Université de Montréal

The Hygiene Hypothesis and the risk of Crohn's disease: a case-control study utilizing prospectively-collected exposure data from an administrative database

par Vicky Springmann

Département de médecine sociale et préventive, École de santé publique de l'Université de Montréal

Mémoire présenté à la l'École de santé publique de l'Université de Montréal, en vue de l'obtention du grade de maîtrise en santé communautaire, option recherche

octobre 2013

Résumé:

La maladie de Crohn (MC) pédiatrique a des conséquences majeures sur la qualité de vie des patients atteints (troubles de croissance, absentéisme scolaire, etc). L'étiologie de la MC est inconnue. La théorie de l'hygiène (TH) stipule que les conditions de vie sanitaires des pays industrialisés préviennent l'exposition antigénique et empêchent le développement de la tolérance immunitaire chez les enfants. Ceci mènerait à une réaction excessive du système immunitaire lors d'expositions subséquentes et engendrerait le développement de maladies inflammatoires chroniques telles la MC.

Objectif: Analyser l'association entre la fréquence, la temporalité et le type d'infections infantiles (indicateurs d'environnements pourvus d'antigènes) et le risque de MC pédiatrique.

Une étude cas-témoin fût réalisée, les cas de MC provenant d'un centre hospitalier tertiaire montréalais. Les témoins, provenant des registres de la *Régie d'assurance maladie du Québec* (RAMQ), furent appariés aux cas selon leur âge, sexe et lieu de résidence. L'exposition aux infections fût déterminée grâce aux codes de diagnostic ICD-9 inscrits dans la base de données de la RAMQ. Un modèle de régression logistique conditionnelle fût construit afin d'analyser l'association entre infections et MC. Des ratios de cotes (RC) et intervalles de confiance à 95% (IC 95%) furent calculés.

Résultats: 409 cas et 1621 témoins furent recrutés. Les résultats de l'analyse suggèrent un effet protecteur des infections infantiles sur le risque de MC (RC: 0,67 [IC: 0,48-0,93], p=0,018), plus particulièrement au cours des 5 premières années de vie (RC: 0.74 [IC: 0,57-0,96], p=0,025). Les infections rénales et urinaires, ainsi que les infections des voies orales et du système nerveux central (virale), semblent particulièrement associées à l'effet protecteur. Les résultats de l'étude appuient la théorie de l'hygiène: l'exposition aux infections infantiles pourrait réduire le risque de MC pédiatrique.

Mots clés : maladie de Crohn pédiatrique, théorie de l'hygiène, infections, étude cas-témoin

Summary:

Crohn's disease (CD) poses specific challenges in the paediatric population (growth failure, depression, etc). The environmental contributors to CD aetiology remain largely unknown. There are suggestions that sanitary living conditions prevailing in developed countries prevent antigen exposure and impede the development of immunological tolerance amongst children, resulting in abnormally heightened immunological responses with subsequent exposures (hygiene hypothesis). Evidence for the hygiene hypothesis in CD aetiology remains unclear.

Objectives: To assess the role of the frequency, timing and type of childhood infections (measures of antigen exposure) on the risk of paediatric CD.

A case-control study was carried out. Confirmed cases of CD were recruited from a tertiary care paediatric hospital. Controls matched to the cases on calendar age, gender, and area of residence, were selected using the provincial health insurance files. Infection exposure was ascertained using ICD-9 diagnostic codes provided by the provincial insurer's administrative databases. Conditional logistic regression analysis was used to assess the relationship between childhood infections and CD. Odds ratios (OR) and corresponding 95% confidence intervals (95% CI) were estimated.

409 cases and 1621 controls were recruited. A diagnosis of infection was associated with reduced risks for paediatric CD (OR=0.67, 95% CI:[0.48-0.93], p=0.018), attributable to infection exposures primarily during the first 5 years since birth [OR=0.74, 95% CI=0.57-0.96, p=0.025]. Infections affecting the kidney and urinary tract, oral tract and viral CNS infections, were most significantly associated with protective effects. Our study provides support for the hygiene hypothesis in CD whereby exposure to infections in early childhood could potentially reduce risks for CD.

Key words: Crohn's disease, hygiene hypothesis, infections, case-control

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CD: Crohn's disease

CI: confidence interval

CNS: central nervous system

GI: gastrointestinal

HL: Hosmer-Lemeshow

IBD: inflammatory bowel disease

IC: indeterminate colitis

ICD: International Classification of Disease

LR: likelihood ratio

MH: Mantel-Haenszel

ML: maximum likelihood

OR: odds ratio

RAMQ: Régie de l'Assurance maladie du Québec

TNF: tumour necrosis factor

UC: Ulcerative Colitis

UK: United Kingdom

Acknowledgements

I would like to thank my thesis supervisors, Dr Devendra Amre and Dr Paul Brassard, for their exceptional guidance. I would also like to extend my gratitude to the CHU Ste-Justine Research Centre and the Ste-Justine Foundation, as well as the Department of Social and Preventive Medicine of the University of Montreal, for their generous financial support. Finally, I would like to thank Paul Col, for his much appreciated help with the SPSS program.

Introduction

The "epidemiologic transition" of the 19th century resulted in an unprecedented decrease of infectious disease incidence [1]. The transition originated from the revolutionary discovery of the role of microorganisms in disease, which led to major improvements in hygienic measures and ultimately, to the development of antibiotics. This scientific evolution changed the face of disease; significantly less afflicted by communicable diseases, life expectancy in developed countries progressed, and thus began the shift from infectious disease predominance to chronic illness.

The epidemiologic transition led to a wide variety of changes in population health. Whereas the decrease in infectious disease incidence is perhaps the most obvious, several more subtle changes have occurred, some for which we are only starting to grasp the impact. The change in the numbers and types of microorganisms in our environment influence the nature of diseases in ways that we do not yet fully understand. For example, the decrease of exposure not only to known pathogens, but to microorganisms endemic to the human body for the past thousands of years (referred to as "heirloom organisms") [1] have altered the relationship between the human gut and its microbiota, resulting in changes for which we can only suspect, but not yet prove, the causal relationship.

One such suspected effect of the improvement of hygienic conditions is the potential association between the change in exposure to microorganisms and the increase in the incidence of Crohn's disease (CD), in adults and children alike. As the rise in Crohn's disease is a recent phenomenon restricted to the most hygienic regions of the world, one cannot help but wonder whether the two events are causally linked.

In the following study, we evaluate the relationship between childhood exposure to antigens and the risk of auto-inflammatory conditions, more specifically Crohn's disease. There is currently no cure for this disease, as its aetiology remains unknown. A popular theory pertaining to risk factors for Crohn's disease is the Hygiene Hypothesis, first proposed by Strachan in the 1980s. It has since then been investigated in the context of multiple autoimmune diseases, most notably for asthma. Strachan's theory implies that the absence of prompting of the immunological system in childhood by a variety of environmental antigens precludes immunological tolerance, leading to an overreaction of the immune system upon subsequent exposure. In accordance with the hygiene hypothesis, we have used administrative data to study the relationship between infectious diseases in childhood and the onset in paediatric Crohn's disease.

Studying the causes of Crohn's disease is of utmost importance, as without knowledge of the mechanism by which disease onset arises, we cannot develop a cure. The incidence of Crohn's disease in Canada is amongst the highest in the world and in Quebec, the incidence is the highest in the country[2, 3]. Crohn's disease is thus a prevalent public health issue in our province, as the need

for lifelong follow-up of patients, as well as medical therapy and surgical intervention, constitute a significant burden on our healthcare system.

Literature Review

Crohn's Disease: an overview

Inflammatory Bowel diseases (IBD)

Inflammatory bowel disease (IBD), a chronic inflammation of the digestive system, comprises of Ulcerative Colitis (UC) and Crohn's disease (CD) [4, 5]. IBD is characterized by symptomatic flare-ups alternating with periods of disease inactivity [6]. Inflammation occurs along the bowel wall in both UC and CD, although the section of the digestive tract primarily affected differs between the two diseases [6]. Whereas the inflammation in UC is typically restricted to the mucosal layer of the lining of the large intestine and the rectum [6], the inflammation in CD is more expansive: it occurs across the gut wall anywhere along the digestive tract (mouth to anus), though more typically along the distal ileum and the colon[6, 7]. The distinction between the two types of IBD is not always obvious, as some patients may present with characteristics of both UC and CD. These ambiguous clinical presentations are referred to IC (indeterminate colitis), and make up 10-15% of cases[5].

Symptoms of CD

CD symptoms alternate between periods of activity and periods of remission. Symptoms typically comprise of "abdominal pain, fever, and clinical signs of bowel obstruction or diarrhoea with passage of blood or mucus, or both" [8], and often compromise the diversity and distribution of the gut microbiota in CD patients [6, 8].

The symptoms are caused by chronic inflammation due to an abnormally permeable gut wall [8]. This allows antigens to cross the epithelial cell lining of the GI (gastrointestinal) tract and to trigger a reaction from the underlying immune system. The permeability of the gut wall is attributable to faulty tight junctions (proteins whose role is to render the space between the epithelial cells of the GI tract impenetrable) [8]. Recurring inflammation of the GI tract causes lesions along its walls, which in turn cause the symptoms associated with CD[7]. Inflammation may also cause more serious damage, such as a narrowing of the GI tract (stricture), swelling of the gut wall (abscesses) or abnormal passages between different regions of the GI tract (fistulas) [7]. There is no evidence that the severity of symptoms are correlated to the degree of injury to the GI tract wall [7].

Diagnosis of CD

Due to the complex nature of the disease, CD cannot be diagnosed by a single test [4]; rather, multiple tests are used in combination with each other [4, 9]. A medical history and physical exam will provide descriptive characteristics of symptoms and their time frame[8]. Blood tests are used to detect inflammation, for example by screening for the inflammation marker C-reactive protein [8, 9]. Biopsies of the GI tract can identify lesions and other pathologies of the intestinal wall [8]. Endoscopy, more specifically ileocolonoscopy, is a more invasive technique commonly used in CD diagnosis. It allows the visualization of lesions, abscesses and fistulas in the GI tract. Endoscopy combined with biopsies is the current gold standard of CD diagnosis [8, 9]. However, the use of imaging techniques both as a diagnostic tool and as a tracking tool for disease progression is becoming increasingly more popular[10]. Imaging techniques, including computed tomography enterography and magnetic resonance enterography, allow specialists to view a cross-sectional image of the bowel, not restricted to the superficial mucosal layer of the gut, as is the case with more traditional endoscopy [10].

Once the CD diagnosis has been established, further classification of the disease according to the Rome, Vienna or Montreal classification systems provides further insight into the determination of the most appropriate course of therapy. These classification systems are based on the anatomical location and the type and severity of intestinal damage, as well as demographic characteristics of the patient [8, 11].

Treatment of CD

There is no cure for CD; treatment objectives comprise of slowing the course of disease and treating the symptoms by repairing damage caused to the GI tract wall (strategy referred to as "mucosal healing") [8, 12]. As disease progression differs from one CD patient to another, careful monitoring of the disease phenotype is crucial in maximizing the benefits of therapy [7, 8, 11]. The disease phenotype, as well as patient characteristics such as age at CD onset, the location and the behaviour of the disease, and medical history, can be used to predict prognosis and to tailor the treatment to the patient. The most recent classification system, the Paris system, was developed to improve treatment of paediatric IBD by including additional patient characteristics when classifying disease phenotype [11].

To control inflammation, medical therapy is used: steroids or anti-TNF (tumour necrosis factor) agents, either as monotherapy or in combination with each other [8]. Fast-active drugs (ex. steroids) are often combined with slower-acting drugs (ex. immunotherapy) [8, 9]. The choice of the type of

medical therapy depends on the nature and the severity of the symptoms, concomitant illnesses and personal factors [8].

When medical therapy fails and damage to the intestinal wall prevails, surgical resection must be considered [7]. The majority of CD patients (70-80%) require surgery within 20 years of diagnosis [7], and the majority of patients undergoing surgery will experience recurrence of the disease [12]. The proportion of patients whose endoscopy results remain normal 10 years post-surgery is less than 5% [7]. The risk of recurrence depends on tobacco use, the extent of the damage to the GI tract wall and surgical history [8].

Biologic therapies are recent developments in CD management. They consist of drugs which target specific components of the inflammation process, including drugs that bind TNF-alpha [9].

Paediatric Crohn's disease

Frequency distribution

The distribution of age at disease onset is not uniform within the CD population; it is bimodal. The first peak occurs in the early twenties and decreases afterwards; a second peak occurs between the ages of 50 and 70. Cases diagnosed before adulthood (<20 years) represent an estimated 25% of all cases [13-17].

Patient and disease characteristics differ between paediatric and adult-onset CD cases. For example, males are more often diagnosed with childhood CD (pre-puberty), whereas the majority of diagnosed adult CD cases are women (though it has been shown that the incidence of CD in males is increasing and eventually might equal that of women) [7, 16]. Additionally, paediatric CD poses challenges unique to this sub-group of the disease population, as it affects patients during a developmental period of their lives [18].

Challenges

Growth Failure: As reported in a 2012 meta-analysis, multiple studies have demonstrated that CD leads to growth failure in children[14]. Growth failure rates, most commonly defined as height below the third percentile, were recorded and ranged from 10-56% at the time of CD diagnosis [14]. This is due to impaired nutritional intake caused by the effects of CD on the GI tract. Paediatric patients are thus unable to meet their caloric needs, impairing growth [18]. Nutritional deficiencies, notably

insufficient Vitamin D levels, lead to bone demineralization [13]. Furthermore, the immunological imbalance underlying CD could perturb the normal release of growth hormones [13].

Quality of Life*:* The chronic nature of the disease, characterized by recurrent bouts of symptoms, gravely affects the quality of life of paediatric cases. The reduction in quality of life manifests itself through "family conflict, trouble socializing with peers, medical adherence problems, and absences from school and extracurricular activities" as well as depression and anxiety [18]. Psychiatric disorders are significantly more likely to arise in paediatric and adolescent CD patients than in the overall population in this age group [13].

CD Risk Factors

The current consensus indicates that the development of CD in an individual is most likely the result of a gene-environment interaction: an environmental triggering factor is thought to induce disease progression in those who are genetically susceptible [19].

Genetic Predisposition

Family history was identified 70 years ago as a risk factor for CD, and since then multiple studies have provided supporting evidence for this risk factor [8]. For example, in a 2005 matched case-control study investigating potential risk factors for IBD, family history was found to be the greatest predictor of disease (OR: 4.6 [95% CI 2.6-8.3]) [20]. Furthermore, twin studies show that monozygotic twins display a concordance rate of 30% to 58%, depending on the study [8, 21-23], which is much higher than the concordance rate of dizygotic twins, estimated around 3% [8, 24].

Ethnic factors have been associated with a higher risk of CD: the prevalence is notably higher in the Caucasian population, as well as in the Ashkenazi Jewish population [6, 17, 22]. In a 2007 Manitoban population-based cohort, comprising of 232 CD cases, Jewish ethnicity was amongst the most significant predictors of CD (OR: 18.5, p=0.008)[24]. This concurs with the results of a Quebec population-based case-control study, in which Jewish ethnicity was more prevalent amongst the cases than the controls (adjusted incidence rate ratio for >20% Jewish descent: 1.70, 95% CI: [1.30-2.21]) [2]. Finally, the cases in a population-based case-control study conducted in New Zealand were significantly more likely than the controls to be of Caucasian ethnicity (adjusted OR: 2.14 95%CI: [1.05-4.38][25].

Familial and ethnic risk factors for CD indicate the presence of a genetic predisposition to the disease, as gene alleles are differentially distributed across populations [22]. A large number of genetic alleles have been associated with IBD, amongst which is the CARD15 polymorphic gene, involved in microbe recognition [22]. A meta-analysis performed by Economou et al in 2004 identified 3 variants of this gene. Each variant was associated with different disease phenotypes regarding the location of inflammation along the GI tract, the severity of the disease, and its clinical manifestation [24, 26]. A gene-dosage effect between CARD15 mutations and CD risk has been observed: individuals homozygous for the mutant allele have greater risks of developing the disease than heterozygous individuals, who in turn are more susceptible to CD than those without the mutant allele [24]. In a case-control study by Brant et al, the CD population attributable risk for CARD15 was estimated at 46.8% [24].

Other genes, involved in immunoregulation, have also been associated with CD: mutations of the ATG16L1 and IRGM genes, part of the pathogen-degradation process, cause disruption in the autophagy pathway [23]. The NOD2 gene plays a role in peptidoglycan recognition (particles found on invading bacteria) and some of its polymorphisms greatly increase the risk of CD [22]. In total, over 160 loci have been associated with IBD (140 of these with CD) [23, 27]. However, the presence of loci explains less than one third of cases of CD, supporting the gene-environment interaction theory [23].

Environmental factors

Of all studies conducted on CD risk factors, the association between smoking and IBD has gathered the most compelling evidence [28]. Curiously, smoking has a protective effect on UC, but is a risk factor for CD [23]. The odds ratio measuring the association between smoking and CD lies between 1.5 and 2.0, as reported in a 2013 literature review by Ng and al [29]. Furthermore, tobacco use increases the severity of CD symptoms, as well as the need for surgery, the recurrence of CD after surgery, and it accelerates the onset of the disease. However the mechanism behind this association is unknown [21, 23, 28]. The effects of tobacco have been demonstrated for adult-onset CD only; second-hand smoking amongst children has not been shown to have the same effect on the course of disease [21].

A similar differential association exists between appendectomies and the risk of UC and CD. Appendectomies significantly reduce the risk of UC, yet most studies show that this procedure constitutes a risk factor for CD [23, 29]. The evidence supporting this relationship is substantially more mitigated than the evidence supporting tobacco-use as a risk factor: some studies have

demonstrated a protective effect of appendectomies on CD incidence, and still others have shown a null association [29]. There is a possibility that appendectomies are a time-dependant risk factor; the appendectomy-CD association is strongest within one year following the procedure, and decreases over time [29]. After 5 years, the risk of CD is no longer significantly higher for those having undergone an appendectomy [21, 29]. It is also hypothesized that the effect observed is due to a misclassification of disease: early-stage CD could be misdiagnosed as appendicitis [21, 29]

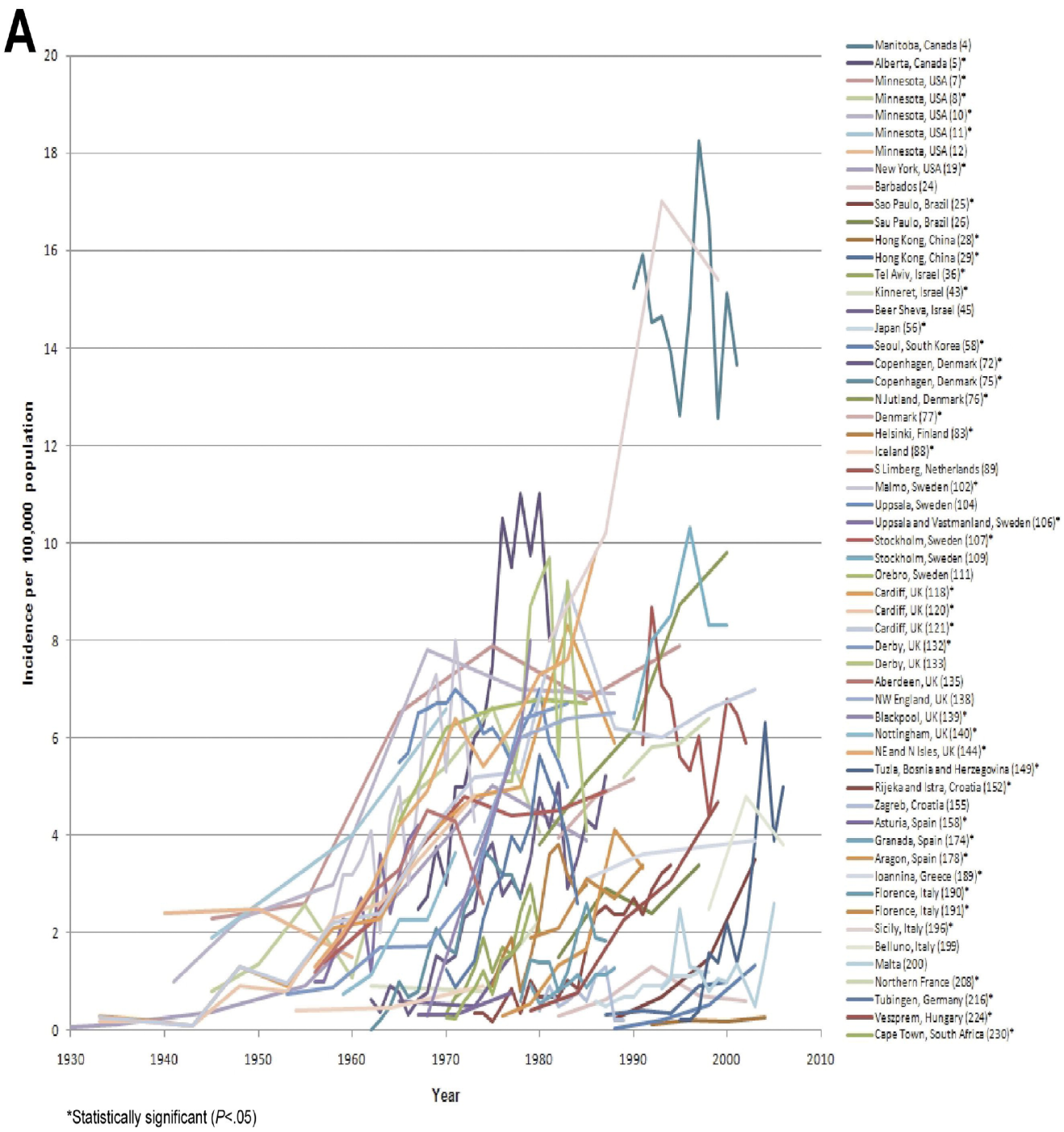
Oral contraceptive are known environmental risk factors for CD [23]. A meta-analysis published in 1995 revealed an adjusted OR of 1.4 for women taking oral contraceptives [29]. This effect seems to increase with the number of years for which oral contraceptives have been taken [29]. Other suspected CD risk factors include oral contraceptives, stress, socioeconomic status, diet and use of antibiotics, whereas breastfeeding and sunlight (vitamin D) are thought to be protective [21, 23, 29, 30]. However, research findings for these factors are thus far inconsistent, and their association with CD has not been clearly ascertained.

Epidemiology of CD

Temporal trends

Over the course of the past century, a significant rise in the global incidence of CD has been reported [3]. The following graph, taken from a 2012 systematic review by Molodecky et al, charts the incidence of CD (per 100,000 individuals), by region, since 1930 [31]. The data used to produce this chart originates from various IBD epidemiological studies, conducted in different regions of the world at different time periods. Inclusion criteria for the studies comprised of a minimum of 10 years of data collection per study, and a minimum of 3 incidence rate time points.

Figure 1: Temporal Trends in the Adult CD Population: Incidence per 100,000 Population, for Different Regions

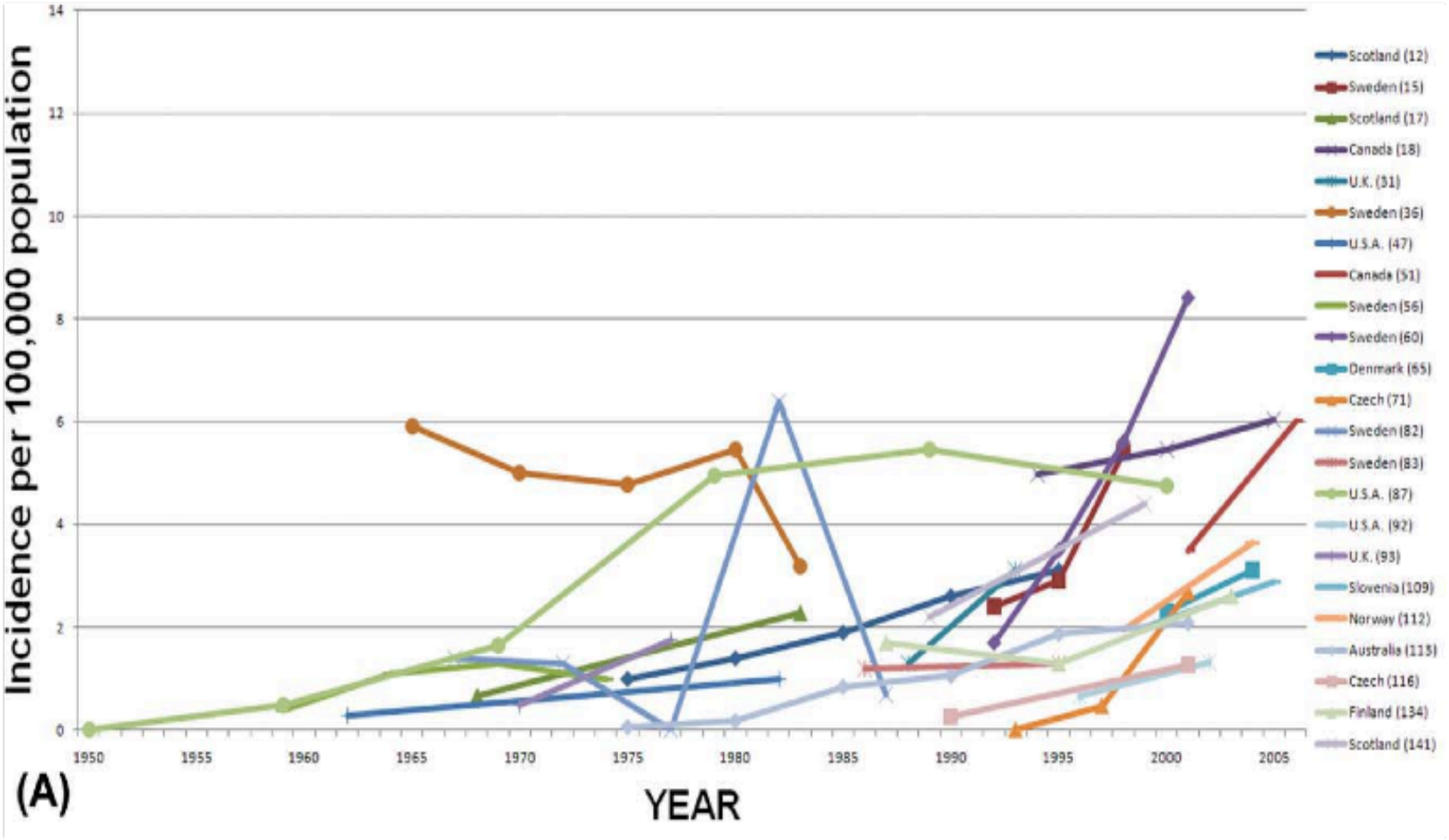


As shown in the legend above (* indicates a p-value of less than 0.05), the majority (75%) of the regional trends show a statistically significant increase in CD incidence over time. The range of increase in CD incidence reported in this meta-analysis is 2.4% to 18.1% [31]. Whereas this trend

could be attributed to the improvement of diagnostic techniques over time (surveillance artefact), longitudinal studies have shown that this is not likely [2].

The prevalence of paediatric CD has also increased over the past 40 years [13]. In a 2011 systematic review of trends in international incidence rates of paediatrics CD, 60% of the studies surveyed reported a significant increase in disease incidence [16]. The following graph depicts the results of this systematic review:

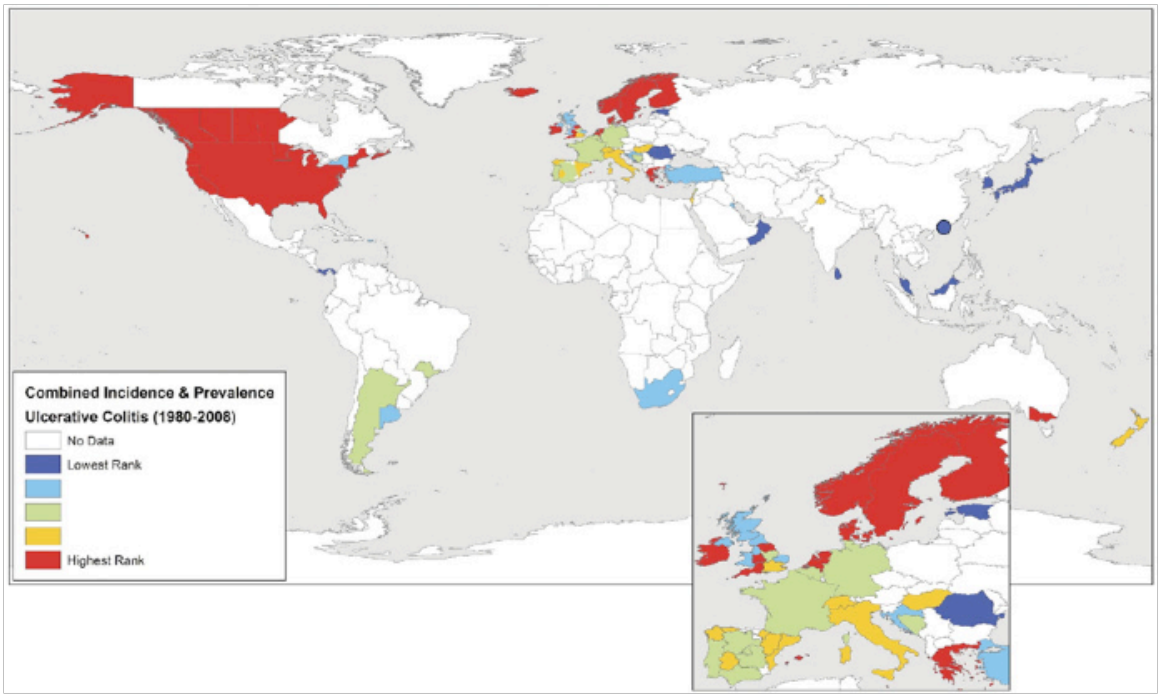
Figure 2: Temporal Trends in the Paediatric CD Population: Incidence per 100,000 Population, for Different Regions



Geographical trends

Below is a world map, taken from the 2012 Molodecky systematic review mentioned above, of worldwide IBD prevalence and/or incidence rates since 1980, by region [31].

Figure 3: Geographical Trends in the Adult CD Population

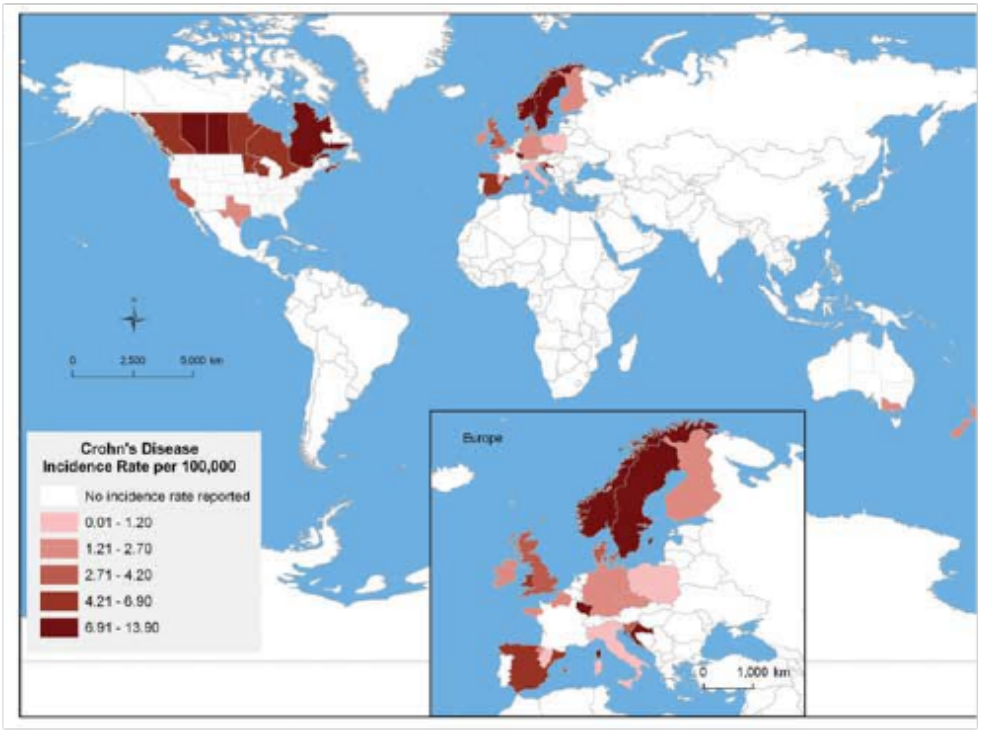


The map demonstrates the wide variation in disease incidence and prevalence between different regions of the world [31]. High incidence rates are concentrated in developed countries, notably, North America, Europe and Australia. Asian, African and South American rates are much lower. These findings concur with those of a previous literature review, conducted by Economou et al in 2008 [3].

It is important to note that most IBD epidemiology studies were conducted in Northern Europe, the United Kingdom and North America [29, 31]. The lack of data collected from other countries could partially explain the discrepancy in incidence rates.

The geographical trends observed in adult CD are mirrored in the paediatric CD population, as demonstrated by this map taken from the Benchimol study (showing CD incidence rates since 1990):

Figure 4: Geographical Trends in the Paediatric CD Population



In Europe, where a significant amount of data has been collected on incident CD cases, a North-South gradient has been observed: incidence of CD is greater in countries North of the Alps (7/100 000) than South (3.9/100 000) [3, 32]. This trend has also been reported within individual countries: regions in northern France and Scotland display higher incidence and prevalence rates than their southern counterparts [29]. In Canada, a West-to-East gradient has been observed (lower rates in British Columbia, with the highest rates in Québec and Nova Scotia) [2]. The cause of these gradients is unknown; hypotheses include differences between rural and urban regions and asymmetrical immigration distribution[2, 31].

The burden of disease in Asia, Southern and Eastern Europe and developing countries remains low, however, it is steadily increasing. This is especially apparent in Asia [29]. In Japan, CD prevalence has increased from 2.9/100 000 in 1986 to 13.5/100 000 in 1998; in South Korea, it has increased from 7.6/100 000 in 1997 to 30.9/100 000 in 2005 [29]. The upward patterns of disease prevalence in these regions mirrors the phenomenon which occurred in the developed countries almost 100 years ago: the UC incidence was the first to increase, followed by the CD incidence [29]. Though the presence of a genetic component of the disease is well established, genetic predisposition alone cannot explain the rapidity of disease expansion. Hence, an underlying environmental triggering factor, linked to industrialization, is most likely at play. Furthermore, children of immigrants from

developing to developed regions display a greater risk of CD than their parents [16, 29], providing additional evidence for an environmental trigger.

Hygiene Hypothesis

Origins

The origins of the hygiene hypothesis date back to 1989, from a study on hay fever risk factors conducted by David Strachan [33]. Strachan investigated the increase in prevalence of hay fever (a type of allergic reaction) in Britain following the industrial revolution [34]. A retrospective cohort study of 17 414 British children was performed, the outcome consisting of self-reported hay fever. Of the multiple exposures studied, "family size and position in the household in childhood" were the most significantly associated with hay fever [34]. To explain his findings, Strachan proposed the following hypothesis: "They could […] be explained if allergic diseases were prevented by infection in early childhood, transmitted by unhygienic contact with older siblings, or acquired prenatally from a mother infected by contact with her older children. Later infection or reinfection by younger siblings might confer additional protection against hay fever." [34] Though the term "hygiene hypothesis" was not used until many years later, Stratchan's theory of post-industrialization hygienic environments as causes of autoimmune disorders has been retained and investigated in multiple diseases.

Biological Mechanism

The proposed mechanism underlying the hygiene hypothesis consists of the role of exposure to infection-causing microorganisms in early childhood as inducers of immunological tolerance, consequently preventing the onset of autoimmune diseases [1, 35]. In the pre-industrialization era, individuals were exposed to a greater range of microbial agents in the environment. Furthermore, a wider variety of microbes made up the body's microbiota since birth, and therefore had to be tolerated. Failure to be exposed to these organisms early on in life due to improved hygienic conditions would preclude immunological tolerance and lead to abnormal immune responses upon subsequent exposure to these antigens [1, 36].

Proxy measures of hygiene

As hygiene is a difficult concept to define, multiple proxy measures have been used in order to quantify exposure to antigens for study purposes:

- Childhood infections (helmintic, *H Pylori,* enteric, bacterial) [21, 32, 36, 37]
- Presence of household pets [21, 38]
- Household size and number of siblings [32, 34, 36]
- Use of antibiotics [32, 39]
- Breastfeeding [32, 36, 39]
- Urban versus rural environment during childhood [32, 39]
- Dental procedures [32]
- Childhood allergies [32]
- Vaccinations [32, 36]
- Access to hot water [40]
- Number of toilets in the household [40]
- Drinking unpasteurized milk [41]
- Attending a day care [42]
- Crowding index of house (number of rooms per inhabitant) [40]

Supporting Evidence for the Hygiene Hypothesis

Since the publication of Strachan's theory, the putative role of the hygiene hypothesis in the aetiology of several autoimmune diseases has been thoroughly investigated [42]. Asthma has been the most extensively studied of these diseases; its prevalence has risen an astounding 75% (in the United States) from 1980 to 1994 [41]. Many observational studies performed in farming communities provide support to the role of hygiene hypothesis in asthma onset; it was demonstrated that children directly exposed to livestock and who drink unpasteurized milk are less likely to develop asthma than those who are not [38]. In urban settings, pet exposure seems to confer protection against paediatric asthma [38]. Furthermore, a meta-analysis by Murk et al (2011) demonstrated that early exposure (during the first year of life) to antibiotics increases the risk of developing asthma (OR: 1.52 CI: 1.30-1.77) [43]. This suggests that early exposure to microbial antigens could play a role in preventing asthma onset.

A steep increase in food allergy rates has followed the increase of asthma prevalence in the developed world. Peanut allergies, notably, are said to currently affect 1-2% of all infants and young children in Canada, the United States, the UK and Australia [44]. Additionally, we have seen a great increase in the prevalence of allergic rhinitis and atopic dermatitis [42]. Multiple studies have shown that rural environments are protective against allergies: children living on farms, especially where livestock is raised, display much lower allergy incidence rates [38]. It has also been shown that attending a day care in the first 6 months of life constitutes a protective factor against atopic

dermatitis [42]. These observational findings in human populations mirror experimental research findings done on animal models, which have reported that early microbial exposure in mice is protective against allergic symptoms [43].

Similar studies have been performed in the context of other auto-inflammatory diseases, investigating the role of early microbial exposure in disease onset. Notably, the evidence supports a possible association between type I diabetes, mostly diagnosed in childhood, and multiple sclerosis [1].

Inflammatory Bowel Disease and the Hygiene Hypothesis

Multiple studies have been performed investigating the hygiene hypothesis in IBD, mainly in the form of case-control studies. In a 2006 Manitoban case-control study, 364 subjects with CD were administered a detailed questionnaire pertaining to past exposures to a variety of hygiene proxy measures, with the objective of identifying risk factors of the disease [28]. The results showed that significantly less CD cases were born outside of Canada than the controls, and were less likely to have siblings. Additionally, the cases were significantly more likely to have been raised in a household with fewer other residents than the controls, and to have a higher birth order (be an older sibling) [28]. CD patients had smaller odds of having been raised on a farm, or to have had a pet in the childhood home [28]. Finally, CD cases were less likely to have drank unpasteurized milk as children [28]. The results of this study support the hygiene hypothesis, as the proxy markers of hygiene were more predominant amongst the cases than the controls.

In parallel, a paediatric case-control study was performed by Amre et al and published in the same issue of the American Journal of Gastroenterology as the Manitoban study. Similarly, a questionnaire was issued to the 287 cases (or to their mothers, for most of these paediatric participants). Opposite results were reported: early childhood infections seemed to increase the risk of developing paediatric CD [40]. Furthermore, the results revealed that the cases in this population were more likely to live with a pet [40]. This second study, unlike the first one, was performed in a hospital setting, and included paediatric CD patients only. The difference in study methods could account for some of the divergence in results [19].

Ambiguity also exists with regards to vaccines as a potential risk factor for CD. In concordance with the hygiene hypothesis, the potential correlation between vaccinations against childhood infectious diseases and the incidence of IBD has been investigated. Notably, a possible association with the measles vaccine has received considerable attention. Whereas a study in 1995 reported

that vaccinated individuals were 3 times more likely to develop IBD, several other studies performed since have found no association between the vaccine and the disease[45].

Conflicting evidence contributes to our ignorance of CD triggering factors. Whereas the hygiene hypothesis is supported by some of the published evidence, other study results point to a different cause of disease. Many believe childhood infections could cause the disease, rather than prevent it [40]. Notably, the *Mycobacterium avium paratuberculosis* bacterium and the measles virus are thought to be potential infection-related triggering factors for CD [28, 36]. Though it has been established that micro-organisms play a role in CD onset, the mechanism through which they act is not confirmed [46]. There are multiple possibilities: the microflora could have an effect on gene expression, invading micro-organisms could elicit abnormal responses from the immune system and cause chronic inflammation, etc [46].

Limitations of published studies investigating the hygiene hypothesis also contribute to the heterogeneity of the results obtained. Firstly, most of the study data collected has been retrospective [36, 39]. Since the onset of disease can occur at any age whereas early childhood exposures are studied (breastfeeding, early childhood infections, early antibiotic use, etc), the critical exposure period is often many years behind. Researchers must therefore rely on event recollection by the participants. The information collected in this manner is very susceptible to recall bias [36, 39]. The need for prospectively collected data has been expressed in many research articles[39, 40]. Additionally, the causes of disease could differ depending on the age of onset. A bimodal age distribution of cases may mean that the critical exposure period for paediatric CD is different from that of adult-onset CD [39]. Lastly, the causes of CD are complex, and probably comprise of interactions between multiple factors [40]. It is thus difficult to isolate the effects of different exposures. One of the advantages of studying the disease in a paediatric population is the absence of many known or suspected confounders within the study population (such as tobacco use and oral contraceptives), which allow us to narrow the number of exposures studied [18].

CD: An important Public Health concern

Incidence in Quebec, Canada and elsewhere

As demonstrated in the *Temporal trends* section, the incidence of CD worldwide is rapidly increasing, making it an "emerging […] global disease." [31] Furthermore, at the moment, Canada maintains one of the highest incidence rates in the world and Quebec, the highest incidence rate in the country. A study conducted in Quebec, using data from the universal health insurance

database, the RAMQ (*Régie de l'Assurance-maladie du Québec),* has estimated that the incidence in this province specifically is 20.2 cases/100 000 person-years. This ranks Québec amongst the top 2 greatest incidence rates in Canada, on par with Nova Scotia [2, 3]. Increasing our knowledge of the disease could lead to more effective treatment, thus reducing the burden of this chronic disease on our healthcare system, while improving the quality of life of those affected. This is especially important in the paediatric population, in order to allow for normal growth and psychological development.

Objectives

Primary Objective

The primary objective of this study was to determine the magnitude and the direction of the potential association between the exposure to childhood infections preceding the CD diagnosis, a proxy measure for hygienic conditions, and the onset of paediatric CD.

Secondary Objectives

- Descriptive analysis of the characteristics (gender, age at diagnosis, urban or rural home environment) of children diagnosed with CD at the Sainte-Justine Hospital.
- Analysis of the frequency (total number) of childhood infections and its possible association with CD
- Analysis of the temporality (during the first year of life, the first five years of life or anytime before CD diagnosis) of childhood infections and its possible association with CD
- Analysis of the type (enteric, respiratory, etc) of childhood infection, and its possible association with the onset of paediatric CD.

Hypothesis

In concordance with the Hygiene Hypothesis, we believe that childhood infections will have a protective effect on the onset of paediatric CD. The scientific literature does not indicate whether a greater number of infectious exposures is expected to be correspondingly more protective. This is why 2 types of analyses will be performed: firstly a comparison of the absence vs presence of childhood infections and their association with paediatric CD, followed by a frequency analysis to investigate whether a greater number of infections is correspondingly more protective.

The results of multiple studies suggest that early microbial contact plays a crucial role in the development of gastrointestinal diseases, notably during the perinatal period [47]. This is due to a period of immunological tolerance building, where the immune system learns to distinguish between self and non-self antigens, and between harmful and benign antigens [35]. We believe that infancy (the first year of life, or the first five years of life) is likely to be influential in determining CD outcome, which is why we will be performing sub-analyses on these specific time periods. Finally, we hypothesize that certain types of infections, notably enteric infections, will have a stronger association with the risk of paediatric CD. The literature suggests that microbes in the gut, specifically, could play a determinant role in CD aetiology [46]. A sub-analysis for different infection types will be performed.

Methods

Study design

A case-control study was conducted, which is a quantitative, quasi-experimental study design [48]. The independent variable of interest was childhood infections and the outcome (dependant variable), paediatric CD.

Study population

Cases

Paediatric CD cases were recruited from the gastroenterology clinic of the Ste-Justine hospital, a paediatric tertiary care centre in Montreal. Included cases were consecutively diagnosed with CD between 1988 and 2005. The diagnosis was confirmed according to standard diagnostic procedures, including clinical data, endoscopy, radiology and histopathology [8]. Patients were 20 years of age or younger at the time of recruitment. Only cases insured by the RAMQ since birth were included. To increase specificity of diagnosis, only cases with a confirmed diagnosis after at least 1 year of follow-up were included. Cases with a diagnosis of UC or IC were excluded.

In the IBD literature, the term "paediatric" has many different definitions; upper age limits of 18, 20 and 21 years have all been previously used to designate paediatric-onset disease [13, 14, 16, 18]. In this study, the inclusion age was based on the age of the patients attending the gastroenterology clinic (0-20 years) of the study hospital.

Controls

The hospital-based cases were matched to population-based controls. This was done through the RAMQ database. The insurance provider was responsible for matching each case to up to 4 controls, to maximize the power of the study and to control for known confounders identified through a literature review. The matching variables comprised of the following:

- **Birth date**: Patients were matched by date of birth to control for the confounding effect of age on CD and on infection exposure.
- **Gender**: As mentioned in the literature review, proportionally more males are diagnosed with paediatric CD than females. Matching was done to control for this potential confounder.

- **Postal code**: Cases and controls were matched on the first three digits (letter, number, letter) of their postal codes, at the time of CD diagnosis, to partially offset differences in environmental exposures attributable to region of residence.
- **RAMQ coverage period**: as mentioned above, cases and controls were matched on the coverage period of the provincial health insurer (participants had to be covered by the RAMQ since birth, to ensure complete exposure information).

Controls were excluded if they had a CD, UC or IC diagnosis.

The sampling method described above is a non-probabilistic, purposeful method [48]. Though the cases are hospital-based and the controls, population-based, they are comparable: if controls were to have developed CD, they would be expected to have been referred to the same tertiary healthcare centre as the cases and thus, have been labelled as a case in the study.

Ascertainment of Exposure

Exposure ascertainment was achieved by utilizing information stored in the RAMQ database to compile childhood infection information.

The RAMQ Database

The RAMQ, the *Régie de l'assurance maladie du Québec,* is the provincial health insurer of Québec. Its mandate comprises of four components: "it informs the public, manages the eligibility of persons, remunerates health professionals and ensures the secure flow of information," reporting directly to the Ministry of Health of Québec . [49]" As Canada boasts a universal health system, the RAMQ insures the quasi-totality of the Québec population: 7.6 million individuals [49] out of a total population of 7.98 million [50]. The RAMQ manages the physician claims database for the province. This database, though administrative, contains a wealth of demographic and medical information for each individual it insures. These advantages make the RAMQ database an attractive choice of source data for epidemiological studies [51].

As part of the billing process and in order to be remunerated by the RAMQ for their services, medical doctors and other health practitioners must submit the following information for each patient visit to a clinic [52]:

- Date of medical visit
- Full name of patient and unique RAMQ identifier
- Full name of health professional and unique RAMQ identifier

- Unique identifier of the general practitioner, in the case of a referral to a specialist
- Address, birth date and gender of the patient
- Diagnostic codes: represent the main purpose of the medical visit, using the ICD (International Classification of Disease) coding system.

ICD codes

The *International Classification of Disease (ICD)* codes were developed by the World Health Organization in 1979, with the objective of standardizing the collection of disease information across different healthcare systems [53]. This standard method of disease classification is now used by all insurance companies, including the RAMQ [53].

The ICD coding system consists of unique numeric codes for "diseases, conditions and injuries [53]." Though the 10th edition of the codes is now the norm, the ICD-9 codes (9th edition) were exclusively in use during the study period (1988-2005). The ICD-9 codes reported by physicians and recorded in the RAMQ database were obtained for this study in order to ascertain exposure status.

The information requested from the RAMQ by our group for the present study was the date of the matching between cases and controls, the date of birth of each participant as well as their gender, the first 3 digits of their postal code, the date of each medical visit since birth and its corresponding ICD-9 codes. A dataset was subsequently assembled by the RAMQ containing the requested information.

Exposure classification using ICD codes

ICD codes are classified by broad categories. Though the coding system includes an infection category (codes 001 to 139), several other infections are classified elsewhere, according to the organ system they affect. Pinner et al, in a 1996 study on the trends in infectious disease mortality in the US, re-classified all ICD-9 codes as infectious diseases, consequences of infections, or non-infectious disease [54]. Simonsen et al further refined this categorization in a 1998 study on the trends of hospitalizations due to infection diseases, separating the infections by type [55]. These categories can be found in appendix I. Simonsen's classification system was used in the present study to determine whether or not a diagnostic of infectious disease was made for each medical visit and if so, which type of infection it was.

Only infections preceding the CD diagnosis were included in the study. This was done to ensure that the exposure preceded the outcome, as stated in the Hill criteria for causality [56]. In addition, to be certain that the exposure truly preceded the onset of disease, any infection occurring in the 2-year period prior to CD diagnosis were excluded. This was done because it is not uncommon for disease diagnosis to lag behind appearance of the symptoms by several months[57]. The CD diagnosis date was referred to as the index date, and the case index date was assigned to each of its 4 controls.

Compilation of infections for the frequency, temporality and type sub-analyses was conducted as follows:

Frequency: The frequency of infections was compiled by adding together the number of visits for which an infection-related diagnosis had been emitted. As infection diagnoses within a short period of time are likely caused by the same antigen exposure, multiple infections occurring within a time span of 15 days were counted as a single episode. A sensitivity analysis was performed to assess the 15-day episode definition, by repeating all analyses using a 30-day time frame as one episode.

Temporality: As stated in the hypothesis, the first years of life are thought to comprise the critical microbial exposure period, as it is during this time that the immune system builds up tolerance. As the window of time for immunological tolerance is not clearly defined in the literature, three different time spans were analyzed: from birth to the first year of life, from birth to 5 years, and from birth to CD diagnosis. The frequency of infections during these 3 different time periods was analyzed separately.

Type: The type of infection was determined by the ICD-9 codes, according to the Simonsen classification of disease (Appendix 1). Infections affecting 5 cases or controls or less were excluded from the analysis.

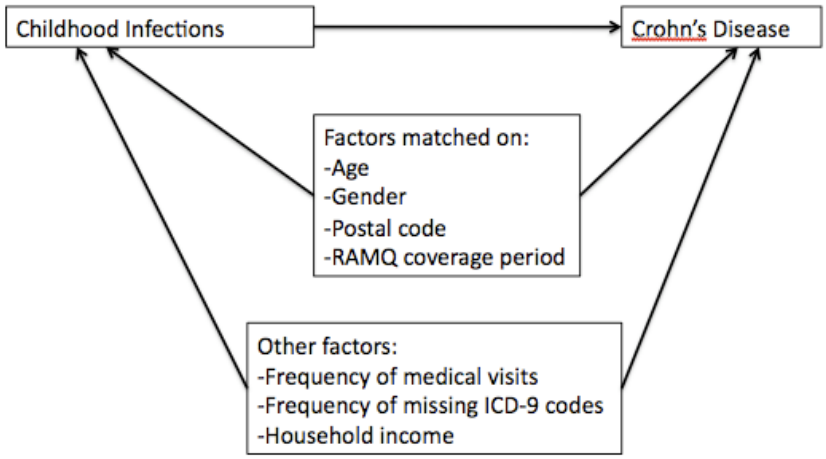
Potential Confounders

As stated in the *Participant Selection* section, the cases and the controls were matched on the following potential confounders: age, gender, and geographical area of residence, as obtained from the first 3 digits of the participants' postal code. The period of coverage by the RAMQ insurer was also considered a potential confounder, as it could influence the number of infections recorded in the database. RAMQ coverage since birth was an inclusion criterion for the study.

Additionally, household income was assessed for potential confounding. Studies have provided support for high socioeconomic status as a risk factor for CD, which is congruent with the observation that disease prevalence is higher in developed countries [29, 30]. To include socioeconomic status in the analysis, family income information was obtained from the Statistics Canada 2001 Census. The average income for specific geographical areas was matched to subjects on the basis of their postal code.

The number of medical visits was considered a possible confounding factor, the rationale being that children who visit a physician more frequently are more likely to be diagnosed with a greater number of infections. Finally, the number of missing ICD-9 codes (physicians omitting the diagnostic code on claim forms) was also assessed for confounding, as missing codes could mask a larger number of infections, or even a CD diagnosis.

Figure 5: Conceptual map of potential confounders



Ethical considerations

In order to obtain confidential information related to health insurance, ethics approval was requested and obtained from the ethics review board of the *Centre hospitalier universitaire Sainte-Justine,* where the study was conducted, as well as from the *Commission d'accès à l'information* (CAI), which monitors requests for access to information for the RAMQ.

Statistical Analysis

Statistical programs

The statistical programs used for the analysis were Epi Info version 3.5.4, SPSS version 20.0 and SAS version 10.1.

Variables in the model

The variables in the original dataset comprised of the following:

Table I: Variables in the Database

Participant characteristics	Characteristics of each medical visit
ID*	Full date of visit
Full date of birth	ICD-9 code
Gender	
First 3 digits of the postal code	
Household revenue	
Index date	
Case-control**	
Stratum number***	
Date of matching	

*Unique identifier for each participant **This dichotomous variable indicated whether the participant was a case or a control ***Each case and its matching controls (forming one matching set) were assigned a number

An urban/rural dichotomous variable was created from the postal code's Forward Sortation Area (first three digits of the postal code). The second digit of Canadian postal codes identifies the route type. A "0" refers to a rural route, whereas digits 1-9 represent an urban route [58]. This rural/urban designation corresponds to the method of mail preparation and delivery, and does not necessarily correlate with Statistic Canada's rural/urban classification system[58].

Descriptive Analysis

The goal of this analysis was to compare the distribution of descriptive and healthcare characteristics between cases and controls. Statistical significance was set at $p<0.05$.

Graphic representation of the distribution for each variable was studied, for all subjects together as well as divided by the case-control variable. To assess whether cases differed significantly from controls on demographic variables, a chi-square test was performed on categorical variables and a t-test for independent samples on continuous variables. Chi-square tests were performed on matched variables (gender, date of birth and urban vs rural) to confirm that the matching had worked properly. P-values approaching 1 were expected for these variables.

The descriptive characteristics assessed at this stage were gender (dichotomous), urban vs rural (dichotomous) and household revenue (continuous, ascertained from Census data)

The healthcare-related variables assessed in the descriptive analysis were index date, age at diagnosis, total number of physician visits and number of missing diagnostic codes. The case's diagnostic (index) date was conferred on its matched controls. The age at diagnosis was obtained by subtracting the diagnosis date from the birth date. The number of visits was obtained by counting the number of individual visits recorded by the RAMQ. The number of missing ICD-9 codes was compared between cases and controls.

The last section of the exploratory analysis assessed the distribution of the exposure variable, childhood infections. The ICD codes were recoded as infection types, based on the Simonsen classifications, and infections were pooled together for each subject. Three time periods were defined, and the sum of all infections calculated for each: from birth to year 1, from birth from year 5, and from birth to diagnosis

Medical episodes for infections were defined as 2 infections having occurred >15 days or >30 days apart. The exploratory analysis was run twice, once for each definition of a medical episode. The LOOP function in SPSS was used in order to identify which infections were to be excluded.

Univariate (Crude) Analysis

Crude measures of association of exposure with CD onset were computed according to the temporality (from birth to diagnosis, from birth to 1 year and from birth to 5 years) as well as infection type (according to the Simonsen categories). At this stage, all variables were categorized as yes/no variables (complete absence of infections vs ≥ 1 infection). This reflects Schlesselman's

proposed analytical method, whereby simple analyses precede and guide future, more complex analyses [59]. As all infections within the 2 years preceding the diagnosis were excluded from the analysis, children diagnosed before the age of 2, as well as their controls, were excluded from the analysis from this point.

The Epi Info program was used to perform the bivariate analysis. The Mantel-Haenszel test was used to obtain estimates of the odds ratios and 95% confidence intervals. The uncorrected MH chi-square p-value for a 2-tailed test was reported.

Multivariate Analysis and Model Building

The Hosmer and Lemeshow method of variable selection was followed in order to build the multivariate model [60]. This method was chosen as it emphasizes the importance of incorporating scientific knowledge to the statistical methods when choosing which variables to include in the model [60]. The steps of the model-building method are outlined below, and detailed in Appendix III.

1. Assessment of independent variables
2. Selection of potential confounders
3. Verification of the effect of potential confounders
4. Assessment of linearity of independent variables
5. Assessment of potential interaction terms

In addition, the following diagnostic measures were performed on the model obtained through the Hosmer Lemeshow model. A detailed description of the application of these measures to the current study can be found in Appendix III.

6. Goodness of fit test
7. Collinearity assessment
8. Extreme Observations assessment

Logistic Regression

Once the final model was obtained, multivariate (adjusted) analyses were carried out to assess the association between exposure and outcome. The measures of association (odds ratios) for the following exposure categories were computed:

Frequency: A categorical infection variable (0 infections, 1-5 infections, more than 5 infections) was created to evaluate the association between frequency of infection and outcome. A trends

analysis was performed to assess dose-response relationships, by calculating the median number of infections for each category of infection frequency, and performing a multivariate logistic regression.

Temporality: Firstly, only infections up to the first year of age were incorporated in the model. Subsequently, infections occurring in first 5 years of age only were taken into consideration. The variable for the number of medical visits (confounder) was re-calculated to include visits in the time frame studied only.

Additionally, in accordance to the hygiene hypothesis, a measure of association was calculated for individuals not exposed to infections in the first year of age, but subsequently exposed. A dichotomous variable was created to identify the individuals who fit this definition, and the odds ratio calculated.

Type of infections: A dichotomous (yes/no) variable for each infection type was used for this sub-analysis. A separate (one for each infection type) adjusted model was fit for each type of infection according to the Simonsen categories, comparing one type of infection with an absence of any other type of infection as the reference category.

The reference category for this analysis was the subset of subjects who had no recorded infection diagnoses. To avoid loss of power, an unconditional logistic regression model adjusting for the original matching variables and other potential confounders was fit.

Please refer to Appendix IV for the frequency of the different infection types amongst the study population.

Sample size

A sufficiently large sample size ensures that the risk of making a type II error is acceptable [59]. Type II errors occur when the null hypothesis (in this case, that childhood infections are not associated with CD onset) is accepted when it should be rejected. Based on estimated sample sizes, the power of this study was calculated. The Quanto program [61] was used to carry out the power calculations. The Canadian prevalence of upper respiratory infections (very common infection among children), was used in the calculations: 23% prevalence for 2-3 year olds, for the 2008-2009 years [62]. For matched analyses and an estimated sample of 400 cases, a power of >80% was calculated for a detection of odds ratios (OR) of >1.6, with ~100% power for detecting an OR of >1.9

Results

The results are presented in article format.

The following manuscript, entitled **Timing, frequency and type of physician-diagnosed infections in childhood and risk for Crohn's disease in children and young adults,** will be submitted to the journal "Gastroenterology". The following is a list of the co-authors and their contributions, in the order in which they appear on the manuscript:

- Vicky Springmann: methodology, data cleaning, statistical analysis, writing of the manuscript
- Paul Brassard: methodology, manuscript edits
- Alfreda Krupoves: coding of the raw data (ICD codes)
- Devendra Amre: study design, methodology, request of data to the *Commission d'Accès à l'Information* and the *Régie d'Assurance Médicale du Québec,* custodian of the database, statistical analysis, manuscript edits

Additional result tables can be found in the appendices:

Appendix II: Steps of the Hosmer-Lemeshow method of model building, as applied to this study

Appendix III: Diagnostic measures applied to the model

Appendix IV: Frequency of infection types amongst cases and controls

Appendix V: Sensitivity analysis

Appendix VI: Crude and adjusted analysis of the hygiene hypothesis variable

Timing, frequency and type of physician-diagnosed infections in childhood and risk for Crohn's disease in children and young adults

Springmann, Vicky (1,2); Brassard, Paul (2,3); Krupoves, Alfreda (4); Amre, Devendra (1,5)

(1) Ste-Justine Hospital Research Centre, Montreal (2) Department of Social and Preventive Medicine, University of Montreal, Montreal (3) Department of Clinical Epidemiology and Community Studies, Jewish General Hospital, Montreal (4) Institut national de santé publique du Québec (5) Department of Paediatrics, University of Montreal, Montreal

Grant Support

This study was funded by the Crohn's and Colitis Foundation of Canada, the Canadian Institute for Health Research and the Fonds de Recherche du Québec – Santé. Ms Springmann was supported by the Ste-Justine Hospital Research Foundation and the Faculty of Graduate and Post-Doctoral Studies of the University of Montreal.

Abbreviations

CD: Crohn's disease

ICD: International Classification of Disease

RAMQ: *Régie de l'assurance maladie du Québec*

Disclosures

No conflicts of interest

Abstract

Background and aims: Recent experimental data show that exposure to microbes during early childhood can confer immunological tolerance and protect against diseases such as Crohn's disease (CD) (the hygiene hypothesis). Epidemiological evidence for this link however, remains controversial. Using prospective data on physician-diagnosed infections, we examined the link between this hypothesis and risk for pediatric CD.

Methods: A case-control study design was used. Pediatric CD cases (<20 yrs) were recruited from a tertiary care pediatric hospital in Montreal and population-based controls were selected using Quebec's provincial medical insurance database and matched for age, gender, geographical location and period of insurance coverage. Exposure to infections was ascertained using prospectively recorded International Classification of Diseases -9 (ICD9) diagnostic codes. The relationship between the timing, frequency and type of infections and CD was assessed using conditional logistic regression analysis.

Results: 409 cases and 1621 controls were included. Adjusted regression analysis suggested that any recorded infection prior to CD diagnosis was associated with reduced risk of CD (OR=0.67, 95% CI=[0.48-0.93], p=0.018). The protective effect was restricted to infections occurring mainly before 5 yrs of age, with increasing number of infections resulting in greater protection (1-5 infections: OR=0.74; ≥6 infections: OR=0.61; p-value for trend=0.039). Observed reduced risks could not be attributed to a single infection type, however, infections affecting the oral tract, kidney and urinary tract and viral CNS infections seemed most protective.

Conclusion: Our study provides support for the hygiene hypothesis whereby exposure to infections in early childhood could potentially reduce risks of CD.

Keywords: Crohn's disease; case-control study; hygiene hypothesis; infections; children

Introduction

Crohn's disease (CD) is an inflammatory bowel disease, characterized by recurrent bouts of inflammation along the digestive tract [1, 2]. The worldwide incidence of CD has significantly increased over the course of the past 80 years, notably in developed countries [3-6]. Of particular concern is the concurrent increase in disease incidence in the pediatric population, as disease evolution in children often results in growth failure, depression and multiple school absences, etc. [7-10].

CD etiology remains unknown. Many genetic loci have been associated with CD [11]; however, they account for less than one third of all CD cases [12]. The rapid rise in disease incidence and the observation that disease prevalence is greatest in developed countries have led to the current belief that CD is caused by interactions between genetic and environmental factors[5]. Further evidence of the role of environmental risk factors in CD etiology is provided by the observation that immigrants from regions of low IBD incidence acquire a higher risk of disease when moving to a country with high CD incidence [1, 2].

Amongst potential environmental triggering factors, the hygiene hypothesis has received considerable attention. This theory, put forward by Strachan in 1989 following his study on risk factors for hay fever[13, 14], suggests that the promotion of infection transmission resulting from living with siblings, for example, is protective against subsequent allergic disorders [14]. The putative mechanism stipulates that the absence of micro-organisms in the environment precludes immunological tolerance, leading to an abnormal immune reaction upon future exposure to microbial agents [15-17]. Since this study, the hygiene hypothesis has been proposed as a possible explanation for a variety of conditions: multiple sclerosis, type 2 diabetes, asthma, allergies and IBD [13, 16].

A study published in 2012 demonstrated the auto-inflammatory effect of the absence of early microbial exposure in mice[18]. Efforts to demonstrate associations between early microbial exposure and CD risk in humans, however, have met with limited success. Some studies suggest that infection exposures can be protective[19-26], whereas others report that infections could enhance risk for CD[27-33]. Given the low population incidence of CD, most previous studies were case-control studies relying on retrospectively collected information on proxy measures of infection. Susceptibility to recall bias, differing definitions of hygiene and infections, and inaccurate estimation of infection exposure are likely to have contributed to the inconsistent

evidence. In order to validly ascertain the relationship between infections and CD, prospectively collected information is thus required.

Similar to other provinces in Canada, under the universal health care insurance program, the provincial health insurance agency of Quebec (RAMQ) prospectively maintains information on all physician visits undertaken by the residents of the province. This database enables the prospective ascertainment of "physician diagnosed" infections. We exploited this database to examine whether the frequency, timing and type of infections during childhood contribute to the development of CD in children and young adults.

Methods

Participant Selection

A case-control study was carried out. Cases of CD <20 years of age, consecutively diagnosed between 1988 and 2005, were selected from the gastroenterology clinic of a tertiary-care pediatric hospital in Montreal, Quebec. The latter is one of the major referral centers for the province of Quebec. Diagnosis of CD was based on established criteria, which included clinical, radiological, endoscopic and histopathological findings [34, 35]. Cases of ulcerative colitis (UC) and indeterminate colitis (IC) were excluded.

For each case, up to 4 controls without a diagnosis of either CD, UC or IC prior to the date of diagnosis of the case (index date), individually matched to the case for birth date, gender, duration of insurance coverage and area of residence, were selected from the RAMQ files. The RAMQ database is a near-complete census of residents in the province. Matching on area of residence was carried out by matching on the first 3 digits of the postal code of the case at the time of his or her diagnosis.

Ascertainment of Exposure

The RAMQ prospectively maintains information on all physician visits undertaken by the residents of the province. The RAMQ's administrative database comprises of information extracted from physician claims forms and includes diagnoses for each visit, in the form of ICD (International Classification of Disease) codes. In order to extract exposure information on childhood infections from this database, we used a classification system devised by Pinner et al (1996) in a previous study describing trends of infectious diseases in the United States, in order to identify the codes representing infections [36]. The 9th edition of the ICD coding system, exclusively in use in Quebec during the study period, was used.

As socio-economic status could be a potential confounder, as a proxy we utilized information on family income. This information was acquired from the 2001 Census carried out by Statistics Canada (www.statscan.ca). The average annual income of families residing in particular geographic regions (based on postal codes) was acquired and linked to the postal codes of the study participants.

This study was approved by the ethics review board of the study hospital, as well as the *Commission d'accès à l'information du Québec*.

Statistical Analysis

The primary purpose of the analysis was to examine whether the frequency of physician diagnosed infections at different time periods during childhood was related to CD. As the diagnosis process for CD often takes several months [37], only infections preceding the CD diagnosis by at least 2 years were included in the analysis, to ensure that the exposure preceded disease onset. Furthermore, infections recurring within 15 days of each other were counted only once. This was done to eliminate the possibility that two visits relating to the same infection were incorrectly recorded as two separate infectious episodes. A sensitivity analysis was done to validate this 15-day cut-off, by counting all infections with the same ICD codes, occurring within 30 days of each other as a single episode. The frequency of infections was compiled by adding the number of visits for which an infection-related diagnosis was emitted. Infections during three different time periods were analyzed: those occurring during the first year of life, those from birth to 5 years, and those from birth to CD onset. Specific infection types were assessed according to the categories created by Pinner et al [36], and infection categories affecting <5 cases or controls were excluded from the analysis.

Descriptive analyses were done using T-tests for continuous variables and chi-square tests for categorical variables. Initially, associations between the frequency of infections at different time periods and the risk of pediatric CD were calculated using the Maentel-Haenszel formula, accounting for the matching variables by carrying out a matched analysis (for dichotomous exposures), and using simple conditional logistic regression for matched data for categorical (>2 categories) and continuous exposures.

Subsequently a multivariate conditional logistic regression analysis for matched data was carried out, to assess the influence of frequency and temporality of infections on CD occurrence. Other potential confounders considered were household income, number of ICD codes missing from physician claims within the RAMQ database and the total number of medical visits for each patient. Model fit was assessed using

methods described by Hosmer & Lemeshow [38]. Odds ratios (OR) and corresponding 95% confidence intervals (95% CI) were estimated.

To assess the influence of specific infections on CD risk, total absence of any infection was used as the reference category, and unconditional logistic regression was carried out controlling for the matching variables (age at diagnosis, gender, area of residence) as well as for income and number of physician visits. As power was limited for this analysis, only models for the presence/absence of specific infections were fit and their influence at different time periods was not assessed.

Two-tailed p-values of <0.05 were considered significant for all analyses.

Results

409 confirmed cases of CD, diagnosed between 1988 and 2005, were included in the study. 394 of these cases were successfully matched with 4 controls; 15 cases were matched with 3 controls. Age at diagnosis ranged from 0.33 to 19.0 years (mean: 11.46, SD± 3.63). The proportion of male cases was slightly higher than for females (51.8% vs 48.2%). The mean income (as established from census data) was significantly higher for the cases than for the controls (Table 1). After excluding children diagnosed during the first 2 years of life to allow for the lag time between exposure and outcome (4 cases and corresponding controls), 405 cases and 1607 controls were included in the final analyses.

Conditional logistic regression analysis adjusting for the matching variables only did not reveal any associations between the presence/absence of infections prior to the index date and risk for CD. Further accounting for other potential confounders showed that exposure to infections prior to the index date was inversely associated with the occurrence of CD (OR=0.67, 95% [CI=0.48-0.93]) (Table 2a). No dose-response effects were evident, as similar effects were seen for those with 1-5 infections (OR=0.67, 95% [CI=0.48-0.94]) and those with >5 infections (OR=0.73, 95% [CI=0.48-1.12]). When the analysis was stratified according to time period of exposure, the observed protective effects were restricted to infections diagnosed during the first 5 years of life only, with increasing number of infections leading to additional protection (OR=0.74, 95% CI=[0.57-0.96] for 1-5 infections and OR=0.61 [0.37-1.01] for >5 infections) (table 2b). When dose-effects were formally tested (by creating an indicator variable representing the median frequency of infection for each infection category and entering it as a continuous variable in the regression model), increasing number of infections during the first 5 yrs of life were associated with decreasing risks for

CD (p-value for trend=0.039)

When associations with specific infections were examined, infections involving the kidney/urinary tracts and the bladder, viral central nervous system infections and oral infections appeared to contribute most strongly to the overall protective effects. Infections of the heart, hepato-billiary diseases, meningitis, postoperative infections, infections in pregnancy and septicemia could not be analyzed owing to infrequent occurrences (< 5 episodes) (Table 3, Figure 1).

A sensitivity analysis considering infections occurring outside a 30-day window as independent infections, compared to the 15-day window, revealed by-and-large similar results (data not shown).

Discussion

Using prospectively collected information on physician-diagnosed infections, we observed that higher infection exposures, particularly during the first 5 years of life, seemed to have a protective effect on the development of CD. These overall protective effects were predominantly due to infections affecting the oral tract, kidney and urinary tract, and viral infections of the central nervous system.

To our knowledge, this study is the largest study to date in children and young adults that examined potential associations between childhood infections and CD. The study population was representative of children and young adults studied worldwide (male: female proportions of 51.8% vs 48.2%and mean age (11.46) [7, 8, 31, 39] [32, 40] allowing their generalization to Caucasian populations.

The hygiene hypothesis, originally formulated by Strachan[38], implies that exposure to poor hygiene or lack of infections experienced during early childhood may play a protective role in atopic disorders. This hypothesis evolved from epidemiological observations based on the inverse relationship between surrogate measures of infection exposure (such as family size, birth order) and hay fever. Protective effects against atopy, allergy and T helper type 1 (Th-1) mediated autoimmune diseases have since been reported with some consistency using various potential indicators of infection exposure and burden.[39-43] The premise underlying the hygiene hypothesis is that early exposure to infections helps establish the immunological balance between pro-inflammatory and tolerance-inducing responses to antigenic stimuli and thus contributes to the maintenance of physiological inflammation from subsequent contact[13, 18].

Though few studies have assessed childhood exposure to infections *per se* as an environmental risk factor, many have evaluated other risk factors commonly associated with the hygiene hypothesis. Such measures include, amongst others, living in a rural environment, owning pets, drinking unpasteurized milk, living in a residence with a high crowding index, attending daycare and having a high number of siblings [5, 14, 15, 32, 41-44]. The association between these proxy measures of hygiene and CD have been inconsistent across studies, most likely due to the heterogeneity of study methods used, potential recall bias due to retrospective ascertainment of exposure and study of prevalent rather than incidence cases. [45]. Most previous studies focusing on environmental risk factors for CD have relied on mailed questionnaires and interviews to assess exposure; however, such methods are prone to recall bias [45, 46]. In the present study, the use of prospectively-collected data, in addition to eliminating the risk of recall bias, insures that the exposure truly preceded the onset of disease.

Notwithstanding the inconsistent epidemiological evidence, a recent experimental study has provided vital clues for the potential link between early microbial exposures and risk for diseases such as IBD and asthma. Using different animal models for assessing age-dependent influence of microbial exposure, Olszak et al (2012) demonstrated that early microbial exposure leads to a decrease in number and functioning of invariant natural killer T-cells (iNKT) (key players in the initiation of mucosal response to exogenous and endogenous microbes); that this early tolerance to iNKT cells generated by the microbes was long-standing and protected from acquiring colitis (and asthma) and that abrogation of the tolerance resulted in increased susceptibility for colitis (and asthma). Although these findings cannot be directly extrapolated to our epidemiological findings (given that they were based on animal models of colitis rather than CD), they provide further impetus to the "hygiene hypothesis" paradigm in susceptibility for human CD[18].

Our study findings should be interpreted in the context of inherent limitations of using administrative databases for epidemiological studies. A proportion of diagnostic codes were missing from the database, as physicians occasionally omit to inscribe the purpose of the medical visit on the claims form. There is a possibility that the missing codes could be differentially related to infection diagnoses, which are likely more difficult to diagnose and code than other medical conditions. Although it was not possible for us to ascertain the later, the frequency of missing diagnostic codes did not differ according to case-control status and was not deemed to be a potential confounder in our study. Another study limitation was that only physician-diagnosed infections were accounted for. This could have led to a potential underestimation of all infection

exposures. We have no reason to believe that such underestimation would be different between cases and controls. There is also the possibility that some physicians may over or under diagnose infections. We matched the cases and controls on geographical area, thereby potentially controlling for differences in medical practices, and limiting potential confounding, if any.

An important point to be noted is that protective effects of infections were observed only in models that controlled for the "number of physician visits" that were higher among cases than controls. This higher frequency among cases was evident even after we excluded all visits that occurred within the 2-yr period prior to diagnosis (during which period cases are expected to visit a physician more often due to symptoms related to the disease). As increased propensity for physician-visits could lead to increased diagnosis of infections in particular, we considered this a potential confounder in the analysis. Statistical modeling suggested that indeed the variable was a potential confounder and that model fits were substantially improved after its controlling. Our accounting for physician visits as a confounder is also consistent with previous reports on other outcomes, which used diagnostic codes to infer infection exposures from administrative databases[47-49].

Conclusion

Our results support the hygiene hypothesis, as infections prior to disease onset were shown to have a protective effect on CD risk. Though this study brings us one step further in the assessment of the validity of the Hygiene Hypothesis in CD, further large prospective studies are needed to confirm these results.

Bibliography

1. Baumgart DC, Sandborn WJ. Crohn's disease. Lancet 2012;380:1590-605.
2. Cosnes J, Gower-Rousseau C, Seksik P, Cortot A. Epidemiology and natural history of inflammatory bowel diseases. Gastroenterology 2011;140:1785-94.
3. Molodecky NA, Soon IS, Rabi DM, Ghali WA, Ferris M, Chernoff G, Benchimol EI, Panaccione R, Ghosh S, Barkema HW, Kaplan GG. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology 2012;142:46-54 e42; quiz e30.
4. Economou M, Pappas G. New global map of Crohn's disease: Genetic, environmental, and socioeconomic correlations. Inflamm Bowel Dis 2008;14:709-20.
5. Frolkis A, Dieleman LA, Barkema H, Panaccione R, Ghosh S, Fedorak RN, Madsen K, Kaplan GG. Environment and the inflammatory bowel diseases. Can J Gastroenterol 2013;27:e18-24.
6. Behr MA, Bruere P, Oxlade O. Global rates of Crohn's disease. Inflamm Bowel Dis 2008;14:1170-2.
7. Benchimol EI, Fortinsky KJ, Gozdyra P, Van den Heuvel M, Van Limbergen J, Griffiths AM. Epidemiology of pediatric inflammatory bowel disease: a systematic review of international trends. Inflamm Bowel Dis 2011;17:423-39.
8. Griffiths AM. Specificities of inflammatory bowel disease in childhood. Best Pract Res Clin Gastroenterol 2004;18:509-23.
9. Kim SC, Ferry GD. Inflammatory bowel diseases in pediatric and adolescent patients: clinical, therapeutic, and psychosocial considerations. Gastroenterology 2004;126:1550-60.
10. Bousvaros A, Sylvester F, Kugathasan S, Szigethy E, Fiocchi C, Colletti R, Otley A, Amre D, Ferry G, Czinn SJ, Splawski JB, Oliva-Hemker M, Hyams JS, Faubion WA, Kirschner BS, Dubinsky MC. Challenges in pediatric inflammatory bowel disease. Inflamm Bowel Dis 2006;12:885-913.
11. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Schumm LP, Sharma Y, Anderson CA, Essers J, Mitrovic M, Ning K, Cleynen I, Theatre E, Spain SL, Raychaudhuri S, Goyette P, Wei Z, Abraham C, Achkar JP, Ahmad T, Amininejad L, Ananthakrishnan AN, Andersen V, Andrews JM, Baidoo L, Balschun T, Bampton PA, Bitton A, Boucher G, Brand S, Buning C, Cohain A, Cichon S, D'Amato M, De Jong D, Devaney KL, Dubinsky M, Edwards C, Ellinghaus D, Ferguson LR, Franchimont D, Fransen K, Gearry R, Georges M, Gieger C, Glas J, Haritunians T, Hart A, Hawkey C, Hedl M, Hu X, Karlsen TH, Kupcinskas L, Kugathasan S, Latiano A, Laukens D, Lawrance IC, Lees CW, Louis E, Mahy G, Mansfield J, Morgan AR, Mowat C, Newman W, Palmieri O, Ponsioen CY, Potocnik U, Prescott NJ, Regueiro M, Rotter JI, Russell RK, Sanderson JD, Sans M, Satsangi J, Schreiber S, Simms LA, Sventoraityte J, Targan SR, Taylor KD, Tremelling M, Verspaget HW, De Vos M, Wijmenga C, Wilson DC, Winkelmann J, Xavier RJ, Zeissig S, Zhang B, Zhang CK, Zhao H, Silverberg MS, Annese V, Hakonarson H, Brant SR, Radford-Smith G, Mathew CG, Rioux JD, Schadt EE, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 2012;491:119-24.
12. Ananthakrishnan AN. Environmental triggers for inflammatory bowel disease. Curr Gastroenterol Rep 2013;15:302.
13. Fishbein AB, Fuleihan RL. The hygiene hypothesis revisited: does exposure to infectious agents protect us from allergy? Curr Opin Pediatr 2012;24:98-102.
14. Strachan DP. Hay fever, hygiene, and household size. BMJ 1989;299:1259-60.
15. Castiglione F, Diaferia M, Morace F, Labianca O, Meucci C, Cuomo A, Panarese A, Romano M, Sorrentini I, D'Onofrio C, Caporaso N, Rispo A. Risk factors for inflammatory bowel diseases according to the "hygiene hypothesis": a case-control, multi-centre, prospective study in Southern Italy. J Crohns Colitis 2012;6:324-9.
16. Rook GA. Hygiene hypothesis and autoimmune diseases. Clin Rev Allergy Immunol 2012;42:5-15.
17. Lashner BA, Loftus EV, Jr. True or false? The hygiene hypothesis for Crohn's disease. Am J Gastroenterol 2006;101:1003-4.
18. Olszak T, An D, Zeissig S, Vera MP, Richter J, Franke A, Glickman JN, Siebert R, Baron RM, Kasper DL, Blumberg RS. Microbial exposure during early life has persistent effects on natural killer T cell function. Science 2012;336:489-93.
19. Duggan AE, Usmani I, Neal KR, Logan RF. Appendicectomy, childhood hygiene, Helicobacter pylori status, and risk of inflammatory bowel disease: a case control study. Gut 1998;43:494-8.
20. Elliott DE, Li J, Blum A, Metwali A, Qadir K, Urban JF, Jr., Weinstock JV. Exposure to schistosome eggs protects mice from TNBS-induced colitis. Am J Physiol Gastrointest Liver Physiol 2003;284:G385-91.

21. Gent AE, Hellier MD, Grace RH, Swarbrick ET, Coggon D. Inflammatory bowel disease and domestic hygiene in infancy. Lancet 1994;343:766-7.
22. Khan WI, Blennerhasset PA, Varghese AK, Chowdhury SK, Omsted P, Deng Y, Collins SM. Intestinal nematode infection ameliorates experimental colitis in mice. Infect Immun 2002;70:5931-7.
23. Moreels TG, Pelckmans PA. Gastrointestinal parasites: potential therapy for refractory inflammatory bowel diseases. Inflamm Bowel Dis 2005;11:178-84.
24. Moreels TG, Nieuwendijk RJ, De Man JG, De Winter BY, Herman AG, Van Marck EA, Pelckmans PA. Concurrent infection with Schistosoma mansoni attenuates inflammation induced changes in colonic morphology, cytokine levels, and smooth muscle contractility of trinitrobenzene sulphonic acid induced colitis in rats. Gut 2004;53:99-107.
25. Summers RW, Elliott DE, Weinstock JV. Is there a role for helminths in the therapy of inflammatory bowel disease? Nat Clin Pract Gastroenterol Hepatol 2005;2:62-3.
26. Feeney MA, Murphy F, Clegg AJ, Trebble TM, Sharer NM, Snook JA. A case-control study of childhood environmental risk factors for the development of inflammatory bowel disease. Eur J Gastroenterol Hepatol 2002;14:529-34.
27. Ekbom A, Adami HO, Helmick CG, Jonzon A, Zack MM. Perinatal risk factors for inflammatory bowel disease: a case-control study. Am J Epidemiol 1990;132:1111-9.
28. Wurzelmann JI, Lyles CM, Sandler RS. Childhood infections and the risk of inflammatory bowel disease. Dig Dis Sci 1994;39:555-60.
29. Van Kruiningen HJ, Joossens M, Vermeire S, Joossens S, Debeugny S, Gower-Rousseau C, Cortot A, Colombel JF, Rutgeerts P, Vlietinck R. Environmental factors in familial Crohn's disease in Belgium. Inflamm Bowel Dis 2005;11:360-5.
30. Thompson NP, Pounder RE, Wakefield AJ. Perinatal and childhood risk factors for inflammatory bowel disease: a case-control study. Eur J Gastroenterol Hepatol 1995;7:385-90.
31. Baron S, Turck D, Leplat C, Merle V, Gower-Rousseau C, Marti R, Yzet T, Lerebours E, Dupas JL, Debeugny S, Salomez JL, Cortot A, Colombel JF. Environmental risk factors in paediatric inflammatory bowel diseases: a population based case control study. Gut 2005;54:357-63.
32. Amre DK, Lambrette P, Law L, Krupoves A, Chotard V, Costea F, Grimard G, Israel D, Mack D, Seidman EG. Investigating the hygiene hypothesis as a risk factor in pediatric onset Crohn's disease: a case-control study. Am J Gastroenterol 2006;101:1005-11.
33. Gilat T, Hacohen D, Lilos P, Langman MJ. Childhood factors in ulcerative colitis and Crohn's disease. An international cooperative study. Scand J Gastroenterol 1987;22:1009-24.
34. Sands BE. From symptom to diagnosis: clinical distinctions among various forms of intestinal inflammation. Gastroenterology 2004;126:1518-32.
35. Lennard-Jones JE. Classification of inflammatory bowel disease. Scand J Gastroenterol Suppl 1989;170:2-6; discussion 16-9.
36. Pinner RW, Teutsch SM, Simonsen L, Klug LA, Graber JM, Clarke MJ, Berkelman RL. Trends in infectious diseases mortality in the United States. JAMA 1996;275:189-93.
37. Heikenen JB, Werlin SL, Brown CW, Balint JP. Presenting symptoms and diagnostic lag in children with inflammatory bowel disease. Inflamm Bowel Dis 1999;5:158-60.
38. Hosmer DW, Lemeshow S. Applied logistic regression. Wiley, 2000.
39. Lopez-Serrano P, Perez-Calle JL, Perez-Fernandez MT, Fernandez-Font JM, Boixeda de Miguel D, Fernandez-Rodriguez CM. Environmental risk factors in inflammatory bowel diseases. Investigating the hygiene hypothesis: a Spanish case-control study. Scand J Gastroenterol 2010;45:1464-71.
40. Pinsk V, Lemberg DA, Grewal K, Barker CC, Schreiber RA, Jacobson K. Inflammatory bowel disease in the South Asian pediatric population of British Columbia. Am J Gastroenterol 2007;102:1077-83.
41. Garn H, Renz H. Epidemiological and immunological evidence for the hygiene hypothesis. Immunobiology 2007;212:441-52.
42. Gearry RB, Richardson AK, Frampton CM, Dodgshun AJ, Barclay ML. Population-based cases control study of inflammatory bowel disease risk factors. J Gastroenterol Hepatol 2010;25:325-33.
43. Wills-Karp M, Santeliz J, Karp CL. The germless theory of allergic disease: revisiting the hygiene hypothesis. Nat Rev Immunol 2001;1:69-75.
44. Okada H, Kuhn C, Feillet H, Bach JF. The 'hygiene hypothesis' for autoimmune and allergic diseases: an update. Clin Exp Immunol 2010;160:1-9.
45. Molodecky NA, Panaccione R, Ghosh S, Barkema HW, Kaplan GG. Challenges associated with identifying the environmental determinants of the inflammatory bowel diseases. Inflamm Bowel Dis 2011;17:1792-9.

46. Bernstein CN, Rawsthorne P, Cheang M, Blanchard JF. A population-based case control study of potential risk factors for IBD. Am J Gastroenterol 2006;101:993-1002.
47. Bremner SA, Carey IM, DeWilde S, Richards N, Maier WC, Hilton SR, Strachan DP, Cook DG. Infections presenting for clinical care in early life and later risk of hay fever in two UK birth cohorts. Allergy 2008;63.
48. Garcia Rodriguez L, Tolosa L, Ruigomez A, Johansson S, Wallander M-A. Rheumatoid arthritis in UK primary care: incidence and prior morbidity. Scand J Rheumatol 2009;38.
49. Cardwell C, McKinney P, Patterson C, Murray L. Infections in early life and childhood leukaemia risk: a UK case-control study of general practitioner records. British Journal of Cancer 2008;99:5.

Table 1: Descriptive Statistics of the Study population

DEMOGRAPHICS			
	Cases (n=409)	**Controls (n=1621)**	**p-value**
Female n (%)	197 (48.2%)	781 (48.2%)	
Rural n (%)	64 (15.6%)	254 (15.7%)	
mean Revenue [SD]	60,261.02 [34,079.06]	55,446.36 [29,395.24]	p=0.004
Mean Age at index date [SD]	11.46 [3.63]	11.48 [3.63]	
Number of medical visits n (%)	76.23 [58.51]	61.60 [46.42]	p<0.001
Missing ICD-9 codes for individual subjects mean [SD]	19.32 [20.27]	21.69 [20.05]	p=0.034

*excluding 2 years prior to diagnosis

Table 2: Association between temporality and frequency of infections and pediatric CD (conditional logistic regression)

2a) FREQUENCY OF INFECTIONS				
	Adjusted for the matching variables*		**Adjusted for the matching variables and other confounders***	
	OR [95% CI]	**p-value**	**OR [95% CI]**	**p-value**
0 infections (ref)				
1-5 infections	0.88 [0.65-1.20]	0.426	0.67 [0.48-0.94]	0.020
>5 infections	1.28 [0.89-1.84]	0.191	0.73 [0.48-1.12]	0.152
≥1 infection	1.04 [0.77-1.41]	0.791	0.67 [0.48-0.93]	0.018
2b) TEMPORALITY OF INFECTIONS				
Infections during first year of life				
	same as above		**same as above***	
	OR [95% CI]	**p-value**	**OR [95% CI]**	**p-value**
0 infections (ref)				
1-5 infections	0.98 [0.74-1.30]	0.984	0.94 [00.70-1.26]	0.663
>5 infections	0.99 [0.10-10.04]]	0.995	0.87 [0.08-9.04]	0.906
Any infection	0.98 [0.74-1.30]	0.909	0.96 [0.71-1.30]	0.796
Infections during first 5 years of life				
	Same as above		**same as above***	
	OR [95% CI]	**p-value**	**OR [95% CI]**	**p-value**
0 infections (ref)				
1-5 infections	0.97 [0.76-1.24]	0.798	0.74 [0.57-0.96]	0.025
>5 infections	1.06 [0.66-1.70]	0.808	0.61 [0.37-1.01]	0.057
Any infection	0.98 [0.77-1.24]	0.845	0.87 [0.67-1.13]	0.284

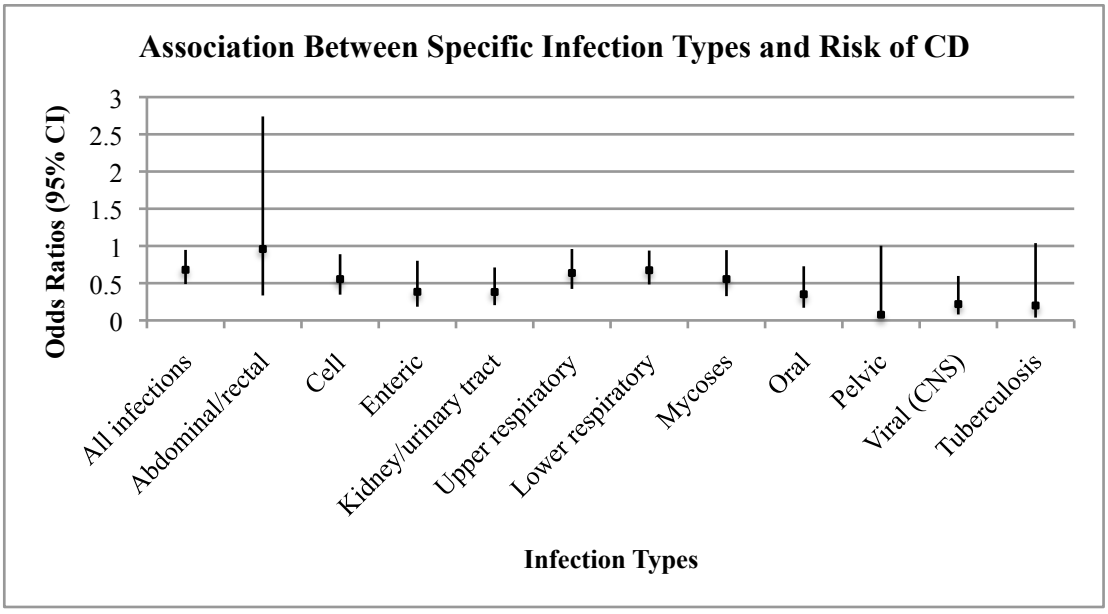
*Adjusted for revenue and number of physician visits upto 2 years prior to index date

Table 3: Association between types of infections and pediatric CD (unconditional logistic regression)

	Crude Analysis		**Adjusted Analysis***	
	OR [95% CI]	**p-value**	**OR [95% CI]**	**p-value**
Abdominal and rectal infections	3.48 [1.99-6.09]	0.000	1.41 [0.65-3.03]	0.382
Cellulitis	1.11 [0.82-1.49]	0.528	0.54 [0.34-0.86]	0.011
Enteric infections	1.09 [0.65-1.81]	0.748	0.42 [0.20-0.86]	0.018
Kidney, urinary tract and bladder infections	0.84 [0.56-1.28]	0.415	0.38 [0.21-0.71]	0.002
Lower respiratory tract infection	1.10 [0.87-1.40]	0.401	0.64 [0.42-0.96]	0.031
Mycosis	1.11 [0.77-1.61]	0.562	0.56 [0.33-0.95]	0.032
Oral infections	0.98 [0.58-1.65]	0.936	0.35 [0.17-0.73]	0.005
Pelvic inflammatory disease	1.19 [0.72-1.97]	0.492	0.51 [0.25-1.07]	0.073
Upper respiratory tract infection	1.17 [0.90-1.52]	0.242	0.67 [0.48-0.93]	0.018
Viral central nervous system infections	0.51 [0.21-1.25]	0.142	0.22 [0.08-0.60]	0.003
Tuberculosis	0.55 [0.12-2.48]	0.436	0.20 [0.04-1.04]	0.055

*Adjusted for revenue and total number of physician visits

Diagram 1: Association between infection types and pediatric CD (adjusted model)



Discussion

This study, to our knowledge, is the largest paediatric case-control study examining potential associations between childhood infections and CD to date. The risk of non-response from eligible participants was eradicated by using data collected from an administrative database.

Our study population was comparable to that of previous paediatric CD studies: a slightly, albeit significant, greater proportion of male cases than female cases was recruited, which is concordant with a trend amongst paediatric studies worldwide[16, 20, 64, 80]. The mean age at diagnosis of the CD cases was 11.46 years, comparable to that of CD cases in studies performed in other provinces: a 2009 study on paediatric CD incidence in Southern Ontario reported a mean age at diagnosis of 13.9 [83], and the results of a paediatric IBD study conducted in the province of British Columbia indicated a mean age of diagnosis of 12.5 [81]. The majority of the study population (about 85%) had a postal code corresponding to an urban mail delivery route. This is reflective of the location and type of care centre from which the cases were recruited: a tertiary-care paediatric hospital in a large city (the population of Montreal in 2001 exceeded 1.8 million [84]).

Results of the statistical analysis

The preliminary univariate analyses did not suggest associations between childhood exposures and CD. The multivariate model however, when adjusted for household income and the total number of medical visits, showed that infections at any time before CD diagnosis seemed to contribute to the protective effect. These associations were probably reflective of protective exposures occurring during the first 5 years of life rather than infections very early (during the first year of life). This could signify that immune stabilization over a longer period of time may be required to confer protection. Interestingly, as proposed for the hygiene hypothesis, our analysis did not reveal that "lower frequency of infections in early childhood followed by infections later on may confer increased risks for CD" (see Appendix VI). It should be noted however that in order to carry out this analysis, the power of the current study was considerably reduced as it entailed having a large number of children who were diagnosed later on during childhood (example after age 10). This reduced power may have resulted in an inability to specifically examine the hygiene hypothesis in its entirety.

The frequency of childhood infections and its association with CD was calculated by breaking up the variable into three categories: no infection, 1-5 infections, and 6 infections and higher. The objective was to assess whether a dose-response relationship existed between the exposure and the outcome. A dose-response relationship was observed for infections occurring before the age of

five; the risk of Crohn's disease increased with the frequency of infections. A trends analysis performed for this age group was significant.

A multivariate logistic regression analysis was carried out to assess the role of specific infection types on risk for CD. The reference category used for this multivariate analysis was a complete absence of any infections (different reference category than for the univariate analysis, which was absence only of that specific type of infection). 382 subjects met the criterion for absence of any recorded infection. As the reference category was changed for this analysis, the matching between cases and controls was broken, and thus an unconditional logistic regression model was fit accounting for the matching variables (age, gender and urban or rural environment) and other potential confounding variables. No specific infection was found to be primarily responsible for the protective effect. However, infections such as viral infections affecting the CNS (central nervous system), infections of the oral tract and those of the kidney and urinary tracts showed the strongest protective effects.

Comparison of study results with those of previous studies

Other studies have assessed childhood infections in the context of CD, using different methodologies, and have reported mixed findings. A case-control study performed by our group (Amre et al) in 2006 used a questionnaire to inquire about "physician-diagnosed infections" at different time points before CD diagnosis in a paediatric population. The results of this study were contrary to those found in the present study: physician-diagnosed infections were identified as risk factors for CD in the multivariate analysis, though these findings were not significant at the p=0.05 level [40]. In another paediatric case-control study performed by Baron et al in France, no association was reported between CD and measles, mumps, rubella or other infections reported in an interview and recorded in the personal health booklet (OR for measles: 1.1; 95%CI: [0.7-1.6]). The results of Lopez-Cerrano's Spanish case-control study reflected our own, as participant-reported infections seemed to confer protection against CD.

Though few studies have assessed childhood exposure to infections as an environmental risk factor for paediatric CD, many have evaluated other risk factors commonly associated with the hygiene hypothesis. Such measures, which include, amongst others, living in a rural environment, owning pets, drinking unpasteurized milk, living in a residence with a high crowding index, sharing a towel and having a high number of siblings, have all been used to ascertain exposure to antigens in the same way that we have used childhood infections as a measure of hygiene [21, 25, 32, 34, 38, 41, 42]. For example, in a New Zealand case-control study, living in a rural area (compared with a city) as a

child significantly decreased the risk of CD (OR: 0.64, 95%CI: [0.46-0.88])[25]. Lopez-Serrano et al observed similar results in Spain: an urban residence during childhood was reported to increase risk of CD (OR: 4.58; 95% CI: [2.17-10]) [80]. Living with a pet during childhood, another example of a measure of antigen exposure was found to be protective by Lopez-Serrano et al (OR: 0.3, 95% CI: [0.2-0.8]) and by Bernstein et al, authors of the Manitoban case-control study (OR: 0.66, 95% CI: [0.46-0.96] for living with a pet cat before the age of 5). Sharing a household with a large number of individuals has been correlated with a greater exposure to antigens. Amre et al reported a protective effect for a higher "crowding index" (ratio of number of inhabitants of a residents to the number of rooms) for paediatric CD (OR: 0.33 (0.13-0.84) [40].

The heterogeneity of epidemiological methods used to study CD risk factors has been identified as a cause of the heterogeneity of results obtained from different studies. Most studies performed on environmental risk factors for CD are registry-based or case-control studies. Amongst case-control studies such as this one, methodology differs greatly. Firstly, the selection of study participants is very diverse: in some studies recruitment of the cases is done through a hospital [40, 80], others from an IBD registry [20] or referrals by gastroenterologists [25]. Secondly, the control selection varies widely: clinical patients with a different illness [40, 80], population controls selected from an electoral roll [25], random telephone dialing [20], health registry [28], etc. These various combinations of participant selection can introduce selection biases (during recruitment) as well as misclassification biases (when ascertaining exposure) [82]. If the cases and the controls are taken from different populations, not only does this affect the internal validity of the study; it also impairs the comparability between the findings of different studies.

Misclassification of exposure can also influence the results of a study. When a questionnaire or interview is used to ascertain exposure, recall bias is most likely the greatest threat to the validity of a study. In the case of our study, possible misclassification bias came from the RAMQ database – both missing ICD-9 codes and potential coding errors. This is especially true in the case of infections, which can be difficult to diagnose. However, this bias is most likely non-differential between the cases and the controls, and would have the effect of diluting the calculated association between childhood exposure and CD rather than accentuating, therefore rendering conservative results.

Study strengths

In the present study, as only hospital and RAMQ records were utilized to create the matching, there were no limitations due to participant motivation or to unequal matching between cases and controls. Studies using questionnaires or interviews for data collection must account for a loss of potential participants due to non-motivation, which may introduce a selection bias. For example, the original study design of Bernstein's 2006 study of IBD risk factors was a matched case-control study, with age, gender and geography set as matching criteria. However, the control respondents of the mailed questionnaire used to ascertain exposure were mostly female and older than the CD patients [28]. This created an imbalance between cases and controls that precluded matching as an option of confounder control [28].

Prospective data collection constitutes one of the greatest strengths of this study. Most studies focusing on environmental risk factors for CD have relied on mailed questionnaires and interviews to assess exposure. However, such methods are prone to recall bias. This is especially true of questionnaires study childhood exposures, which are administered to adult participants. In the current study, all exposure information was entered prospectively as part of the physician claims process. This eliminated the potential for differential recollection between cases and controls. As most insurance providers worldwide utilize the ICD coding system, this novel use of administrative databases and ICD codes to establish environmental exposure levels could be a potential solution to ensure comparability between environmental exposures in different parts of the world.

Another advantage of prospective data collection is the assurance that the exposure truly precedes the onset of disease. By utilizing the administrative database, we were able to exclude all infections which occurred after the diagnosis of CD, as well as exclude infections within 2 years preceding the diagnosis of CD. This approach ensured that the potential bias associated with reverse causality was limited to a large extent.

Study Limitations

As ascertainment of exposure was reliant on physician billing requests, only physician-diagnosed infections were considered in this study. However, many infections may not be captured in administrative databases, for the following reasons:

- Child was not brought to a medical clinic upon presentation of symptoms of infections
- Physician did not diagnose an infection

- Physician diagnosed the infection, but did not enter the ICD-9 code on the claims form accordingly

By matching the cases and controls on geographical area, the differences in practice were partially offset by insuring, for example, that a case did not reside in a significantly more remote locale than its control and thus had the same opportunity of obtaining a correctly entered ICD coding information for an episode of infection.

A large proportion of diagnostic codes were missing, as physicians occasionally omit to inscribe the purpose of the medical visit on the claims form. Approximately 30% of all ICD-9 codes were missing, possibly obscuring important exposure information. There is a potential that these missing codes could be differentially related to infection diagnoses, which are likely more difficult to diagnose and code than trauma-related medical visits, for example. The proportion of missing ICD-9 codes was significantly higher for the cases than for the controls, however our analysis revealed that this variable did not confound the exposure-disease association, and hence final multivariate models did not include this variable.

The frequency of medical visits was significantly higher for cases than for controls. A possible reason for this discrepancy is that some of the visits are associated with CD symptoms. The variable for the number of medical visits was confirmed as a confounder by a Wald test and likelihood ratio test, and was included in the multivariate analysis.

A limitation of our study was the absence of household income data, as this information is not stored in the RAMQ database. Full postal codes can be correlated to socioeconomic status[85]; however, only the first 3 digits of the postal code were available for this study, precluding the use of the postal codes for this purpose. Many epidemiological studies have identified a high household income as a risk factor for CD [29 80]. A proxy measure of household income, the average familial income of the geographical area of residence as reported in the 2001 Statistics Canada Census [86], was used for adjustment purposes in the final model. The average household revenue was significantly higher for the cases than the controls. However, conclusions should not be drawn from these results as they represent estimates rather than participant-specific data.

Meaning of the study: possible mechanisms

Our research findings support the hygiene hypothesis, as childhood infections were protective against paediatric CD. A possible biological mechanism for the protective role of antigen exposure in childhood involves T cells [25]. TH2 cells (type of T helper cells) require prompting by antigens

early in childhood in order to build immunological tolerance and preclude allergic reactions later on in life [41]. Our study attempted to measure the presence or absence of prompting by infectious pathogens.

It is possible that paediatric and adult-onset CD are induced by different environmental factors. A bimodal age distribution for a disease could suggest that different causes underlie the incidence for each peak in the distribution [87]. As IBD is thought to be due to gene-environment interactions, the possibility that trigger factors for genetically pre-disposed individuals differ between paediatric and adult-onset cases should not be disregarded. Absence of antigen exposure could contribute to CD aetiology in children, as supported by our study, but it is possible that the protective effect of antigen exposure is limited to this population. We therefore suggest that studies pertaining to CD aetiology be separated by age category according to the bimodal nature of age at diagnosis.

Unanswered questions and future research

Due to the nature of the data collection method, we would expect the results to have good external validity, as the data collection was done prospectively and undifferentially between cases and controls [82]. We suggest that this study, utilizing administrative insurance data to compile childhood exposure, be replicated in other populations. As ICD-9 codes are used worldwide, employing them in a standardized manner to assess associations between CD and childhood infections would provide a means of comparison between risk factors in different populations.

Furthermore, large studies are needed that combine both environmental and genetic risk factors, in order to determine the impact of each, and the interaction between them. This type of information is not available in administrative databases, but could be collected in the context of a cohort study. For example, Pinsk et al indirectly assessed genetic components of IBD in their 2007 study conducted in British Columbia by comparing the incidence of IBD amongst those of South-Asian decent, prevalent in the area, and the rest of the IBD population [81]. Twin studies have also been conducted to "match" for genetic factors, such as the 2012 study by Ng et al, which included British twins in the UK, where at least one of the twins had been diagnosed with IBD [88]. Combining a twin-control method with our method of prospective data collection on exposure would incorporate the genetic component in the assessment of environmental risk factors and improve internal validity of the study.

Conclusion

The objective of the present study was to assess the association between childhood infections (timing, frequency and type) and the risk of paediatric CD. This was done through a case-control study utilizing the RAMQ database, by matching hospital-confirmed cases with population controls. The information on childhood infections was extracted from the RAMQ database, using ICD-9 diagnostic codes entered by physicians as part of the billing process. A major strength of this study was the use of prospectively collected exposure information. Major caveats comprised of the absence of genetic information or family history of IBD, as well as the unverified validity of the ICD-9 codes in the database.

The results obtained concurred with our hypothesis, as childhood infections were associated with a lower risk of paediatric CD, particularly in the first 5 years of life. The results suggest that there is no dose-response effect corresponding to increased frequency of infections. We had hypothesized that enteric infections would be more protective than other infection types, as these are the regions predominantly affected by CD. This was not the case.

As a whole, our study provides further evidence supporting the popular hygiene hypothesis, whereby early exposure to antigens is protective against IBD as it likely confers immunological tolerance and decreases the risk of autoimmune diseases [88]. Assessing the validity of the hygiene hypothesis has important public health implications, as this theory leads to questioning on the role of hygiene, and potentially vaccine, as preventative measures of disease. As indicated by Rook in a 2011 literature review, the objective, pending of validity of the hygiene hypothesis, is not to purposely expose ourselves to pathogenic micro-organisms, but to further our understanding of the mechanisms leading to increasingly more prevalent chronic autoimmune diseases in order enable the development of more efficient treatment. Additionally, better understanding of the gene-environment interactions in these diseases could potentially lead to preventive measures, such as vaccines and probiotics, which would act to stimulate the immunoregulatory system with non-pathogenic micro-organisms early in childhood and induce immunological tolerance[1]. Thus, further research is needed to validate our findings in other populations, and to further investigate the specific types of infections which might be most prone to instigate protection against autoimmune disorders.

Bibliography

1. Rook GA. Hygiene hypothesis and autoimmune diseases. Clin Rev Allergy Immunol 2012;42:5-15.
2. Lowe AM, Roy PO, M BP, et al. Epidemiology of Crohn's disease in Quebec, Canada. Inflamm Bowel Dis 2009;15:429-35.
3. Economou M, Pappas G. New global map of Crohn's disease: Genetic, environmental, and socioeconomic correlations. Inflamm Bowel Dis 2008;14:709-20.
4. Ferkolj I, Gangl A, Galle PR, Vucelic B. Pathogenesis and Clinical Practice in Gastroenterology: Springer London, Limited; 2008.
5. Geboes K, Colombel JF, Greenstein A, et al. Indeterminate colitis: a review of the concept--what's in a name? Inflamm Bowel Dis 2008;14:850-7.
6. Cohen RD. Inflammatory Bowel Disease: Diagnosis and Therapeutics: Humana Press; 2011.
7. Cosnes J, Gower-Rousseau C, Seksik P, Cortot A. Epidemiology and natural history of inflammatory bowel diseases. Gastroenterology 2011;140:1785-94.
8. Baumgart DC, Sandborn WJ. Crohn's disease. Lancet 2012;380:1590-605.
9. Morrison G, Headon B, Gibson P. Update in inflammatory bowel disease. Aust Fam Physician 2009;38:956-61.
10. Al-Hawary M, Zimmermann EM. A new look at Crohn's disease: novel imaging techniques. Curr Opin Gastroenterol 2012;28:334-40.
11. Vermeire S, Van Assche G, Rutgeerts P. Classification of inflammatory bowel disease: the old and the new. Curr Opin Gastroenterol 2012;28:321-6.
12. De Cruz P, Kamm MA, Prideaux L, Allen PB, Moore G. Mucosal healing in Crohn's disease: a systematic review. Inflamm Bowel Dis 2013;19:429-44.
13. Kim SC, Ferry GD. Inflammatory bowel diseases in pediatric and adolescent patients: clinical, therapeutic, and psychosocial considerations. Gastroenterology 2004;126:1550-60.
14. Abraham BP, Mehta S, El-Serag HB. Natural history of pediatric-onset inflammatory bowel disease: a systematic review. J Clin Gastroenterol 2012;46:581-9.
15. Shikhare G, Kugathasan S. Inflammatory bowel disease in children: current trends. J Gastroenterol 2010;45:673-82.
16. Benchimol EI, Fortinsky KJ, Gozdyra P, Van den Heuvel M, Van Limbergen J, Griffiths AM. Epidemiology of pediatric inflammatory bowel disease: a systematic review of international trends. Inflamm Bowel Dis 2011;17:423-39.
17. Karlinger K, Gyorke T, Mako E, Mester A, Tarjan Z. The epidemiology and the pathogenesis of inflammatory bowel disease. Eur J Radiol 2000;35:154-67.
18. Bousvaros A, Sylvester F, Kugathasan S, et al. Challenges in pediatric inflammatory bowel disease. Inflamm Bowel Dis 2006;12:885-913.
19. Lashner BA, Loftus EV, Jr. True or false? The hygiene hypothesis for Crohn's disease. Am J Gastroenterol 2006;101:1003-4.
20. Baron S, Turck D, Leplat C, et al. Environmental risk factors in paediatric inflammatory bowel diseases: a population based case control study. Gut 2005;54:357-63.
21. Frolkis A, Dieleman LA, Barkema H, et al. Environment and the inflammatory bowel diseases. Can J Gastroenterol 2013;27:e18-24.
22. Cho JH, Brant SR. Recent insights into the genetics of inflammatory bowel disease. Gastroenterology 2011;140:1704-12.

23. Ananthakrishnan AN. Environmental triggers for inflammatory bowel disease. Curr Gastroenterol Rep 2013;15:302.
24. Brant SR, Wang MH, Rawsthorne P, et al. A population-based case-control study of CARD15 and other risk factors in Crohn's disease and ulcerative colitis. Am J Gastroenterol 2007;102:313-23.
25. Gearry RB, Richardson AK, Frampton CM, Dodgshun AJ, Barclay ML. Population-based cases control study of inflammatory bowel disease risk factors. J Gastroenterol Hepatol 2010;25:325-33.
26. Economou M, Trikalinos TA, Loizou KT, Tsianos EV, Ioannidis JP. Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: a metaanalysis. Am J Gastroenterol 2004;99:2393-404.
27. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 2012;491:119-24.
28. Bernstein CN, Rawsthorne P, Cheang M, Blanchard JF. A population-based case control study of potential risk factors for IBD. Am J Gastroenterol 2006;101:993-1002.
29. Ng SC, Bernstein CN, Vatn MH, et al. Geographical variability and environmental risk factors in inflammatory bowel disease. Gut 2013;62:630-49.
30. Green C, Elliott L, Beaudoin C, Bernstein CN. A population-based ecologic study of inflammatory bowel disease: searching for etiologic clues. Am J Epidemiol 2006;164:615-23; discussion 24-8.
31. Molodecky NA, Soon IS, Rabi DM, et al. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology 2012;142:46-54 e42; quiz e30.
32. Castiglione F, Diaferia M, Morace F, et al. Risk factors for inflammatory bowel diseases according to the "hygiene hypothesis": a case-control, multi-centre, prospective study in Southern Italy. J Crohns Colitis 2012;6:324-9.
33. Fishbein AB, Fuleihan RL. The hygiene hypothesis revisited: does exposure to infectious agents protect us from allergy? Curr Opin Pediatr 2012;24:98-102.
34. Strachan DP. Hay fever, hygiene, and household size. BMJ 1989;299:1259-60.
35. Berin MC, Mayer L. Can we produce true tolerance in patients with food allergy? J Allergy Clin Immunol 2013;131:14-22.
36. Koloski NA, Bret L, Radford-Smith G. Hygiene hypothesis in inflammatory bowel disease: a critical review of the literature. World J Gastroenterol 2008;14:165-73.
37. de Silva HJ, de Silva NR, de Silva AP, Jewell DP. Emergence of inflammatory bowel disease 'beyond the West': do prosperity and improved hygiene have a role? Trans R Soc Trop Med Hyg 2008;102:857-60.
38. Garn H, Renz H. Epidemiological and immunological evidence for the hygiene hypothesis. Immunobiology 2007;212:441-52.
39. Gearry RB, Dodgshun AJ. The "hygiene hypothesis" in IBD. J Crohns Colitis 2012;6:869; author reply 70.
40. Amre DK, Lambrette P, Law L, et al. Investigating the hygiene hypothesis as a risk factor in pediatric onset Crohn's disease: a case-control study. Am J Gastroenterol 2006;101:1005-11.
41. Wills-Karp M, Santeliz J, Karp CL. The germless theory of allergic disease: revisiting the hygiene hypothesis. Nat Rev Immunol 2001;1:69-75.

42. Okada H, Kuhn C, Feillet H, Bach JF. The 'hygiene hypothesis' for autoimmune and allergic diseases: an update. Clin Exp Immunol 2010;160:1-9.
43. Murk W, Risnes KR, Bracken MB. Prenatal or early-life exposure to antibiotics and risk of childhood asthma: a systematic review. Pediatrics 2011;127:1125-38.
44. Prescott S, Nowak-Wegrzyn A. Strategies to prevent or reduce allergic disease. Ann Nutr Metab 2011;59 Suppl 1:28-42.
45. Davis RL, Bohlke K. Measles vaccination and inflammatory bowel disease: controversy laid to rest? Drug Saf 2001;24:939-46.
46. Bernstein CN. Epidemiologic clues to inflammatory bowel disease. Curr Gastroenterol Rep 2010;12:495-501.
47. Rautava S, Luoto R, Salminen S, Isolauri E. Microbial contact during pregnancy, intestinal colonization and human disease. Nat Rev Gastroenterol Hepatol 2012;9:565-76.
48. Polgar S, Thomas SA. Introduction to research in the health sciences. 5 ed. United Kingdom: Elsevier; 2008.
49. Régie de l'Assurance Maladie du Québec: Clientèles. Gouvernement du Québec, 2003. (Accessed March, 2012, at http://www.ramq.gouv.qc.ca/fr/.)
50. Population, Québec et Canada, 1851-2001. Institut de la Statistique Québec, 2012. (Accessed March 2012, at http://www.stat.gouv.qc.ca/.)
51. Monfared AA, Lelorier J. Accuracy and validity of using medical claims data to identify episodes of hospitalizations in patients with COPD. Pharmacoepidemiol Drug Saf 2006;15:19-29.
52. Manuel des médecins spécialistes: Régie de l'assurance maladie du Québec; 2013.
53. Richards C. Coding basics : medical billing and reimbursement fundamentals. 1st ed. Florence, KY: Delmar; 2009.
54. Pinner RW, Teutsch SM, Simonsen L, et al. Trends in infectious diseases mortality in the United States. JAMA 1996;275:189-93.
55. Simonsen L, Conn LA, Pinner RW, Teutsch S. Trends in infectious disease hospitalizations in the United States, 1980-1994. Arch Intern Med 1998;158:1923-8.
56. Rothman KJ. Epidemiology : an introduction. 2nd ed. New York, NY: Oxford University Press; 2012.
57. Heikenen JB, Werlin SL, Brown CW, Balint JP. Presenting symptoms and diagnostic lag in children with inflammatory bowel disease. Inflamm Bowel Dis 1999;5:158-60.
58. How Postal Codes Map to Geographic Areas. Statistics Canada, 2007. (Accessed May, 2013, at http://www.statcan.gc.ca/pub/92f0138m/2007001/4054931-eng.htm.)
59. Schlesselman JJ, Stolley PD. Case-control studies : design, conduct, analysis. New York: Oxford University Press; 1982.
60. Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. New York: Wiley; 2000.
61. Gauderman W, Morrison J. QUANTO In. 1.1 ed; 2006:A computer program for power and sample size calculations for genetic-epidemiology studies.
62. Étude : Tendances récentes des infections des voies respiratoires supérieures, des infections de l'oreille et de l'asthme chez les enfants. Statistique Canada, 2010. (Accessed March, 2012, at http://www.statcan.gc.ca/daily-quotidien/101117/dq101117b-fra.htm.)

63. Behr MA, Bruere P, Oxlade O. Global rates of Crohn's disease. Inflamm Bowel Dis 2008;14:1170-2.
64. Griffiths AM. Specificities of inflammatory bowel disease in childhood. Best Pract Res Clin Gastroenterol 2004;18:509-23.
65. Duggan AE, Usmani I, Neal KR, Logan RF. Appendicectomy, childhood hygiene, Helicobacter pylori status, and risk of inflammatory bowel disease: a case control study. Gut 1998;43:494-8.
66. Elliott DE, Li J, Blum A, et al. Exposure to schistosome eggs protects mice from TNBS-induced colitis. Am J Physiol Gastrointest Liver Physiol 2003;284:G385-91.
67. Gent AE, Hellier MD, Grace RH, Swarbrick ET, Coggon D. Inflammatory bowel disease and domestic hygiene in infancy. Lancet 1994;343:766-7.
68. Khan WI, Blennerhasset PA, Varghese AK, et al. Intestinal nematode infection ameliorates experimental colitis in mice. Infect Immun 2002;70:5931-7.
69. Moreels TG, Pelckmans PA. Gastrointestinal parasites: potential therapy for refractory inflammatory bowel diseases. Inflamm Bowel Dis 2005;11:178-84.
70. Moreels TG, Nieuwendijk RJ, De Man JG, et al. Concurrent infection with Schistosoma mansoni attenuates inflammation induced changes in colonic morphology, cytokine levels, and smooth muscle contractility of trinitrobenzene sulphonic acid induced colitis in rats. Gut 2004;53:99-107.
71. Summers RW, Elliott DE, Weinstock JV. Is there a role for helminths in the therapy of inflammatory bowel disease? Nat Clin Pract Gastroenterol Hepatol 2005;2:62-3.
72. Feeney MA, Murphy F, Clegg AJ, Trebble TM, Sharer NM, Snook JA. A case-control study of childhood environmental risk factors for the development of inflammatory bowel disease. Eur J Gastroenterol Hepatol 2002;14:529-34.
73. Ekbom A, Adami HO, Helmick CG, Jonzon A, Zack MM. Perinatal risk factors for inflammatory bowel disease: a case-control study. Am J Epidemiol 1990;132:1111-9.
74. Wurzelmann JI, Lyles CM, Sandler RS. Childhood infections and the risk of inflammatory bowel disease. Dig Dis Sci 1994;39:555-60.
75. Van Kruiningen HJ, Joossens M, Vermeire S, et al. Environmental factors in familial Crohn's disease in Belgium. Inflamm Bowel Dis 2005;11:360-5.
76. Thompson NP, Pounder RE, Wakefield AJ. Perinatal and childhood risk factors for inflammatory bowel disease: a case-control study. Eur J Gastroenterol Hepatol 1995;7:385-90.
77. Gilat T, Hacohen D, Lilos P, Langman MJ. Childhood factors in ulcerative colitis and Crohn's disease. An international cooperative study. Scand J Gastroenterol 1987;22:1009-24.
78. Sands BE. From symptom to diagnosis: clinical distinctions among various forms of intestinal inflammation. Gastroenterology 2004;126:1518-32.
79. Lennard-Jones JE. Classification of inflammatory bowel disease. Scand J Gastroenterol Suppl 1989;170:2-6; discussion 16-9.
80. Lopez-Serrano P, Perez-Calle JL, Perez-Fernandez MT, Fernandez-Font JM, Boixeda de Miguel D, Fernandez-Rodriguez CM. Environmental risk factors in inflammatory bowel diseases. Investigating the hygiene hypothesis: a Spanish case-control study. Scand J Gastroenterol 2010;45:1464-71.
81. Pinsk V, Lemberg DA, Grewal K, Barker CC, Schreiber RA, Jacobson K. Inflammatory bowel disease in the South Asian pediatric population of British Columbia. Am J Gastroenterol 2007;102:1077-83.

82. Molodecky NA, Panaccione R, Ghosh S, Barkema HW, Kaplan GG. Challenges associated with identifying the environmental determinants of the inflammatory bowel diseases. Inflamm Bowel Dis 2011;17:1792-9.
83. Grieci T, Butter A. The incidence of inflammatory bowel disease in the pediatric population of Southwestern Ontario. J Pediatr Surg 2009;44:977-80.
84. Population Totale. Ville de Montréal. (Accessed April, 2013, at http://ville.montreal.qc.ca/portal/page?_pageid=6897,67887840&_dad=portal&_schema=PORTAL.)
85. Hamel D, Pampalon R, Gamache P. Guide d'utilisation du programme d'assignation de l'indice de défavorisation 2006; 2009.
86. 2001 Census of Canada. Statistics Canada. (Accessed March, 2012, at http://www12.statcan.gc.ca/english/census01.)
87. de W, de LJ, Baanders-Vanhalewijn EA. On the bimodal age distribution of mammary carcinoma. Br J Cancer 1960;14:437-48.
88. Ng SC, Woodrow S, Patel N, Subhani J, Harbord M. Role of genetic and environmental factors in British twins with inflammatory bowel disease. Inflamm Bowel Dis 2012;18:725-36.
89. Kleinbaum DG, Klein M, Pryor ER. Logistic regression : a self-learning text. 3rd ed. New York: Springer; 2010.
90. Lesson 3: Logistic Regression Diagnostics. Institute for Digital Research and Education. (Accessed Februrary, 2013, at http://www.ats.ucla.edu/stat/stata/webbooks/logistic/chapter3/statalog3.htm.)

Appendices

Appendix I

Figure 6: Simonsen infection categories for ICD-9 codes[55]

Table 1. Groups of Infectious Diseases and Related *ICD-9* Codes*

Infectious Disease Grouping	*ICD-9* Codes
Tuberculosis	010-018, 137
Meningitis	027.0, 036, 320.0-321.3, 321.8
Septicemia	038
HIV and AIDS	042-044, 279.1
Hepatobiliary disease	070, 095.3, 573.1-573.2, 576.1
Selected perinatal infections	090, 770.0, 771
Mycoses	110-118
Infections of the heart	093, 391, 392.0, 393, 394.1, 395.0-395.2, 397.1, 397.9, 398, 421, 422.0, 424.9
Upper respiratory tract infections	032.0-032.3, 034.0, 098.6, 101, 460-465, 473.0-474.0, 475
Lower respiratory tract infections	022.1, 031.0, 033, 095.1, 466, 480-487, 510, 511.1, 513, 517.1
Abdominal and rectal infections (appendicitis, peritonitis, and abscess of the intestine)	095.2, 098.7, 540-542, 566, 567.0-567.2, 569.5
Kidney, urinary tract, and bladder infections	095.4, 099.4, 590, 595.0, 597, 598.0, 599.0
Cellulitis	680-686
Enteric infections	001-009, 022.2
Infection and inflammatory reaction to prosthetic devices (including cardiac, vascular, neurologic, and orthopedic devices)	996.6
Postoperative infection	998.5
Viral CNS infections	045-049
Pelvic inflammatory disease	614.0-614.5, 616.0-616.1, 616.3-616.4
Oral infections	522.4-522.5, 522.7, 523.0, 523.3-523.4, 527.3, 528.0-528.3
Osteomyelitis	730
Infections in pregnancy	634.0, 635.0, 636.0, 637-0, 638.0, 639.0, 646.5, 646.6, 647, 655.3, 658.4, 659.3, 670, 675

*ICD-9 *indicates* International Classification of Diseases, Ninth Revision; *HIV, human immunodeficiency virus; AIDS, acquired immunodeficiency syndrome; and CNS, central nervous system.*

Appendix II

Steps of the Hosmer-Lemeshow method of model building, as applied to this study:

Step 1: Assessment of independent variables (univariate analysis)

A combination of scientific reasoning and statistical interpretation must be used to determine which variables will be retained in the final multivariate model [60]. The main exposure variable (childhood infections) is included in the model regardless of statistical significance in the preliminary analyses. The other variables are confounders, meaning they must influence the association between the main exposure variable (childhood infections) and the study outcome (CD). It was decided a priori that the revenue variable would be included in the final model regardless of the results of the model building, as it has been identified in the scientific literature as a confounder. Additionally, the number of medical visits and the number of missing ICD-9 codes were assessed for confounding using simple logistic regression in combination with descriptive statistics, was used to evaluate confounding by the above-mentioned variables.

The following table shows the results of the simple logistic regression (conditional logistic regression for matched observations). The revenue variable was divided into quartiles and the number of medical visits and number of missing codes, kept as continuous variables. Potential confounders were used as the single exposure in the regression, with CD onset as the outcome. The number of medical visits and missing ICD-9 codes was calculated from birth up to 2 years prior to CD diagnosis.

Table II: Simple Logistic Regression for Assessment of Confounding

Potential Confounding Variable	OR	95% IC	p-value
Revenue	1.110	0.984-1.253	0.090
# medical visits	1.004	1.001-1.006	0.003
# missing ICD-9 codes	1.005	0.999-1.011	0.127

Step 2: Selection of potential confounders

The results of the univariate analysis were assessed to identify the potential confounders to retain at this stage of the model building. As suggested by Hosmer and Lemeshow, variables with a p-value of less than 0.25 were kept for further consideration [60]. At this stage, all three variables were retained in the model, as all three differed significantly between cases and controls in the

descriptive analysis, and all three were associated with the study outcome in the simple logistic regression in Step 1 (at a significance level of $p<0.25$). In addition, household revenue was a pre-determined confounder.

The full logistic model, at this stage, was the following:

$$\log itP(X) = \alpha + \beta(INF) + \gamma_1(REV) + \gamma_2(VIS) + \gamma_3(MCO)$$

where INF=childhood infections, REV=revenue, VIS=number of medical visits and MCO=number of missing codes.

Step 3: Verification of the effect of potential confounders

At this stage, the variables retained in step 2 were tested using a Wald test, confirmed with a likelihood ratio test [89]. This was done to ensure that the variables contribute sufficiently to the model to justify their presence [60].

Wald test: To obtain the Wald test statistic, a logistic regression model including the potential confounder must be computed. The coefficient of the potential confounder is then divided by its standard error. The result is squared, and corresponds to a chi-square statistical test, with one degree of freedom. Only one variable at a time can be assessed using this test [89].

Likelihood Ratio (LR) test: This statistical test is used to compare two models. A full model must first be created, containing the potential confounding variable(s) being assessed. A second reduced model is created, excluding the variables being tested. The difference between the maximum likelihood ratio statistics for both models is computed, and a chi-square test performed. The number of degrees of freedom is equal to the number of parameters that differ between the two models. The null hypothesis for LR tests is that the difference between the two models is not significant, and thus that the potential confounder tested does not influence the model sufficiently to be considered a confounder.

Table III: Results of Wald and ML Tests for Potential Confounders

Potential Confounding Variable	Wald test		ML test	
	Wald statistic	p-value	Difference of log likelihood	p-value
Revenue	0.085	0.770	2.905	0.088
# medical visits	29.243	0.000	30.201	0.000
# missing ICD-9 codes	2.468	0.116	2.415	0.120

Based on the Wald and ML tests, the revenue and the number of medical visits were kept in the model. The Wald and ML tests for these variables were significant, indicating an important impact of the variables on the association between the exposure (infections) and the outcome (CD diagnosis). The variable for the number of missing ICD codes was dropped from the model, as the statistical tests show that this variable did not seem associated with the outcome, CD.

The "preliminary main effects model":

$$\log itP(X) = \alpha + \beta(INF) + \gamma_1(REV) + \gamma_2(VIS)$$

Step 4: Assessment of linearity of continuous variables

At this stage, the variables (both confounders and main exposure variable) were assessed for linearity, to decide whether they should be kept as continuous variables or not (an assumption of the logistic model). The continuous variables were plotted against their logit. The curves were assessed for linearity, and the non-linear variables were transformed when necessary.

The Hosmer-Lemeshow procedure for creating the logit plot was carried out in SPSS, which consists of plotting the regression coefficients of the percentiles of the continuous variables against the middle values of the percentiles [60].

The results of the linearity test were confirmed using the LINKTEST function in STATA. Once the non-linear variables were categorized, the LINKTEST was used to verify that the model was well-specified, meaning that the variables used were predictors of the outcome, and that no major predictor was omitted [90].

The following linearity graphs were generated:

Figure 7: Assessment of the Linearity Assumption for Revenue

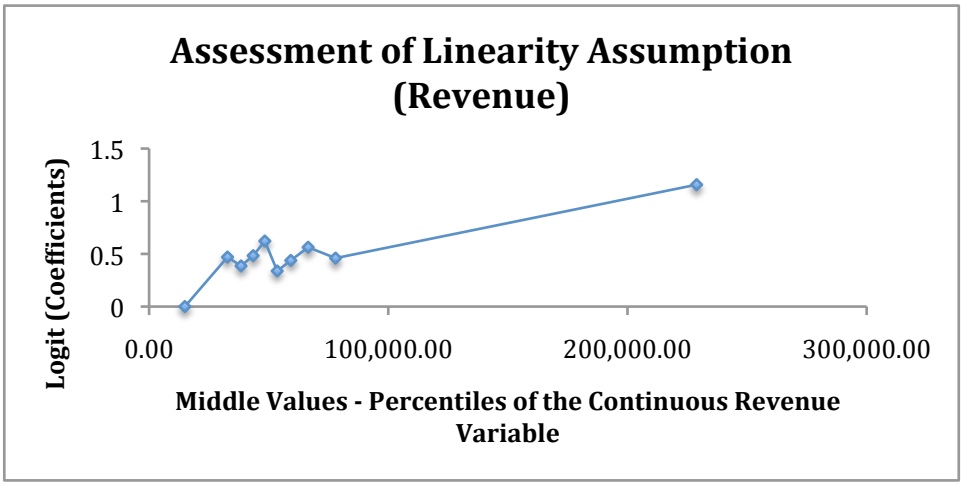


Figure 8: Assessment of the Linearity Assumption for the Number of Visits

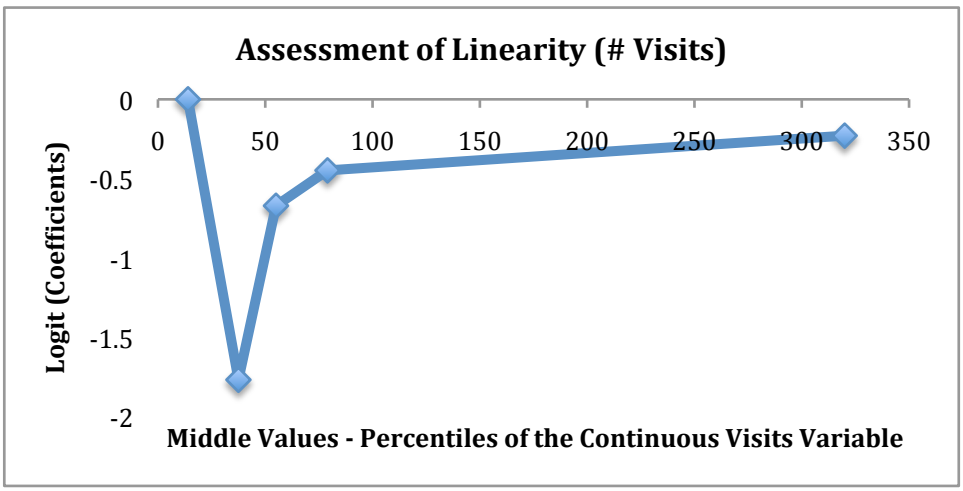
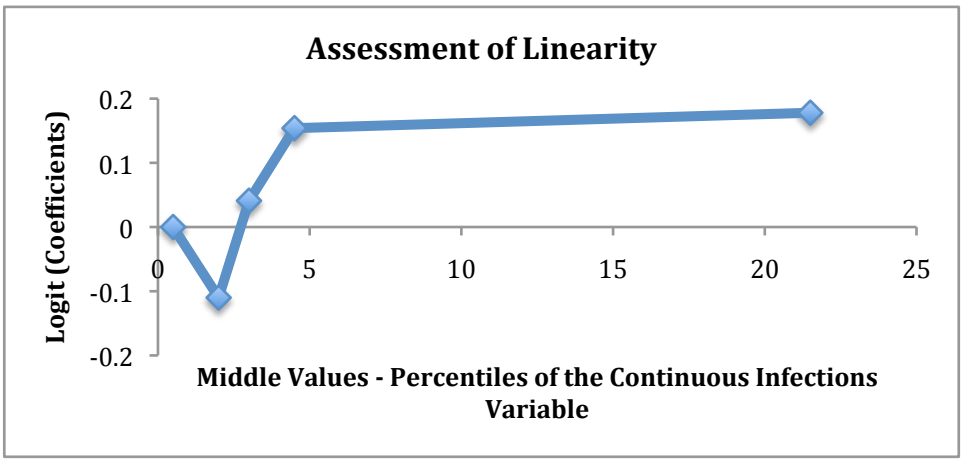


Figure 9: Assessment of the Linearity Assumption for the Number of Infections



All graphs display non-linear curves, indicating that the variables should not be kept as continuous. The revenue and number of visits variables were re-coded as quartiles. The Infections variable was converted to a categorical variable (0 infections, 1-5 infections, >5 infections).

The linearity assumption for these new categorical variables and the specificity of the resulting model were then assessed, using the LINKTEST function in STATA. This test yielded a significant _hat value (0.039) and a non significant _hatsqu value (0.132). As the _hat value represents the predictive power of the model, a significant result points to a well-specified model.

Model:

$$\log itP(X) = \alpha + \beta(D:INF) + \gamma_1(Q:REV) + \gamma_2(Q:VIS)$$

where D=dichotomized variable; Q:quartiles

Step 5: Assessing potential interaction terms

Interactions between variables signify that the effect of one variable is not the same, depending on the value of a second variable. According to the Hosmer-Lemeshow method of model building, a list of potential interaction terms should be drafted *a priori,* before being statistically tested. The interaction terms should be reasonable from a clinical point of view [60].

From the variables possibly included in the model (childhood infections, the number of medical visits and the number of missing ICD-9 codes), there were no variables for which it might have made sense clinically to assess interaction.

Appendix III

Diagnostic Measures Applied to the Model

Goodness of Fit

Once the model has been determined, it must be assessed to verify that it adequately fits the observations. Ideally, the model would predict the data collected perfectly [89]. As this is not feasible, we have used the log-likelihood chi-square test to compare our model with the "empty" model (containing only the intercept) [90].

The STATA log-likelihood statistic for the model was 50.65. The p-value for the statistic was 0.0000. The model is highly significant, which means that the model fit the data well.

Collinearity

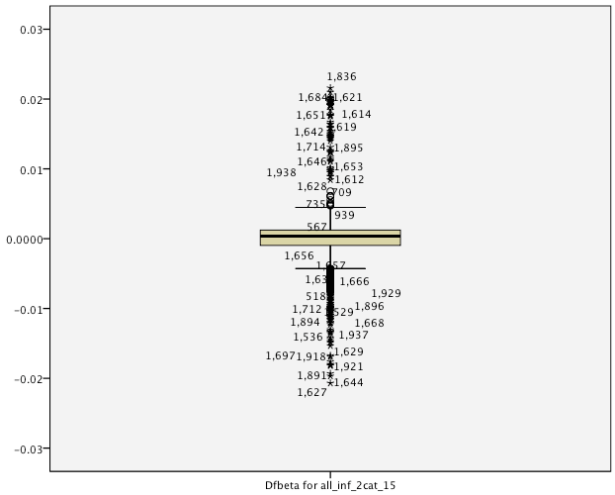
Multicollinearity occurs when a variable in the model can be predicted from another variable in the model [89]. This causes unreliable regression coefficients, and large variances. In order to assess multicollinearity, the variances of the coefficients were compared. An abnormally large variance for one of the coefficients points to possible multocollinearity. As no coefficient displayed an abnormally high variance, no multicollinearity issues between variables were found.

Extreme Observations

In this final stage of the model building, the observations were assessed to detect those with extreme values, which could have had a significant impact on the coefficients of the variables in the model [89]. The Df beta values (a measure of the change that occurs in the coefficients in the model if the observation were removed) for each observation were saved for the model, and plotted as a histogram and stem-and-leaf plot for each independent variable. The outliers were identified and investigated.

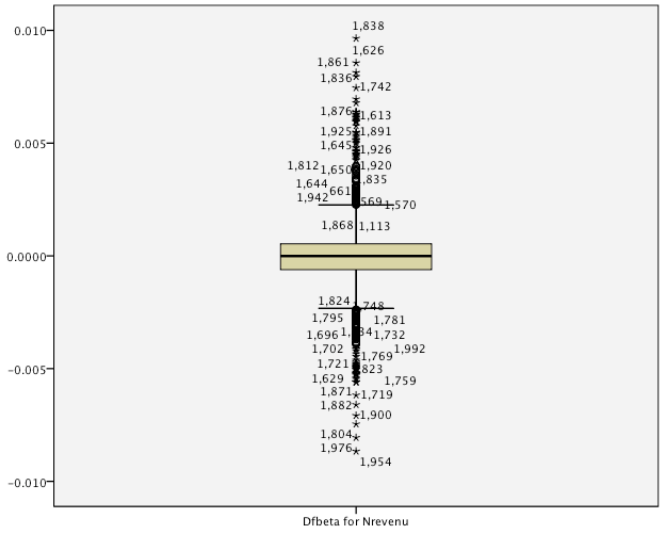
The df beta values were plotted for each for each variable as histograms, stem-and-leaf plots and box plots.

Figure 10: Box Plot of Df Beta Values for the Infection (Yes/No) Variable



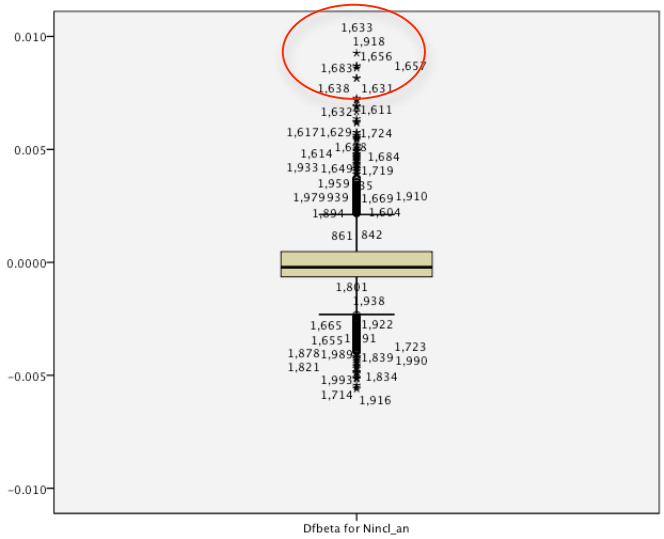
The infections variable did not seem to have any extreme values.

Figure 11: Box Plot of Df Beta Values for the Revenue Variable



The Revenue variable also did not seem to contain extreme observations. Though the plot showed that the observations were scattered across a wide range of values, this is to be expected from a household income variable, and thus no observations were excluded.

Figure 12: Box Plot of the Df Beta Values for the Number of Visits Variable



From this box plot, it appeared that 4 observations in particular could be more influential on the coefficients than the rest (where the number of visits exceeded 400). The 4 observations were identified. The logistic regression was run, omitting these four cases. The purpose was to evaluate whether these extreme observations significantly influenced the regression coefficient:
The coefficients for the variable # visits were compared, between a model including the strata with the extreme observations, and a model without.

Table IV: Comparison Between the Coefficients of the Number of Visits Variable

Quartile	B (all observations)	B (strata with extreme observations excluded)	Difference between the coefficients
1	.811	0.819	1.0%
2	.936	0.925	1.2%
3	1.288	1.271	1.3%

As the extreme variables influenced the regression coefficient for the # visits variable only by around 1%, the strata containing the extreme observations were retained, to avoid losing power.

Appendix IV

Table V: Frequency of Infection Types Amongst Cases and Controls

Infection type	Definition of medical episode	Case/Control	0 infection n (%)	> 1 infection n (%)
Abdominal and Rectal infections	15 days	Cases	385 (94.1%)	24 (5.9%)
		Controls	1593 (98.3%)	28 (1.7%)
	30 days	Cases	401 (98.0%)	8 (2.0%)
		Controls	1607 (99.1%)	14 (0.9%)
Cellulitis	15 days	Cases	338 (82.6%)	71 (17.4%)
		Controls	1359 (83.8%)	262 (16.2%)
	30 days	Cases	338 (82.6%)	71 (17.4%)
		Controls	1366 (84.3%)	255 (15.7%)
Enteric infections	15 days	Cases	389 (95.1%)	20 (4.9%)
		Controls	1548 (95.5%)	73 (4.5%)
	30 days	Cases	390 (95.4%	19 (4.6%)
		Controls	1549 (95.6%)	72 (4.4%)
Heart infections	15 days	Cases	409 (100%)	0 (0%)
		Controls	1616 (99.7%)	5 (0.3%)
	30 days	Cases	409 (100%)	0 (0%)
		Controls	1616 (99.7%)	5 (0.3%)
Hepato-billiary infections	15 days	Cases	408 (99.8%)	1 (0.2)
		Controls	1619 (99.9%)	2 (0.1%)
	30 days	Cases	408 (99.8%)	1 (0.2)
		Controls	1619 (99.9%)	2 (0.1%)
Kidney and urinary infections	15 days	Cases	381 (93.2%)	28 (6.8%)
		Controls	1490 (91.9%)	131 (8.1%)
	30 days	Cases	382 (93.4%)	27 (6.6%)
		Controls	1493 (92.1%)	128 (7.9%)
Lower	15 days	Cases	279 (68.2%)	130 (31.8%)

respiratory infections		Controls	1139 (70.3)	482 (29.8%)
	30 days	Cases	282 (68.9%)	127 (31.0%)
		Controls	1145 (70.6%)	476 (29.4%)
Infections of the Meninges	15 days	Cases	409 (100%)	0 (0%)
		Controls	1619 (99.9%)	2 (0.1%)
	30 days	Cases	409 (100%)	0 (0%)
		Controls	1619 (99.9%)	2 (0.1%)
Mycoses infections	15 days	Cases	369 (90.2%)	40 (9.8%)
		Controls	1477 (91.1%)	144 (8.9%)
	30 days	Cases	370 (90.5%)	39 (9.5%)
		Controls	1480 (91.3%)	141 (8.7%)
Oral infections	15 days	Cases	391 (95.6%)	18 (4.4%)
		Controls	1548 (95.5%)	73 (4.5%)
	30 days	Cases	391 (95.6%)	18 (4.4%)
		Controls	1548 (95.5%)	73 (4.5%)
Pelvic infections	15 days	Cases	388 (94.9%)	21 (5.1%)
		Controls	1550 (95.6%)	71 (4.4%)
	30 days	Cases	389 (95.1%)	20 (4.9%)
		Controls	1551 (95.7%)	70 (4.3%)
Post-operational infections	15 days	Cases	408 (99.8%)	1 (0.2%)
		Controls	1621 (100%)	0 (0%)
	30 days	Cases	408 (99.8%)	1 (0.2%)
		Controls	1621 (100%)	0 (0%)
Infections during pregnancy	15 days	Cases	408 (99.8%)	1 (0.2%)
		Controls	1620 (99.9%)	1 (0.1%)
	30 days	Cases	408 (99.8%)	1 (0.2%)
		Controls	1620 (99.9%)	1 (0.1%)
Septicaemia	15 days	Cases	408 (99.8%)	1 (0.2%)
		Controls	1621 (100%)	0 (0%)

	30 days	Cases	408 (99.8%)	1 (0.2%)
		Controls	1621 (100%)	0 (0%)
Tuberculosis	15 days	Cases	407 (99.5%)	2 (0.5%)
		Controls	1607 (99.1%)	14 (0.9%)
	30 days	Cases	407 (99.5%)	2 (0.5%)
		Controls	1607 (99.1%)	14 (0.9%)
Upper respiratory infections	15 days	Cases	112 (27.4%)	297 (72.6%)
		Controls	487 (30.0%)	1134 (70.0%)
	30 days	Cases	112(27.4%)	297 (72.6%)
		Controls	492 (30.4%)	1129 (69.6%)
Viral-CNS infections	15 days	Cases	403 (98.5%)	6 (1.5%)
		Controls	1576 (97.2%)	45 (2.8%)
	30 days	Cases	403 (98.5%)	6 (1.5%)
		Controls	1581 (97.5%)	40 (2.5%)

Appendix V

Sensitivity analysis:

Results of crude and adjusted analyses using a medical episode definition of infections separated by a minimum of 30 days:

Table VI: Results of the Sensitivity Analysis for the Frequency and Temporality of Infections

A) Frequency of Infections				
	Crude Analysis		**Adjusted Analysis***	
	OR point estimate [95% CI]	**p-value**	**OR point estimate [95% CI]**	**p-value**
0 infections (ref)				
1-5 infections	0.88 [0.65-1.20]	0.429	0.67 [0.48-0.94]	0.018
>5 infections	1.34 [0.92-1.94]	0.126	0.77 [0.50-1.18]	0.1231
Any infection	0.96 [0.71-1.30]	0.791	0.68 [0.49-0.95]	0.023
B) Temporality of Infections				
Infections during first year of life				
	Crude Analysis		**Adjusted Analysis***	
	OR point estimate [95% CI]	**p-value**	**OR point estimate [95% CI]**	**p-value**
0 infections (ref)				
1-5 infections	0.99 [0.75-1.31]	0.958	0.83 [0.62-1.11]	0.219
>5 infections**	N/A	N/A	N/A	N/A
Any infection	0.98 [0.74-1.30]	0.909	0.96 [0.71-1.30]	0.796
Infections during first 5 years of life				
	Crude Analysis		**Adjusted Analysis***	
	OR point estimate [95% CI]	**p-value**	**OR point estimate [95% CI]**	**p-value**
0 infections (ref)				
1-5 infections	0.97 [0.76-1.23]	0.788	0.74 [0.57-0.96]	0.024
>5 infections	1.09 [0.67-1.76]	0.732	0.61 [0.37-1.03]	0.064
Any infection	0.98 [0.76-1.25]	0.845	0.88 [0.67-1.14]	0.324

*Adjusted for revenue and total number of physician visits (dependant on time period assessed)

**Only 4 subjects in this category

Table VII: Results of the Sensitivity Analysis for the Type of Infection

	Crude Analysis		Adjusted Analysis*	
	OR [95% CI]	**p-value**	**OR [95% CI]**	**p-value**
Abdominal and rectal	2.23 [0.93-5.35]	0.0612	0.96 [0.34-2.74]	0.937
Cell	1.14 [0.85-1.55]	0.3950	0.56 [0.35-0.89]	0.014
Enteric	1.04 [0.62-1.75]	0.8711	0.38 [0.18-0.80]	0.011
Kidney and urinary tract	0.83 [0.54-1.27]	0.3873	0.38 [0.21-0.71]	0.002
Lower respiratory	1.09 [0.86-1.38]	0.4864	0.64 [0.42-0.96]	0.031
Mycoses	1.11 [0.76-1.62]	0.5796	0.56 [0.33-0.94]	0.030
Oral	0.98 [0.58-1.65]	0.9359	0.35 [0.17-0.73]	0.005
Pelvic	1.15 [0.67-1.93]	0.5984	0.08 [0.23-1.00]	0.051
Upper respiratory	1.19 [0.91-1.55]	0.1945	0.67 [0.48-0.94]	0.019
Viral - CNS	0.58 [0.24-1.41]	0.230	0.25 [0.09-0.68]	0.019
TBC	0.55 [0.12-2.48]	0.436	0.20 [0.04-1.04]	0.007

*Adjusted for revenue and total number of physician visits

Appendix VI

Crude and adjusted analysis of the hygiene hypothesis variable

Result of the analysis for the hygiene hypothesis

A dichotomous variable was created:

1: no infections in the first year of life, with subsequent exposure to childhood infections

0: all other participants

Table VIII: Results of the Hygiene Hypothesis Sub-analysis

	Crude Analysis		**Adjusted Analysis**	
	OR [95%CI]	p-value	OR [95%CI]	p-value
0 (ref category)				
1	0.99 [0.78-1.25]	0.913	0.93 [0.73-1.19]	0.572

Analysis for the hygiene hypothesis variable, using the 30-day medical episode definition

	Crude Analysis		**Adjusted Analysis**	
	OR [95%CI]	p-value	OR [95%CI]	p-value
0 (ref category)				
1	0.99 [0.78-1.25]	0.913	0.93 [0.73-1.19]	0.572