

Diagnostic misclassification reduces the ability to detect linkage in inflammatory bowel disease genetic studies

M S Silverberg, M J Daly, D N Moskovitz, J D Rioux, R S McLeod, Z Cohen, G R Greenberg, T J Hudson, K A Siminovitch, A H Steinhart

Abstract

Background—Linkage data have now identified several inflammatory bowel disease (IBD) susceptibility loci but these data have not been consistently replicated in independent studies. One potential explanation for this is the possibility that patients enrolled in such studies may have been erroneously classified with respect to their diagnosis.

Aims—To determine the rate and type of misclassification in a large population of individuals referred for participation in an IBD genetics study and to examine the effect of diagnostic misclassification on the power to detect linkage.

Methods—The medical records of 1096 patients entered into an IBD genetics programme were reviewed using standardised diagnostic criteria. The original patient reported diagnoses were changed, if necessary, based on review, and the reasons for the change in diagnosis were recorded. To evaluate the effect of misclassification on linkage results, simulations were created with Gensim and analysed using Genehunter to evaluate a model for IBD inheritance.

Results—Sixty eight of 1096 (6.2%) individuals had a change in diagnosis from that originally reported. The majority of changes were patients with either Crohn's disease or ulcerative colitis who were determined not to have IBD at all. The principal reasons for changes to the original diagnosis were discordance between the patients' subjective reports of diagnosis and actual clinical history, endoscopic, or pathological results; a change in disease pattern over time; and insufficient information available to confirm the original diagnosis. A 10% misclassification rate resulted in 28.4% and 40.2% loss of power to detect a true linkage when using a statistical model for a presumed IBD locus with λ_s values of 1.8 and 1.3, respectively.

Conclusions—Diagnostic misclassification occurs in patients enrolled in IBD genetic studies and frequently involves assigning the diagnosis of IBD to non-affected individuals. Even low rates of diagnostic misclassification can lead to significant loss of power to detect a true linkage, particularly for loci with modest effects as are likely to be found in IBD.

(Gut 2001;49:773–776)

Keywords: inflammatory bowel disease; Crohn's disease; ulcerative colitis; genetics; linkage analysis

Despite being associated with distinct cytokine profiles and specific pathological features, Crohn's disease (CD) and ulcerative colitis (UC) are both associated with recurrent inflammation and ulceration of the gastrointestinal tract and thus share many symptoms, clinical characteristics, and treatment strategies. Distinction of the two disorders, which at the current time is based on specific clinical manifestations, pathological features, and anatomical distribution of disease, can sometimes be difficult. Moreover, in 10–15% of patients with colonic inflammation, overlap in the clinical and histological manifestations make distinction between CD and UC indiscernible, resulting in the classification of such cases as indeterminate colitis (IDC).¹ Despite the similarities between CD and UC, it is important to accurately classify patients with inflammatory bowel disease (IBD) as prognosis and treatment strategies may differ in significant ways.² Also, correct classification of CD and UC will affect the interpretation of epidemiological data regarding incidence and prevalence of these disorders. Appropriate and consistently applied criteria to classify patients with IBD may be particularly important for studies attempting to elucidate potential genetic and environmental determinants of IBD.

In view of epidemiological data revealing an important contribution of genetic factors to the pathogenesis of IBD, several groups have carried out genome wide scans for the purpose of localising IBD susceptibility genes.^{3–9} The results of such studies have identified several chromosomal regions linked to IBD but these data have not been consistently replicated in all groups. There are many possible explanations for such discrepancies, including: differing admixtures of CD and UC; varying ethnic and geographic origins of the populations studied; small sample sizes; type I error; type of genetic markers used; and the presence of genetic heterogeneity and incomplete penetrance.¹⁰ Moreover, as the relative contribution to the overall genetic risk for IBD attributable to each of the susceptibility loci identified thus far is

Abbreviations used in this paper: IBD, inflammatory bowel disease; CD, Crohn's disease; UC, ulcerative colitis; IDC, indeterminate colitis; ASP, affected sibling pair; LOD, logarithm of odds; MLS, multipoint LOD score.

Department of
Medicine, Mount Sinai
Hospital Inflammatory
Bowel Disease Centre,
University of Toronto,
Toronto, Canada
M S Silverberg
D N Moskovitz
G R Greenberg
K A Siminovitch
A H Steinhart

Department of
Surgery, Mount Sinai
Hospital Inflammatory
Bowel Disease Centre,
University of Toronto,
Toronto, Canada
R S McLeod
Z Cohen

Whitehead
Institute/MIT Center
for Genome Research,
Cambridge, MA, USA
M J Daly
J D Rioux
T J Hudson

Montreal Genome
Centre, McGill
University Health
Centre, Montreal,
Canada
T J Hudson

Correspondence to:
Dr A H Steinhart, Room
445, Mount Sinai Hospital,
600 University Avenue,
Toronto, Ontario M5G 1X5,
Canada.
hsteinhart@mtsinai.on.ca

Accepted for publication
29 March 2001

low, the potential exists that diagnostic misclassification, even in a small proportion of patients, has a significant impact on results of linkage analyses and thus leads to failure to replicate a previously identified locus or inability to detect additional novel loci of potential relevance. To address this possibility, we have examined the frequency with which such misclassification of IBD occurs, the reasons for misclassification, and the impact of classification errors on the ability to detect linkage using a statistical model of a simulated genome wide scan for IBD.

Methods

The study population was ascertained from a database consisting of IBD patients and their family members recruited at the Inflammatory Bowel Disease Centre of Mount Sinai Hospital, Toronto, Canada, where an IBD genetics programme has been in progress since 1995.¹¹ This programme was established principally to identify susceptibility genes for IBD but also to examine genotype-phenotype relationships in IBD. Patients were derived mainly from IBD patients followed at Mount Sinai Hospital but also by referral from gastroenterologists and surgeons in the Toronto area. Recruitment into this IBD genetics programme initially focused on families with multiple affected members but subsequently individuals without a family history were also recruited. Patients providing informed consent to enter into the IBD genetics programme were asked to complete a questionnaire detailing information about their illness, and DNA was extracted from peripheral blood samples taken from affected individuals and unaffected relatives, as described previously.¹¹ Questions asked included, but were not limited to: type of inflammatory bowel disease (UC, CD, IDC); anatomical distribution; previous surgical procedures; and presence of extraintestinal manifestations. After completing the questionnaire, consent was obtained from patients to acquire all relevant clinical records, including consultation, radiology, endoscopy, surgery, and pathology reports, from any previous physicians providing care for the patients. For all affected individuals, the records were then reviewed by at least one of the five authors (MSS, RSM, ZC, GRG, AHS). Diagnoses of CD, UC, and IDC were subsequently determined from the documentation obtained using standard criteria. The diagnosis originally provided by the patient or referring physician was either confirmed or changed based on application of the standardised criteria.

Table 1 Distribution (number (%)) of diagnostic changes and reasons for change

	Type of diagnostic change				
	UC→CD or CD→UC Group 1	UC or CD →IDC Group 2	UC or CD →nil Group 3	IDC→UC or IDC→CD Group 4	IDC→nil Group 5
Reason for change in diagnosis					
Discordance (reason 1)	5 (38)	5 (28)	11 (41)	1 (25)	4 (58)
Change in disease pattern (reason 2)	6 (46)	4 (24)	4 (15)	1 (25)	1 (14)
Data entry error (reason 3)	1 (8)	4 (24)	3 (11)	1 (25)	1 (14)
Insufficient information (reason 4)	1 (8)	4 (24)	9 (33)	1 (25)	1 (14)

SUBJECTS

Individuals analysed for this report were identified by determining those patients in the database for whom a change in diagnosis was made at any point after the original diagnosis provided by the patient or referring physician had been entered. For analysis of the cases where the original diagnosis was altered, five groups were created (table 1): those patients where the diagnosis was changed from CD to UC or where the diagnosis was changed from UC to CD (group 1); patients whose diagnosis changed from UC or CD to IDC (group 2); patients whose diagnosis changed from UC or CD to no diagnosis of IBD (group 3); patients who were originally classified as having IDC but changed to UC or CD (group 4); and patients originally diagnosed with IDC but ultimately deemed not to have IBD (group 5). Qualitative analysis of the reasons for diagnostic change was undertaken for each patient and a primary reason for change was assigned.

STATISTICAL MODELLING AND DEFINITIONS

Simulations were created with Gensim⁹ (computer program designed by MJD) and analysed using Genehunter.¹² A multiplicative model¹³ for IBD inheritance with allele frequencies and penetrances designed to produce between 1/1000 to 1/2000 affected individuals in the population was applied (approximating the actual prevalence of IBD). Three sets of simulations were performed, all using a simulated 10 centiMorgan map. The first model evaluated 1000 sets of 100 affected sibling pairs (ASP) for a locus with a λ_s value (relative risk to sibling of an affected individual) of 3.5, the second model evaluated 1000 sets of 200 ASP for a locus with a λ_s value of 1.8, and the third model included 1000 replicates of 500 ASP for a locus with a λ_s value of 1.3. Subsequent simulations within each model were performed with identical parameters but varying the number of ASP linked to the presumed locus so as to simulate misclassification of patients. This was accomplished by generating 5%, 10%, or 20% of ASP in a random fashion such that these ASP would not be linked to the presumed locus. The number of ASP used was increased for each model as the locus utilised became weaker; this was done to correct for the fact that the expected logarithm of odds (LOD) score would be less for a locus with a weaker effect and increasing the sample size would result in a roughly equivalent expected LOD score. The average multipoint LOD score (MLS) was determined for each of the simulations (a measure of the degree of linkage

to presumed locus) and averaged. The number of simulations that achieved an MLS of at least 3.6, the level of genome wide significance,¹⁴ was determined as well as the percentage of simulations that did not achieve this threshold at varying levels of simulated classification errors (that is, number of patients not linked to the presumed locus).

Results

The medical records of 1096 IBD affected study participants were reviewed for the analysis. Of these individuals, 68 (6.2%) required a change in diagnosis from that originally reported: 35 (51%) had an original diagnosis of UC, 22 (32%) had an original diagnosis of CD, and 11 (17%) were originally diagnosed as having IDC. Change in diagnosis from UC to CD, or from CD to UC, was made in 13 of these 68 subjects (group 1), 17 individuals had their diagnosis changed from UC or CD to IDC (group 2), 27 with an original diagnosis of UC or CD were subsequently reclassified as unaffected (group 3), four patients with an original diagnosis of IDC were recategorised as UC or CD (group 4), and seven patients with an original diagnosis of IDC were reclassified as unaffected (group 5) (table 1).

After review of the clinical, endoscopic, radiological, surgical, and pathological reports of all affected individuals, reasons for change of diagnosis were assessed and shown to fall into four general categories: discordance between the patients' reported history and objective information obtained in the form of endoscopic, surgical, radiological, or histological results (reason 1); change in disease pattern over time such that reports reviewed closer to the time of study entry were different than those that may have been available to the original diagnosing physician (reason 2); data entry error (that is, no change in diagnosis should have been documented) (reason 3); and insufficient information available to confirm the originally reported diagnosis (reason 4). The distribution of these reasons for each of the five groups of diagnostic change are summarised in table 1.

Using Gensim and Genehunter to model various misclassification rates as described above, loss of power to detect linkage of genome wide significance is demonstrated when patients are incorrectly classified and thus not likely to be linked to a presumed susceptibility locus. For a locus with a relatively strong genetic effect ($\lambda_s=3.5$), a 10% error rate in classification led to a loss of power of 12% to detect a significant linkage. For loci contributing more modest genetic risks, such as one with $\lambda_s=1.8$ or 1.3, the same 10% error rate led to 28.4% and 40.2% loss of power, respectively, to detect significant linkage (table 2).

Discussion

In the current study, 6.2% of 1096 patients with presumed IBD required a change in diagnosis when their clinical records were reviewed and consistent diagnostic criteria applied. When computer data entry errors were eliminated, the actual diagnostic misclassification

Table 2 Simulations created with Gensim and analysed with Genehunter. Successive simulations were performed with increasing proportion of affected sibling pairs generated randomly (that is, not linked to presumed locus)

% Linked	Avg MLS	No sims MLS >3.6	% Loss
<i>(a) Presumed locus with $\lambda_s=3.5$</i>			
100%	6.28	900	NA
95%	5.67	882	2.0
90%	5.29	792	12.0
80%	4.17	603	33.0
<i>(b) Presumed locus with $\lambda_s=1.8$</i>			
100%	4.84	828	NA
95%	4.49	645	22.1
90%	4.21	593	28.4
80%	3.50	428	48.3
<i>(c) Presumed locus with $\lambda_s=1.3$</i>			
100%	3.28	368	NA
95%	3.06	296	19.6
90%	2.76	220	40.2
80%	2.39	158	57.1

λ_s , relative risk in sibling versus general population risk.

% Linked, per cent of affected sibling pairs linked to presumed locus.

Avg MLS, average multipoint LOD score seen in 100 (a), 200 (b), or 500 (c) affected sibling pairs after 1000 replicates.

No sims MLS >3.6, number/1000 simulations where the multipoint LOD score reached a threshold of 3.6.

% Loss, per cent fewer simulations that reached a threshold of 3.6.

NA, not applicable.

rate based on clinical criteria was 58/1096 (5.3%). Importantly, the thorough chart review revealed 30 of these 58 individuals to have no evidence of IBD at all. These results are similar to a previous analysis investigating diagnostic change in a population of IBD patients.² The results of the current study also identify several reasons underlying the need for a change in diagnosis. The most frequent source of misclassification reflects discordance between the patients' history and more objective criteria reported at surgery or endoscopy, or by review of tissue pathology. In addition, amendments to diagnosis occurred secondary to a change in disease pattern over time.^{15 16} Such alterations complicate attempts to correctly classify IBD patients, particularly for genetic studies. Lack of sufficient information in the clinical records to confirm a diagnosis was another common reason for alteration in diagnosis. Many of these patients, for whom classification changed from UC or CD to no IBD, did not meet histological criteria for a diagnosis of IBD or lacked ongoing symptomatology. In this group of patients it therefore appears that the original report of IBD should not have been made as criteria for IBD based on histological evidence were not met and features of chronic inflammation were frequently absent. The tendency of IBD to relapse is thought to be a critical diagnostic criterion, especially in the absence of confirmatory histological data.¹⁷ Other studies have suggested that specific histological criteria must exist before a diagnosis of IBD is made.¹⁸

While the majority of diagnostic changes in this report involved patients ultimately deemed not to have IBD, a proportion included those with one type of IBD (that is, CD, UC, or IDC) requiring reclassification to another form of IBD. Insufficient numbers of patients within these categories were available to draw conclusions regarding the reasons for these changes but clearly this type of change can have a significant impact on interpretation of IBD

genetic studies as evidence exists suggesting some susceptibility loci are particularly relevant to CD or UC rather than to IBD as a whole.^{19,20}

In the context of this study, patients were recruited for the purpose of elucidating susceptibility loci for IBD. Thus we performed this analysis to determine what impact, if any, patient misclassification would have on the ability of multipoint linkage analysis to detect chromosomal loci that may be truly linked to disease susceptibility. The modelling carried out in this report demonstrates that even an error rate as low as 10% can lead to a 40% risk that the threshold for significant linkage would not be reached and a false negative result would be obtained. The negative impact of misclassification errors on the power to detect a true linkage is even more pronounced when the presumed locus is one with a relatively modest effect. As demonstrated in table 2, when a λ_s value of 1.8 was employed, even a 5% classification error rate (less than that found in this report) led to 22% loss of power to detect a true linkage; when a λ_s value of 1.3 was used, similar to the λ_s of the *IBD1* susceptibility locus on chromosome 16,³ a 5% or 10% misclassification rate resulted in 20% and 40% loss of power, respectively. Thus loci contributing to IBD susceptibility are likely to be particularly sensitive to even a small degree of misclassification of patients.

With available data now indicating the existence of at least four confirmed IBD susceptibility loci, all with relatively small estimated contributions to the total genetic risk for disease, these results underscore the importance of rigorous review of clinical records of all patients entered into IBD genetic studies and the consistent application of diagnostic criteria to these individuals. The difficulties in replicating IBD susceptibility loci and in interpreting the multitude of candidate gene association studies clearly demonstrate the importance of attempting to maintain homogeneity with respect to diagnosis in patient populations studied. It is acknowledged that the overlap between the clinical, endoscopic, and pathological appearances of CD and UC may sometimes make distinction between the two conditions difficult. However, consistent application of predetermined criteria may be extremely useful in these studies.

The authors would like to thank Brenda O'Connor for her assistance in completing the work for this project. This work was supported in part by a grant from the Crohn's and Colitis

Foundation of Canada and from research contracts from Bristol Myers Squibb, Millennium Pharmaceuticals Inc., and Affymetrix. MS Silverberg is supported by a fellowship from the Canadian Association of Gastroenterology, the Canadian Institutes of Health Research, and Axcan Pharma, and by a Research Initiative Award from AstraZeneca. TJ Hudson is a clinician scientist of the Canadian Institutes of Health Research. KA Siminovitch is a senior scientist of the Canadian Institutes of Health Research.

- Podolsky DK. Inflammatory bowel disease. *N Engl J Med* 1991;325:928-37.
- Marcello PW, Schoetz DJ, Roberts PL, et al. Evolutionary changes in the pathologic diagnosis after the ileoanal pouch procedure. *Dis Colon Rectum* 1997;40:263-9.
- Hugot JP, Laurent-Puig P, Gower-Rousseau C, et al. Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature* 1996;379:821-3.
- Satsangi J, Parkes M, Louis E, et al. Two stage genome-wide search in inflammatory bowel disease provides evidence for susceptibility loci on chromosomes 3, 7 and 12. *Nat Genet* 1996;14:199-202.
- Cho JH, Nicolae DL, Gold LH, et al. Identification of novel susceptibility loci for inflammatory bowel disease on chromosomes 1p, 3q, and 4q: evidence for epistasis between 1p and IBD1. *Proc Natl Acad Sci USA* 1998;95:7502-7.
- Hampe J, Schreiber S, Shaw SH, et al. A genomewide analysis provides evidence for novel linkages in inflammatory bowel disease in a large European cohort. *Am J Hum Genet* 1999;64:808-16.
- Ma Y, Ohmen JD, Li Z, et al. A genome-wide search identifies potential new susceptibility loci for Crohn's disease. *Inflamm Bowel Dis* 1999;5:271-8.
- Duerr RH, Barmada MM, Zhang L, et al. High-density genome scan in Crohn's disease shows confirmed linkage to chromosome 14q11-12. *Am J Hum Genet* 2000;66:1857-62.
- Rioux JD, Silverberg MS, Daly MJ, et al. Genomewide search in Canadian families with inflammatory bowel disease reveals two novel susceptibility loci. *Am J Hum Genet* 2000;66:1863-70.
- Risch N. Searching for genes in complex diseases: lessons from systemic lupus erythematosus. *J Clin Invest* 2000;105:1503-6.
- McLeod RS, Steinhart AH, Siminovitch KA, et al. Preliminary report on the Mount Sinai Hospital inflammatory bowel disease genetics project. *Dis Colon Rectum* 1997;40:553-7.
- Kruglyak L, Daly MJ, Reeve-Daly MP, et al. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* 1996;58:1347-63.
- Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996;273:1516-17.
- Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995;11:241-7.
- Greenstein AJ, Lachman P, Sachar DB, et al. Perforating and non-perforating indications for repeated operations in Crohn's disease: evidence for two clinical forms. *Gut* 1988;29:588-92.
- Meucci G, Bortoli A, Riccioli FA, et al. Frequency and clinical evolution of indeterminate colitis: a retrospective multi-centre study in northern Italy. *Eur J Gastroenterol Hepatol* 1999;11:909-13.
- Surawicz CM, Haggitt RC, Husseman M, et al. Mucosal biopsy diagnosis of colitis: acute self-limited colitis and idiopathic inflammatory bowel disease. *Gastroenterology* 1994;107:755-63.
- Le Berre N, Heresbach D, Kerbaol M, et al. Histological discrimination of idiopathic inflammatory bowel disease from other types of colitis. *J Clin Pathol* 1995;48:749-53.
- Parkes M, Barmada MM, Satsangi J, et al. The IBD2 locus shows linkage heterogeneity between ulcerative colitis and Crohn's disease. *Am J Hum Genet* 2000;67:1605-10.
- Cavan J, the IBD International Genetics Consortium. International collaboration provides convincing linkage replication in complex disease through analysis of a large pooled data set: Crohn's disease and chromosome 16. *Am J Hum Genet* 2001;68:1165-71.