**Statistical Methods**

We used modified segregation analysis to fit a range of genetic models to the observed colorectal cancer family histories for the proband and their first-degree relatives. Information on second-degree relatives and other relatives of the proband were not used for this analysis as data were incomplete. Individuals were assumed to be at risk of colorectal cancer from birth until the earliest of the following: diagnosis of colorectal cancer or any other cancer (except skin cancer), first polypectomy, death, the earlier of last known age at baseline interview or age 80 years. Individuals known to have died but with unknown age at death were censored at age 70 years. Relatives reported to have had colorectal cancer with unknown age at diagnosis were assigned an age at diagnosis as their age of death (if deceased) or their age at the time of the proband’s baseline interview. Relatives with no age information were treated as lost to follow-up with age censored at zero years.

Models of Susceptibility

The model we used was an extension of the mixed model of Morton and MacLean (1974) (1). incorporating both major gene and polygenic components. In our model, the colorectal cancer incidence *λi*(*t*,*k*) for individual *i* at age *t* in sex group *k* (*k* = 1 for males*, k* = 2 for females) was assumed to depend on genotype according to a parametric survival analysis model *λi*(*t*,*k*)=*λ*0(*t*,*k*) exp(*Gi*+*Pi(t)*), where *λ*0(*t*,*k*) is the sex-specific baseline incidence at age *t* . *Gi* is the natural logarithm of the relative risk associated with the major genotype and *Pi*(*t*) is the polygenic component for age *t*.

The major genotype was defined by six components representing each of the genes: *MLH1, MSH2, MSH6, PMS2, MUTYH* and one representing the hypothetical unidentified major genes, here given the name *UNIDENTIFIED\_MAJOR\_GENES*. As the probability of having a mutation in more than one of *MLH1, MSH2, MSH6, PMS2, MUTYH* was very small, to reduce the computation time, we coded these genes into one locus with seven alleles: *MLH1* positive, *MSH2* positive, *MSH6* positive, *PMS2* positive, *MUTYH* positive, *UNIDENTIFIED\_MAJOR\_GENES* positive and a normal allele. For simplicity, the hierarchical order in which mutations were assumed to be dominant over other alleles was as follows: *MLH1*, *MSH2*, *MSH6*, *PMS2*, *MUTYH*, *UNIDENTIFIED\_MAJOR\_GENES* and normal alleles. These assumptions are not critical to the model because carriers in multiple genes are rare. Thus, *Gi* takes on the following potential risk categories: *MLH1* carriers, *MSH2* carriers, *MSH6* carriers, *PMS2* carriers, *MUTYH* biallelic carriers, *MUTYH* monoallelic carriers, *UNIDENTIFIED\_MAJOR\_GENES* heterozygotes, *UNIDENTIFIED\_MAJOR\_GENES* homozygotes and non-carriers. We fitted models in which the unidentified major genes component was autosomal with a normal and a mutant allele unlinked to mutations in the MMR genes or *MUTYH*. These assumptions are not critical to the model because carriers in multiple genes are rare. This is consistent with observation, since in this data set, there were no probands observed with mutations in more than one MMR gene. We have also previously observed that having *MUTYH* monoallelic mutation does not significantly alter the risk of colorectal cancer in addition to having a MMR mutation (2, 3). We also fitted models in which the average relative risk for the unidentified major genes componentwas assumed to be age dependent. As the number of MMR and *MUTYH* mutation carriers in our dataset was too small to obtain reliable colorectal cancer risk estimates for mutation carriers, we used the published age-, sex- and country-specific incidence rates for *MLH1* and *MSH2* mutation carriers (4), and published age- and sex-specific incidence rates for *MSH6, PMS2* and *MUTYH* mutation carriers (2, 5, 6).

The polygenic component for age *t*, *Pi*(*t*), was assumed to be normally distributed with zero mean and variance σ2*p*(*t*) . Because the standard polygenic model is not amenable to “peeling” (7-9), we used as an approximation the hypergeometric polygenic model (HPM) (10, 11). *P* was approximated by *P*=(*R*-*N*)σ*p*/(*N*/2)1/2, where *R* has a binomial distribution Bin(2*N*, 1/2). *N*, the number of loci used in the HPM, was set to 3 (see Appendices in Antoniou et al. (2001) for further details of the model(12)). We also fitted models where the variance of the polygenic ‘modifying’ component was allowed to take a different value σ2*m* for MMR and for *MUTYH* carriers.

To compute the baseline colorectal cancer incidence *λ*0(*t*), we constrained the overall incidence of colorectal cancer to agree with the national age and sex specific incidences (1998-2002) separately for Australia, Canada and USA (13). Other cancers were ignored in this model. Published incidences are reported in 5-year intervals, which can result in large variation in incidences between adjacent age intervals, so we smoothed the population incidences using locally weighted regression (LOWESS)(14) with a bandwidth of 0.2.

We assumed that the sensitivity of the mutation testing of probands for MMR genes and *MUTYH* was 80% (15), and we examined the effect of varying this sensitivity. For relatives, we assumed the mutation screening for the proband’s mutation (i.e. predictive testing) was 100% sensitive and specific.

Estimation

The genetic models were specified in terms of colorectal cancer incidence for MMR gene and *MUTYH* mutation carriers, the frequency (*q*A) of the putative high risk allele ‘‘A’’ of the unidentified major genes, the average relative risk of colorectal cancer for carriers of mutations in theunidentified major genes, the number of polygenic loci assumed in the HPM (*N*), and the variances of the polygenic and modifying components (σ2*p* and σ2*m*). Maximum likelihood estimation was used to estimate the parameters. The estimates we present are the values that were the most likely (i.e. most consistent) with the data. Maximum likelihood is the optimal method for making such estimates, and provides confidence intervals (CIs). We adjusted for ascertainment by maximizing the likelihood of each pedigree conditioned on the colorectal cancer status of the proband and his or her age of diagnosis (but not mutation carrier status as this information was not known at the time of recruitment). The variances of the parameter estimates were obtained by inverting the observed information matrix. To allow for the restricted ranges of the parameter values and to provide estimates likely to be more nearly normally distributed, we used transformed values of the parameters as a basis for calculating CIs: the high-risk allele frequencies *q*A were transformed using the logit function log[*p*/(1–*p*)], while for relative risk and σ2*p* we used the log transformation.

Goodness of Fit

The relative goodness of fit for nested models was tested by the likelihood ratio test. The Akaike’s Information Criterion(16) [AIC=-2 log-likelihood + 2x(no. of parameters)] was used to assess goodness of fit between non-nested models (1).

Estimation of Familial Relative Risk

The familial relative risk of the polygenic component was estimated using exp(σ2 /2) (17).

Expected versus Observed Number of Affected Relatives

The goodness of fit of the models was assessed by comparing the observed number of affected relatives with that expected under each model. For each model, we computed the predicted number of probands with one parent with colorectal cancer, one sibling with colorectal cancer, two siblings with colorectal cancer and one parent and one sibling with colorectal cancer. The probability of observing each of these events for a family was computed from the ratio of likelihoods for two pedigrees: the likelihood for a family in which the predicted event occurred and the likelihood of the affected proband (12). The expected versus observed number of affected relatives under each fitted model was assessed using the Pearson χ2 goodness of fit statistic Σ*j*[(O*j*-E*j*)2/E*j*] where O*j* is the observed number of probands or families with a particular characteristic and E*j* is the corresponding predicted numbers under each model.

Expected versus Observed number of Mutation Carriers

The expected number of probands with MMR and *MUTYH* mutation carriers for families that had undergone mutation testing based on their cancer family history was computed using Bayes theorem (18). For example, *P*(*MLH1* carrier | family history) = *P*(*MLH1* carrier, family history) / Σ*P*(mutation carrier status, family history) = *L1*/(*L0*+*L1*+*L2*+*L3*+*L4*+*L5*), where *Li* is the likelihood of observing the family with the proband carrying mutation *i* (=0, 1, 2, 3, 4, 5 for mutation negative, *MLH1*, *MSH2*, *MSH6*, *PMS2* and *MUTYH*, respectively). These probabilities were then added over all families to compute the total number of probands expected to carry a mutation for each gene (12).

Statistical Computation

Segregation analyses were performed using an optimised version(19) of the pedigree analysis software MENDEL version 3.2 (20). Incidence smoothing and other descriptive statistics were calculated using STATA 13.0 (21).

**References**

1. Elston R. Models for discrimination between alternative modes of inheritance. In: Gianola D, Hammond F, editors. Advances in statistical methods for genetic improvement of livestock. Berlin: Springer; 1990. p. 41-55.

2. Win AK, Dowty JG, Cleary SP, Kim H, Buchanan DD, Young JP, et al. Risk of colorectal cancer for carriers of mutations in MUTYH, with and without a family history of cancer. Gastroenterology. 2014;146:1208-11.e1-5.

3. Win AK, Cleary SP, Dowty JG, Baron JA, Young JP, Buchanan DD, et al. Cancer risks for monoallelic MUTYH mutation carriers with a family history of colorectal cancer. Int J Cancer. 2011;129:2256-62.

4. Dowty JG, Win AK, Buchanan DD, Lindor NM, Macrae FA, Clendenning M, et al. Cancer risks for MLH1 and MSH2 mutation carriers. Hum Mutat. 2013;34:490-7.

5. Baglietto L, Lindor NM, Dowty JG, White DM, Wagner A, Gomez Garcia EB, et al. Risks of Lynch Syndrome Cancers for MSH6 Mutation Carriers. J Natl Cancer Inst. 2010;102:193-201.

6. Senter L, Clendenning M, Sotamaa K, Hampel H, Green J, Potter JD, et al. The Clinical Phenotype of Lynch Syndrome Due to Germ-Line PMS2 Mutations. Gastroenterology. 2008;135:419-28.

7. Cannings C, Thompson E, Skolnick M. Probability functions on complex pedigrees. Adv Appl Prob. 1978;10:26-61.

8. Hasstedt S. A mixed-model likelihood approximation on large pedigrees. Comput Biomed Res. 1982;15:295-307.

9. Lalouel J. Probability calculations in pedigrees under complex models of inheritance. Hum Hered. 1980;30:320-3.

10. Lange K. An approximate model of polygenic inheritance. Genetics. 1997;147:1423-30.

11. Fernando RL, Stricker C, Elston RC. The finite polygenic mixed model: An alternative formulation for the mixed model of inheritance. Genetics. 1994;88:573-80.

12. Antoniou AC, Pharoah PDP, McMullan G, Day NE, Ponder BAJ, Easton D. Evidence for further breast cancer susceptibility genes in addition to *BRCA1* and *BRCA2* in a population-based study. Genet Epidemiol. 2001;21:1-18.

13. Curado MP, Edwards B, Shin HR, Storm H, Ferlay J, Heanue M, et al., editors. Cancer Incidence in Five Continents, Vol. IX. Lyon, France: International Agency for Research on Cancer; 2007.

14. Royston P. The use of cusums and other techniques in modelling continuous covariates in logistic regression. Stat Med. 1992;11:1115-29.

15. Palomaki GE, McClain MR, Melillo S, Hampel HL, Thibodeau SN. EGAPP supplementary evidence review: DNA testing strategies aimed at reducing morbidity and mortality from Lynch syndrome. Genet Med. 2009;11:42-65.

16. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Control. 1974;19:716-23.

17. Pharoah PDP, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BAJ. Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet. 2002;31:33-6.

18. Antoniou AC, Pharoah PPD, Smith P, Easton DF. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. Br J Cancer. 2004;91:1580-90.

19. Lee AJ, Cunningham AP, Kuchenbaecker KB, Mavaddat N, Easton DF, Antoniou AC. BOADICEA breast cancer risk prediction model: updates to cancer incidences, tumour pathology and web interface. Br J Cancer. 2014;110:535-45.

20. Lange K, Weeks D, Boehnke M. Programs for pedigree analysis: MENDEL, FISHER, and dGENE. Genet Epidemiol. 1988;5:471-2.

21. StataCorp. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP; 2013.