Estimating risks of common complex diseases across genetic and environmental factors: the example of Crohn disease

C M Lewis, S C L Whitwell, A Forbes, J Sanderson, C G Mathew, T M Marteau

Background: Progress has been made in identifying mutations that confer susceptibility to complex diseases, with the prospect that these genetic risks might be used in determining individual disease risk.

Aim: To use Crohn disease (CD) as a model of a common complex disorder, and to develop methods to estimate disease risks using both genetic and environmental risk factors.

Methods: The calculations used three independent risk factors: CARD15 genotype (conferring a gene dosage effect on risk), smoking (twofold increased risk for smokers), and residual familial risk (estimating the effect of unidentified genes, after accounting for the contribution of CARD15). Risks were estimated for high-risk people who are siblings, parents and offspring of a patient with CD.

Results: The CD risk to the sibling of a patient with CD who carries two CARD15 mutations is approximately 35%, which represents a substantial increase on the population risk of 0.1%. In contrast, the risk to a non-smoking sibling of a patient with CD who carries no CARD15 mutations is 2%. Risks to parents and offspring were lower.

Conclusions: High absolute risks of CD disease can be obtained by incorporating information on smoking, family history and CARD15 mutations. Behaviour modification through smoking cessation may reduce CD risk in these people.

Crohn disease (CD) is a severe inflammatory disorder of the gastrointestinal tract. The disorder can affect any part of the gut, but primarily the ileum and colon are affected. Symptoms vary between patients and include abdominal pain, diarrhoea and weight loss. The inflammation is controlled by drugs (steroids, immune modulators) and surgery.

Symptoms vary between patients and include abdominal pain, diarrhoea and weight loss. The inflammation is controlled by drugs (steroids, immune modulators) and surgery.

The prevalence of CD varies across geographic region and time, with incidence and prevalence rates increasing recently in the developed world. Prevalence rates are between 54 and 214 per 100 000 population, and the onset of CD generally occurs between 15 and 30 years of age.

The aetiology of CD is complex. Monozygotic twins show approximately 50–60% disease concordance, with much lower rates in dizygotic twins (~10%), highlighting the role of both environmental and genetic components in the development of CD

The strongest and best replicated environmental risk factor is smoking, which increases both the risk of developing CD and the severity of CD after diagnosis. Other environmental factors that may play a role in CD are appendectomy, oral contraception, diet and domestic hygiene, but the evidence for each of these is much weaker than for smoking, and many contradictory studies exist (reviewed by Lofthus).

Familial studies confirm that first degree relatives have a substantially increased risk of CD compared with the general population. The lack of a consistent pattern of inheritance in families and the failure of linkage studies to identify strong linkage to any single region indicate that genetic susceptibility to CD is multifactorial, with many genes contributing a small effect towards underlying disease risk. Major progress on the identification of genes contributing to CD has been made recently. The first susceptibility gene to be identified was CARD15 (NOD2), three mutations increase risk of CD in a gene dosage model, with CD risk increasing with the number of mutations (0, 1, 2) carried.

However, these mutations are also present in 12% of the general Western population, and therefore do not form a very specific test for CD risk. Two further genes have been identified through genome-wide association studies: IL23R and ATG16L1, and both findings have been independently replicated.

In this paper, we present a model for estimating an individual’s risk of developing CD, incorporating genetic and environmental risk factors. The model combines CARD15 genotypes with smoking and family history to calculate CD risk. Although no single factor forms a test of high predictive value, risk factors in combination with each other may raise risks of CD substantially. This provides the potential to identify population subgroups that are at high risk of developing CD, with a view to intervening to prevent disease.
2. Define a gene, $G$, with three genotypes ($AA$, $AB$, $BB$), where $A$ is the wildtype allele having frequency $p$, and $B$ is a mutation increasing risk of disease and having frequency $q$ (= 1 – $p$). Assume an arbitrary model for disease risk, with the genotype relative risks for disease risk by genotype being $g_0$ (= 1), $g_1$, and $g_2$, where

$$P(D|AB) = g_1 P(D|AA),$$

and

$$P(D|BB) = g_2 P(D|AA).$$

Let the disease penetrances be $f_0$, $f_1$, and $f_2$, where $f_0 = P(D|AA)$, and similarly for $f_1$ and $f_2$, so that $g_2 = f_2/f_0$ and $g_1 = f_1/f_0$. Penetrances can be calculated directly from genotype counts in a case control study using Bayes rule,

$$P(D|AA) = P(AA|D)P(D)/P(AA),$$

where $P(D)$ is disease prevalence, and $P(AA)$ is the population frequency of genotype $AA$. $P(AA)$ is estimated from control genotype frequency if controls are a random population sample of people, or if the disease is rare. If controls are selected to be unaffected and the disease is common, then $P(AA)$ is determined from the weighted case and control frequencies.

Alternatively, the summary measures of risk presented in meta-analysis of genetic association studies may be used to estimate penetrances, enabling risk calculations to be based on more robust estimates than a single case-control study. Risks are usually presented as genotype relative risks ($g_1$, $g_2$) for genotypes $AB$ and $BB$, relative to the wildtype $AA$ genotype.

Expressing the disease prevalence as $K$ solving for $g_1 = f_1$

used, with

$$P(D|AB) = g_1 P(D|AA),$$

and

$$P(D|BB) = g_2 P(D|AA).$$

For the sibling of a case, the increase in risk unaccounted for by gene $G$ is $\lambda_{SG'G}$. Piecing together both genetic and environmental contributions to risk, assuming they act independently, gives the probability of disease in the sibling of a case as:

$$P(D|G, E, sib affected) = P(D|G)P(D|E) \lambda_{SG'}/P(D) = GRR(G)/OR(E)\lambda_{SG'} K$$

where GRR($G$) and OR($E$) are the genotype relative risk and odds ratios of disease conferred by exposure or non-exposure to the environmental risk factor.

### Application to CD

#### Smoking

Smoking is a well-established risk factor for CD, which increases the risk and severity of CD. A series of epidemiological papers estimated the CD risk associated with smoking, matching patients with CD to controls by relevant factors such as age, sex and birth year, then calculating the odds ratio of CD in smokers compared with never-smokers. Many of these papers considered small series of cases and controls, and some used imprecise temporal data of current smoking status (especially in controls), rather than smoking at age of diagnosis of CD in the matched case. A meta-analysis of seven studies estimated a pooled odds ratio (OR) of 2.0 (95% CI 1.65 to 2.47) for CD in smokers compared with never-smokers. No consistent effect on CD risks was seen when data were analysed by sex, by smoking duration, or numbers of cigarettes smoked per day; however, cigarette smoking may specifically increase risk in late-onset CD patients (age of diagnosis >40 years).

Several studies also calculated CD risks in former smokers. The meta-analysis showed an increased risk of CD compared with lifetime non-smokers, with an odds ratio of 1.80 (95% CI 1.33 to 2.51), only slightly lower than the risk for current smokers. Detailed assessment of the time since smoking cessation in the ex-smokers suggests that CD risks remain raised immediately after smoking cessation, and reduce after 2–4 years. A more recent case control study of CD and both genetic and epidemiological risk factors estimated an OR for CD of 3.3 in smokers, and 1.7 in ex-smokers (95% CI 1.1 to 2.7). However, no study has systematically examined the risks of CD as age, sex and birth year, then calculating the odds ratio of CD in the matched case. A meta-analysis of seven studies estimated a pooled odds ratio (OR) of 2.0 (95% CI 1.65 to 2.47) for CD in smokers compared with never-smokers. No consistent effect on CD risks was seen when data were analysed by sex, by smoking duration, or numbers of cigarettes smoked per day; however, cigarette smoking may specifically increase risk in late-onset CD patients (age of diagnosis >40 years).

Similar expressions using genotype relative risks ($1$, $g_1$, $g_2$) in place of penetrances ($f_0$, $f_1$, $f_2$) can be derived, and then no assumption of disease prevalence is needed in calculating $\lambda_{SG'}$. The sibling relative risk, $\lambda_\sigma$, can also be extended to multiple genes, provided no epistasis exists between the genes modelled and between the genes and the residual family history. For estimating disease risk to parents or offspring, $\lambda_\tau$ and $\lambda_\sigma$ are used, with

$$\lambda_{G\sigma} = \lambda_{A\sigma} = 1 + \nu D/\nu E.$$

### Disease risk calculation

To model the disease risk of an individual based on their genotype at $G$, exposure to the environmental factor $E$, and their family history, we assume independence between the environmental factor, genotype at gene $G$ and residual family history (after accounting for $G$). For the environmental factor $E$, $P(D|E) = r P(D|E')$ and $P(D) = P(D|E) P(E) + P(D|E') (1 - P(E))$, therefore $P(D|E) = r K/(rs + (1-s))$.

$$\lambda_{G} = g_{0}/f_{0}$$

are given above for $g_{0} = (AA, AB, BB)$, using genotype relative risks and disease prevalence.

$$\lambda_{SG'} = P(D|G)P(D|E) \lambda_{SG'}/P(D) = GRR(G)/OR(E)\lambda_{SG'} K$$
familial relative risks\textsuperscript{4–8} (table 1). CD twin studies were omitted because of the low numbers of CD probands.\textsuperscript{2,3} Other studies could not be used because CD relatives were classified only as first or second degree, without information on the precise relationship to the proband.\textsuperscript{28}

Kuster et al\textsuperscript{a} performed a segregation analysis of 265 probands with CD from Dusseldorf, Germany, finding 13 siblings with CD (from a total of 453 siblings), and 4 parents with CD (from 530 parents). Satsangi et al\textsuperscript{b} studied 433 adult CD patients in Oxford, UK, showing that 33 siblings, 18 parents and 4 offspring were affected with inflammatory bowel disease. Of the 33 siblings, 20 were affected with CD, and assuming that the ratio of relatives affected with CD to those affected with ulcerative colitis remains constant across all first degree relatives, we estimate that 11.5 parents and 2.4 offspring were affected with CD. Peeters et al\textsuperscript{c} ascertained 640 probands with CD in Belgium, showing that 13.6\% had a first degree relative with CD, and classifying the affected relatives into siblings (n = 57/1728), parents (n = 21/1280) or offspring (n = 17/835). Probert et al\textsuperscript{d} collected family history information on 424 patients with CD from Leicester, UK, finding 19 of 984 siblings affected with CD, 8 of 825 parents and 8 of 493 offspring.

Calculating the sibling relative risk requires a value for CD prevalence. This has been estimated in several UK studies, with rates per 100 000 population of 76 in Leicestershire\textsuperscript{e} and 147 in NE Scotland.\textsuperscript{29} The incidence of CD is increasing in all Western populations, so more recent estimates are higher, but these studies are from the same time period as the familial studies in table 1. In this study, we assumed 100 patients with CD per 100 000 of population, giving a disease prevalence of K = 0.001.

The sibling relative risk (\(S\)) is defined as the ratio of the observed risk to siblings of patients with CD to the population prevalence (so \(S = K\) provides a baseline of no increased familial risk). Study-specific sibling relative risks ranged from 19\% to 33\%, and weighting by the number of probands per study gives a mean sibling relative risk of 27.2 (table 1). Similar calculations from these papers show the estimated relative risk for parents to be \(O = 12.8\) (from four studies), and relative risk for offspring of probands to be \(N = 16.6\) (excluding the German study, which did not study the offspring of probands). The lower relative risks of parents and offspring may be due to the greater shared environment of siblings compared with offspring and parents. Alternatively, risks may reflect the lower prevalence of CD relevant for the parental cohort, and the young age of the offspring, who may not have lived through the full risk period for CD.

CARD15 risks

CARD15 (NOD2) was localised in 2001, with the identification of three different mutations (R702W, G908R, L1007fs) conferring increased risk of CD.\textsuperscript{9–11} Other, much rarer mutations in CARD15 exist, with frequency <0.1\%,\textsuperscript{30,31} but are found in only a small proportion of patients with CD. In common with other studies, these variants are not considered here.

Many research groups have published estimates of the population frequency and genotype relative risks of these CARD15 mutations, and a meta-analysis of 42 studies showed that people carrying one mutation had 2.3-fold increased odds of CD (95\% CI 2.00 to 2.86), and people carrying two mutations had a 17.1-fold increased risk of CD (95\% CI 10.7 to 27.2).\textsuperscript{13} Differences across populations exist, with no mutations present in Asian populations,\textsuperscript{32} and a trend across Europe, with higher frequencies of the 1007fs frameshift in central and southern Europe than in northern Europe.\textsuperscript{33} Differences in risk conferred by each mutation were detected in the meta-analysis and a large French family study,\textsuperscript{13,34} but few individual studies have sufficient power to detect this effect.

In this study, we used data from our large UK study of 1639 patients with CD and 1808 controls, genotyped for the 3 CARD15 mutations as described by Prescott et al.\textsuperscript{14} Individuals were categorised by the number of CD mutations carried (0, 1, 2), as mutations occur only rarely on the same haplotype, and without discriminating between the different mutations (table 2). Genotype relative risks are 2.25 (95\% CI 1.88 to 2.68) for people carrying one mutation and 9.25 (95\% CI 5.44 to 15.7) for people carrying two mutations, compared with those carrying no CARD15 mutations. Using these figures in the equations for the sibling relative risk due to CARD15 above gives \(S_{5,CARD15} = 1.16\).

RESULTS

The estimated risks of developing CD, based on the model developed above, are shown in tables 3 and 4. The parameter estimates for the risk calculation are summarised in table 3; in addition to the values derived and discussed above, we assumed that the smoking prevalence in the UK is 24\%.\textsuperscript{15} The CD risks for the sibling, parent or offspring of a patient with CD, based on CARD15 mutation status (0, or 2 mutations or ungenotyped) and smoking status (smoker, non-smoker or unknown), are given in table 4. The baseline relative risk for the

<table>
<thead>
<tr>
<th>Reference</th>
<th>Population</th>
<th>CD probands, n</th>
<th>CD siblings, n</th>
<th>Total siblings, n</th>
<th>CD risk, %</th>
<th>Sibling relative risk*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kuster et al (1989)\textsuperscript{a}</td>
<td>Germany</td>
<td>265</td>
<td>13</td>
<td>453</td>
<td>2.87</td>
<td>28.7</td>
</tr>
<tr>
<td>Probert et al (1993)\textsuperscript{b}</td>
<td>UK</td>
<td>424</td>
<td>19</td>
<td>984</td>
<td>1.93</td>
<td>19.3</td>
</tr>
<tr>
<td>Peeters et al (1996)\textsuperscript{c}</td>
<td>Belgium</td>
<td>640</td>
<td>57</td>
<td>1728</td>
<td>3.3</td>
<td>33.0</td>
</tr>
<tr>
<td>Satsangi et al (1994)\textsuperscript{d}</td>
<td>UK</td>
<td>433</td>
<td>20</td>
<td>782</td>
<td>2.56</td>
<td>25.6</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>2003</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Average risk</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.67</td>
<td>26.7</td>
</tr>
<tr>
<td><strong>Average, weighted by study size</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.72</td>
<td>27.2</td>
</tr>
</tbody>
</table>

*Using a population prevalence of 0.001 for all studies.

Estimates of sibling relative risk for each study, mean value across studies, and a mean value weighting by study size (number of probands) are given.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Population</th>
<th>CD probands, n</th>
<th>CD siblings, n</th>
<th>Total siblings, n</th>
<th>CD risk, %</th>
<th>Sibling relative risk*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases, n (%)</td>
<td>1639</td>
<td>1128</td>
<td>403 (24.6)</td>
<td>108 (6.6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controls, n (%)</td>
<td>1808</td>
<td>1546 (85.5)</td>
<td>246 (13.6)</td>
<td>16 (0.9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotype</td>
<td>1</td>
<td>22.5</td>
<td>9.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>relative risk</td>
<td>(95% CI)</td>
<td>1.88 to 5.44</td>
<td>2.68 to 15.7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

www.jmedgenet.com
sibling of a patient with CD is a sibling relative risk of \( \lambda_S = 27.2 \), which gives a 2.7% risk of developing CD (assuming a disease prevalence of 0.001). A person who carries no mutations in \( CARD15 \) has the risk of CD reduced slightly to 2.3%, person with a single \( CARD15 \) mutation has a slightly increased risk of 5.3%, and people with two mutations have a greatly increased risk of 21.7%. Siblings who also smoke increase their risks still further to 3.8% (0 mutation), 8.5% (1 mutation), or 35.0% (2 mutations). Non-smokers have risks equal to approximately half the risks for siblings who smoke. Without information on \( CARD15 \), the risks for non-smoking and smoking siblings of patients with CD are 2.2% and 4.4%, respectively. Adding \( CARD15 \) mutation status therefore provides considerable resolution of these risks, and permits the identification of a cohort of people (smokers who carry two \( CARD15 \) mutations and have a sibling with CD) who have a high probability of developing CD.

For siblings of a patient with CD, a non-smoker with no \( CARD15 \) mutations decreases their risk by approximately one-third, whereas a smoker with two \( CARD15 \) mutations increases their risk 13-fold, compared with risks of 2.7% based solely on family history information. A non-smoker with a single \( CARD15 \) mutation is at approximately the same risk of developing CD as a smoker who carries no \( CARD15 \) mutations (as the OR for smoking and for one \( CARD15 \) mutations are 2 and 2.24 respectively). The major contribution to disease risk is carrying two \( CARD15 \) mutations (OR = 9.24), and the residual family history risk, after accounting for \( CARD15 \) (\( \lambda_S, CARD15 = 23 \)). Similar patterns of risk are seen for the parents and offspring of patients with CD. A program for risk calculation is available on request.

**DISCUSSION**

We have developed a model to calculate individual-specific risks for complex genetic diseases, based on our current knowledge of environmental and genetic risk factors. This model shows how inclusion of genotype information from only a single gene and environmental risk factor can substantially change disease risks. For an individual with all risk factors present, a moderately high disease risk was obtained (34% for smoking siblings of a patient with CD, with two \( CARD15 \) mutations).

Few complex genetic diseases have sufficient information on both genetic and environmental factors to estimate disease risks. In age-related macular degeneration, Maller et al. estimated risks of up to 50% from single-nucleotide polymorphism genotypes at three genes. Smoking is also a major risk factor for this condition and, in contrast to CD, an interaction with the \( LOC387715 \) locus exists, with risks for smokers who also carry the high-risk genotypes being much higher than predicted by the marginal risks conferred by smoking and \( LOC387715 \). A population study of CD in Manitoba showed that people with a family history of CD who also smoked and carried two \( CARD15 \) mutations had a much increased risk of CD, with an OR of 237 (95% CI 63 to 1054). The methodology of the Manitoba study was different from the current study, with risk information assessed from a series of healthy controls (n = 336) and patients with CD (n = 232), giving an OR of 13.9 for people carrying two \( CARD15 \) mutations, 3.0 for smokers, and 6.2 for the presence of CD in a first degree relative. Despite the disparity of the designs and the specific risk estimates calculated, both this study and the Manitoba study illustrate how combining risks across several factors of modest effect can produce high predicted disease risks for specific subgroups of the population.

One limitation of our model is that it uses no information on the age of the affected and unaffected siblings, or the time since diagnosis of the patient with CD. Siblings are correlated in age at diagnosis, with most affected siblings diagnosed within 10 years of each other. Therefore a sibling who remains unaffected with CD more than a decade after the diagnosis of their sibling has lived through the majority of their at-risk period, and must be considered at lower risk. Joint distributions of age at diagnosis of CD in family members could be used to develop further methodology. For example, age-dependent methods for risk estimation were developed in the REVEAL study of Alzheimer’s disease, incorporating \( APOE \) genotypes and family history and age, using Kaplan–Meier survival methods. Our risk model assumes that risk factors are independent, with multiplicative relationship across risks conferred. In CD, a joint study on \( CARD15 \) and variants in IL23R and on chromosome 5q31 showed that risks across genes were multiplicative, and smoking prevalence is similar in CD patients with and without \( CARD15 \) mutations, implying the lack of any interaction between these risk factors. Current studies therefore support the assumption...
of independent contribution to risk from each genetic and environmental factor. The model can be expanded to multiple genes (eg IL23R and ATG16L1) and additional environmental risk factors, provided the assumption of multiplicative risks across factors holds.

The method for risk estimation developed here is flexible in allowing risk factors estimated from independent studies to be combined. This is a valuable property for CD, as few studies with adequate power have produced consistent findings. The utility in screening smokers at risk who have the potential for early diagnosis has been much debated.44-44 Such screening is considered to be of limited value because of the low predictive value of the test (14% of the UK population carries at least one mutation in CARD15). However, the utility in screening smokers at risk who have the potential to reduce their risk is unknown. The risk estimates published here are being used to provide smokers who are first degree relatives of patients with CD with estimates of the likelihood that they will develop CD, together with information about the potentially risk-reducing effects of smoking cessation. Of interest is whether providing precise disease risk estimates based on DNA analysis together with information about reducing these risks motivates smoking cessation, a question we are investigating in a randomised controlled trial.

ACKNOWLEDGEMENTS

We thank Natalie Prescott and Sheila Fisher for data on CARD15 genotypes. We acknowledge use of DNA from the British 1958 Birth Cohort collection, funded by the Medical Research Council (grant G0000994) and The Wellcome Trust (grant 068545/Z/02).

Authors’ affiliations

C M Lewis, C G Mathew, Department of Medical and Molecular Genetics, Division of Genetics and Molecular Medicine, King’s College London School of Medicine, Guy’s Campus, London, UK

C M Lewis, Social, Genetic and Developmental Psychiatry Research Centre, Institute of Psychiatry, King’s College London, London, UK

S C L Whitwell, T M Mateau, Psychology and Genetics Research Unit, King’s College London School of Medicine, Guy’s Campus, London, UK

A Forbes, Department of Gastroenterology, Institute for Digestive Diseases, University College London Hospitals Trust, London, UK

J Sanderson, Department of Gastroenterology, Guy’s and St Thomas’ NHS Foundation Trust, London, UK

Funding: This research was supported by the Wellcome Trust (076024 to CML; 072029 to CGM) and the Medical Research Council (G0000274 to TMM for Risk communication in preventative medicine: optimising the impact of DNA risk information). Guy’s and St Thomas’ NHS Trust in conjunction with KCL and University College Hospital in conjunction with UCL, receive funding support from the UK National Institute for Health Research.

Competing interests: None declared.

REFERENCES


21 Camp NJ. Genomewide transmission/disequilibrium testing – consideration of the genetic complexity and effects of disease phenotype in Inflammatory Bowel Disease. Gastroenterology 2007;132:1657-64.

22 Camp NJ. Genomewide transmission/disequilibrium testing – consideration of the genetic complexity and effects of disease phenotype in Inflammatory Bowel Disease. Gastroenterology 2007;132:1657-64.


27 Brant SR, Wang MH, Rawsthorne P, Sargent M, Datta LW, Nauvet EJ, Shugart YY, Bernstein CN. A population-based case-control study of CARD15 and


43 Rubin DT. To test or “Nod-2” test: What are the questions? The balanced viewpoint. *Inflamm Bowel Dis* 2005;11:510–12.

Estimating risks of common complex diseases across genetic and environmental factors: the example of Crohn disease

C M Lewis, S C L Whitwell, A Forbes, J Sanderson, C G Mathew and T M Marteau

doi: 10.1136/jmg.2007.051672

Updated information and services can be found at:
http://jmg.bmj.com/content/44/11/689

These include:

References
This article cites 43 articles, 4 of which you can access for free at:
http://jmg.bmj.com/content/44/11/689#BIBL

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Topic Collections
Articles on similar topics can be found in the following collections

- Inflammatory bowel disease (16)
- Open access (198)

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/