 – 0.088)
	40	7.04	(3.3 – 15.0)	0.12	(0.058 – 0.26)
	50	7.04	(3.3 – 15.0)	0.33	(0.15 – 0.69)
	60	7.04	(3.3 – 15.0)	0.65	(0.31 – 1.4)
	70	7.04	(3.3 – 15.0)	1.1	(0.52 – 2.3)
	80	7.04	(3.3 – 15.0)	1.4	(0.66 – 3.0)

Table 4a (hMLH1 families). See Table 4 for legend.

Age	CRC (male)		CRC (female)		EC (female)		CRC+EC (female)	
30	0.28	(0.062 – 1.3)	0.36	(0.062 – 2.1)	0.11	(0.046 – 0.27)	0.47	(0.12 – 1.9)
40	1.4	(0.54 – 3.8)	1.6	(0.60 – 4.2)	1.4	(0.56 – 3.3)	2.9	(1.3 – 6.4)
50	6.3	(2.0 – 19.0)	6.3	(2.0 – 18.0)	10.0	(4.2 – 23.0)	16.0	(7.0 – 33.0)
60	15.0	(4.4 – 46.0)	14.0	(4.0 – 41.0)	40.0	(19.0 – 72.0)	48.0	(24.0 – 79.0)
70	22.0	(6.6 – 61.0)	18.0	(5.4 – 52.0)	66.0	(35.0 – 93.0)	72.0	(41.0 – 95.0)
80	24.0	(7.3 – 64.0)	20.0	(6.0 – 54.0)	82.0	(50.0 – 99.0)	86.0	(56.0 – 99.0)

	MC (male)		MC (female)		CRC+EC+MC (male)		CRC+EC+MC (female)	
30	0.076	(0.016 – 0.36)	0.43	(0.091 – 2.0)	0.36	(0.10 – 1.3)	0.90	(0.31 – 2.6)
40	0.44	(0.093 – 2.1)	1.2	(0.26 – 5.7)	1.9	(0.80 – 4.3)	4.1	(2.0 – 8.1)
50	2.0	(0.42 – 9.0)	3.4	(0.73 – 15.0)	8.1	(3.1 – 20.0)	19.0	(9.4 – 35.0)
60	7.2	(1.6 – 30.0)	8.1	(1.8 – 33.0)	22.0	(8.0 – 51.0)	52.0	(29.0 – 80.0)
70	19.0	(4.4 – 64.0)	16.0	(3.6 – 56.0)	37.0	(15.0 – 75.0)	76.0	(49.0 – 95.0)
80	38.0	(9.5 – 89.0)	26.0	(6.2 – 76.0)	53.0	(21.0 – 91.0)	89.0	(66.0 – 99.0)

Penetrance estimates from MSH2 families

Table 2b (hMSH2 families). See Table 2 for legend.

Cancer site	Polynomial	Coefficient	Estimate ± standard error
CRC	$a_0 + a_1 t + a_2 t^2$		
		a_0	2.9 ± 0.51
		a_1	-2.3 ± 0.91
		a_2	-1.2 ± 0.70
EC	b_0 (constant)		
		b_0	3.1 ± 0.70
MC	c_0 (constant)		
		c_0	1.7 ± 0.57

Table 3b (hMSH2 families). See Table 3 for legend.

Tumor	Age	Relative risk		Absolute risk (%)	
CRC (males)	30	47.2	(23.7 – 93.9)	0.12	(0.062 – 0.24)
	40	50.8	(29.8 – 86.3)	0.46	(0.27 – 0.78)
	50	29.6	(18.1 – 48.3)	1.3	(0.79 – 2.1)
	60	9.3	(5.5 – 16.1)	1.3	(0.76 – 2.3)
	70	1.6	(0.57 – 4.5)	0.49	(0.17 – 1.4)
	80	0.15	(0.021 – 1.05)	0.066	(0.0094 – 0.47)
CRC (females)	30	47.2	(23.7 – 93.9)	0.11	(0.057 – 0.23)
	40	50.8	(29.8 – 86.3)	0.53	(0.31 – 0.90)
	50	29.6	(18.1 – 48.3)	1.1	(0.68 – 1.8)
	60	9.3	(5.5 – 16.1)	0.94	(0.55 – 1.6)
	70	1.6	(0.57 – 4.5)	0.31	(0.11 – 0.87)
	80	0.15	(0.021 – 1.05)	0.047	(0.0067 – 0.33)
EC	30	21.4	(11.0 – 41.8)	0.0060	(0.0031 – 0.012)
	40	21.4	(11.0 – 41.8)	0.064	(0.033 – 0.13)
	50	21.4	(11.0 – 41.8)	0.49	(0.25 – 0.94)
	60	21.4	(11.0 – 41.8)	1.2	(0.64 – 2.4)
	70	21.4	(11.0 – 41.8)	1.3	(0.69 – 2.6)
	80	21.4	(11.0 – 41.8)	1.6	(0.83 – 3.2)
MC (males)	30	5.3	(3.1 – 9.0)	0.010	(0.0058 – 0.017)
	40	5.3	(3.1 – 9.0)	0.047	(0.027 – 0.080)
	50	5.3	(3.1 – 9.0)	0.21	(0.12 – 0.36)
	60	5.3	(3.1 – 9.0)	0.61	(0.36 – 1.1)
	70	5.3	(3.1 – 9.0)	1.4	(0.87 – 2.5)
	80	5.3	(3.1 – 9.0)	2.2	(1.3 – 3.8)
MC (females)	30	5.3	(3.1 – 9.0)	0.031	(0.018 – 0.053)
	40	5.3	(3.1 – 9.0)	0.093	(0.054 – 0.16)
	50	5.3	(3.1 – 9.0)	0.25	(0.14 – 0.42)
	60	5.3	(3.1 – 9.0)	0.49	(0.29 – 0.84)
	70	5.3	(3.1 – 9.0)	0.82	(0.48 – 1.4)
	80	5.3	(3.1 – 9.0)	1.1	(0.62 – 1.8)

Table 4b (hMSH2 families). See Table 4 for legend.

Age	CRC (male)		CRC (female)		EC (female)		CRC+EC (female)	
30	0.89	(0.14 – 5.6)	1.2	(0.14 – 9.7)	000.027	(000.067 – 0.11)	1.2	(0.15 – 9.5)
40	3.5	(1.1 – 11.0)	4.0	(1.1 – 14.0)	0.32	(0.82 – 1.3)	4.3	(1.3 – 13.)
50	12.0	(4.5 – 28.0)	12.0	(4.6 – 29.0)	2.5	(6.3 – 9.5)	14.0	(6.2 – 30.)
60	23.0	(10.0 – 48.0)	21.0	(9.2 – 44.0)	11.0	(30.0 – 38.0)	30.0	(15.0 – 53.)
70	30.0	(13.0 – 57.0)	25.0	(12.0 – 50.0)	22.0	(62.0 – 64.0)	42.0	(22.0 – 70.)
80	31.0	(14.0 – 60.0)	26.0	(12.0 – 52.0)	33.0	(98.0 – 80.0)	51.0	(26.0 – 82.)

	MC (male)		MC (female)		CRC+EC+MC (male)		CRC+EC+MC (female)	
30	000.057	(000.019 – 0.17)	0.32	(0.11 – 0.98)	0.95	(0.16 – 5.4)	1.5	(000.29 – 8.0)
40	0.33	(0.11 – 1.00)	0.92	(0.30 – 2.8)	3.8	(1.3 – 11.0)	5.1	(1.9 – 14.)
50	1.5	(0.49 – 4.4)	2.6	(0.86 – 7.6)	13.0	(5.5 – 29.0)	16.0	(8.0 – 31.)
60	5.5	(1.8 – 16.0)	6.2	(2.1 – 18.0)	27.0	(14.0 – 50.0)	34.0	(20.0 – 56.)
70	15.0	(5.2 – 39.0)	12.0	(4.2 – 33.0)	40.0	(22.0 – 65.0)	49.0	(29.0 – 73.)
80	30.0	(11.0 – 66.0)	20.0	(7.3 – 50.0)	52.0	(30.0 – 78.0)	61.0	(37.0 – 85.)

Details of statistical methods

Estimation of the penetrances was based on retrospective likelihood. For genetic data this means the maximization of the probability of the genotypes given the phenotypes. It can be formulated in terms of prospective likelihood by means of the Bayes rule:

$$\mathcal{L}_R = \mathcal{P}\left[\mathbf{G}|\mathbf{P}\right] = \frac{\mathcal{P}\left[\mathbf{G},\mathbf{P}\right]}{\mathcal{P}\left[\mathbf{P}\right]} = \frac{\mathcal{P}\left[\mathbf{P}|\mathbf{G}\right]\mathcal{P}\left[\mathbf{G}\right]}{\sum_{G^*\in\mathcal{G}}\mathcal{P}\left[\mathbf{P}|\mathbf{G}^*\right]\mathcal{P}\left[\mathbf{G}^*\right]} \tag{1}$$

Here $\mathbf{P}$ and $\mathbf{G}$ indicate the phenotypic and genotypic information, respectively, of the members of a pedigree. $\mathcal{G}$ contains all possible genotypes in the pedigree. However, it turns out that is is convenient to restrict the genetic model to two alleles of one gene. If for these genes the disease (mutant) alleles are very rare one can assume that if the disease allele is present in the family for one gene, all the other genes carry the wild type. This condition requires restriction of the likelihood (1) to genotypes in which at least one diseased person carries the mutation segregating in the family, called genotype restricted likelihood [1]:

$$\mathcal{L}_{GR} = \mathcal{P}\left[\mathbf{G}|\mathbf{P},\mathcal{G}_0\right] = \frac{\mathcal{P}\left[\mathbf{P}|\mathbf{G}\right]\mathcal{P}\left[\mathbf{G}\right]}{\sum_{G^*\in\mathcal{G}_0}\mathcal{P}\left[\mathbf{P}|\mathbf{G}^*\right]\mathcal{P}\left[\mathbf{G}^*\right]} \tag{2}$$

$\mathcal{G}_0$ contains only genotypes for which at least one diseased person carries the mutant allele of the gene of interest. MENDEL calculates probabilities $\mathcal{P}\left[\mathbf{P},\mathbf{G}\right]$ efficiently. Summation over possible genotypes is obtained by providing missing values for genotype data.

Because of an independency assumption $\mathcal{P}\left[\mathbf{P}|\mathbf{G}\right]$ is the product of individual penetrances $\mathcal{P}\left[P_i|G_i\right]$. These individual penetrances have to include all phenotypic data that are relevant for ascertainment of the family. We assume that these data are the observation time T and the time of occurrence of any HNPCC specific cancer. This would require a multivariate failure time model which would include all dependencies between the events. By modelling the data only until the first event occurs, i.e. a competing risks model, these dependencies need not to be modelled. This is admissible in genotype restricted likelihood if further events do not change the probability that the family is ascertained.

Let T_i denote the possibly unobserved time at which event i of k events occurred. The cause-specific hazard function of an individual of genotype g for event i is

$$\lambda_{ig}(t) = \lim_{h\to 0+}\frac{1}{h}\mathcal{P}\left[T_i\in[t,t+h)|T_j\geq t \text{ for all } j=1,\ldots,k, G=g\right]. \tag{3}$$

The probability to experience no event is

$$S_g(t) = e^{-\int_0^t \lambda_g(s)\,\mathrm{d}s} \tag{4}$$

with $\lambda_g(t) = \sum_{i=1}^k \lambda_{ig}(t)$.

Let (T,d) denote the phenotypic data from an individual. If no event has occurred in that individual, T is the age at the end of the observation time and $d=0$. If an event has occurred let T denote the age at occurrence d the type of the first event. Using the indicator expression $(d=i)$, which is one if d is equal to i and zero otherwise, the likelihood of the observation (T,d) is given by

$$S_g(T)\prod_{i=1}^{k}\lambda_{ig}(T)^{(d=i)} \tag{5}$$

The reasoning behind (5) is that if not event has occurred, the probability if given by (4). Otherwise, if the first event occurred at T, there has been not occurred an event up to T with probability given by (4) and an event of type d occurred instantaneously afterwards with probability (3). The product of (5) over all persons of a family gives expression $\mathcal{P}[\mathbf{P}|\mathbf{G}]$ in (2).

From estimates of λ_{ig} the incidence of event i as first event is

$$I_{ig}(t) = \int_0^t S_g(s)\lambda_{ig}(s)\,\mathrm{d}s. \tag{6}$$

This incidence depends on the set of competing events. In order to maintain comparability to the literature, where competing events are not considered, we ignore competing events and calculate the incidence

$$F_{ig}(t) = 1 - e^{\int_0^t \lambda_{ig}(s)\,\mathrm{d}s} \tag{7}$$

F_{ig} is analogous to single-risk analysis because in the estimation of λ_{ig} only data up to the first of competing events are used.

Whereas in classical competing risks analysis there are no events after the first event, here a competing event effectively acts as a censoring event. In the context of HNPCC it has to be assumed that this censoring event carries information about the risk of the index event i. If, for example , after the diagnosis of one HNPCC-related cancer, a person would be at higher risk for another HNPCC-related cancer, the incidence in (7) would underestimate the incidence, although it would be higher than in (6). However, death from a competing cancer risk would no longer constitute informative censoring, as diagnosis of that cancer would have cause censoring of the subject before. On the other hand, in penetrance studies that take into account only one cancer risk, death from a competing risk constitutes an informative censoring event, thereby possibly causing the underestimation of the incidence of the cancer under study.

HNPCC families do not contain much information about hazard rates of non-carriers of disease mutations, especially at young ages. Therefore the age-dependent population incidences from a cancer registry were used in non-carriers, assuming that annual incidences approximate the limit 3. Age-related incidences of the population that were derived from non-competing risk models were justified because multiple cancers were very rarely diagnosed at early ages and still rarely at older ages. Assuming that the disease mutations are rare, population hazard rates are nearly equal to carrier hazard rates. Carrier hazard rates were modelled to be proportional to non-carrier hazard rates by a relative hazard function of polynomial shape.

The calculations were performed on an annual grid. As age-related incidences from the cancer registry were given in five year's intervals, non-mutation-carrier hazards were smoothed by calculating a moving average of hazard rates

$$\tilde{\lambda}(t) = \frac{1}{K}\sum_{h=-5}^{5} \lambda(t-h)\kappa(h), \tag{8}$$

with a kernel of triangular shape $\kappa(h) = 6 - |h|$ and $K = \sum_{h=-5}^{5} \kappa(h)$.

References

1. 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 Published Online: 26 Apr 2004.