The Role of Infections in the Etiology of Childhood Acute Lymphoblastic Leukemia

by

Jeremiah Hwee

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Department of Public Health Sciences
Dalla Lana School of Public Health
University of Toronto

The Role of Infections in the Etiology of Childhood Acute Lymphoblastic Leukemia

Jeremiah Hwee
Doctor of Philosophy
Department of Public Health Sciences
Dalla Lana School of Public Health
University of Toronto
2018

Abstract

Childhood acute lymphoblastic leukemia (ALL) is the most common childhood cancer. However, the etiology of childhood ALL is uncertain. An infectious trigger for ALL is hypothesized based on evidence from biological and epidemiologic studies. The goal of the dissertation was to assess the relationship between prior infections and the development of childhood ALL. In a systematic review and meta-analysis, no overall relationship between prior infections and childhood ALL could be identified (odds ratio, OR=1.10, 95% confidence interval, CI 0.95-1.28). The systematic review showed most studies that used self-reported data to measure infections were susceptible to recall bias. Thus, administrative data may be particularly useful in furthering our understanding of the infectious etiology of ALL. Using electronic medical records as the reference standard, a study was conducted to assess the criterion validity of administrative databases to identify infectious syndromes in children aged 0-18 years from Ontario, Canada. Administrative billings codes for an infection (respiratory, skin and soft tissue, gastrointestinal, urinary tract or otitis externa) demonstrated moderate sensitivity (0.74, 95%CI 0.70-0.77), and high specificity (0.95, 95%CI 0.93-0.96), positive predictive value (0.87, 95%CI 0.84-0.90), and negative predictive value (0.88, 95%CI 0.86-0.89). Finally, the association between prior infections and the development of childhood ALL,

using health administrative data, was conducted applying the findings from the validation study to define and measure infections. Overall, having >2 infections per year increased odds of ALL by 43% compared to children with ≤0.25 infections per year in Ontario. Infections occurring between 1 to 1.5 years of life may be a critical period as having an infection in this window increased the odds of ALL by 20%. Certain infections such as respiratory and invasive infections may be more important than other infections in the development of ALL. The accumulated insights from each study was used to inform subsequent objectives, resulting in a unified dissertation that found infections have a role in the etiology of childhood ALL. Future work should extend the empirical study investigate the critical period between 1 to 1.5 years by collecting detailed infection data and other exposures that begin around this period.

Acknowledgements

Thank you to everyone that has helped me get here. This wouldn't be possible without you! I would especially like to thank my co-supervisors, Dr. Jason D. Pole, who has always provided thoughtful feedback, supported, encouraged and mentored me throughout this work and in my career; and to Dr. Lillian Sung, thank you for constantly challenging me to improve, your exceptional guidance, support and big picture thinking. To my committee members Dr. Rinku Sutradhar, thank you for challenging me, your mentorship and open-mindedness; and to Dr. Jeffrey C. Kwong, thank you for your thoughtful feedback and diligence throughout. You have all inspired me in your own ways and I am grateful to have worked with and learned from each of you. This could not have been possible without you.

Thank you to the Canadian Institutes for Health Research, the Pediatric Oncology Group of Ontario, and the University of Toronto Open Fellowship for supporting this work and my PhD studies.

To my fellow colleagues, thank you! Our studying, writing and extracurricular activities made this a fun journey. I have learned so much from each of you.

To my family, thank you for your constant love, support and the home cooked meals throughout my dissertation. Last, but not least, I would like to thank my lovely wife, Dr. Shirley Hu. Your love, humour, positivity, encouragement and support through this journey pushed me to be better. I could not have done this without you. You are my inspiration.

Thank you all!

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1 : Introduction and Objectives

Cancer is the leading cause of disease-related death among children 1-14 years of age in North America.[1,2] Leukemia accounts for 32% of all cancers in Canada among children aged 0-14 years.[2,3] As of 2015, there are almost 8,000 5-year pediatric leukemia cancer survivors in Canada.[4] Acute lymphoblastic leukemia (ALL) accounts for over 80% of leukemias and is the most common childhood cancer in high-income countries, including Canada.[2,3,5-7] Peak incidence for childhood ALL occurs between the ages of 1 to 4 years.[5,8] In Canada, there were ~230 incident ALL cases each year from 2006 to 2010.[3] More importantly, there is a rapidly growing cohort of survivors of childhood ALL and this number is expected to continue growing.[9] The remarkable breakthroughs in research in the past 50 years have yielded a 5-year survival rate of 91% for children diagnosed with ALL in Canada.[3]

While childhood ALL is a rare disease and treatment outcomes are excellent, children with ALL treated with contemporary standard-risk protocols have slightly higher rates of chronic medical conditions in the survivorship period compared to their siblings.[10] The authors from the Childhood Cancer Survivor Study, a cohort of children from North America, reported a median follow-up of 18.4 years from 5 years after diagnosis. The authors found ALL survivors were at a 60% increased risk of having multiple health conditions and at a 2-fold increased risk of having severe or life-threatening chronic conditions compared to their siblings. Survivors reported poor functional status twice as frequently as siblings (8% vs. 4%, respectively). In Ontario, hospital admissions and duration of stay in hospital among childhood ALL survivors are more than 10 times higher than the general population 3 years after the diagnosis of ALL.[11]

Despite the burdens of childhood ALL, the etiology of childhood ALL is largely unknown. The etiology of childhood ALL likely arises from interactions between exogenous and/or endogenous exposures, genetic susceptibility, and chance.[12,13] Genetic causes of ALL account for a small proportion of cases. Studies on twins have indicated strong genetic risk link to childhood ALL,[14] and that the concordance rate for childhood ALL was 10%.[15] Thus a discordance of 90% in twin studies of disease suggests a postnatal promotional exposure or other event is necessary

for disease emergence.[15] While certain genetic changes may occur that suggest the disease may be present in utero, these changes are insufficient for disease emergence.[15]

One possible promotional exposure is early life infections. There are two key hypotheses related to infections and the development of ALL. In 1988, Kinlen proposed the 'population mixing' hypothesis to describe the observed increased cases of childhood ALL following an influx of migrants into rural areas.[16,17] In Kinlen's hypothesis, the mixing of susceptible and infected individuals would create a localized epidemic of an underlying infection. The epidemic is caused by the increased level of contact between susceptible and infected individuals and this mixing may in turn produce the rare response of childhood ALL. The hypothesis suggests a direct pathological role of a specific infection, presumed to be viral, and that a protective effect may be acquired from previous exposure. Kinlen and others have found evidence to support the 'population mixing' hypothesis.[16-20] In Kinlen's meta-analysis of 17 studies, he found rural population-mixing was associated with excess childhood leukemia (relative risk = 1.57, 95% confidence interval 1.44-1.72) in children aged 0-14 years.[17]

Also introduced in 1988, Greaves' "delayed infection" hypothesis for childhood leukemia suggests a two-hit model.[12,21,22] The hypothesis emphasizes the timing of exposure and the child's immune system. The first hit occurs in utero through a genetic mutation that produces preleukemic clones at a rate of 1% of the normal population.[23,24] However, only a small proportion of preleukemia carriers will progress to leukemia. In a small number of preleukemia carriers, it is the absence of exposure to infections in early life, and a postnatal secondary genetic event caused by a delayed, stress-induced infection (second hit) on the developing, "unprepared" immune system that may increase the risk of developing childhood ALL. While the mechanisms differ, both hypotheses suggest ALL is a rare response to one or more common infections.

Previous studies that have assessed the association between early exposure to infections and childhood ALL are conflicting. A history of infections has been found to reduce the risk of ALL,[25-29] increase the risk of ALL,[30,31] or have no association with ALL.[32-42] Differences between the studies may arise from methodological differences and our limited understanding of the underlying mechanism. Studies using a history of infections were typically self-report questionnaire-based,[26,29,32,34,35,37-39,41,42] and likely suffered from recall bias.[43] In the United Kingdom Childhood Cancer Study, the authors assessed the concordance of maternal recall of

infections in the first year of life compared to general practitioners medical records and found mothers of cases consistently under-reported infections, more so than mothers of controls.[44] About 1 in 3 mothers of cases who took their child to a general practitioner with an infectious illness did not report doing so at the time of interview. Mothers of controls had slightly better recall of infections. The authors concluded the poor recall is likely due to the challenge of recalling *mild illnesses* that occurred 5 to 6 years prior to the interview.[44] Indeed, evidence suggest mothers interviewed when the child is 30 to 33 months of age, serious health events are reported accurately.[45] However, common childhood illness and minor complaints, such as respiratory infections and otitis media infections were not accurately recalled.[44,46-48] Even chronic diseases were not well reported by maternal recall and milder chronic diseases were underreported.[49,50] This may explain some of the differences in the findings from studies that used self-reported measures and found children with ALL had fewer prior infections[25-29] to the studies that used administrative data or medical records data and found children with ALL had more prior infections.[30,31]

Greaves' hypothesis has also been tested using indirect measurement of infectious exposures. For example, day-care attendance has been found to increase the risk of exposure to infections and has been used as a proxy for infections. A meta-analysis found day-care attendance reduced the risk of childhood ALL.[51] Other indirect measurement include having older siblings,[25] birth order,[52] contact with pets or farm animals, and caesarean section have shown inconclusive findings.[25] These indirect measurements of infectious exposures are difficult to obtain on large population samples. The dissertation will focus on direct measurements of infections.

Studies that use administrative data or medical records to assess history of infections are less likely to be affected by recall bias. These studies found children with ALL had more infections in childhood compared to controls,[30,31] found no difference,[33,36] or reported a protective effect.[53] The discrepancy in conclusions between these studies that used administrative data or medical records may be due to missing information on important confounders, such as ethnicity, parental occupation, maternal age, birthweight, and parity.[35,54,55] Parental smoking and exposure to pollution are other confounders not typically captured.[56-59] The heterogeneity in the exposure definitions between the studies may explain some of the difference. Alternatively, studies that used administrative data may have high levels of exposure misclassification, for example, misclassification of infectious syndrome diagnoses was reported to be as high as 30%.[36] Without

explicit validation of the administrative data to identify infections in the studies using those data, it is difficult to quantify the potential misclassification bias.[60]

Alternatively, there may be another explanation for the difference in the study outcomes that suggests a different, co-existing mechanism that focuses on an already altered immune function at birth. In the United Kingdom Childhood Cancer Study, the authors reported children with ALL had an increased number of infections with increasing indices of infectious exposures (such as parity and social activity outside the home), a phenomenon not seen in the healthy controls.[44] In a subsequent study from the same research group, the authors found children with ALL had fewer social contacts and concluded overall exposure to infections were likely lower than healthy controls.[61] This suggests an alternative mechanism influencing ALL risk; children with ALL may have an altered immune system at birth leading to more infections and this could help explain the differences between studies. There is accumulating evidence that children with ALL may have an altered congenital responder status to infection, resulting in a functionally aberrant clinical presentation of occasional infections (that is, a greater propensity to need clinical care when contracting infections).[62] Recent genetic studies reported children with ALL were severely deficient in Interleukin-10, a critical cytokine responsible for regulating the intensity and duration of immune responses to infections.[63-65] Children with lowered expression of Interleukin-10 may be at a higher risk of ALL because their immune systems are less able to prevent overactive inflammatory responses to pathogenic infections.[62] Cautious interpretations of the findings should be used given ALL affects immunity and the immune function markers were not obtained prior to the development of ALL. However, it is possible that this deficiency and the biological stress from the postnatal infection may concede a growth advantage for the preleukemic clones to quickly expand, increasing the opportunity for the second genetic mutation required for the development of ALL.[12,62]

The ascertainment methods most studies use to measure infections can be categorized into three broad groups: self-reported measures, administrative or medical records data, and laboratory investigations. The advantages of self-reported measures are that they are often done in primary data collection studies and investigators can determine the number of variables to collect, and the timing and frequency of data collection. The disadvantages include recall bias and the costly nature of protracted periods of data collection for large cohorts required.

Definitive laboratory investigation provide the most accurate measure of infectious disease classification and are considered the reference standard – in addition with other clinical information.[66] The major benefit to using this measure is to identify the putative infectious agent(s) in the development of childhood ALL. However, it is still unable to yield information on the timing, severity of disease (if used by itself), exposure to untested agents, and most importantly, most clinicians do not test for most organisms because it is unlikely to change patient management. Another limitation of laboratory investigations is the cost associated with each test that could limit study size, and variability in quality control.[67-69]

In summary, self-reported measures may include important confounders but can be susceptible to recall bias and is costly for lengthy periods of data collection, and laboratory investigations can provide classification of potential putative infectious agents but are not feasible for large population-based studies. Healthcare administrative data are passively collected for administrative purposes rather than for research but can be a rich source for population-based research. Using administrative data to study infections would be advantageous in this scenario by addressing several weaknesses of self-reported measurements and laboratory investigations such as recall, accuracy of the information, temporality, and cost. The date and reason for the visit are often captured in the administrative databases and allows for assessment of the time and type of infection.[70] The ability to capture health care visits from birth to diagnosis of a childhood disease provides an efficient method to conduct large studies of rare diseases over protracted periods of time.

Known causes of ALL include ionizing radiation and chemotherapy from the treatment of other cancers.[13,71] Down syndrome has been shown to be associated with infections and the development of ALL and is considered as a genetic confounder of the prior infections and childhood ALL relationship.[72,73] Non-modifiable confounders for infections and childhood ALL include sex and ethnicity/race. Males have high incidence rates of childhood ALL and a higher susceptibility to many childhood infections.[74,75] Hispanic and non-Hispanic white children have high incidence rates of childhood ALL, while black and American Indian/Alaska Native children have the lowest incidence rates.[74] Ethnicity/race has also been shown to be associated with infections.[76] Modifiable risk factors and confounders include a larger set of variables including pesticides, parental occupation, parental smoking, and air pollution, some of which have been

studied extensively. Exposure to any pesticides in the home at time of conception, during pregnancy and after birth was associated with increased odds of ALL.[77] Parental occupational exposure to any pesticides around the time of conception was found to be associated with an increased odds of ALL.[78] Parental smoking before conception, during and after pregnancy was associated with childhood ALL.[58] Further, a dose response relationship has been identified between childhood ALL and parental smoking before conception or after birth. Parental smoking has been shown to increase the odds of respiratory infections and respiratory related infections.[79] Three meta-analyses have been conducted on ambient exposure to traffic pollution and childhood ALL risk. The studies showed exposure to traffic density and traffic related pollution in the postnatal period was associated with increased odds of childhood ALL.[56,57,80] Air pollution has also demonstrated to be associated with infections in children.[59] However, with the exception of Down syndrome and sex, these confounders are not captured in administrative databases and were unable to be accounted for in the current thesis.

1.1 Key Areas of Uncertainty

The overarching goal of this dissertation is to better understand the role of infections in the etiology of childhood ALL. To our knowledge, there was no current systematic summary of the literature on the association between a history of infections and childhood ALL that incorporated the recent developments in the field and considered biases such as potential for recall bias. Also absent in the literature were answers to whether a history of mild or severe infections played a role in the development of childhood ALL and these uncertainties led to the conduct of project 1, "*A Systematic Review and Meta-analysis of the Association Between Childhood Infections and the Risk of Childhood Acute Lymphoblastic Leukemia*". In order to address recall bias, administrative data may be particularly useful to furthering the understanding of the link between infections and ALL. However, in order to conduct such a study, it would be important to first evaluate the criterion validity of administrative data for the purpose of identifying infections and thus, these issues led to the conduct of project 2 "*Use of physician billing claims to identify infections in children: a population-based validation study of administrative data from Ontario, Canada*". The results were used to inform project 3 "*Rate of infections and the association with childhood acute lymphoblastic leukemia: a population-based case-control study*" which assessed the difference in

the rate of prior infections and the association between the development of childhood ALL, beyond an exploratory analysis.[43] Insights gained from the systematic review were used to identify types of infections that may be important in the infectious etiology of childhood ALL. No study has used a life course approach to assess whether the infections and the development of childhood ALL follows a critical period model, and this was identified as an important component in our understanding of the etiology of ALL.[81] Therefore, the overall aim was to determine the role of infections in the etiology of childhood ALL by filling the identified gaps in knowledge while addressing the limitations identified in the literature.

1.2 Dissertation Objectives

Objective 1: To perform a systematic review and meta-analysis to determine whether a history of infections increases the odds of childhood ALL in children aged 0 to 19 years compared to cancer-free children, and to assess the association between the frequency, severity, timing of infections and examine specific infectious agents and syndromes.

Objective 2: Using electronic medical records containing primary physician records from April 1, 2012 to March 31, 2014 as the reference standard, to determine the criterion validity of health administrative databases to identify infectious syndromes in a pediatric patient population aged 0-18 years visiting primary care physicians in Ontario.

Objective 3: To assess whether children diagnosed with ALL between the ages of 2 and 14 years have a higher rate of infections compared to cancer-free children from Ontario, Canada between 1995 and 2014, whether different types of infections and severity of infections are associated with the development of ALL, and to assess whether infections and the development of childhood ALL follows a critical period model.

Chapter 2 : Data Sources

2.1 Data Sources Overview

In this chapter, I present a general overview of the data sources used in Objectives 2 and 3. The data sources were accessed and held at the Institute for Clinical Evaluative Sciences (ICES). ICES is an independent not-for-profit research institute that holds Ontario's health-related data, including coded patient records, clinical and administrative databases, and population-based health surveys. The uniqueness of the available data holdings at ICES is the ability to link population-based health information at the patient level to provide a complete health services use profile of each individual, while ensuring the privacy and confidentiality of personal health information. ICES is named as a prescribed entity under Ontario's privacy legislation. Under this designation, ICES can, without patient consent, receive and use health information for the purposes of health-related research and health system analysis and evaluation.[82]

This dissertation used data and information from the Government of Canada (Immigration, Refugees, and Citizenship Canada Permanent Resident Database), Ontario's Ministry of Health and Long-Term Care (MOHLTC; Registered Persons Database, Ontario Health Insurance Plan Database, Canadian Institute for Health Information Discharge Abstract Database, National Ambulatory Care Reporting System database) and the Pediatric Oncology Group of Ontario (POGO; POGO Networked Information System). The opinions, results, views, and conclusions reported in the dissertation are those of the author and do not necessarily reflect those of the Government of Canada, MOHLTC, or the POGO. No endorsement by the Government of Canada, MOHLTC, or POGO is intended or should be inferred.

2.2 Administrative Data Sources

Ontario is a large, diverse and multicultural province in Canada with over 13 million residents. Residents in Ontario have a universal public health insurance plan called the Ontario Health Insurance Plan (OHIP) with a single payer, the Government of Ontario. The Government of Ontario pays for all medically necessary services across providers and hospitals. The administrative data in Ontario captures information on health services and are collected by the

province for payment or funding purposes, for example physician remuneration. Linkages between databases are possible by using an encoded unique personal identifier generated from the resident's OHIP Health Card Number. Table 2.1 outlines the variables obtained from each database for Objectives 2 and 3.

2.2.1 Registered Persons Database

The Registered Persons Database (RPDB) is a population-based registry that is maintained by MOHLTC. The database is used to manage the publicly funded health care system and contains current and historical listings of unique health card numbers for health insurance eligible residents. The database also contains demographic information on Ontario residents including date of birth, sex, postal code, date of death (if applicable) and captures changes in health insurance eligibility. The Census and geography data from Statistics Canada are linked to RPDB to determine variables like neighbourhood socioeconomic status and the Ontario Marginalization Index (ON-Marg; adapted from the Canadian Marginalization Index for the Ontario population).[83] The person's corresponding dissemination area was used to determine the ON-Marg value. It is worth noting that socioeconomic status and ON-Marg are at the neighbourhood level and not the individual level. While using neighbourhood level income as a proxy for individual level income may subject the study to ecological fallacy, evidence suggests there is no difference in risk estimates when using individual level income compared to neighbourhood level income.[84]

2.2.2 Ontario Health Insurance Plan (OHIP) Database

The OHIP database contains all billing claims made by physicians (and other health care providers) for insured services for eligible residents. The database includes over 95% of all fee-for-service physician billing claims submitted to OHIP for reimbursement and excludes the activity of physicians under a limited number of alternative funding models.[85] Physicians that work in Community Health Centres or Health Service Organizations are not required to submit "shadow billings", meaning they are not required to submit billings as if they were billing fee-for-service like the other physicians.[86] Every visit to a physician by a patient is captured as a distinct service rendered by the physician and includes information on the physician that provided the service, the patient that received the service, type of service provided, diagnostic information, date that it occurred, associated fee code, location of service performed (office or emergency room) and the total fee paid to the physician. It is worth noting that physicians can only submit 1 claim for 1

problem per patient per day, however, a patient can be seen by many physicians in one day. For example, a patient may see their family physician at the physician's office for multiple health concerns, however, the physician may only submit 1 claim for one of the patient's health concerns. This is a limitation of this administrative database and the impact from this limitation was assessed in Objective 2.

2.2.3 Canadian Institute for Health Information (CIHI)

The Canadian Institute for Health Information (CIHI) Discharge Abstract Database (DAD) contains information that has been abstracted from the hospital medical records. The patient-level data include all acute- and chronic-care hospitals, rehabilitation hospitals, and day surgery clinics across Ontario. Each row of data in CIHI-DAD represents a distinct hospitalization. Data elements include patient demographics, clinical data on the patient's diagnoses, procedures that were performed, physicians that cared for the patient, information on the institution from which services were performed, patient's length of stay, and the patient's disposition at discharge.

The National Ambulatory Care Reporting System (NACRS) contains data for all hospital-based and community-based ambulatory care for emergency departments, outpatient and community-based clinics, and day surgeries. NACRS contains many of the data elements found in CIHI-DAD. In February 2000, the Government of Ontario mandated the collection of emergency department services activities using the NACRS Minimum Data Set developed by CIHI.[87] For the purposes of the studies in this dissertation, NACRS was used to obtain emergency department visits from 2001 onwards.

Both CIHI datasets capture data from all hospitals in Ontario and the data are cleaned prior to being used for secondary purposes. Ontario has mandated all publicly funded hospitals to submit emergency department visits and inpatient data to CIHI. The data quality procedures for the CIHI datasets can be found elsewhere.[88,89] Prior to 2001, both CIHI datasets use International Classification of Diseases Ninth Revision (ICD-9); from 2001 onward, CIHI started using the Tenth Revision, ICD-10.

2.2.4 Electronic Medical Record Administrative data Linked Database (EMRALD)

The Electronic Medical Record Administrative data Linked Database (EMRALD) consists of all clinically relevant information from family physician electronic medical records (EMRs) and

can be linked to the administrative databases held at ICES. EMRALD contains data on over 400,000 patients (with 17.8% aged <18 years) who receive primary care from over 350 family physicians who are distributed throughout Ontario and use Practice Solutions® EMR. EMRALD contains clinical information such as laboratory results, prescriptions, blood pressures and anthropometric measures, and the presence of medical conditions recorded by physicians. Physicians participate in EMRALD on a voluntary basis and are required to have had their EMR a minimum of two years to ensure that the EMR is adequately populated. Compared to all Ontario family physicians, EMRALD physicians are more likely to be female (56.0% vs. 41.4%), younger, and Canadian medical graduates (89.3% vs. 74.1%), respectively.[90] Compared to Ontario's population age distribution, EMRALD has a smaller proportion of pediatric patients. The pediatric population in EMRALD is more likely to be of higher socioeconomic status and live in rural areas compared to the overall pediatric population from Ontario.[91] EMRALD will be used as the reference standard for Objective 2's validation study.

2.2.5 ICES Physician Database

The IPDB is a database created and maintained by ICES that contains information on practicing physicians in Ontario. The IPDB amalgamates information from the OHIP Corporate Provider Database, OHIP database on physician billings and the Ontario Physician Human Resource Data Centre database. The IPDB includes demographic information about each physician (including age and sex), practice location, physician speciality, services provided, where the physician was trained and year of graduation.

2.2.6 Pediatric Oncology Group of Ontario Networked Information System (POGONIS)

The Pediatric Oncology Group of Ontario Networked Information System (POGONIS) captures information on the demographics, cancer-specific characteristics, prognostic factors, treatments, outcomes, and complications for children and adolescents diagnosed with cancer in the 5 tertiary care pediatric hospitals in the province. This registry captures 98% of all cancers in children under 15 years in Ontario, Canada when compared to the Ontario Cancer Registry.[92] The pediatric cancer registry began capturing data in 1985 as a standardized paper registration form.[8] In 1995, the registry was converted to an electronic networked database and expanded to include key outcomes and standardized treatment information on all registered cases. In the early 2000's

all cases registered from 1985 to 1994 inclusive were reviewed by trained data managers and their treatment and outcome data were collected, thereby ensuring that all cases captured have a similar detail of data available for analysis. Funded, dedicated data managers actively collect POGONIS-standardized data at each tertiary hospital using hospital chart review, internal hospital information systems, and direct connections with the patient's health care team. POGONIS uses the International Classification of Childhood Cancer nomenclature to map the diagnosis code data element.[93,94] Data quality procedures are routinely conducted, and data can be linked to other administrative databases for research purposes.

Like ICES, the Pediatric Oncology Group of Ontario is a "prescribed entity" under *Ontario Personal Health Information Protection Act*, which authorizes the collection, use and disclosure of personal health information for the purposes of analysis or compiling of statistical information. The data are used for management, evaluation or monitoring of the allocation of resources, or planning for all or part of the health system, including the delivery of services. Strict privacy and security specifications must be followed as outlined by the office of Ontario's Information and Privacy Commissioner. The designation permits POGO to establish linkages between POGONIS and other administrative databases such as RPDB, OHIP and CIHI DAD and NACRS. POGONIS is transferred to ICES annually to be linked to other databases and is available to researchers conducting research on childhood cancer.[95]

2.2.7 Immigration, Refugees and Citizenship Canada (IRCC) Permanent Resident Database

The Immigration, Refugees, and Citizenship Canada (IRCC) Permanent Resident Database contains records from IRCC and is maintained and provided by the Government of Canada. IRCC is responsible for overall management of Canada's immigration system and maintains historical records of immigrants arriving in land and seaports. The Ontario portion of the IRCC database contains individual-level demographic information of immigrants arriving in land or seaports in Ontario from 1985 to 2012. Socio-demographic information for all legal immigrants to Ontario, Canada include country of birth, citizenship, country of last permanent residence, and mother tongue. Over 2.9 million immigrant residents landed in Ontario over the period from 1985 to 2010.[96] Landed immigrants become eligible for health insurance after a 3-month waiting period.

Table 2.1 Data sources for variables used in Objectives 2 and 3

Data Source	**Objective 2 Variables**	**Objective 3 Variables**
Registered Persons Database	Age, sex, place of residence (urban or rural), ON-Marg	Age, sex, place of residence (urban or rural), ON-Marg
Electronic Medical Record Administrative data Linked Database	Reference standard, date of visit, diagnosis in medical chart, chronic conditions and illnesses	
ICES Physicians Database	Physician age, sex, speciality, medical training location, practice location, and graduation year	
Ontario Health Insurance Plan Database	Date of service, diagnosis code	Date of service, infection and other diagnosis codes
National Ambulatory Care Reporting System		Date of service, infection and other diagnosis codes
Discharge Abstract Database		Date of service, infection and other diagnosis codes
Pediatric Oncology Group of Ontario Networked Information System		Date of diagnosis, ALL diagnosis
Immigration, Refugees, and Citizenship Canada Permanent Resident Database		Immigration status, date of immigration

Objective 2 is titled "Use of physician billing claims to identify infections in children: a population-based validation study of administrative data from Ontario, Canada". Objective 3 is titled "Rate of infections and the association with childhood acute lymphoblastic leukemia: a population-based case-control study". ON-Marg represents Ontario Marginalization Index. ALL represents childhood acute lymphoblastic leukemia.

Chapter 3 : A Systematic Review and Meta-analysis of the Association Between Childhood Infections and the Risk of Childhood Acute Lymphoblastic Leukemia

This is a post-peer-review, pre-copyedit version of an article published in *British Journal of Cancer*. The final authenticated version is available online at: http://dx.doi.org/10.1038/bjc.2017.360

Reference

Hwee J, Tait C, Sung L, Kwong JC, Sutradhar R, Pole JD. A systematic review and meta-analysis of the association between childhood infections and the risk of childhood acute lymphoblastic leukaemia. *British Journal Of Cancer*. 2017;118:127.

3.1 Abstract

Background: To determine whether childhood infections were associated with the development of childhood acute lymphoblastic leukemia (ALL).

Methods: We included studies that assessed any infection in childhood prior to the diagnosis of ALL in children aged 0-19 years compared to children without cancer. The primary analysis synthesized any infection against the odds of ALL, and secondary analyses assessed the frequency, severity, timing of infections and specific infectious agents against the odds of ALL. Subgroup analyses by data source were investigated.

Data Synthesis: In our primary analysis of 12,496 children with ALL and 2,356,288 children without ALL from 38 studies, we found any infection was not associated with ALL (odds ratio (OR)=1.10, 95%CI 0.95-1.28). Among studies with laboratory confirmed infections, the presence of infections increased the odds of ALL by 2.4-fold (OR=2.42, 95%CI 1.54-3.82). Frequency, severity and timing of infection was not associated with ALL.

Conclusions: The hypothesis put forward by Greaves and others about an infectious etiology are neither confirmed nor refuted and the overall evidence remains inadequate for good judgement. The qualitative difference in the subgroup effects require further study, and future research will need to address the challenges in measuring infectious exposures.

3.2 Introduction

The etiology of childhood acute lymphoblastic leukemia (ALL) is largely unknown, and likely arises from interactions between exogenous and/or endogenous exposures, genetic susceptibility and chance. Genetic causes of ALL account for a small proportion of cases, and while the disease is usually initiated in utero, other promotional exposures are probably necessary for disease emergence.[15] There are two key hypotheses on infections and the development of ALL. Kinlen proposed the 'population mixing' hypothesis to describe the observed increased rates of childhood ALL following an influx of migrants into rural areas.[16,17] Briefly, the mixing of rural, isolated individuals with the influx of mostly urban individuals into a rural area would create a localized epidemic of an underlying infection due to the increased level of contact between susceptible and infected individuals; that may produce the rare response of ALL. Studies from Kinlen and others have found evidence to support the hypothesis.[16-20] The hypothesis suggests a direct pathological role of a specific infection, presumed to be viral, in the development of ALL and that a protective effect may be acquired from previous exposure. Currently, there is limited molecular evidence that implicates a specific infection.[97,98] Greaves' 'delayed infection' hypothesis for childhood ALL suggests a two-hit model that emphasizes the timing of exposure and the child's immune system.[12,21] The first hit occurs in utero through one's genetic makeup that produces a pre-leukemic clone. In a small number of pre-leukemia carriers, it is the absence of exposure to infections in early life, and a postnatal secondary genetic event caused by a delayed, stress-induced infection (second hit) on the developing, "unprepared" immune system that may increase the risk of childhood ALL. While the mechanisms differ, both hypotheses suggest ALL is a rare response to one or more common infections acquired through personal contact.

The difficulties in measuring exposure to infectious agents and subsequent responses make it challenging to directly test the hypotheses, especially since no specific leukemogenic agent has been identified. Several previous epidemiological studies have used a history of infections as an indicator for early exposure to infections. Establishing the timing of the infections is critical to testing the hypotheses, however, birth cohort studies are not feasible given the rarity of childhood ALL. Thus, most studies used a case-control design and interviews to measure infections. Assessing a history of infections through interviews can be problematic due to the potential for recall bias and misclassification of children who had asymptomatic infections.[99] Other methods for measuring infections such as using administrative data overcome these limitations, but may

lack information on important confounders. Other than narrative summaries,[100-103] no study has attempted to synthesize and quantitatively pool studies examining the relationship using a history of infections or tried to explain the differences between the studies. The aim of this systematic review and meta-analysis was to assess the relationship between childhood infections and the development of childhood ALL by summarizing the findings for an overall measure of infections, the frequency, severity, timing of infections, and examining specific infectious agents and syndromes.

3.3 Methods

The Meta-analysis of Observational Studies in Epidemiology (MOOSE) was developed as a guideline for the reporting of meta-analyses of observational studies in epidemiology and was used for the current study.[104]

3.3.1 Data Sources and Searches

We performed electronic searches from inception to February 21, 2017 in Ovid MEDLINE, MEDLINE In-Process and Other Non-Indexed Citations, EMBASE, Web of Science (Science Citation Index Expanded, Social Sciences Citation Index, Conference Proceedings Citation Index for both Science and Social Science & Humanities), and Scopus. Supplementary Table 3.1 (Appendix A) shows the search strategies used. Text words used included *acute lymphoblastic leukemia, acute leukemia, infection, virus,* and *bacteria*. We limited the search to subjects 0-19 years old and did not restrict the search by language. References of the included studies were searched, and the first 4 pages of a Google search using the same key words were used to search for grey literature.

3.3.2 Study Selection

We defined the inclusion and exclusion criteria a priori as studies of any design excluding editorials, reviews and case reports. Studies were included if: 1) the primary exposure of interest included a prior history of any infection before the diagnosis of childhood ALL; 2) the primary outcome of interest was defined as clinically diagnosed ALL in children aged ≤19 years; 3) comparisons were made against a control or comparison group; and 4) testing samples must have been collected and assessed prior to treatment, if laboratory investigations were used to determine past infections. Infections must have been reported by the parent or guardian or obtained through other data sources such as medical records.

We excluded studies based on the following order: 1) definition for infections was not at the individual level, for example, at an ecological level that examines infections aggregated for a region; 2) definition for infections that examined population mixing; 3) infections were not explicitly infections during childhood (e.g., infections during pregnancy); 4) outcomes was not childhood ALL in children aged ≤19 years; 5) absence of a comparison group; 6) it was a review article; and 7) duplicate publication with the same study population. When more than one publication from a study was available, the most recent version, or the version with the exposure or outcome of interest that was closest to the objectives of this review was included. Studies were not restricted by publication status, and relevant studies in other languages were translated.

Two reviewers (JH and CT) independently evaluated the titles and abstracts of publications identified by the search strategy, and any publication thought to be potentially relevant by either reviewer was retrieved in full. Final inclusion of studies in the systematic review was determined by agreement of both reviewers. Agreement between reviewers was evaluated using the kappa statistic (κ). Strength of agreement was defined as slight (κ=0.00 to 0.20), fair (κ=0.21 to 0.40), moderate (κ=0.41 to 0.60), substantial (κ=0.61 to 0.80), or almost perfect (κ=0.81 to 1.00).[105]

3.3.3 Data Extraction and Quality Assessment

Data extraction was conducted in duplicate (JH and CT) using a standard form, which collected information on: the primary exposure of "common infections", defined as any infection occurring from birth to the diagnosis of ALL; secondary exposures of infection frequency, severity of infections; and study design, region, publication era, and source of controls. In studies that used laboratory investigations for identification of infectious agents, we extracted IgG antibody estimates to represent past infections, and if that wasn't available, the polymerase chain reactions (PCR) method was extracted to assess for the presence of the agent. We extracted infections occurring in the first year of life or similar time-windows in cases with multiple time-windows, as we felt this best represented early exposure to infections. We extracted infection frequency levels for common infections, and defined severity based on admission to hospital. The adjusted models that incorporated the most confounders for our primary outcome ALL were extracted. Authors were contacted for further information regarding results that were not presented. Five authors were contacted,[106-110] and 3 responded with no additional information.[107-109]

Study quality was assessed using the Meta Quality Appraisal Tool (MetaQAT).[111] and the Critical Appraisal Skills Programme (CASP) for case-control,[112] and cohort studies.[113] Two reviewers (JH and CT) assessed each study. For case-control studies, we considered CASP scores of 1-3, 4-6, and 7-9 to be high, moderate, and low-risk of bias respectively; for cohort studies, we considered CASP scores of 1-4, 5-8, and 9-11 to be high, moderate and low-risk of bias respectively.

3.3.4 Data Synthesis and Analysis Methods

Our analysis combined data at the study level. Our primary analysis sought to assess exposure to common infections versus no common infections (referent group) on the risk of developing ALL, relying on each study's definition. The most frequent infection was used when studies did not report a common infection variable. We used the adjusted odds ratio (OR) or rate ratio (RR) to calculate a pooled overall effect, and assumed OR and RR were equivalent due to the rarity of the outcome[114]; ORs or RRs <1 suggest infections are protective against ALL. If a study presented multiple frequency categories, we used the lowest versus the highest category, a method commonly used in meta-analyses.[115] The method described by Greenland was used to calculate the variance using the reported 95% confidence intervals (CI).[114] We calculated a crude OR for studies not reporting one, and to facilitate the calculation we added 0.5 to all cells if one of the four cells reported a zero.[116] In secondary analyses, we used the different exposure levels of infection to compute a regression slope.[117] If an exposure level was defined using a range, we used the midpoint of the range (e.g., 1-3 infections was assigned a frequency of 2), and if the level was ≥4, we assigned a frequency of 4. For infection severity, a dichotomous variable (yes versus no) was used to determine the relationship with ALL. Post hoc analyses examining timing of infections in the first year of life compared to infections that occurred after the first year of life, and putative infectious agents was conducted if ≥3 studies reported the agent.

As we anticipated heterogeneity between the studies, we used an inverse variance weighted average, random-effects model where the Wald-type tests and confidence intervals were estimated under a normal distribution.[118] We investigated potential sources of heterogeneity using subgroup analyses and mixed-effects meta-regression. To examine the association of study-level characteristics and infection effect, we fitted mixed-effects meta-regression models to the natural logarithm of the OR. The natural logarithm of the OR was assumed to have a normal distribution,

and a method of moments based estimator to estimate model variables. The mixed-effects model included fixed effects for the covariates, and a random intercept term was specified to model residual heterogeneity not accounted for by the covariates. We corrected for multiple testing using a Bonferroni correction that divides the p-value by the number of tests.[119] Because of methodological differences,[62] we tested for interactions to assess the differences between studies that used administrative/medical records, self-reported, and laboratory investigation data.[120] We stratified infections in the first year of life by self-reported data and administrative/medical records data. We explored clinical heterogeneity by conducting a subgroup analysis limiting cases of ALL to B-cell precursor ALL.[62] We also explored the extent to which region (North America, Europe, Asia, or other), publication era (≤1999, 2000-2009, ≥2010), source of controls (general population, general practitioner list, or hospital controls), and risk of bias influenced the magnitude of the average effect estimate in the meta-analysis. Publication bias was assessed by funnel plot and the Egger's test.[121,122] The meta-analysis was performed using the metafor package in R, version 3.3.[123]

3.4 Results

Titles and abstracts of 9,445 records were reviewed, and 314 full-text articles were retrieved (Figure 3.1). There were 39 studies that satisfied the inclusion criteria,[25-39,41,42,97,106-110,124-139] and of those, 38 were included in the meta-analysis. One study did not report infections and the effect estimate could not be calculated.[124] The reviewers had strong agreement on the articles for inclusion (κ=0.85, 95%CI 0.75-0.95). Characteristics of the included studies are presented in Table 3.1. The exposure definitions are presented in Supplementary Table 3.2 (Appendix A). The reviewers had moderate agreement on the judgement of the risk of bias for each study (κ=0.50, 95%CI 0.28-0.72). Thirteen studies were judged as being low-risk of bias, 7 as being moderate-risk of bias, and 19 as being high-risk of bias (Supplementary Table 3.3a-b; Appendix A). We found evidence of publication bias (bias coefficient=1.19, 95%CI 0.30-2.08; Supplementary Figure 3.1; Appendix A).

Our analysis included 12,496 children with ALL and 2,356,288 children without ALL. There was no association between infections and ALL, OR=1.10, 95%CI 0.95-1.28; p=0.187 (Figure 3.2). We observed considerable heterogeneity between the studies (I^2=76.5%; Q-statistic p<0.001). The trend analysis included 13 studies and we did not find frequency of infections to be associated

with ALL (OR=1.00, 95%CI 0.95-1.05; p=0.967). In the 4 studies that assessed the infection severity, the combined average effect of hospitalizations for infections was not associated with ALL (OR=1.22, 95%CI 0.85-1.75; p=0.239). Infections that occurred in the first year of life was not associated with ALL (OR=0.99, 95%CI 0.85-1.16, p=0.920; Supplementary Figure 3.2). Infections that occurred after the first year of life suggested an association with ALL (OR=1.45, 95%CI 0.71-2.96, p=0.313), but did not differ compared to infections in the first year of life (interaction effect OR=0.69, 95%CI 0.32-1.43, p=0.314) (Supplementary Figure 3.2; Appendix A). Parvovirus B19 (OR=2.69, 95%CI 1.16-6.22, p=0.020) was found to be associated with ALL (Figure 3.2). No associations were observed for human herpesvirus-6 (OR=0.89, 95%CI 0.42-1.87, p=0.752), however Epstein-Barr virus (OR=1.39, 95%CI 0.83-2.33, p=0.208), cytomegalovirus (OR=1.95, 95%CI 0.64-5.96, p=0.242), influenza (OR=1.97, 95%CI 0.97-3.98, p=0.061), and herpes simplex virus (OR=2.04, 95%CI 0.66-6.23, p=0.214) showed a strong association with ALL but lacked precision. Varicella, rubella, mumps, measles, and pertussis were not associated with ALL (Supplementary Figure 3.3; Appendix A).

3.4.1 Subgroup, and Sensitivity Analyses

After applying the Bonferroni correction, the p-value to indicate statistical significance for the additional analyses was <0.005. The data sources for the studies can be found in Table 3.1. Among the studies that used self-reported data, we found no association between infections and ALL (OR=0.89, 95%CI 0.79-1.00, p=0.049; I^2=50.5%). Among studies that used administrative/medical record data, we found no association between infections and ALL (OR=1.00, 95%CI 0.61-1.63, p=0.994; I^2=90.8%). Among studies that used laboratory data, we found infections to be associated with ALL (OR=2.42, 95%CI 1.54-3.82, p<0.001, I^2=54.2%). The interaction effect showed no difference between self-reported and administrative/medical records data sources (OR=0.89, 95%CI 0.54-1.48, p=0.656). Infections identified through laboratory data increased the risk of ALL compared to infections captured through self-reported data (interaction effect OR=2.73, 95%CI 1.71-4.36, p<0.001), but not administrative/medical records data sources (interaction effect OR=2.43, 95%CI 1.24-4.75, p=0.009). Among studies that used self-reported data, every additional infection reduced the odds of ALL by 4% (OR=0.96, 95%CI 0.94-0.98; p<0.001). Whereas among studies that used administrative/medical records data, every additional infection increased the odds of ALL by 11% (OR=1.11, 95%CI 1.07-1.15; p<0.001). We found self-reported and administrative/medical records data sources qualitatively differed in the

frequency of infections (interaction effect OR=0.86, 95%CI: 0.83-0.90, p<0.001). Severity of infections remained unchanged in studies with self-reported data (OR=1.51, 95%CI 0.86-2.65; p=0.158; I^2=70.2%). Among self-reported studies, infections in the first year of life suggested a protective effect against ALL (OR=0.88, 95%CI: 0.80-0.98, p=0.017). No association was found between infections in the first year of life and ALL among administrative/medical records data (OR=0.93, 95%CI 0.55-1.56, p=0.775), and did not differ from self-reported studies (interaction effect OR=0.95, 95%CI 0.56-1.62, p=0.862).

The results from our primary analysis remained unchanged when we restricted the analysis to B-cell precursor ALL or B-cell common ALL (OR=0.87, 95%CI 0.77-0.98, p=0.022). In the meta-regression models that assessed included data source, region, publication era, source of controls, and risk of bias. Data source and region accounted for the largest proportion of heterogeneity between the studies (R^2=47.2%, see Supplementary Table 3.4; Appendix A). Stratification by risk of bias indicated studies of low-risk of bias showed similar results to our main analysis (OR=0.92, 95%CI 0.76-1.10, p=0.349), while studies of moderate-to-high-risk of bias suggested infections increased the risk of ALL (OR=1.45, 95%CI 1.12-1.86, p=0.005). Compared to studies of moderate-to-high-risk of bias, studies of low-risk of bias were more likely to suggest infections were protective against ALL (OR=0.63, 95%CI 0.46-0.87, p=0.004).

3.5 Discussion

In this systematic review of 39 studies, we found no association between any common infections, frequency, severity of infections, and timing of infections and childhood ALL. We did however, find a qualitative difference in our subgroup analyses; infections increased the odds of developing ALL by 2.4-fold in studies with laboratory investigations. Further, infections identified through laboratory investigations increased the odds of ALL by 2.7-fold and 2.4-fold compared to infections identified through self-reported and administrative/medical records data, respectively. Among studies that used self-reported data, we found each additional infection reduced the odds of ALL by 4%, and this differed significantly from studies that used administrative/medical records data that suggested each additional infection increased the odds of ALL by 11%. The heterogeneity between the studies remained a challenge and could partly be explained by differences in the data sources.

We failed to demonstrate an association in our primary analysis, but found associations in our secondary and subgroup analyses by data source. There are 3 plausible explanations for the observed findings. First, the apparent results may be a chance finding from multiple testing. Second, the ascertainment of infections from parental recall has been shown to under-report childhood infections and may be inaccurate in both the timing and occurrence of infections, compared to medical records.[46,99] Despite these potential issues, studies that confirmed the self-reported infections with medical records for accuracy and completeness still found an inverse association.[25,34] Whereas studies that used medical records were void of recall bias, they were often unable to include other important confounders, such as ethnicity, parental occupation, maternal age, birthweight, and parity.[35,54,55,140] Finally, the findings from the laboratory studies must be interpreted with caution due to the study quality, and smaller sample sizes and larger effect sizes as shown by the asymmetry of the funnel plot.

The mutational mechanisms of ALL point to three potential pathways: 1) anomalies in lineage-specific factors (ETV6-RUNX1, IKZF1, and PAX5); 2) flaws in receptor protein tyrosine kinases and their down-stream pathways; and 3) epigenetic modifiers.[141] Recent developments in genome and mouse model studies may change our initial understanding of the etiology of ALL as new studies have generated new hypotheses with respect to identifying potential infectious candidates.[98,142] The presence of parvovirus B19 IgG antibodies is associated with the presence of ETV6-RUNX1,[126] and is associated with certain class II HLA alleles that are risk factors for the development of childhood ALL. Furthermore, parvovirus B19 has certain characteristics similar to other oncoviruses, that is, its DNA genome persists indefinitely in human tissues following acute infection, causing mild or no disease, and upregulates pro-inflammatory cytokines associated with ALL onset.[143] The results from the small laboratory studies will require confirmation in larger population studies. Since half of 15 year old adolescents have specific antiparvovirus B19 antibodies,[144] the measurement of the clinical syndromes caused by parvovirus B19 may be preferred to assess manifestations of the pathogen. Parvovirus B19 infection may provide only a subset of an oncogenic hit in a multistep carcinogenesis process.

The qualitative differences in our findings supports the hypothesis of an alternative pathway for ALL development. Recent qualitative reviews have attempted to explain the positive association between infections and ALL, and suggested studies that used medical records or

administrative data may be capturing children with an earlier than expected altered immune system. These children may respond differently to infections, have a greater propensity to seek medical care when infections are contracted, and/or have a stronger immune response.[62,141] The sensitivity to infections may be due to a lack of immunomodulation from lower levels of anti-inflammatory cytokine interleukin-10 in newborns who later go on to develop ALL.[63]

As in previous reviews, there continues to be substantial heterogeneity among the studies, however our review focuses on specific objectives and highlights the recent developments of the field.[12,100-103] There are several limitations of this study. The heterogeneity between the studies in the definition of infections, the time-period to observe the infections and the evidence of publication bias was a challenge. We decided to use *any common infection* as our main exposure variable in the primary analysis because we felt it to be the most appropriate measure that reflects the hypotheses from Kinlen and Greaves.[12,16] The heterogeneity likely stems from the unknown etiology of ALL, and one that requires further research. The limitations with laboratory investigation studies is the inability to disentangle temporality. The presence of the infectious agent was assessed after a diagnosis of ALL was made and it is unknown if the agent was present before or after the onset of ALL. It is unclear if the infection occurred before the onset of ALL, or if the potentially reduced immune function because of ALL contributed to the contraction of specific infections. Further, the laboratory studies were appraised as high-risk of bias, often small, and may not be generalizable. Despite the differences in the risk of bias amongst the included studies, our conclusions were unchanged after we stratified the analysis to the 13 studies with a low-risk of bias. Another limitation was the quality of reporting in the studies included in the review. Most studies clearly reported their findings, but studies published earlier tended to have incomplete reporting.

Costs and feasibility are the usual barriers to establishing new large pregnancy and birth cohorts,[145] research groups have instead combined existing cohorts to study childhood cancers,[81,146] and other diseases.[147] The increased power may help to identify high risk or vulnerable, and understudied populations. The next step should focus on the measurement of infections and infectious exposures. The use of linked administrative data provides a large population for study with accurate information on the timing of physician diagnosed infections, frequency and severity of infections as answers to these questions remain elusive. Enhancing the

administrative data with surveys to obtain other infectious exposures such as day-care attendance, breastfeeding, or by applying emerging technologies that detect and quantify the pathogen burden with greater speed, accuracy and simplicity [148] in a subset sample would improve the accuracy and strengthen the measurement of infections. Day-care attendance has been found to increase the risk of exposure to infections, and has been used as a proxy for infections. A meta-analysis found day-care attendance reduced the risk of childhood ALL.[51] Breastfeeding has been found to reduce the risk of ALL through its immunologically active components, antibodies and other elements that influence the development of the infant's immune system.[149-151] The challenge will be to disentangle the mechanistic pathways of the infectious etiology hypothesis by combining different measurements of infectious exposures to determine the total, direct, and indirect effect of infections on the risk of developing childhood ALL.

An infectious etiology of ALL is suggestive in our study, however, the challenges in measuring infections must be addressed. Parvovirus B19 as a putative causal infectious agent for childhood ALL needs to be tested in larger cohorts, and the rather substantial point estimates from influenza, cytomegalovirus and herpes simplex virus warrant a follow-up in larger studies. Whether children with ALL have a dysregulated immune function present at birth requires further investigation. Only one study conducted an exploratory assessment on a key aspect of Greaves' hypothesis, the timing of the infections in early life.[43] Our future research aims to provide further insight on the timing of infections and the risk of developing childhood ALL. The use of administrative data or medical records with linked laboratory data would overcome the challenges facing studies that used self-reported and laboratory investigation data, and would be ideal to evaluate the association between childhood ALL and the timing and frequency of infections. The review has highlighted knowledge gaps surrounding the relationship between childhood ALL and severity of infections. The causal association of infections will need to be tested in conjunction with other identified risk factors to quantify the direct, indirect, interaction and mediated effect of infections on ALL risk. These will be critical research questions in discovering the causes of childhood ALL and will be the foundation for future studies that can combine epidemiologic, genetic and environmental factors.

Figure 3.1 Study selection flow diagram

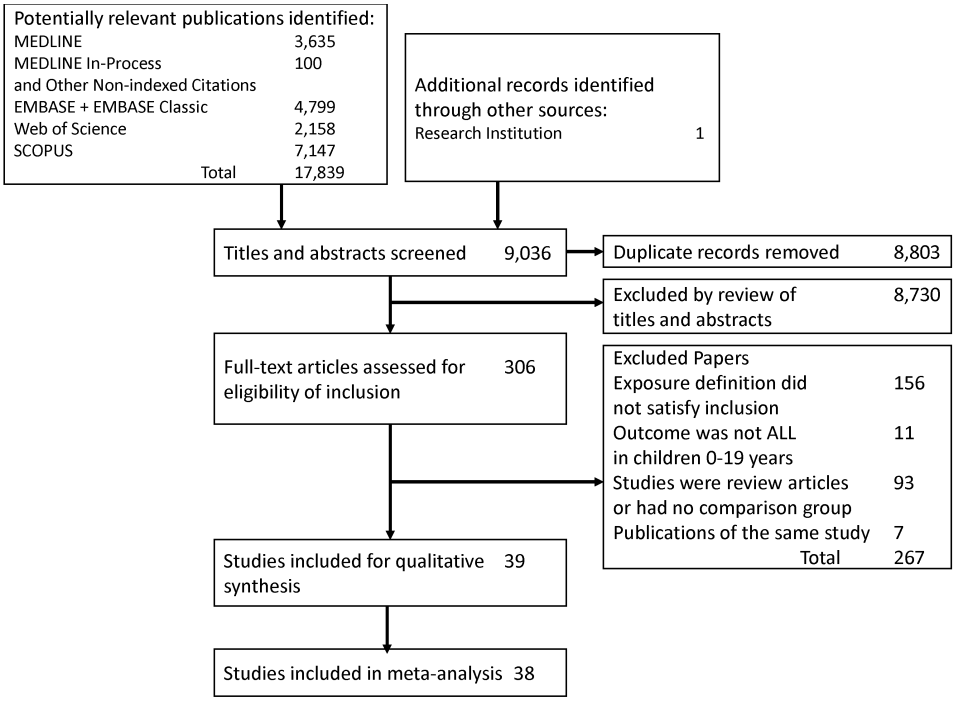
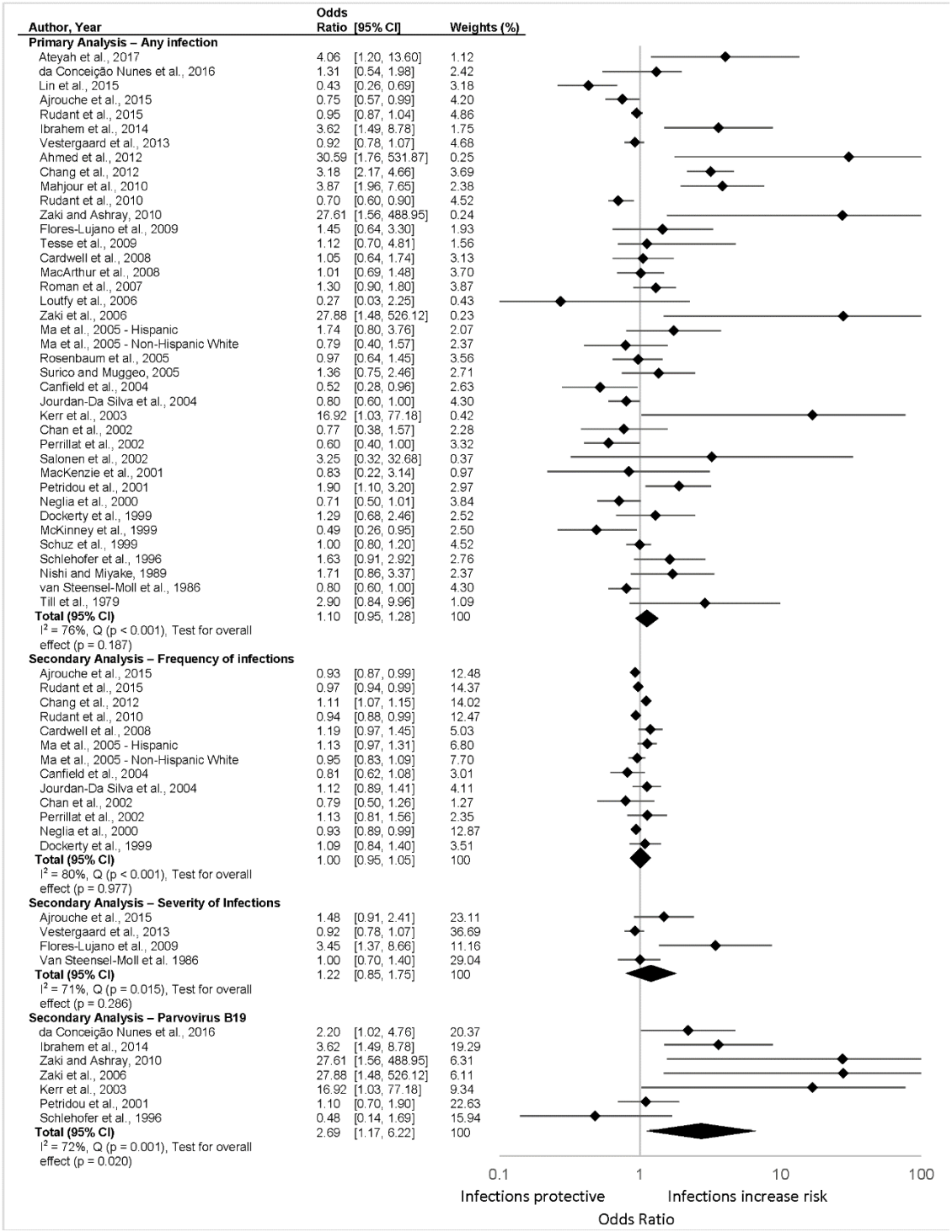


Figure 3.2 Random effects model examining the association between common infections and odds of childhood acute lymphoblastic leukemia



CI represents confidence interval. Common infections are reported as a two-class variable, or highest vs lowest in more than 2 categories. The secondary analysis for frequency of infections is a combined maximum likelihood effect estimate that estimates a trend from summarized dose-response data. The presence of parvovirus B19 was measured as a dichotomous variable, presence of IgG antibodies versus no IgG antibodies for parvovirus B19. All other studies, the reference was no infections.

Table 3.1 Characteristics of included studies and associated references

Table 1. Characteristics of the included studies and associated references					
Study Design	**Case ascertainment**	**Control selection**	**Data source and collection**	**Selected exposure definition**	**Matching variables**
Ateyah et al. 2017					
CC	45 ALL cases Single hospital	40 controls without cancer Same hospital as cases	Laboratory investigation	EBV anti-VCA IgG	1:1 on age and sex
Conceicao Nunes et al. 2016					
CC	60 ALL cases Single hospital	120 controls without cancer Same hospital as cases	Laboratory investigation	EBV anti-VCA IgG	1:2 on age and sex
Ajrouche et al. 2015					
CC	617 cases National cancer registry	1225 controls without cancer Population controls	Self-report: interviews	Common infections	1:M on age and sex
Lin et al. 2015					
Co	62 ALL cases National cancer registry	564 573 children without cancer from national administrative database	Administrative database	Enterovirus infection	1:1 on sex, age, urbanization level, parental occupation, and index year of enterovirus infection
Rudant et al. 2015*					
CC	4641 ALL cases National, clinical cancer, general physician registries, and hospitals	7971 controls without cancer Birth, general physician registries, hospitals, population quotas	Self-report: interviews, or questionnaires	Common infections	-
Ibrahem et al. 2014					
CC	40 ALL cases Single hospital	60 healthy controls from same region	Laboratory investigation	Parvovirus B19 IgG	Age and sex
Vestergaard et al. 2013					
Co	815 ALL cases National cancer registry	1 777 314 children without cancer from national database	Administrative data	Hospitalization for infections	-
Ahmed et al. 2012					
CC	54 ALL cases Single hospital	20 controls without leukemia Single hospital	Laboratory investigation	EBV PCR	-
Chang et al. 2012					
CC	1039 ALL cases National cancer registry	4140 controls without cancer National administrative database	Administrative data	Common infections	1:M on date of birth, sex, time of case diagnosis
Mahjour et al. 2010					
CC	90 ALL cases Single hospital	90 controls without ongoing cancer from single hospital	Laboratory investigation	HSV IgG	1:1 on age and sex
Rudant et al. 2010					
CC	634 ALL cases National cancer registry	1494 controls without cancer Population controls	Self-report: interviews	Common infections	1:M age and sex
Zaki and Ashray 2010					

CC	40 acute leukemia Single hospital	20 healthy controls from same hospital	Laboratory investigation	Parvovirus B19 IgG	Age and sex
Flores-Lujano et al. 2009					
CC	45 ALL cases with Down syndrome from 6 select cancer institutions in Mexico City	218 controls with Down syndrome without leukemia Specialized institutions exclusively for Down syndrome	Self-report: interview	Common infections	-
Tesse et al. 2009					
CC	40 ALL cases from single hospital	40 healthy controls from same hospital	Laboratory investigation	EBV IgG	1:1 on ethnic origin and socioeconomic status
Cardwell et al. 2008					
CC	112 ALL cases National population-based medical records from general physician offices	2125 controls without leukemia Same database as cases	Medical records: Chart abstraction	Common infections	1:M on physician practice, sex, date of birth
MacArthur et al. 2008					
CC	351 ALL cases Population-based cancer registries and oncology centres	399 controls without cancer Provincial health insurance registration database	Self-report: interviews	Varicella	1:1 on age, sex, area of residence
Roman et al. 2007					
CC	425 ALL cases National population-based medical records from general physician offices	1031 controls without cancer Same database as cases	Medical records: Chart abstraction	Common infections	1:M on region of residence at diagnosis, sex, month and year of birth
Loutfy et al. 2006					
CC	68 ALL cases Single hospital	20 controls Siblings of cases	Laboratory investigation	EBV anti-VCA IgG	-
Zaki et al. 2006					
CC	20 acute leukemia Single hospital	20 healthy controls from same hospital	Laboratory investigation	Parvovirus B19 IgG	Age and sex
Ma et al. 2005					
CC	294 ALL cases Hospital-based network registry covering 35 counties in Northern and Central California	376 controls without cancer Random selection from statewide birth files	Self-report: interview	Stratified by non-Hispanic white and Hispanic; Common infections	1:1 and 1:2 on child's date of birth, sex, mother's race, Hispanic status, mother's county of residence
Rosenbaum et al. 2005					
CC	255 ALL cases Institutional cancer registry at 4 major centres serving 31 counties	760 controls State live birth registry	Self-report: questionnaire	Colds	1:M on sex, year of birth, race
Surico and Muggeo 2005					
CC	82 ALL cases Single hospital	196 controls without cancer From same hospital as cases	Laboratory investigation	EBV anti-VCA IgG and EBNA IgG latent infection	1:2 on age, sex and comparable socioeconomic status

Jourdan-Da Silva et al. 2004					
CC	393 ALL cases National cancer registry	530 controls without leukemia or lymphoma Population controls	Self-report: questionnaire	Common infections	1:M on age, sex and region of residence
Canfield et al. 2004					
CC	97 ALL cases with Down syndrome Children's Oncology Group registration files	173 controls with Down syndrome without leukemia From the same physician practice as the cases	Self-report: interview	Common infections	1:M on age
Kerr et al. 2003					
CC	16 acute leukemia	23 controls with diseases requiring cerebral spinal fluid extraction	Laboratory investigation	Parvovirus B19 PCR	-
Chan et al. 2002					
CC	80 ALL cases Clinical database	228 controls without leukemia Regional controls	Self-report: interviews	Common infections	-
Perrillat et al. 2002					
CC	219 ALL cases Hospital records from 4 cities in France	237 controls without cancer Controls from the same hospital, and same catchment area of the hospital	Self-report: interview	Repeated common infections	1:M on sex, age, hospital, hospital catchment area, ethnicity
Salonen et al. 2002					
CC	40 acute leukemia	39 hospital controls	Laboratory investigation	HHV-6 IgG	1:1 on age, sex and season
MacKenzie et al. 2001					
CC	27 ALL cases	28 children with other cancers	Laboratory investigation	EBV PCR	-
Petridou et al. 2001					
CC	94 ALL cases Clinical database of participating centres	94 controls Hospital controls for non-infectious reason	Laboratory investigation	Parainfluenza 1, 2 and 3 IgG	1:1 on sex, age, hospital, time-period
Neglia et al. 2000					
CC	727 ALL cases Clinical database of participating centres	637 controls Random digit dialing of residents	Self-report: Interviews	Ear infection	1:M on age at diagnosis, race, telephone area code
Schuz et al. 1999					
CC	884 ALL cases National cancer registry	2566 controls without cancer Population-based registration files	Self-report: interview and questionnaire	Common infections	1:M on age and sex
McKinney et al. 1999					
CC	124 ALL cases National cancer registry	236 controls without cancer Population-based general practice registration files	Medical records: chart abstraction	Common infections	1:M on age, sex, health board area of residence
Dockerty et al. 1999					
CC	97 ALL cases National cancer registry	303 controls without cancer National birth records	Self-report: interview	Common infections	1:M on age and sex

Schlehofer et al. 1996					
CC	118 ALL cases National cancer registry	187 controls Hospital controls from participating sites	Laboratory investigation and self-report: questionnaire	Varicella	1:M on age, sex
Nishi et al. 1989					
CC	63 ALL cases 9 hospitals in Hokkaido Prefecture, Japan	126 healthy controls Same hospitals located in areas where the index case resided	Self-report: interview	Measles	1:M on age, sex, district residence at diagnosis
McKinney et al. 1987					
CC	148 ALL cases Epidemiological study database	342 controls Same hospital admission records and general practitioner lists as cases	Self-report: interview Medical chart: abstraction where possible	Common infections	1:M on age, sex
van Steensel-Moll et al. 1986¥					
CC	492 ALL cases Study Group national registry	480 controls without cancer Randomly drawn from municipal registration files from same region as cases	Self-report: questionnaire	Hospitalizations for infections	1:1 on age, sex, , place of residence at diagnosis
Till et al. 1979					
CC	54 ALL cases Single hospital	121 controls without leukemia Ascertained from parent's suggested friends or neighbours for matching	Self-report: questionnaire, and interview	Common infections	1:M on age

*Only selected sites contributed early infection information and the presented information is based on those sites that contributed data.
¥Not included in primary analysis but was included in the secondary analysis examining severe infections. Selected exposure definition represents the infection definition used in the primary analysis. CC represents case-control and Co represents cohort studies. 1:M represent frequency matching. EBV represents Epstein-Barr virus. EBNA represents Epstein-Barr nuclear antigen. HSV represents herpes simplex virus. VCA represents viral capsid antigen. PCR represents polymerase chain reaction.

Chapter 4 : Manuscript titled Use of physician billing claims to identify infections in children: a population-based validation study of administrative data from Ontario, Canada

4.1 Abstract

Background: Few studies have validated the use of administrative data for identifying infections in pediatric populations.

Methods: Pediatric patients aged <18 years were randomly sampled from the Electronic Medical Record Administrative data Linked Database (EMRALD). Using physician diagnoses from the electronic medical record (EMR) as the reference standard, we determined the criterion validity of physician billing claims in administrative data for identifying infectious disease syndromes from 2012 to 2014. Diagnosis codes were assessed by infection category (respiratory, skin and soft tissue, gastrointestinal, urinary tract and otitis externa) and for all infections combined. Sensitivity analyses assessed the performance if patients had more than one reason to visit the physician.

Results: We analysed 2,139 patients and found 33.3% of all visits were for an infection, and respiratory infections accounted for 67.6% of the infections. When we combined all infection categories, sensitivity was 0.74 (95%CI 0.70-0.77), specificity was 0.95 (95%CI 0.93-0.96), positive predictive value (PPV) was 0.87 (95%CI 0.84-0.90), and negative predictive value (NPV) was 0.88 (95%CI 0.86-0.89). For respiratory infections, sensitivity was 0.77 (95%CI 0.73-0.81), specificity was 0.96 (95%CI 0.95-0.97), PPV was 0.85 (95%CI 0.81-0.88), and NPV was 0.94 (95%CI 0.92-0.95). Similar performance was observed for skin and soft tissue, gastrointestinal, urinary tract, and otitis externa infections, but with lower sensitivity. Performance measures were highest when the patient visited the physician with only one health complaint.

Conclusions: We found when using linked EMR data as the reference standard, administrative billing codes are reasonably accurate in identifying infections in a pediatric population.

4.2 Introduction

Healthcare administrative data provide a rich source of population-based information. However, since the data are passively collected for administrative purposes rather than for research, validation studies are necessary to determine the accuracy of these data for identifying diseases. Infections are the most frequent reason reported for seeking healthcare in children and adolescents aged <18 years, accounting for the majority of emergency department and physician office visits.[152-156] Using administrative data to study infections would be advantageous, allowing large populations of children to be studied efficiently. However, few studies have validated the use of healthcare administrative data for identifying infections in pediatric populations.[157]

Ontario is Canada's most populous province, with a population of 13.9 million as of 2016, including 2.6 million residents aged <18 years.[158] Because of the single-payer healthcare system, almost all encounters with the system are captured in province-wide administrative databases. The data are accurate for identifying other pediatric diseases such as diabetes and asthma, as well as the receipt of immunizations.[159-161] Our objective was to assess the criterion validity of administrative data for identifying infections compared to electronic medical records (EMR) data as the reference standard.

4.3 Methods

4.3.1 Study Design, Population, and Setting

We conducted a validation study of infectious disease billing codes submitted by physicians compared to the reference standard of infections documented in a primary care EMR. We sampled a random cohort of Ontario residents aged <18 years who were under the care of family physicians who share their practice's EMR data with the Electronic Medical Record Administrative data Linked Database (EMRALD). Patient visits between April 1, 2012 and March 31, 2014 were randomly chosen for extraction and verification. During our sampling, we restricted the cohort to one visit per patient to minimize the impact of multiple visits for the same illness.

We used an intermediate-prevalence estimate to determine the sample size for the infectious syndromes with the goal to validate any infection. The estimated annual prevalence of otitis media infections in a pediatric population was 11.5% in Ontario.[162] Using the binomial

distribution, we needed 2,044 patients, with 235 patients with otitis media infections to obtain a specificity of 90% and a lower 95% confidence interval of 80%.[163]

4.3.2 Data Sources and Covariates

EMRALD is an advantageous data source for validating infection codes because it consists of all clinically relevant information from EMRs that can be linked to physician billing records within administrative databases. It has been used to validate other diseases.[164,165] EMRALD contains data for >400,000 patients who receive their primary care from a convenience sample of >350 family physicians distributed throughout Ontario who use the PS Suite® EMR. EMRALD contains clinical information such as a cumulative patient profile, progress notes, laboratory results, and prescriptions. Physicians participate in EMRALD on a voluntary basis, and are required to have had their EMR for ≥2 years to ensure it is adequately populated.

The Registered Persons Database contains basic demographic information on all individuals covered by provincial health insurance in Ontario (virtually the entire population) and was used to identify patient age, sex, and place of residence at the time of the physician office visit (index date). The child's postal code was linked to Canadian census data to determine rural residence (communities with <10,000 residents).[166] Postal code was also used to ascertain the quintile of neighbourhood material deprivation as derived from the Ontario Marginalization Index, with 1 being the least deprived and 5 being the most deprived.[83] The Ontario Health Insurance Plan (OHIP) database contains information on all physician billing claims, including diagnosis codes. Only one billing claim with an associated diagnosis code is processed for each service provided to the patient in the primary care setting. The diagnosis codes in OHIP are limited to 3 digits and is a truncated version of the International Classification of Diseases versions 7, 8 and 9, but also includes OHIP specific codes.[167] The ICES Physician Database contains information on all physicians practicing in Ontario and was used to obtain physician characteristics and specialization at the index date.

4.3.3 Abstraction of EMR Chart Data

An abstraction manual and structured data collection form were created to identify and collect information about the infections by anatomic region and specific infectious syndromes. We selected a group of clinical syndromes that accounted for the majority of physician office visits for infections (Table 4.1). These infections were chosen a priori based on the knowledge gained from

a systematic review and meta-analysis of common infections in children and the association with the development of childhood acute lymphoblastic leukemia.[168] We thought these infections would account for the majority of infection-related physician visits. We hierarchically defined each visit to assess whether the visit was for an infection, the corresponding anatomical region, and the specific infectious syndrome. Anatomic regions were respiratory, skin and soft tissue, gastrointestinal, urinary tract and otitis externa infections. The physician's diagnosis must have reported one of the syndromes listed in Table 4.1 to be categorized as an infection. A diagnosis was not inferred if none was explicitly stated. The abstractor was blinded to the submitted diagnostic billing codes. We also abstracted any complex chronic conditions that impact health services utilization,[169,170] and other chronic conditions from the cumulative patient profile. Since the abstractor did not have clinical experience (JH), and only one abstractor was used, we piloted the abstraction manual prior to full abstraction to clarify ambiguous situations, such as consultations with multiple diagnoses or complaints, and to measure the validity of the abstractor to correctly abstract the diagnoses from the medical charts. Diagnoses were abstracted verbatim from the medical charts to minimize subjective classifications. The results from the pilot were reviewed by co-authors with clinical experience (Drs. Sung and Kwong) to verify the validity and deemed them to be valid. If multiple diagnoses were made, all were kept and compared to the corresponding billing code.

4.3.4 Statistical Analysis

Duplicate abstraction of a random sample of 200 patient visits was performed (JH) to assess intra-rater reliability. We calculated Cohen's kappa, which measures the reliability of a single data collector who is presented with the same scenario interpreting the data and recording the same value.[105] We compared the demographic characteristics of the included and excluded patients using standardized differences and χ^2 test for categorical variables, and one-way ANOVA test for mean age.[171] A standardized difference >0.10 indicates a potential imbalance in the prevalence of a variable between included and excluded patients. Diagnoses of infections in EMRALD were used as the reference standard and linked to the OHIP database. We calculated sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for OHIP infection diagnosis codes occurring on the same day as the patient's physician office visit. We also examined discordant results between EMRALD and OHIP to ascertain the nature of the discordance. A binomial distribution was used for the performance measures to calculate 95% confidence intervals

(CI). We performed three sensitivity analyses to assess the performance measures based on: (1) if only one diagnosis was made, or a patient visited the physician for only one health complaint; (2) if multiple diagnoses were made at the time of the visit or a patient visited the physician for multiple complaints; and (3) patient characteristics stratified by age group, sex, rural versus urban residence, and presence of asthma and complex chronic conditions.

4.4 Results

We identified 48,744 eligible patients of 251 physicians practising in 39 different clinics in EMRALD, and successfully abstracted data from 2,438 randomly sampled patients. After linkage to the administrative databases and applying the exclusions, 2,139 patients remained for analysis. We excluded 35 patients due to data quality concerns, such as being ineligible for OHIP at index date, and 264 patients due to the visit date on the EMR and the billing date in OHIP not aligning. Intra-rater reliability was almost perfect [k=0.97 (95%CI 0.94-1.00)].

Characteristics of the patients and physicians in the study cohort are summarized in Table 4.2. We observed a difference in rural residence, and in age groups 0 to <2 and 2 to 5 years between included and excluded patients (Supplementary Table 4.1; Appendix B). There were 2,185 unique OHIP billing claims in our cohort, and of those 1,669 (76.4%) EMR visit notes contained 1 diagnosis and 490 (22.4%) EMR visit notes contained multiple diagnoses. We found 33.3% of the visits in the EMR were for an infection. In mutually inclusive categories, respiratory infections accounted for 22.5% of all visits, skin and soft tissue infections for 8.3%, gastrointestinal infections for 2.0%, urinary tract infections for 1.3%, and otitis externa infections for 0.9%.

When we combined all infection categories, sensitivity was 0.74 (95%CI 0.70-0.77), specificity was 0.95 (95%CI 0.93-0.96), PPV was 0.87 (95%CI 0.84-0.90), and NPV was 0.88 (95%CI 0.86-0.89) (Table 4.3). Respiratory infections performed similarly with a sensitivity of 0.77 (95%CI 0.73-0.81), specificity of 0.96 (95%CI 0.95-0.97), PPV of 0.85 (95%CI 0.81-0.88), and NPV of 0.94 (95%CI 0.92-0.95). However, lower sensitivity was observed for skin and soft tissue, gastrointestinal, urinary tract, and otitis externa infections (0.42-0.53, Table 4.3). Specific infectious syndromes had sensitivity ranging from 0.32 to 1.00, PPV ranging from 0.50 to 1.00, specificity ranging from 0.96 to 1.00, and NPV ranging from 0.94 to 1.00 (Table 4.4). The sensitivity analyses suggested that almost all categories of infectious syndromes performed better if only one diagnosis was made or patients visited the physician for only one issue. Additional

sensitivity analyses stratified by age group, sex, rural versus urban residence, asthma, and complex chronic conditions had similar performance to our primary analysis (Supplementary Table 4.2; Appendix B).

4.5 Discussion

Overall, we found that using linked EMR data as the reference standard, administrative billing codes are valid to identify infections in a pediatric population. The approach of measuring infections using administrative data performed best when the patient visited the physician with only one health complaint or if only one diagnosis was made. Administrative data performed well in capturing any infection and respiratory infections, while skin and soft tissue, gastrointestinal, urinary tract, and other ear infections maintained high specificity, but had lower sensitivity. Performance characteristics were similar among children with chronic diseases and complex chronic conditions. These results suggest administrative data can accurately capture infections with minimal risk of including false positives.

Other validation studies of administrative data to measure infections have shown consistent findings with our study.[157,172-178] These studies assessed hospitalizations or emergency room visits for respiratory infections, respiratory syncytial virus, rotavirus, pneumonia, skin infection, Clostridium difficile infection, and urinary tract infections. They found poor-to-high sensitivity (0.45% to 0.99), moderate-to-high specificity (0.69 to 1.00), poor-to-high PPV (0.55 to 1.00), and had to trade-off higher sensitivity for lower specificity or vice versa by expanding the number of International Classification of Diseases (ICD) diagnosis codes, the number of data fields, or the diagnosis types.[157,172-178] Our estimates for any infection, respiratory infection, and specific infectious syndromes such as otitis media and conjunctivitis performed well compared to these studies. The lower sensitivity observed for the other infections types such as gastrointestinal, urinary tract and otitis externa infections are likely due to the small number of events, and this is shown with the wide confidence intervals. The lower sensitivity for skin and soft tissue infections are likely due to the difficulties in determining the differences and causes between skin allergies and skin infections.

We found infections accounted for 33.3% of all visits to a physician, respiratory infections accounted for 67.6% of those infections, skin and soft tissue infections represented a 25.0% of the visits for an infection, gastrointestinal infections represented 5.9%, urinary tract infections

represented 3.8%, and otitis externa represented 2.6%. Infections continue to represent one of the most frequent reasons to seek healthcare in children and adolescents aged <18 years.[152-156]

Our study had several limitations. First, only one abstractor without clinical experience was used. However, our pilot demonstrated that one abstractor was able to abstract the diagnoses from the medical charts accurately and reliably. Second, our reference standard relied on the physician's clinical judgement and completeness of documentation. Third, we did not use laboratory confirmation to identify specific infectious agents. It is not known how well the syndromic data correlate with microbiological test results. However, a study in an emergency department setting demonstrated that respiratory syndrome diagnosis counts were associated with positive viral tests for infectious respiratory agents, and showed that the rate of respiratory syncytial virus and influenza virus was positively associated with respiratory syndrome counts (rate ratio = 1.51, 95%CI 1.10-2.07).[179]

The data available through EMRALD are from a voluntary sample of physicians in Ontario who all use one type of EMR system and practice under some type of primary care reform model of care, and therefore may not be entirely representative of all physicians in the province. In a 2011 study examining the impact of implementation of EMR in the EMRALD physician population, the authors found EMRALD physicians to be younger, more likely to be female, to be a Canadian medical graduate and to participate in patient-enrolment models compared to the general physician population in Ontario.[180] However, this likely reflects the characteristics of physicians who have adopted EMR software and trends in the primary care workforce. Ontario has been undergoing a primary care reform for more than a decade where the new primary care models require 'rostering' of patients (patient-enrollment models) and the physician acts as the their most responsible physician.[181] Although patients rostered in EMRALD are more likely to live in rural areas and be of higher socioeconomic status, the age, sex, presence of chronic conditions and measures of comorbidity are similar to rostered patients in Ontario.[90] The differences in physician characteristics between EMRALD and Ontario are unlikely to bias the internal validity of the study. While our findings provide insight into the validity of administrative data to identify infectious syndromes in Ontario, they may not be generalizeable to Ontario specialists or family physicians not participating in EMRALD, or to other jurisdictions where physician billing practices or disease classification systems may differ.

Our study demonstrates the diagnostic performance of a viable method to identify syndromic conditions for the use of syndrome-based burden of disease estimates using healthcare administrative data. Future priorities could include the development of a surveillance system using EMR data as demonstrated in other studies.[182] Other priorities could include investigations of factors, needs and healthcare barriers that contribute to inequalities in healthcare in vulnerable populations. For example, infectious diseases in children contribute substantially to healthcare utilization in primary care physician offices and at emergency departments. The associated annual cost for emergency department visits for infections was almost $10 billion in the United States in 2011.[183] However, the proportion of healthcare utilization for infections was disproportionally higher in children of lower socioeconomic status in the emergency department, but was lower in primary care offices.[156,183] Studies that address the potential needs, factors, and barriers to healthcare utilization are required to inform decision-makers of the most cost-effective, impactful population-based preventive interventions, and for resource planning.

Table 4.1 The infections of interest from the electronic medical records and the corresponding Ontario Health Insurance Plan (OHIP) physician billing claim diagnosis codes

Infections	**OHIP diagnosis code**
Respiratory infections	
Upper respiratory infections or common cold	460
Otitis media	381, 382
Conjunctivitis	372
Streptococcal sore throat	034
Acute sinusitis	461
Acute tonsillitis	463
Acute laryngitis or croup	464
Pertussis or whooping cough	033
Infectious mononucleosis	075
Lower respiratory infections	486, 487, 466
Pneumonia	486
Influenza	487
Acute bronchitis	466
Skin and soft tissue infections	
Warts	078
Impetigo	684
Chalazion or sty	373
Cellulitis	682
Chicken pox or varicella	052
Dental carries or dental abscess	521, 525
Boils	680
Herpes simplex	054
Ringworm	110
Candidiasis or thrush	112
Gastroenteritis or viral diarrhea	009
Pinworm	127
Urinary tract infections	590, 595, 599
Otitis externa infection	380

Table 4.2 Patient and physician characteristics of study cohort

Characteristic	EMRALD patients, n (%)
Number of patients	2139
Female	1039 (48.6)
Age, average (SD)	6.7 (5.4)
0 to < 2	530 (24.8)
2 to 5	509 (23.8)
6 to 9	384 (18.0)
10 to 14	488 (22.8)
15 to 18	228 (10.7)
Rural residence	410 (19.2)
Material deprivation	
1 least	613 (28.7)
2	453 (21.2)
3	408 (19.1)
4	366 (17.2)
5 most	294 (13.8)
Chronic conditions or illnesses*	
Complex chronic conditions	77 (3.6)
Allergies	27 (1.3)
Asthma or reactive airways	203 (9.5)
Behavioral and emotional disorders with onset usually occurring in childhood and adolescence	144 (6.7)
Mood disorders	21 (1.0)
Pervasive and specific developmental disorders	48 (2.2)
Physician Characteristics	
Number of physicians	259
Female	145 (56.0)
Age, average (SD)	44.0 (10.7)
<35 years	71 (26.7)
35 to 44 years	85 (32.0)
45 to 54 years	58 (21.8)
55 to 75 years	52 (19.6)
Rural practice	26 (10.0)
Family physician or general practitioner	255 (98.5)
Canadian medical graduate	230 (88.8)
International medical graduate	29 (11.2)
Years of practice, average (IQR)	17.0 (7 to 26)

*Chronic conditions were identified through the electronic medical record's cumulative patient profile; behavioural and emotional disorders, mood disorders and pervasive disorders were also identified through the cumulative patient profile as well as the diagnosis on the progress notes and were categorized based on International Classification of Disease-10 diseases categories. Material deprivation had 5 missing patients. SD represents standard deviation.

Table 4.3 Performance measures of the Ontario Health Insurance Plan physician billing claims for identifying infectious syndromes compared to electronic medical records

Classification of infection	% infection in EMR	% infection in AD	Sensitivity [95% CI]	Specificity [95% CI]	PPV [95% CI]	NPV [95% CI]
Performance of the different infections based on anatomic region, n=2185						
Any infection	33.3	28.1	74 (70-77)	95 (93-96)	87 (84-90)	88 (86-89)
Respiratory infection	22.5	20.5	77 (73-81)	96 (95-97)	85 (81-88)	94 (92-95)
Skin and soft tissue infection	8.3	4.8	49 (41-56)	99 (99-100)	86 (77-92)	96 (95-96)
Gastrointestinal infection	2.0	1.3	53 (38-69)	100 (99-100)	82 (63-94)	99 (99-99)
Urinary tract infections	1.3	1.0	50 (31-69)	100 (99-100)	64 (41-83)	99 (99-100)
Otitis externa infection	0.9	0.5	42 (20-67)	100 (100-100)	67 (35-90)	99 (99-100)
Performance of different infections based on anatomic regions - Only 1 diagnosis was made at the visit, n=1669						
Any infection	30.4	27.4	79 (76-83)	95 (94-96)	88 (84-91)	91 (90-93)
Respiratory infection	20.3	20.1	84 (80-88)	96 (95-97)	85 (81-89)	96 (95-97)
Skin and soft tissue infection	7.3	5.0	57 (47-66)	99 (98-99)	82 (72-90)	97 (96-97)
Gastrointestinal infection	1.7	1.2	55 (36-74)	100 (99-100)	80 (56-94)	99 (99-100)
Urinary tract infections	0.7	0.8	73 (39-94)	100 (99-100)	62 (32-86)	100 (99-100)
Otitis externa infection	0.6	0.4	50 (19-81)	100 (100-100)	83 (36-100)	100 (99-100)
Performance of different infections based on anatomic regions - Multiple diagnoses was made at the visit, n=490						
Any infection	44.9	30.8	61 (54-67)	94 (90-96)	89 (83-93)	75 (70-79)
Respiratory infection	31.0	22.0	62 (54-70)	96 (93-98)	87 (79-93)	85 (81-88)
Skin and soft tissue infection	12.2	4.1	33 (22-47)	100 (99-100)	100 (83-100)	91 (89-94)
Gastrointestinal infection	2.7	1.6	54 (25-81)	100 (99-100)	88 (47-100)	99 (97-100)
Urinary tract infections	3.5	1.8	35 (14-62)	99 (98-100)	67 (30-93)	98 (96-99)
Otitis externa infection	1.8	1.2	33 (7-70)	99 (98-100)	50 (12-88)	99 (97-100)

EMR=electronic medical records, AD=administrative data, PPV=positive predictive value, NPV=negative predictive value.

Table 4.4 Performance measures of the Ontario Health Insurance Plan physician billing claims for identifying specific infectious syndromes compared to electronic medical records

Classification of infectious syndrome	**% infection in EMR**	**% infection in AD**	**Sensitivity [95% CI]**	**Specificity [95% CI]**	**PPV [95% CI]**	**NPV [95% CI]**
Upper respiratory infection + conjunctivitis + otitis media	18.9	17.8	75 (71-80)	96 (94-96)	80 (75-84)	94 (93-95)
Upper respiratory infection (Pharyngitis, sinusitis, tonsillitis, laryngitis, or streptococcal sore throat)	13.6	11.9	69 (63-74)	97 (96-98)	79 (73-84)	95 (94-96)
Otitis media	4.7	4.4	72 (62-80)	99 (98-99)	77 (67-85)	99 (98-99)
Conjunctivitis	1.4	1.6	77 (58-90)	99 (99-100)	68 (49-83)	100 (99-100)
Strep throat	2.2	1.0	32 (19-47)	100 (99-100)	71 (48-89)	99 (98-99)
Bronchitis	0.6	0.7	64 (35-87)	100 (99-100)	56 (30-80)	100 (99-100)
Croup or laryngitis	0.8	0.4	41(18-67)	100 (100-100)	88 (47-100)	100 (99-100)
Tonsillitis	0.5	0.6	70 (35-93)	100 (99-100)	50 (23-77)	100 (100-100)
Sinusitis	0.5	0.5	73 (39-94)	100 (100-100)	67 (35-90)	100 (100-100)
Infectious mononucleosis	0.5	0.2	40 (12-74)	100 (100-100)	100 (40-100)	100 (99-100)
Lower respiratory infection (unspecified lower respiratory infection, pneumonia, influenza, or acute bronchitis)	3.1	2.4	62 (49-73)	100 (99-100)	81 (67-90)	99 (98-99)
Pneumonia	2.0	1.3	60 (44-75)	100 (100-100)	90 (73-98)	99 (99-100)
Warts	2.7	2.1	69 (55-80)	100 (99-100)	87 (74-95)	99 (99-100)
Impetigo	1.0	0.7	59 (36-79)	100 (100-100)	87 (60-98)	100 (99-100)
Chalezon or stye	0.6	0.4	54 (25-81)	100 (100-100)	88 (47-100)	100 (99-100)
Cellulitis	0.5	0.5	55 (23-83)	100 (100-100)	60 (26-88)	100 (99-100)
Gastroenteritis, viral diarrhea, or viral gastritis	1.7	1.2	59 (42-75)	100 (99-100)	81 (62-94)	99 (99-100)
Urinary tract infections	1.3	1.0	50 (31-69)	100 (99-100)	64 (41-83)	99 (99-100)

Infectious syndromes with ≤10 events from the electronic medical record are not reported. EMR=electronic medical records, AD=administrative data, PPV=positive predictive value, NPV=negative predictive value.

Chapter 5 : Manuscript titled Rate of infections and the association with childhood acute lymphoblastic leukemia: a population-based case-control study

5.1 Abstract

Introduction: The etiology of childhood acute lymphoblastic leukemia (ALL) is uncertain, however, an infectious trigger for ALL is hypothesized. We assessed the association between the rate, type, severity and critical exposure period for prior infections and the development of ALL.

Methods: We conducted a matched case-control study using administrative databases to evaluate the association between the rate of infections and childhood ALL diagnosed between the ages of 2-14 years from Ontario, Canada between 1995 and 2014. We matched 10 controls to each ALL case on date of birth, sex, and location of residence. We used a validated measure for infections to determine the rate of infections among the study cohort. Odds ratios were estimated using adjusted conditional logistic regression models. The mean number of infections over time was calculated using the mean cumulative function.

Results: In 1,600 cases of ALL, and 16,000 matched cancer-free controls aged 2-14 years, having >2 infections/year increased the odds of childhood ALL by 43% (OR=1.43, 95%CI 1.13-1.81) compared to children with ≤0.25 infections/year. Having >2 respiratory infections/year increased odds of ALL by 28% (OR=1.28, 95%CI 1.05-1.57) compared to children with ≤0.25 respiratory infections/year. Having an invasive infection increased the odds of ALL by 72% (OR=1.72, 95%CI 1.31-2.26). Having an infection between the age of 1 to 1.5 years increased the odds of ALL by 20% (OR=1.20, 95%CI 1.04-1.39). The cumulative incidence of infections was slightly higher for children with ALL compared to cancer-free controls. Both cases and controls had decreasing recurrence rates for infections over time.

Conclusions: Infections in childhood may be an important factor in the development of childhood ALL. This study indicated that infections between the ages of 1 to 1.5 years may be an important time period and which types of infection may play a larger role than others.

5.2 Introduction

Childhood acute leukemia is the most common cancer in children, with ~230 new cases annually in Canada.[3] Childhood acute lymphoblastic leukemia (ALL) accounts for 80% of all leukemias in high income countries and peak incidence occurs between 2 and 5 years of age. [3,184] However, the etiology of childhood ALL is mostly unknown. ALL may be present in utero and may arise from an interaction between exogenous and/or endogenous exposures, genetic susceptibility, and chance. With genetic causes accounting for only a small proportion of ALL cases,[15] other promotional exposures are likely necessary for disease emergence.

Kinlen and Greaves have both hypothesized that infections may play a role in the development of ALL. Kinlen proposed the 'population mixing' hypothesis to describe the observed increased rates of childhood ALL following an influx of mostly urban migrants into a rural area with an isolated population. The contact between infected and susceptible individuals create a localized epidemic of an underlying viral infection that may produce the rare response of ALL.[16,17]

Greaves' 'delayed infection' hypothesis for childhood ALL suggests a two-hit model that emphasizes the child's immune system and the timing of infectious exposure. The hypothesis describes a prenatal initiation of pre-leukemic clones as the first hit, followed by postnatal promotion, secondary mutation and overt disease as the second hit. In a small number of pre-leukemic carriers, it is the absence of infectious exposure in early life, and a postnatal secondary genetic event caused by a delayed, stress-induced infection (second hit) on the developing, "unprepared" immune system that leads to the development of ALL. The latency period after initiation can be variable, ranging from a few months to 15 years.[12] While the mechanisms differ between the hypotheses, both suggest ALL is a rare response to one or more infections early in life.

The objectives of this study are to assess whether Ontario children diagnosed with ALL between 1995 and 2014 and the ages of 2 and 14 years have a higher rate of infections prior to the development of ALL compared to cancer-free children, and whether different types of infections, severity and critical exposure period for infections are associated with the development of ALL.

5.3 Methods

The study was approved by University of Toronto's Health Sciences Research Ethics Board. The Institute for Clinical Evaluative Sciences (ICES) is named as a prescribed entity under Ontario's privacy legislation. Under this designation, ICES can receive and use health information without consent for the purposes of health-related research and health system analysis and evaluation.[82] Individual-level patient health information was linked across multiple databases using unique coded identifiers to create a complete health services profile for each subject.

5.3.1 Study Design, Population, and Setting

We conducted a matched case-control study to evaluate the association between the rate of infections and the development of childhood ALL in children from Ontario, Canada aged 2-14 years at the time of diagnosis between 1995 and 2014. Children diagnosed with ALL before 1 year of age were excluded because they are linked to genetic factors and not theorized to have an infectious etiology.[12] We started our inclusion of diagnosed ALL at age 2 to ensure cases and controls had at least 1 year of observation prior to the diagnosis. The case's diagnosis date was used as the index date for the matched controls. We matched without replacement 10 controls with no previous cancers to each case of ALL on the case's date of birth, sex, location of residence (urban or rural) at the beginning of the observation period. Matching on the date of birth allows for equal length of observation look-back periods which permits subsequent analyses. We only matched on the length of the observation look-back period if not born in Ontario. For example, a case not born in Ontario would be matched with controls using the same observation look-back period defined as the time between the start of the case's OHIP coverage to 1 year prior to the index date. Age, sex and location of residence have been shown to be associated with childhood ALL.[5,17]

5.3.2 Data Sources and Covariates

The Pediatric Oncology Group of Ontario Networked Information System (POGONIS) captures information on the timing and definitive diagnosis of cancer, staging, and demographic information on subjects that are diagnosed and or treated at one of 5 tertiary care hospitals in Ontario that treat children with cancer. This registry captures 98% of all cancers in children under 15 years in Ontario, Canada.[92] POGONIS classifies childhood cancers based on morphology, and

uses diagnosis codes that map onto the International Classification of Diseases for Oncology (ICD-O).[92]

The Registered Persons Database contains basic demographic information on all individuals covered by provincial health insurance in Ontario (virtually the entire population) and was used to identify each patient's date of birth, sex, and location of residence. The child's postal code was linked to Canadian census data to determine rural residence (communities with <10,000 residents).[166] Postal code was also used to ascertain the four dimensions of the Ontario Marginalization Index (ON-Marg), a comprehensive measure of socioeconomic status.[83] The four dimensions include the quintile of neighbourhood dependency, material deprivation, ethnic concentration, and residential instability. We used ON-Marg dimensions at the start of the observation period (i.e. at birth or start of OHIP eligibility), or if missing, at the first available year. The Ontario Health Insurance Plan (OHIP) database contains information on all physician billing claims, including diagnosis codes for infections. We used the Canadian Institute for Health Information (CIHI) Discharge Abstract Database (DAD) and the National Ambulatory Care Reporting System (NACRS) to identify hospitalizations and emergency room visits for infections, respectively. Both CIHI datasets use International Classification of Diseases Ninth Revision (ICD-9) before 2001 and Tenth Revision (ICD-10) after 2001. The Immigration, Refugees and Citizenship Canada (IRCC) Permanent Resident Database contains immigration data for Ontario's permanent residents who landed from 1985 to 2012

An a priori causal diagram of the relationship between prior infections and the development of childhood ALL was constructed to identify covariates and confounders to consider for the analysis (Supplementary Figure 5.1; Appendix C). ON-Marg dimensions and immigrant status were considered as covariates, since these factors could affect access, use of health services, contraction and identification of infections. Down syndrome was defined as those with one of the following codes at any point during the observation period: an OHIP diagnosis code of 758, a DAD or NACRS main diagnosis code 758 for ICD-9, or Q90 for ICD-10. Down syndrome was considered a confounder because of its association with ALL and infections.[72,73] The IRCC Permanent Resident Database was used to obtain immigration status at the index date. Recent immigrants were defined as children who landed in Ontario within 5 years of the index date.

Immigrant status was also considered as a confounder due to different infectious disease patterns and certain ethnic groups having lower ALL incidence rates.[185,186]

5.3.3 Outcome and Exposure Definitions

POGONIS was used to identify children diagnosed with ALL, defined as ICD-O morphology codes 9821 for ICD-O-2, and 9835, 9836, and 9837 for ICD-O-3. First primary cancers of ALL were included as cases. History of infection from birth up to 1 year prior to the index date was identified using OHIP, NACRS, and DAD. OHIP was used to identify ambulatory care visits and emergency room visits before 2001 (OHIP codes), and NACRS for emergency room visits after 2001 (ICD-10 codes). DAD was used to obtain hospitalizations for infections.

We selected a group of clinical syndromes that accounted for the majority of physician office visits for infections (Table 5.1). We hierarchically defined each visit to assess whether the visit was for an infection followed by the corresponding anatomical region. Anatomical regions included respiratory, skin and soft tissue, gastrointestinal, urinary tract, otitis externa, and invasive infections (Supplementary Table 5.1; Appendix C). In our previous validation work validating health administrative data diagnostic codes against primary care electronic medical records, we found any infection (a combination of all anatomical regions) had a sensitivity of 0.74 (95% confidence interval, CI 0.70-0.77), specificity was 0.95 (95%CI 0.93-0.96), positive predictive value (PPV) was 0.87 (95%CI 0.84-0.90), and negative predictive value (NPV) was 0.88 (95%CI 0.86-0.89). The administrative data performed well in capturing any infection and respiratory infections, while skin and soft tissue, gastrointestinal, urinary tract, and other ear infections maintained high specificity (range 0.99 to 1.00) but had lower sensitivity (range 0.42 to 0.53).

The anatomic region-specific infections have only been validated in a primary care setting but will also be used for hospitalizations and emergency room visits for infections. We defined hospitalizations using discharge records that listed any infection as the most responsible diagnosis for the hospitalization. The most responsible diagnosis or main diagnosis codes were used to identify the infections within the CIHI DAD and NACRS data sources, respectively. Infections occurring within the previous 365 days of the diagnosis date or index date were excluded to prevent lag-time bias. We applied episode lengths to avoid counting visits for the same infection multiple times for our recurrent events modeling. We defined episode length as the amount of time that must have elapsed between visits for the same infection in the health administrative databases to

be considered separate events in an individual. The episode length for respiratory infections was 21 days,[70,187,188] skin and soft tissue infection was 30 days,[70] gastrointestinal infection was 14 days,[70,189] urinary tract infection was 30 days,[70] otitis externa was 30 days,[70] and invasive infections was 3 years.[70]

5.3.4 Statistical Analysis

The rate of infections for each individual was calculated using the number of infections (numerator) divided by the total observation period in days (denominator). We categorized the rate of infections as ≤0.25 infection per year, >0.25 to 0.50 infection per year, >0.50 to 1 infection per year, >1 to 2 infections per year, and >2 infections per year. These categories were chosen to account for the higher rate of visits to physicians for infections in younger ages and a lower rate at older ages.[154,162,190-192] Peak incidence of infections in this cohort occurs between the ages of 0 to 4 years with a mean of 4.9 respiratory infections per year which declines to a mean of 2.8 infections per year for children aged 5-19 years.[154] In another study, children under 18 years of age averaged 3 episodes of viral respiratory infections in the past year, but only 31.7% of children visited a physician for the infection.[191]

Descriptive analyses were first conducted on the matched cases and controls. The distributions of the ON-Marg dimensions, presence of Down syndrome, immigrant status, and rate of infection between cases and controls were compared using chi-squared tests, and mean age and length of immigration among immigrants were compared using t-tests. Conditional logistic regression, accounting for the case-control matched set, was performed to generate odds ratios (OR) and 95%CI estimating the odds of ALL associated with the rate of all infections categorized as described, rate of respiratory infections, and by the presence (yes or no) of infections corresponding to the anatomical regions skin and soft tissue, gastrointestinal, urinary tract, otitis externa and invasive infections.

Adjusted models included the ON-Marg dimensions (dependency, material deprivation, ethnic concentration, and residential instability), Down syndrome, and immigrant status. We used a model building strategy most applicable to etiologic research to obtain valid estimates of an exposure-disease relationships that tests for and accounts for confounding and effect modification, and interaction terms were tested using the likelihood ratio tests.[193] Once this group of covariates and confounders were identified by the a priori causal model (Supplementary Figure 5.1; Appendix

C), they were included into the final model. The approach uses a hierarchically well-formulated model to assesses interactions of included variables and rate of infections and assesses empirical confounders (if variable changes the effect estimate for infections by more than 10%, and whether it impacts the precision of the effect estimate). If a variable was not an empirical confounder, it was still included into the final model.

In a subgroup restricted to matched sets where the observation began at birth, we conducted a critical exposure period analysis to examine the time period when having an infection has the strongest effect on the development of ALL.[194] The exposure periods were defined as having an infection (yes or no) in the time periods at age 0 to 1 year, 1 to 1.5 years, 1.5 to 2 years, 2 to 2.5 years, 2.5 to 3 years, 3 to 3.5 years and 3.5 to 4 years. Each exposure period was treated as a separate binary covariate, whether the child had or did not have an infection during that period. The critical exposure period analysis used a joint model approach that included all periods under one model, adjusted for ON-Marg and Down syndrome. Using a joint model adjusts for the other exposure periods.[195] The correlation coefficients for all exposure period variables were inspected using a correlation matrix and were determined to be sufficiently low signifying that issues of multicollinearity were not indicated (the correlation with the largest magnitude between exposure period variables was -0.17).

To address potential residual confounding, we conducted sensitivity analyses on confounders (i.e., Down syndrome and immigrant status) by removing from the cohort matched sets that contained a child with Down syndrome or an immigrant child. To assess the effect of including individuals without complete exposure histories, a sensitivity analysis was conducted on a cohort of matched sets with observation periods starting at birth to assess infections and ALL with complete exposure information from birth onwards. We conducted another sensitivity analysis restricted to our validated infection definitions of infections diagnosed in physician offices to assess robustness of the findings and potential effect of using non-validated measures of infections.

We used a mean cumulative function approach under a recurrent event modeling framework to assess cumulative exposure to infections over time for various groups of individuals. The model allowed the depiction of the mean cumulative number of infectious disease events over

time, starting from birth, and whether the intensity of infectious diseases increases or decreases with time.[196] Analyses were conducted using SAS Enterprise 7.4® and R version 3.1®.

5.4 Results

In our analysis, 100% of the eligible cases were matched to 10 controls which included 1,600 ALL cases and 16,000 matched cancer-free controls (Figure 5.1). The median age at index date was 4 years (interquartile range 3-8), 43.1% were females, 12.0% lived in rural areas, most of the cases of ALL were diagnosed between 2005 and 2010 (30.0%), and cases and controls had similar ONMarg characteristics (Table 5.2). None of the variables were confounders based on the applied modeling approach. Cases were more likely to have Down syndrome (4.2% vs. 0.5%) and to be immigrants (3.4% vs. 0.4%) compared to controls, respectively. Controls that are immigrants have been in Ontario for longer than cases that are immigrants.

5.4.1 Rates of infection

In the cohort, 47.8% had >2 infections per year. By anatomical region, 38.3% had >2 respiratory infections per year, and throughout the observation period, 43.0% had a gastrointestinal infection, 35.6% had a skin or soft tissue infection, 12.7% had a urinary tract infection, 11.7% had otitis externa, and 2.5% had an invasive infection. Having >2 infections/year increased the odds of childhood ALL by 43% (OR=1.43, 95%CI 1.13-1.81; Table 5.3) compared to children with ≤0.25 infections/year. Being an immigrant child increased the odds of developing ALL by ~15-fold (OR=14.68, 95%CI 9.30-23.16). Having Down syndrome increased the odds of developing ALL by ~9-fold (OR=8.85, 95%CI 6.31-12.40). The ON-Marg dimensions of dependency, material deprivation, ethnic concentration, and residential instability were not associated with the development of ALL. A global test of interactions and separate individual interaction terms of included variables and the rate of infections were assessed using the likelihood ratio test, and no interaction model demonstrated evidence of an interaction on the multiplicative scale (data not shown).

5.4.2 Types and Timing of infections

Having >2 respiratory infections/year increased the odds of childhood ALL by 28% (OR=1.28, 95%CI 1.05-1.57; Table 5.4) compared to children with ≤0.25 respiratory

infections/year. Having an invasive infection increased the odds of ALL by 72% (OR=1.72, 95%CI 1.31-2.26). Having any hospitalization for an infection suggested an increase in the odds of ALL by 11% (OR=1.11, 95% CI 0.99-1.25). No associations were found for the other infection types.

In the critical period analysis, keeping the matching design and using a restricted subgroup of 1,268 cases matched to 12,680 controls where the observation started at birth, having an infection between the age of 1 to 1.5 years increased the odds of developing ALL by 20% (OR=1.20, 95%CI 1.04-1.39) compared to not having an infection within the same exposure period, after controlling for ON-Marg dimensions, Down syndrome and the other exposure periods (Figure 5.2). Exposure to infections in any other period was not associated with ALL.

5.4.3 Sensitivity analyses

To assess residual confounding, keeping the 1:10 matching design restricted to children without Down syndrome and among non-immigrants, we found stronger results compared to our primary analysis and demonstrated similar findings (Supplementary Table 5.2; Appendix C). In the sensitivity analysis restricting to matched sets with observation periods starting at birth to assess the effect of including individuals without complete exposure histories, the association between the rate of infections and ALL was stronger. Children with >2 infections/year had 67% increased odds of childhood ALL (OR=1.67, 95%CI 1.23-2.28) compared to children with ≤0.25 infections/year (Table 5.5).

In our sensitivity analysis that was restricted to our previously validated definition of infections in a primary care setting, we found similar results to our primary analysis (Supplementary Table 5.3; Appendix C).

5.4.4 Mean cumulative number of infections

Figure 5.3 illustrates the mean cumulative number of infections over time (and the corresponding 95% CI) for children with ALL and cancer-free controls. Although this was a case-control study design, we were able to use this mean cumulative function method since the observation look-back period was the same within a matched set. Figure 5.3 demonstrates that the cumulative incidence of infections was slightly higher for children with ALL and stayed higher compared to cancer-free controls (throughout the observation period/over time). Both children

with ALL and cancer-free controls had decreasing recurrence rates for infections. Similar patterns in the mean cumulative number of infections over time were observed when examining children diagnosed with ALL between ages 2 and 5 years and their matched controls (Figure 5.4a), and children without Down syndrome (Figure 5.4b) and non-immigrants (Figure 5.4c). Figure 5.4a shows cases and controls begin to diverge around the ages 1 to 2 years, suggesting cases begin to experience more infections around this time.

5.5 Discussion

In our study of children aged 2-14 years, we found having >2 infections/year increased the odds of childhood ALL by 43% compared to children with ≤0.25 infections/year, and over time the rate of infections in cases was higher than that of controls. The association between the rate of infections and ALL was even stronger among a cohort of matched cases and controls when the observation period started at birth. Certain types of infections are more likely to be associated with the odds of ALL than others, specifically respiratory and invasive infections. Finally, an infection between the ages of 1 and 1.5 years may be a critical period for infections in the development of ALL. This study does not confirm but presents evidence that suggests children who develop ALL may have dysregulated immune function that is already present in early childhood and leads them to have more clinically severe infections to the extent that medical care may be required.[13] Further biological testing is needed to confirm the findings.

The main strengths of our study are the population-based matched case-control study sample with complete longitudinal observation from birth to disease for cases and controls and the use of a validated method to measure infections that also include the date of the physician visit. Other studies that used administrative data or medical records have also suggested infections increased the odds of ALL,[30,31] although the literature is inconclusive.[168] Some studies that used self-reported measures to ascertain infection history found infections in childhood reduced the odds of ALL.[25,26,41,42] These differences across studies may be due to heterogeneity in the definition and measurement of infections. Other studies that used administrative data did not use a validated method to measure infections, and thus it would be difficult to quantify the degree of misclassification. Second, studies that used self-reported measures to ascertain infection history are subject to recall bias and suggested mothers of cases under-reported childhood infections more

than mothers of controls; further they may be inaccurate in both the timing and occurrence of infections compared to medical records.[44,46,47]

A previous study that also examined the rate of infections and the development of ALL after the age of 2 years found cases had a higher rate of infections in the first year of life compared to controls.[43] The authors also found cases consistently had a higher rate of infections. Our results demonstrated the rate of infections between cases and controls follow a similar pattern, with cases having a slightly higher rate than controls throughout. This supports the notion that children who develop ALL may have a dysregulated immune system at birth that leads to a greater propensity to require medical care during infections.[13] However, we did not find a dose-response relationship between the rate of infections and childhood ALL.

To our knowledge, this is the first study to examine whether infections and the development of ALL followed a critical period model. A previous study took an exploratory approach to examine the distribution in the rate of infections by time to ALL diagnosis and age, and found the rate of infections was markedly higher in cases during the 5 months preceding the diagnosis of ALL.[43] Unlike other studies, we found infections that occur after the first year of life to be more important than infections occurring in the first year of life.[30,31,43] The timing coincides with the start of daycare for most children in our population and we can not rule out the effect of daycare attendance on the observed effect.[197] Nonetheless, using a life course approach allowed us to examine the relationship between the timing of infections and ALL while also adjusting for the different exposure periods.

Previous studies showed inconclusive evidence with respect to the association between severity of infections and childhood ALL.[25,28,36,37] These studies assessed the relationship between hospitalizations and childhood ALL, and two studies showed a positive association with ALL.[25,37] We showed that children with ALL are more likely to have invasive infections, and may be more likely to be hospitalized for an infection. Children who later develop ALL were found to have a lack of immunomodulation from lower levels of anti-inflammatory cytokine interleukin-10 which may cause sensitivity and a higher susceptibility to infections.[63] Interleukin-10 has emerged as a key immunoregulator during infection with viruses, bacteria, fungi, protozoa, and helminths. The removal of the cytokine results in the onset of severe immune responses.[198] Among the children with ALL, our observed higher exposure rate of invasive infections and hospitalizations for

infections could be attributed to the lack of immunomodulation in children with ALL. Certain interleukein-10 polymorphisms have also been shown to be associated with cancer[199,200] and childhood ALL.[201]

Interleukin-10 has also been found to predict risk for respiratory infections in children,[202] and lower levels of interleukin-10 was found in children with severe Mycoplasma pneumoniae pneumonia.[203] Interleukin-10 is also associated with severity of respiratory syncytial virus bronchiolitis,[204,205] and with other infections and diseases not considered in this study.[206] This may explain why our study found children with ALL were more likely to have respiratory infections compared to children without cancer.

There are limitations to the data sources used in this study that need to be considered. While we validated the definition for infections used in this study, that validation was within a primary care setting. We are unsure of the accuracy of the data to capture diagnoses of infections in the emergency department and during hospitalizations. However, when we restricted our analysis to infections within the primary care setting (OHIP dataset), we found similar results, and the majority of visits for infections among children occur in the primary care setting.[152-154] We were unable to capture those with an infection but did not seek medical care. Barriers to access to care and reasons for avoiding medical care have been reported elsewhere,[207,208] and even under a universal healthcare system, there were differences in the access to health services for children.[209] However, the analysis restricted to matched sets with observation periods starting at birth does provide insight into having access to the healthcare system since birth. The analysis suggested a stronger relationship between the rate of infections and the development of childhood ALL. We were unable to account for other potential confounders such as ethnicity,[35] daycare attendance,[51] traffic emissions and genetic factors.[12,141] We were however able to address residual confounding due to Down syndrome and immigrant children by conducting an analysis that removed children with Down syndrome and immigrant children and observed similar findings. While cases may be more likely to be immigrants, and controls that are immigrants are more likely to have been in Ontario for longer, the sensitivity analysis that removed immigrants suggests these differences had minimal impact on the infections and ALL relationship. Further, the rate of infection and immigrant interaction term was not significant during the model building process, and additional interaction testing using length of time since immigration and other infection types increases the

chance of type 1 error.[210] The length of time since immigration is unlikely to effect the rate of infection, rather it may impact other unmeasured factors such as stress-level, behaviour and socio-cultural constructs. Since immigrant status was not a matching factor, a stratified analysis using the small number of immigrants are insufficiently powered to test for additional associations. The large magnitude of the OR and wide confidence intervals are also likely due to small numbers or chance, since only 117 cases and controls were immigrants. The immigration data were unable to identify immigrants who landed in different provinces and entered Ontario afterwards. However, between 1991 to 2006 landing years, 91% of the immigrants in Ontario who filed for taxes had originally landed in Ontario,[211] thus we can be confident that we correctly captured most immigrants. Another limitation was the use of the joint model which may not be as sensitive in capturing effect estimates if the critical windows did not align with the predefined windows. However, since we were interested in a cumulative measure and infections are often sporadic, using a distributed lag model would not be appropriate. The distributed lag model assumes the exposure and outcome to vary smoothly throughout the time period.[195] Finally, our results may not be generalizable to other populations with different baseline characteristics.

Overall, the present study found infections increased the odds of developing childhood ALL, the ages of 1 to 1.5 years may be an important time period for the impact of infections, and certain infections may be more important than others in the development of ALL. Future studies will need to combine relevant epidemiologic, biological, and environmental risk factors to elucidate the important individual and joint effects in the etiology of childhood ALL.

Figure 5.1 Study flow diagram

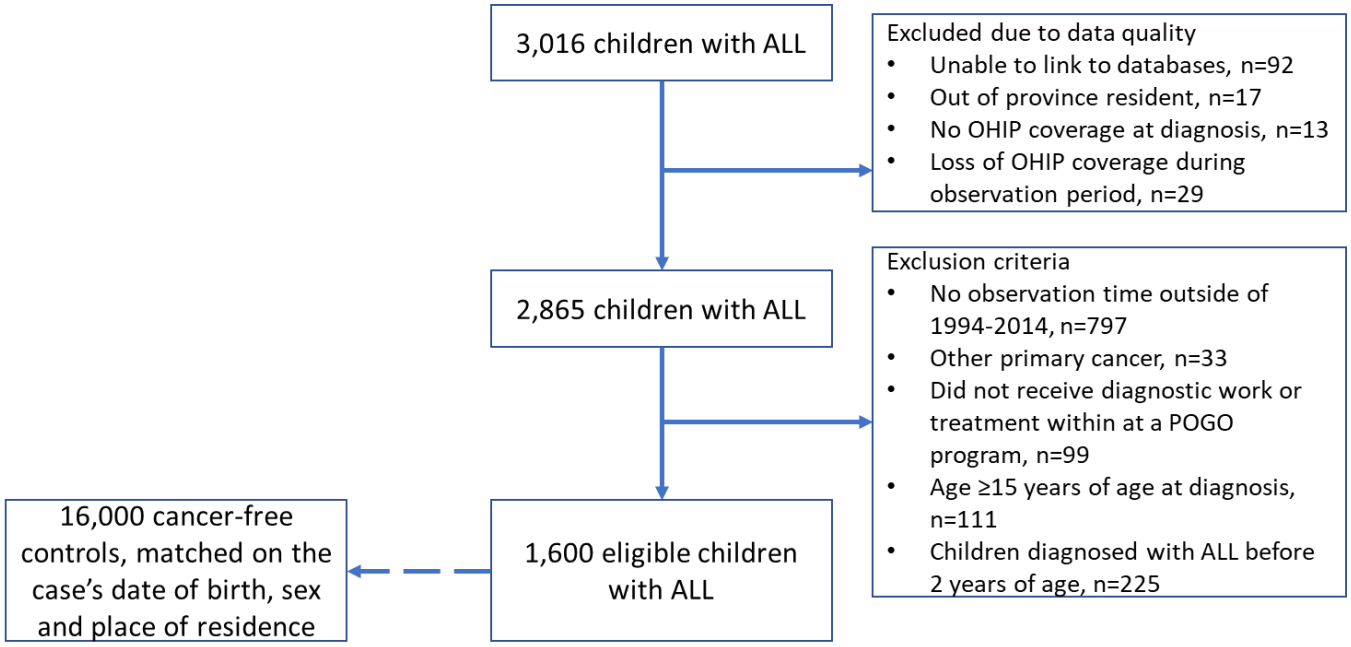
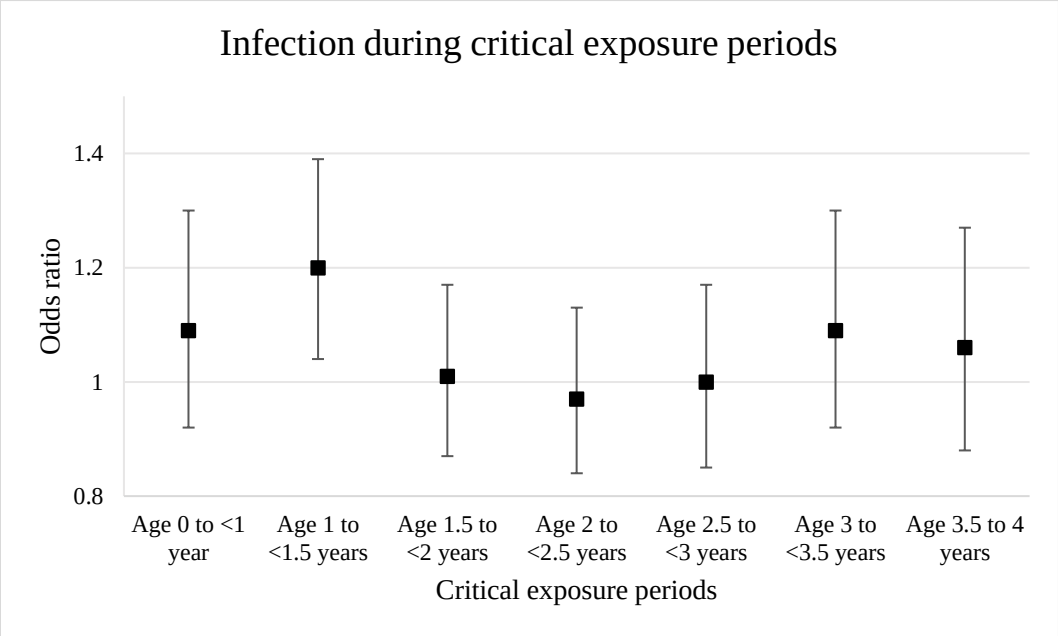


Figure 5.2 Critical exposure period analysis examining infections in each of the exposure periods, restricted to a birth cohort



The point estimates are odds ratios and bands are 95% confidence intervals from an adjusted model that included each of the exposure periods and controlled for Ontario Marginalization Index dimensions and Down syndrome. The reference category was no infection during that exposure period. The cohort consisted of 13,948 children, maintaining the 1:10 case and control matched pairs.

Figure 5.3 Mean cumulative number of infections over time (along with 95% confidence intervals) for children with acute lymphoblastic leukemia and cancer-free matched controls

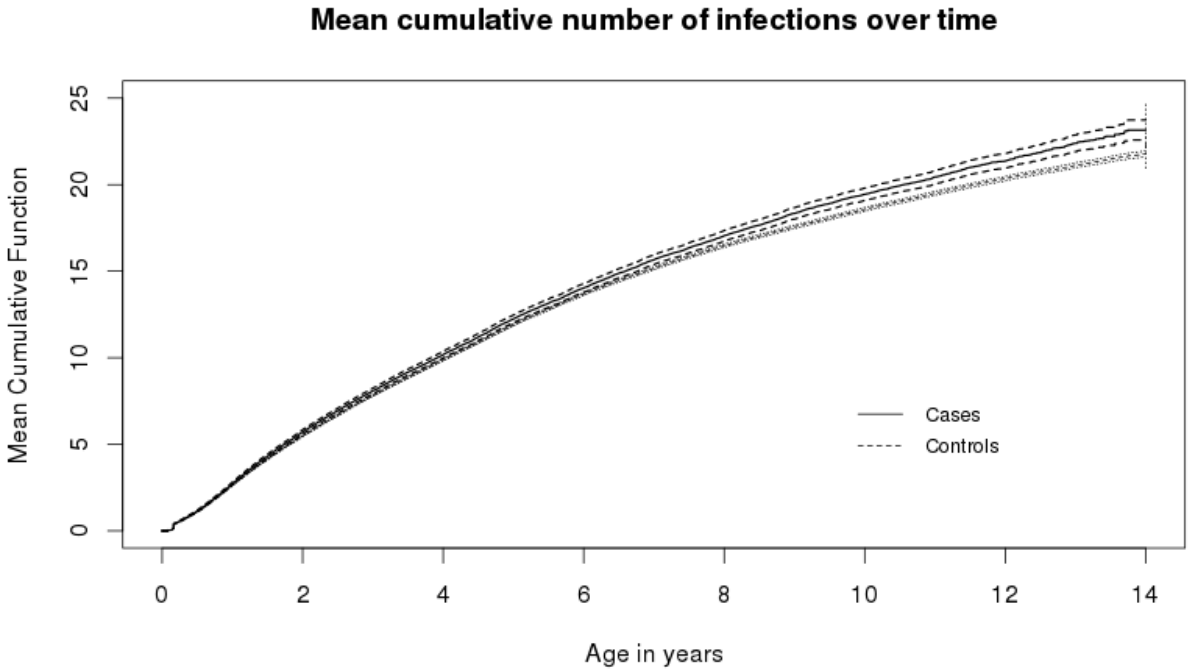
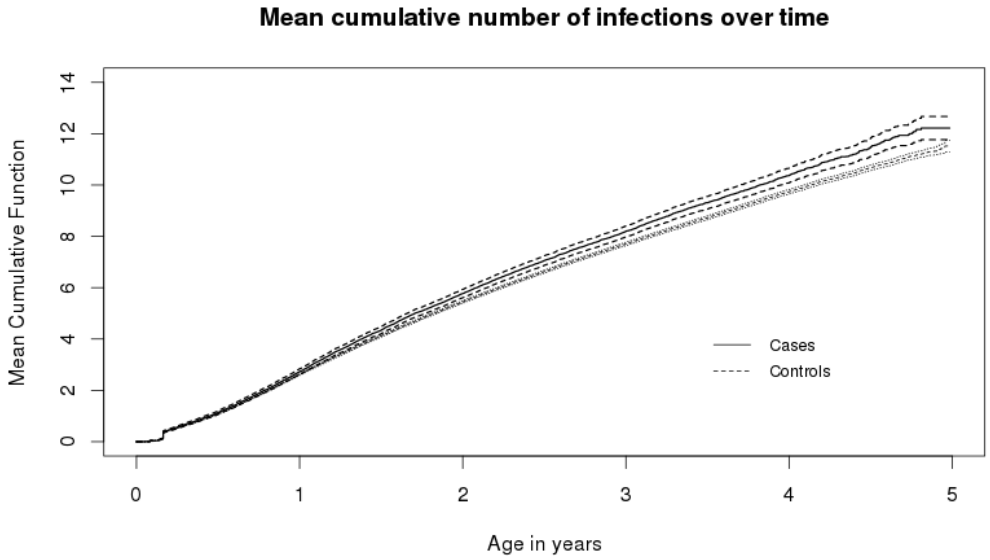
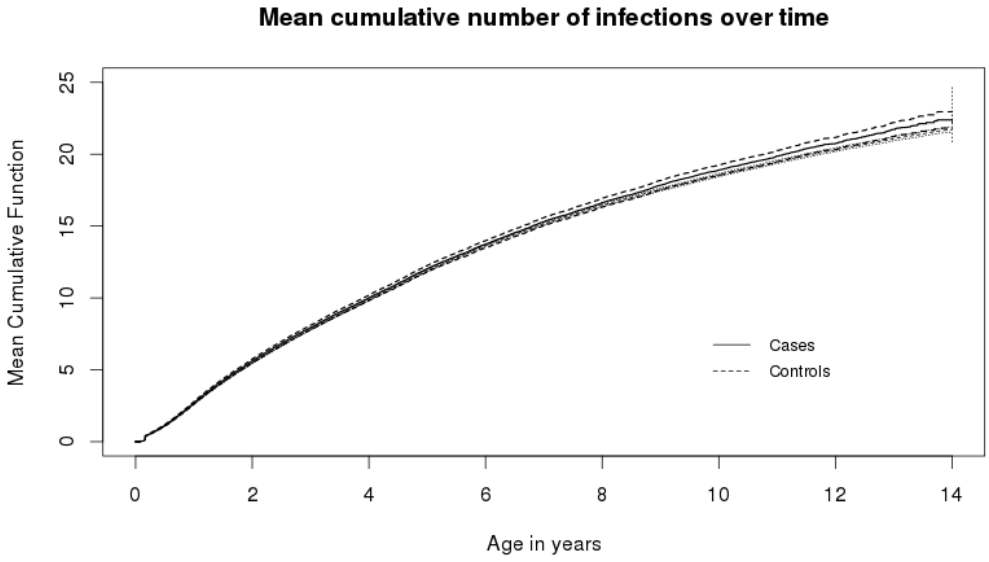


Figure 5.4 Mean cumulative number of infections over time (along with 95% confidence intervals) for children with acute lymphoblastic leukemia and cancer-free matched controls, diagnosed between 2-5 years of age (a), without Down syndrome (b), and non-immigrants (c)

a. Age group 2-5 years



b. Without Down syndrome



c. Non-immigrants

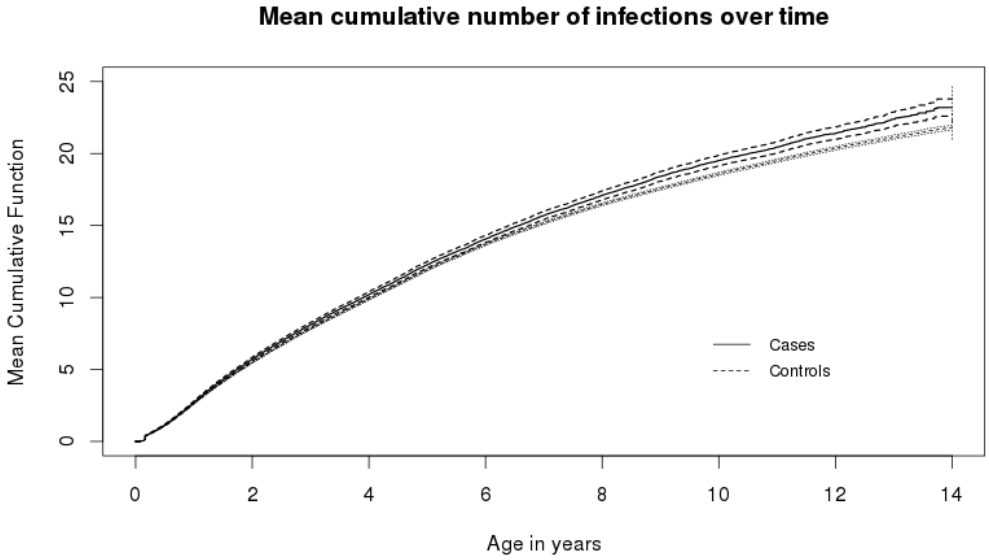


Table 5.1 Definitions of infections the corresponding Ontario Health Insurance Plan (OHIP) physician billing claim diagnosis codes, Canadian Institute for Health Information National Ambulatory Care Reporting System Metadata and the Discharge Abstracts Database

Infections	**OHIP diagnosis codes**	**CIHI DAD or NACRS diagnosis codes (ICD-9)**	**CIHI DAD or NACRS diagnosis codes (ICD-10)**
Any infection (all infection types below)			
Respiratory infections	033, 034, 075, 372, 381, 382, 383, 460, 461, 463, 464, 466, 486, 487	033, 034, 075, 372, 38100, 38101, 38102, 38103, 3811, 3814, 3820, 3824, 3829, 3830, 3839, 460, 461, 462, 463, 464, 465, 466, 480, 481, 482, 483, 484, 485, 486, 487, 488, 490	A37, A38, B27, H10, H65, H66, H70, J00, J01, J02, J03, J04, J05, J06, J09, J10, J11, J12, J13, J14, J15, J18, J20, J21, J22, P23
Gastrointestinal infections	005, 009, 127	005, 009, 127	A04, A05, A08, A09, B80
Otitis externa	380	3801, 3802	H60, H62
Skin and soft tissue infections	052, 054, 078, 112, 117, 373, 521, 682, 684, 685, 686	052, 054, 0743, 0780, 0781, 0784, 112, 117, 373, 5210, 680, 681, 682, 684, 685, 6868, 6869	B00, B01, B02, B07, B37, B49, H000, H001, H010, L01, L02, L03, L05, L089, K029
Urinary tract infections	595, 599	595, 5990	N30, N390
Invasive infections	036, 038, 047, 049, 320, 321, 323	036, 038, 047, 049, 320, 321, 323	A39, A40, A41, A86, A87, G00, G039, G04

OHIP represents the Ontario Health Insurance Plan, CIHI represents the Canadian Institute for Health Information, DAD represents the Discharge Abstract Database, NACRS represents the National Ambulatory Care Reporting System Metadata, ICD-9 and ICD-10 represents the International Statistical Classification of Diseases and Related Health Problems. Any infection is a combination of all the infections from each anatomical region.

Table 5.2 Patient characteristics of the cases of childhood acute lymphoblastic leukemia and the matched cancer-free controls, matched on date of birth, sex, and rural residence among children aged 2-14 years from Ontario, Canada between 1993-2014

Patient characteristic	**Categories**	**ALL cases, n (%)**	**Control, n (%)**	**p-value**
N		N=1,600	N=16,000	
Age at index (years)	Mean (SD)	5.7 ± 3.5	5.7 ± 3.5	0.842
	Median (IQR)	4 (3-8)	4 (3-8)	0.772
Follow-up time (years)	Mean ± SD	4.5 ± 3.0	4.5 ± 3.0	1.000
	Median (IQR)	4.9 (2.2-6.1)	4.9 (2.2-6.1)	
Sex	Female	690 (43.1%)	6,900 (43.1%)	1.000
Index period	1993-1998	354 (22.1%)	3,540 (22.1%)	1.000
	1999-2004	428 (26.8%)	4,280 (26.8%)	
	2005-2010	480 (30.0%)	4,800 (30.0%)	
	2011-2014	338 (21.2%)	3,380 (21.1%)	
Rural residence at start of observation	Yes	192 (12.0%)	1,920 (12.0%)	1.000
Dependency quintile at start of observation	1 - Least deprived	448 (28.0%)	4,290 (26.8%)	0.863
	2	343 (21.4%)	3,509 (21.9%)	
	3	311 (19.4%)	3,091 (19.3%)	
	4	267 (16.7%)	2,745 (17.2%)	
	5 - Most deprived	218 (13.6%)	2,198 (13.7%)	
	Missing	13 (0.8%)	167 (1.0%)	
Material deprivation quintile at start of observation	1 - Least deprived	307 (19.2%)	3,150 (19.7%)	0.546
	2	321 (20.1%)	2,945 (18.4%)	
	3	297 (18.6%)	3,100 (19.4%)	
	4	290 (18.1%)	2,996 (18.7%)	
	5 - Most deprived	372 (23.3%)	3,642 (22.8%)	
	Missing	13 (0.8%)	167 (1.0%)	
Ethnic concentration quintile at start of observation	1 - Least deprived	221 (13.8%)	2,355 (14.7%)	0.245
	2	284 (17.8%)	2,745 (17.2%)	
	3	305 (19.1%)	2,887 (18.0%)	
	4	298 (18.6%)	3,298 (20.6%)	
	5 - Most deprived	479 (29.9%)	4,548 (28.4%)	
	Missing	13 (0.8%)	167 (1.0%)	
Residential instability quintile at start of observation	1 - Least deprived	326 (20.4%)	3,273 (20.5%)	0.882
	2	320 (20.0%)	3,115 (19.5%)	
	3	280 (17.5%)	2,935 (18.3%)	
	4	348 (21.8%)	3,391 (21.2%)	
	5 - Most deprived	313 (19.6%)	3,119 (19.5%)	
	Missing	13 (0.8%)	167 (1.0%)	

Down syndrome at index date	Yes	67 (4.2%)	75 (0.5%)	**<0.001**
Immigrant at index date	Yes	55 (3.4%)	62 (0.4%)	**<0.001**
Length of time since immigration in years (landing to index date)	Mean ± SD	3.8 ± 2.3	7.3 ± 2.7	**<0.001**
	Median (IQR)	3 (2-5)	7 (5-9)	**<0.001**
Rate of any infection				**0.008**
	≤0.25 infection per year	101 (6.3%)	1,286 (8.0%)	
	>0.25 to 0.50 infection per year	89 (5.6%)	826 (5.2%)	
	>0.50 to 1 infection per year	198 (12.4%)	2,056 (12.8%)	
	>1 to 2 infections per year	389 (24.3%)	4,240 (26.5%)	
	>2 infections per year	823 (51.4%)	7,592 (47.5%)	

ALL represents acute lymphoblastic leukemia. SD represents standard deviation. IQR represents interquartile range. Ontario Marginalization Index dimensions: dependency quintile, material deprivation quintile, ethnic concentration quintile, residential instability quintile was taken at start of observation, or if missing at the first available year.

Table 5.3 Association between rate of infections and ALL in children aged 2-14 years from Ontario, Canada between 1993-2014

	Univariate model estimates		**Adjusted model estimates**	
Parameters	**OR**	**95% CI**	**OR**	**95% CI**
Rate of any infection				
≤0.25 infection per year	Ref		Ref	
>0.25 to 0.50 infection per year	**1.41**	**(1.03-1.91)**	**1.39**	**(1.02-1.92)**
>0.50 to 1 infection per year	1.26	(0.97-1.63)	1.29	(0.99-1.68)
>1 to 2 infections per year	1.21	(0.95-1.53)	1.24	(0.97-1.59)
>2 infections per year	**1.44**	**(1.15-1.81)**	**1.43**	**(1.13-1.81)**
Immigrant (Ref: no)	**14.46**	**(9.24-22.64)**	**14.68**	**(9.30-23.16)**
Down syndrome (Ref: no)	**9.15**	**(6.56-12.77)**	**8.85**	**(6.31-12.40)**
Dependency quintile				
1: Least marginalized	Ref		Ref	
2	0.94	(0.81-1.08)	0.94	(0.81-1.10)
3	0.96	(0.82-1.12)	0.96	(0.81-1.13)
4	0.93	(0.79-1.09)	0.96	(0.80-1.14)
5: Most marginalized	0.95	(0.80-1.12)	0.95	(0.77-1.16)
Missing	0.73	(0.41-1.31)	0.73	(0.25-2.10)
Material deprivation quintile				
1: Least marginalized	Ref		Ref	
2	1.12	(0.95-1.32)	1.11	(0.94-1.32)
3	0.98	(0.83-1.16)	0.96	(0.81-1.15)
4	0.99	(0.84-1.17)	0.96	(0.79-1.15)
5: Most marginalized	1.05	(0.89-1.23)	*	
Missing	0.79	(0.44-1.42)	0.96	(0.46-2.04)
Ethnic concentration quintile				
1: Least marginalized	Ref		Ref	
2	1.11	(0.92-1.33)	1.10	(0.91-1.33)
3	1.14	(0.94-1.37)	1.13	(0.92-1.38)
4	0.97	(0.80-1.18)	0.93	(0.75-1.15)
5: Most marginalized	1.14	(0.95-1.36)	*	
Missing	0.82	(0.46-1.47)	1.10	(0.53-2.28)
Residential instability				
1: Least marginalized	Ref		Ref	
2	1.03	(0.88-1.21)	1.04	(0.88-1.23)
3	0.96	(0.81-1.13)	0.99	(0.82-1.18)
4	1.03	(0.88-1.21)	1.02	(0.86-1.23)
5: Most marginalized	1.01	(0.86-1.19)	*	
Missing	0.78	(0.43-1.39)	1.09	(0.52-2.29)

ALL represents acute lymphoblastic leukemia. Cases and controls were matched on date of birth, sex, rural residence at start of observation. Univariate models are univariate conditional logistic regression models. Adjusted models are conditional logistic regression models, and includes confounders immigrant status, down syndrome, and the covariates dependency, material deprivation, ethnic concentration, and residential instability. OR represents odds ratio. CI represents confidence interval. *Parameters have been set to 0 since the variables are a linear combination of other ONMarg dimensions (dependency, material deprivation, ethnic concentration, and residential instability) shown in the model.

Table 5.4 Association between rate of infections and ALL in children aged 2-14 years from Ontario, Canada between 1993-2014, by infection type

		ALL cases		**Controls**		**Univariate model estimates**		**Adjusted model estimates**	
Physician diagnosed infections		**n**	**%**	**n**	**%**	**OR**	**95% CI**	**OR**	**95% CI**
N		1,600		16,000					
Recurrent infections									
Rate of respiratory infections per year									
≤0.25 infection		148	9.3	1,736	10.9	Ref		Ref	
>0.25 to 0.50 infection		99	6.2	1,059	6.6	1.11	(0.85-1.46)	1.10	(0.83-1.46)
>0.50 to 1 infection		262	16.4	2,631	16.4	1.19	(0.96-1.47)	1.19	(0.96-1.49)
>1 to 2 infections		433	27.1	4,485	28.0	1.16	(0.94-1.42)	1.17	(0.95-1.45)
>2 infections		658	41.1	6,089	38.1	**1.31**	**(1.07-1.59)**	**1.28**	**(1.05-1.57)**
Number of children with one infection									
Gastrointestinal	No	890	55.6	9,136	57.1	Ref		Ref	
	Yes	710	44.4	6,864	42.9	1.07	(0.96-1.19)	1.07	(0.96-1.19)
Skin or soft tissue	No	998	62.4	10,342	64.6	Ref		Ref	
	Yes	602	37.6	5,658	35.4	**1.12**	**(1.00-1.25)**	1.11	(0.99-1.25)
Urinary tract	No	1,388	86.8	13,975	87.3	Ref		Ref	
	Yes	212	13.3	2,025	12.7	1.06	(0.90-1.24)	1.02	(0.87-1.20)
Otitis externa	No	1,413	88.3	14,122	88.3	Ref		Ref	
	Yes	187	11.7	1,878	11.7	1.00	(0.85-1.17)	1.00	(0.84-1.18)
Invasive	No	1,533	95.8	15,620	97.6	Ref		Ref	
	Yes	67	4.2	380	2.4	**1.81**	**(1.39-2.37)**	**1.72**	**(1.31-2.26)**
Hospitalization for an infection	No	1,093	68.3	11,403	71.3	Ref		Ref	
	Yes	507	31.7	4,597	28.7	**1.16**	**(1.04-1.31)**	1.11	(0.99-1.25)

ALL represents acute lymphoblastic leukemia. Cases and controls were matched on date of birth, sex, rural residence at start of observation. Univariate models are univariate conditional logistic regression models. Adjusted models are conditional logistic regression models, and includes confounders immigrant status, down syndrome, and covariates dependency, material deprivation, ethnic concentration, and residential instability. OR represents odds ratio. CI represents confidence interval. There were not enough infections in the gastrointestinal, skin or soft tissue, urinary tract, otitis externa anatomical regions, in hospitalizations for an infection, and therefore a binary (yes or no) outcome was used.

Table 5.5 Association between rate of infections and ALL in children aged 2-14 years from Ontario, Canada between 1993-2014, restricted to matched sets of cases and controls with the observation period starting from birth

	Univariate model estimates		**Adjusted model estimates**	
Parameters	**OR**	**95% CI**	**OR**	**95% CI**
N=13,948				
Rate of any infection				
≤0.25 infection per year	Ref		Ref	
>0.25 to 0.50 infection per year	**1.45**	**(1.00-2.25)**	1.47	(0.97-2.21)
>0.50 to 1 infection per year	**1.56**	**(1.11-2.19)**	**1.53**	**(1.09-2.14)**
>1 to 2 infections per year	**1.46**	**(1.06-2.00)**	**1.44**	**(1.05-1.98)**
>2 infections per year	**1.74**	**(1.28-2.37)**	**1.67**	**(1.23-2.28)**
Down syndrome (Ref: no)	**8.12**	**(5.64-11.69)**	**7.85**	**(5.44-11.33)**
Dependency				
1: Least marginalized	Ref		Ref	
2	0.95	(0.81-1.12)	0.93	(0.78-1.10)
3	1.02	(0.86-1.21)	1.00	(0.83-1.20)
4	0.95	(0.79-1.14)	0.96	(0.79-1.18)
5: Most marginalized	0.99	(0.81-1.21)	0.98	(0.78-1.22)
Missing	0.66	(0.33-1.32)	0.64	(0.19-2.11)
Material deprivation				
1: Least marginalized	Ref		Ref	
2	1.19	(0.99-1.43)	1.19	(0.99-1.44)
3	0.99	(0.82-1.20)	1.00	(0.82-1.22)
4	1.00	(0.83-1.21)	0.99	(0.80-1.21)
5: Most marginalized	1.04	(0.87-1.24)	*	
Missing	0.70	(0.35-1.41)	1.02	(0.45-2.34)
Ethnic concentration				
1: Least marginalized	Ref		Ref	
2	1.09	(0.88-1.34)	1.07	(0.86-1.33)
3	1.15	(0.93-1.41)	1.16	(0.93-1.45)
4	0.96	(0.78-1.19)	0.96	(0.76-1.22)
5: Most marginalized	1.06	(0.87-1.29)	*	
Missing	0.71	(0.35-1.42)	1.13	(0.50-2.53)
Residential instability				
1: Least marginalized	Ref		Ref	
2	0.97	(0.81-1.16)	0.97	(0.81-1.18)
3	0.96	(0.81-1.16)	1.00	(0.81-1.22)
4	0.99	(0.83-1.19)	1.03	(0.84-1.22)
5: Most marginalized	0.97	(0.81-1.16)	*	
Missing	0.66	(0.33-1.32)	1.06	(0.46-2.40)

ALL represents acute lymphoblastic leukemia. There were 1,268 cases and 12,680 controls matched on date of birth, sex, rural residence at start of observation. Univariate models are univariate conditional logistic regression models. Adjusted models are conditional logistic regression models, and includes confounder down syndrome, and the covariates dependency, material deprivation, ethnic concentration, and residential instability. OR represents odds ratio. CI represents confidence interval. *Parameters have been set to 0 since the variables are a linear combination of other ONMarg dimensions (dependency, material deprivation, ethnic concentration, and residential instability) shown in the model.

Chapter 6 : Discussion

6.1 Summary of Key Findings

The results from this dissertation add to the existing evidence on the association between infections and childhood ALL, addresses knowledge gaps, and identifies future research directions. In this dissertation, infections were found to be associated with childhood ALL and may play a role in the etiology of the disease. The dissertation raises additional questions on the critical exposure period for infections and the type of infection that increases the odds of ALL that will need to be answered in future research.

6.1.1 Chapter 3: A Systematic Review and Meta-analysis of the Association Between Childhood Infections and the Risk of Childhood Acute Lymphoblastic Leukemia

In Chapter 3, our systematic review and meta-analysis of 39 studies found no association between number, frequency, severity, and timing of prior infections to the development of childhood ALL. A qualitative difference in our subgroup analyses showed differences in the relationship between prior infections and the development of childhood ALL based on the type of data used to ascertain infections. The interpretation of the subgroup findings must be made with caution because of the nature of subgroup analyses. In this specific instance, since the overall effect was nonsignificant, the chance of one subgroup-specific test being significant is at least 7%.[210] Infections increased the odds of developing ALL by 2.4-fold in studies with laboratory investigations and this was significantly different compared to studies using self-reported and administrative/medical records data to capture infections prior to childhood ALL. The study highlighted the challenges in measuring infections, gaps in the literature, and insights into the expected findings given the type of data used to measure infections.

6.1.2 Chapter 4: Manuscript titled Use of physician billing claims to identify infections in children: a population-based validation study of administrative data from Ontario, Canada

In Chapter 4, we found the billing codes to be generally valid to identify infections in children aged 0 to 18 years when compared to an EMR reference standard. Administrative data performed well in capturing any infection and respiratory infections, while skin and soft tissue,

gastrointestinal, urinary tract, and other ear infections maintained high specificity, but had lower sensitivity. The results suggest administrative data can accurately capture infections with minimal risk of including false positives and is a viable method to identify infectious syndromic conditions for the use of syndrome-based disease estimates.

6.1.3 Chapter 5: Manuscript titled Rate of infections and the association with childhood acute lymphoblastic leukemia: a population-based case-control study

In Chapter 5, we used administrative data from Ontario, Canada to assess the relationship between the rate of infections and the odds of childhood ALL. Having >2 infections per year increased the odds of ALL by 43% compared to children with ≤0.25 infections per year, and over time the rate of infections in the cases of ALL was higher than controls. In the critical exposure period of 1 to 1.5 years of age, having an infection increased the odds of childhood ALL by 20%. The association between the rate of infections and ALL was even stronger among a cohort of matched cases and controls with observation periods starting from birth. Certain types of infections are more likely to be associated with the risk of ALL than others, that is, respiratory and invasive infections increased the odds of ALL. This study suggests children who develop ALL have more infections than controls with no cancer.

6.2 Methodological Considerations

Methodological strengths and limitations pertaining to each of the individual objectives are discussed in the respective chapters. Here, I discuss the methodological considerations that span multiple components of the dissertation, including measuring infections using administrative data.

6.2.1 Measuring Infections Using Administrative Data

Routinely collected electronic administrative data offer the advantage of identifying many infectious diseases in large populations at low cost. However, applying the identification criteria for diseases to an entire population requires considerations. The approach taken to measure infections using administrative data has several implications. In general, administrative data definitions are restricted to patients who interact with the healthcare system for a disease. In a study from the Netherlands, the authors compared the association between symptoms such as colds/flu, respiratory tract problems, and fever to general practitioner consultations.[212] Ear

problems, fever, and respiratory tract problems often triggered a visit to a physician, but despite the frequency of colds/flus, physicians were consulted 10% of the time. The consultation rates were higher for younger children and for boys.[212] Others have reported up to 20% of illnesses experienced by children at home are brought to a physician office.[213] However, studies have shown that overall health and certain conditions such as mental illness have been associated with more health services use.[208,214] Predictors of high health services use include the child's health needs such as the number of acute or recurring illnesses or whether the child was on medications, and maternal patterns of health care use such as the amount of health services the mother used in the previous years.[208] A child's age and consultation with other health care professionals were also associated with health services for a child.[214] Even under a universal healthcare system such as Ontario, there were differences in the use of health services for children that depend on the number of local physicians in the area and socioeconomic status.[209] In the context of the studies in this dissertation, I may be underestimating the number of infections among our study populations. From this perspective, it is possible that ascertainment of disease may be linked to disease severity, with less severe diseases being poorly ascertained in administrative databases. Alternatively, there may be an unmeasured factor that is associated with ALL that leads the parents to take their children to see the physician.

Second, I was only able to validate infections in children in the primary care setting. However, a study conducted in Ottawa, Canada assessed the criterion validity of administrative data for identifying hospitalizations for respiratory syncytial virus infection among children in Ontario.[172] The chart review data was linked to Ontario's administrative data and used to evaluate the diagnostic accuracy of algorithms of RSV-related ICD-10 codes within provincial hospitalization and emergency department databases. The best algorithm, based on hospitalization data, resulted in sensitivity of 97.9% (95%CI:95.5– 99.2%), specificity of 99.6% (95%CI:98.2–99.8%), PPV of 96.9% (95%CI:94.2–98.6%) and NPV of 99.4% (95%CI:99.4–99.9%). This suggests hospital discharge data from Ontario may be able to accurately capture certain infectious diseases. The accuracy with which emergency department visits accurately capture infections in Ontario is less certain, and evidence suggests administrative data have different levels of accuracy and may require further assessment.[60] However, in a study from Boston, United States, the authors found routinely collected administrative data for syndromic definitions for respiratory infections strongly correlated with virologic test results that suggested accurate detection of disease.[179] In

another study that assessed the criterion validity of International Classification of Disease diagnostic codes for identifying respiratory infections in emergency room visits, the authors found similar patterns to our study with specificity >0.97, and sensitivity ranging from 0.56 to 0.87.[176] Unlike the adult population, very few validation studies have been conducted on a pediatric population.[157] For the purposes of the dissertation, I have shown that administrative data from primary care visits were able to reasonably identify patients with infections and rule in patients with infections. Further, the administrative data maintained its performance across different ages and patients with different diseases.

Third, I was unable to identify and explore infectious pathogens in the studies that used administrative data. There are two broad categories of approaches for testing of infectious pathogens that are relevant to consider, those that were tested for and the results contained elsewhere, second, those that were not tested. For instances that infectious agents were tested, microbiology data to identify pathogens for the healthcare encounters were not available. Syndromic approaches to identifying infectious diseases are commonly used within Canada and internationally.[215,216] There are opportunities to overcome this limitation for future research. Public Health Ontario collects microbiology data on reportable communicable diseases such as influenza, and provides an opportunity to identify certain pathogens and link them to administrative databases for patient-level analyses.[217] The limitation to this approach is that Public Health Ontario focuses data collection on reportable diseases and some non-reportable respiratory viruses not on the reportable disease list. In the second approach to testing of infectious pathogens during physician visits, most patients with infections are unlikely to be lab tested because it generally does not change management of the patient and is therefore unlikely to be a data source for more common infections occurring in children.[218]

Temporality and reverse causality concerns were present in many of the included studies in the systematic review and meta-analysis chapter. Most studies did not account for the potential for reverse causality, such that ALL may cause a child to have more infections prior to the diagnosis of ALL. One way of assessing this problem is to create models with various lag-times between the development of ALL and infections. The empirical study in Chapter 5 utilized a 1-year lag-time to address reverse causality, and any infection that occurred within 1-year of the diagnosis date was not included. Extending the lag-time beyond 1-year is possible but may be

suboptimal and unnecessary according to evidence from the literature. A study from the United Kingdom suggested ALL may impact the susceptibility to infections at 5 months prior to the diagnosis of ALL.[43] Further, a validation study from Ontario demonstrated that the diagnostic interval from the initial physician visit to the diagnosis of childhood ALL were short, the intervals median was 2 days (interquartile range 1 to 3).[219] These studies provide evidence to suggest the 1-year lag period used in Chapter 5 was an appropriate and optimal to account for reverse causality.

Administrative data often does not include other important potential confounders, but this is not specific to measuring infections. A brief discussion on the confounders of the infection and childhood ALL relationship has already been discussed in the Introduction. In this scenario, the calculation of the E-value may be helpful in assessing the minimum strength of association that an unmeasured confounder would need to have with both the exposure and the outcome to fully explain away a specific exposure-outcome association – conditional on the measured covariates.[220] The observed odds ratio of 1.43 could be explained away by an unmeasured confounder that was associated with both infections and ALL by an odds ratio of 2.21-fold each, above and beyond the measured confounders, but weaker confounding could not do so. Further, the unmeasured confounder requires the lower 95% CI to be greater than an odds ratio of 1.51.

Finally, the use of the administrative data in Ontario, Canada may not be generalizable to other jurisdictions. However, Ontario is Canada's largest province with over 13.6 million residents as of 2014 and more than half of the visible minorities in Canada reside in Ontario.[221,222] Without explicit testing, we are unsure how the criterion validity for infections hold for other populations with different characteristics, but this limitation does not affect the internal validity of the studies.

6.3 Future Work

Future work should expand on the results from this dissertation to investigate other potential exposures around the ages 1 to 1.5 years, for example, to obtain data on day-care attendance. A meta-analysis has demonstrated that day-care attendance reduced the risk of childhood ALL,[51] however no study has considered the interaction between day-care attendance and physician diagnosed infections on the development of ALL.

Healthcare administrative data can be a rich source for population-based research that can be used to efficiently study rare diseases and using administrative data would be advantageous for studying vulnerable populations. The date and reason for the visit are often captured in the administrative databases, allowing for assessment of the time and type of infection.[70] Particularly useful for etiology studies is the ability for administrative data to be used to create cohorts of individuals that can be followed longitudinally for potential outcomes and covariates. Most administrative data, including the data used in this dissertation, are often missing information on other confounders in the relationship of interest, such as parental smoking status, ethnicity and race, parental occupation, and other environmental data such as pesticide.[141] Conducting an observational study that uses additional measures to obtain the confounder information, such as surveys could address this data limitation.

Future work may include assessing the association between childhood ALL and reportable and non-reportable respiratory infectious diseases using individual level data from Public Health Ontario. Since respiratory infections were found to be associated with childhood ALL, the Public Health Ontario data could be used to investigate whether certain reportable respiratory diseases such as influenza may be associated with childhood ALL. Ecological studies have suggested an association with seasonal variation in birth month and childhood ALL.[223-225] A recent ecological study demonstrated the exposure to- and timing of the influenza and respiratory syncytial virus seasons are associated with the development of childhood ALL.[226]

The use of a negative control (exposure and outcome) approach could be useful in a future study to detect both suspected and unsuspected sources of spurious associations such as potential confounders. The purpose of a negative control is to reproduce a situation that cannot involve the hypothesized causal mechanism, but is likely to involve the same sources of bias that may be present in the original association.[227] For example, a negative exposure analysis would provide evidence that the relationship between infections and ALL is real, and not driven by some other factor that may lead to cases being taken to the physician for infections more often than controls. If an exposure not known or thought to be associated with childhood ALL is assessed, it would be expected that there would be no association. Similarly, in a negative outcome analysis, it should also be expected that infections would not be associated with the negative outcome.

Costs and feasibility are the usual barriers to creating large pregnancy and birth cohorts.[145] An innovative approach to overcome the costs and feasibility in studying rare diseases is to combine already established cohorts.[81,146] Another way to circumvent the costly creation of birth cohorts is to use administrative data to follow infants to disease outcomes. Coordinated efforts from different governments, organizations, institutes, universities, and others should proceed with enriching the administrative data with genetic, clinical, social, political, and behavioural data. The data could be used to answer research questions inside and outside of health. With the Government of Canada's recent emphasis and funding in *Harnessing Big Data* projects in health research through building digital infrastructure that is more open and creating equitable access across Canada, researchers in Canada now have a window of opportunity to answer previously unfeasible research questions and to be world leaders in big data research.[228] In the context of this dissertation, it would address the limitation of administrative data in terms of the lack of data availability for other confounders, such as genetics, ethnicity, day-care attendance and environmental exposures. Combining the potential of big data with the developments in causal inference methodology would allow researchers to assess the relative magnitude of different pathways and mechanisms by which an exposure may affect an outcome.[229]

6.4 Conclusions

Prior infections have been shown to be associated with the development of childhood ALL in Ontario. Through three distinct research aims, the overall goal of this dissertation was to assess the relationship between prior infections in the development of childhood ALL. The key results from the dissertation are that the association between prior infections and childhood ALL in the previous studies may depend on the way infections were ascertained. The use of administrative data could overcome the limitations identified in the systematic review and meta-analysis. I found administrative data could reasonably identify infections in a pediatric population in Ontario. Finally, a higher rate of infections, respiratory and invasive infections, and having an infection between the ages of 1 to 1.5 years were associated with the development of childhood ALL. Together, the results from the dissertation demonstrated infections *have a role* in the etiology of childhood ALL. Future research should attempt to address knowledge gaps identified and take

advantage of the developments and opportunities in *big data* to better understand the mechanisms and pathways in the etiology of childhood ALL.

References

1. Statistics Canada. Table 5.5 Leading causes of death of children and youth, by age group, 2006 to 2008. 2016; http://www.statcan.gc.ca/pub/11-402-x/2012000/chap/c-e/tbl/tbl05-eng.htm. Accessed February 1, 2018, 2018.

2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians.* 2018;68(1):7-30.

3. *Canadian Cancer Society's Advisory Committee on Cancer Statistics. Canadian Cancer Statistics 2017.* Toronto, ON2017.

4. Chan V, Nathan PC, Taccone MS, et al. Provincial health system planning for the increasing prevalence of childhood cancer survivors in Ontario: A population-based analysis in Ontario. . Paper presented at: POGO Symposium on Childhood Cancer2017; Toronto, Canada.

5. Canadian Cancer Society. Canadian Cancer Statistics 2008: Special Topic Childhood Cancer. Canadian Cancer Society National Cancer Institute of Canada. Toronto2008.

6. Parkin DM, Stiller CA, Draper GJ, Bieber CA. The international incidence of childhood cancer. *Int J Cancer.* 1988;42(4):511-520.

7. Institute NC. *SEER Cancer Statistics Review, 1975-2012.* Bethesda, MDNovember 2014.

8. Greenberg M, Barnett H, Williams J. *Atlas of Childhood Cancer in Ontario, 1985-2004.* Toronto: Pediatric Oncology Group of Ontario;2015.

9. Phillips SM, Padgett LS, Leisenring WM, et al. Survivors of Childhood Cancer in the United States: Prevalence and Burden of Morbidity. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology.* 2015;24(4):653-663.

10. Essig S, Li Q, Chen Y, et al. Risk of late effects of treatment in children newly diagnosed with standard-risk acute lymphoblastic leukaemia: a report from the Childhood Cancer Survivor Study cohort. *The Lancet Oncology.* 2014;15(8):841-851.

11. Greenberg MLN, P. C.; Agha, M.; Hodgson, D.; Pole, J.D.; Greenberg, C. *Chapter 5: Health Service Utilization. In: Atlas of Childhood Cancer in Ontario, 1985-2004.* TorontoJanurary 2015.

12. Greaves M. Infection, immune responses and the aetiology of childhood leukaemia. *Nat Rev Cancer.* 2006;6(3):193-203.

13. Wiemels J. Perspectives on the causes of childhood leukemia. *Chem Biol Interact.* 2012;196(3):59-67.

14. Kadan-Lottick NS, Kawashima T, Tomlinson G, et al. The risk of cancer in twins: a report from the childhood cancer survivor study. *Pediatric blood & cancer.* 2006;46(4):476-481.

15. Greaves MF, Maia AT, Wiemels JL, Ford AM. Leukemia in twins: lessons in natural history. *Blood.* 2003;102(7):2321-2333.

16. Kinlen L. Evidence for an infective cause of childhood leukaemia: comparison of a Scottish new town with nuclear reprocessing sites in Britain *The Lancet.* 1988;332(8624):1323-1327.

17. Kinlen LJ. An examination, with a meta-analysis, of studies of childhood leukaemia in relation to population mixing. *British Journal of Cancer.* 2012;107(7):1163-1168.

18. Alexander FE, Boyle P, Carli PM, et al. Spatial clustering of childhood leukaemia: summary results from the EUROCLUS project. *British Journal of Cancer.* 1998;77(5):818-824.

19. Kinlen L. Childhood leukaemia and ordnance factories in west Cumbria during the Second World War. *British Journal of Cancer.* 2006;95(1):102-106.

20. Kinlen L, Doll R. Population mixing and childhood leukaemia: Fallon and other US clusters. *British Journal of Cancer.* 2004;91(1):1-3.

21. Greaves MF. Aetiology of acute leukaemia. *Lancet.* 1997;349(9048):344-349.

22. Greaves MF. Speculations on the cause of childhood acute lymphoblastic leukemia. *Leukemia.* 1988;2(2):120-125.

23. Mori H, Colman SM, Xiao Z, et al. Chromosome translocations and covert leukemic clones are generated during normal fetal development. *Proceedings of the National Academy of Sciences.* 2002;99(12):8242-8247.

24. Zuna J, Madzo J, Krejci O, et al. ETV6/RUNX1 (TEL/AML1) is a frequent prenatal first hit in childhood leukemia. *Blood.* 2011;117(1):368-369; author reply 370-361.

25. Ajrouche R, Rudant J, Orsi L, et al. Childhood acute lymphoblastic leukaemia and indicators of early immune stimulation: The Estelle study (SFCE). *British Journal of Cancer.* 2015;112:1017-1026.

26. Jourdan-Da Silva N, Perel Y, Mechinaud F, et al. Infectious diseases in the first year of life, perinatal characteristics and childhood acute leukaemia. *Br J Cancer.* 2004;90(1):139-145.

27. McKinney PA, Juszczak E, Findlay E, Smith K, Thomson CS. Pre- and perinatal risk factors for childhood leukaemia and other malignancies: A Scottish case control study. *British Journal of Cancer.* 1999;80(11):1844-1851.

28. van Steensel-moll HA, Valkenburg HA, van Zanen GE. Childhood leukemia and infectious diseases in the first year of life: a register-based case-control study. *American journal of epidemiology.* 1986;124(4):590-594.

29. Canfield KN, Spector LG, Robison LL, et al. Childhood and maternal infections and risk of acute leukaemia in children with Down syndrome: A report from the Children's Oncology Group. *British journal of cancer.* 2004;91(11):1866-1872.

30. Chang JS, Tsai CR, Tsai YW, Wiemels JL. Medically diagnosed infections and risk of childhood leukaemia: A population-based case-control study. *International journal of epidemiology.* 2012;41(4):1050-1059.

31. Roman E, Simpson J, Ansell P, et al. Childhood Acute Lymphoblastic Leukemia and Infections in the First Year of Life: A Report from the United Kingdom Childhood Cancer Study. *American journal of epidemiology.* 2007;165(5):496-504.

32. Rudant J, Lightfoot T, Urayama KY, et al. Childhood acute lymphoblastic leukemia and indicators of early immune stimulation: a Childhood Leukemia International Consortium study. *American journal of epidemiology.* 2015;181(8):549-562.

33. Cardwell CR, McKinney PA, Patterson CC, Murray LJ. Infections in early life and childhood leukaemia risk: A UK case-control study of general practitioner records. *British Journal of Cancer.* 2008;99(9):1529-1533.

34. Dockerty JD, Skegg DC, Elwood JM, Herbison GP, Becroft DM, Lewis ME. Infections, vaccinations, and the risk of childhood leukaemia. *British Journal of Cancer.* 1999;80(9):1483-1489.

35. Ma X, Buffler PA, Wiemels JL, et al. Ethnic difference in daycare attendance, early infections, and risk of childhood acute lymphoblastic leukemia. *Cancer Epidemiology Biomarkers and Prevention.* 2005;14(8):1928-1934.

36. Vestergaard TR, Rostgaard K, Grau K, Schmiegelow K, Hjalgrim H. Hospitalisation for infection prior to diagnosis of acute lymphoblastic leukaemia in children. *Pediatric Blood and Cancer.* 2013;60(3):428-432.

37. Flores-Lujano J, Perez-Saldivar ML, Fuentes-Panana EM, et al. Breastfeeding and early infection in the aetiology of childhood leukaemia in down syndrome. *British Journal of Cancer.* 2009;101(5):860-864.

38. Chan LC, Lam TH, Lau YL, et al. Is the timing of exposure to infection a major determinant of acute lymphoblastic leukaemia in Hong Kong? *Paediatric and Perinatal Epidemiology.* 2002;16(2):154-165.

39. Schuz J, Kaletsch U, Meinert R, Kaatsch P, Michaelis J. Association of childhood leukaemia with factors related to the immune system. *British Journal of Cancer.* 1999;80(3-4):585-590.

40. Urayama KY, Ma X, Selvin S, et al. Early life exposure to infections and risk of childhood acute lymphoblastic leukemia. *International journal of cancer Journal international du cancer.* 2011;128(7):1632-1643.

41. Perrillat F, Clavel J, Auclerc MF, et al. Day-care, early common infections and childhood acute leukaemia: a multicentre French case-control study. *Br J Cancer.* 2002;86(7):1064-1069.

42. Rudant J, Orsi L, Menegaux F, et al. Childhood acute leukemia, early common infections, and allergy: The ESCALE Study. *American journal of epidemiology.* 2010;172(9):1015-1027.

43. Crouch S, Lightfoot T, Simpson J, Smith A, Ansell P, Roman E. Infectious Illness in Children Subsequently Diagnosed With Acute Lymphoblastic Leukemia: Modeling the Trends From Birth to Diagnosis. *American journal of epidemiology.* 2012;176(5):402-408.

44. Simpson J, Smith A, Ansell P, Roman E. Childhood leukaemia and infectious exposure: a report from the United Kingdom Childhood Cancer Study (UKCCS). *Eur J Cancer.* 2007;43(16):2396-2403.

45. D'Souza-Vazirani D, Minkovitz CS, Strobino DM. Validity of maternal report of acute health care use for children younger than 3 years. *Arch Pediatr Adolesc Med.* 2005;159(2):167-172.

46. McKinney PA, Alexander FE, Nicholson C, Cartwright RA, Carrette J. Mothers' reports of childhood vaccinations and infections and their concordance with general practitioner records. *Journal of public health medicine.* 1991;13(1):13-22.

47. Alho OP. The validity of questionnaire reports of a history of acute otitis media. *American journal of epidemiology.* 1990;132(6):1164-1170.

48. Kvestad E, Kværner KJ, Røysamb E, Tambs K, Harris JR, Magnus P. The reliability of self-reported childhood otitis media by adults. *International Journal of Pediatric Otorhinolaryngology.* 2006;70(4):597-602.

49. Miller JE, Gaboda D, Davis D. Early childhood chronic illness: comparability of maternal reports and medical records. *Vital and health statistics Series 2, Data evaluation and methods research.* 2001(131):1-10.

50. Perrin EC, Newacheck P, Pless IB, et al. Issues involved in the definition and classification of chronic health conditions. *Pediatrics.* 1993;91(4):787-793.

51. Urayama KY, Buffler PA, Gallagher ER, Ayoob JM, Ma X. A meta-analysis of the association between day-care attendance and childhood acute lymphoblastic leukaemia. *International journal of epidemiology.* 2010;39(3):718-732.

52. Schuz J, Luta G, Erdmann F, et al. Birth order and risk of childhood cancer in the Danish birth cohort of 1973-2010. *Cancer causes & control : CCC.* 2015;26(11):1575-1582.

53. Lin JN, Lin CL, Lin MC, et al. Risk of leukaemia in children infected with enterovirus: A nationwide, retrospective, population-based, Taiwanese-registry, cohort study. *The Lancet Oncology.* 2015;16(13):1335-1343.

54. Dockerty JD, Draper G, Vincent T, Rowan SD, Bunch KJ. Case-control study of parental age, parity and socioeconomic level in relation to childhood cancers. *International journal of epidemiology.* 2001;30(6):1428-1437.

55. Hjalgrim LL, Rostgaard K, Hjalgrim H, et al. Birth Weight and Risk for Childhood Leukemia in Denmark, Sweden, Norway, and Iceland. *JNCI: Journal of the National Cancer Institute.* 2004;96(20):1549-1556.

56. Boothe VL, Boehmer TK, Wendel AM, Yip FY. Residential traffic exposure and childhood leukemia: a systematic review and meta-analysis. *American journal of preventive medicine.* 2014;46(4):413-422.

57. Filippini T, Heck JE, Malagoli C, Del Giovane C, Vinceti M. A review and meta-analysis of outdoor air pollution and risk of childhood leukemia. *Journal of environmental science and health Part C, Environmental carcinogenesis & ecotoxicology reviews.* 2015;33(1):36-66.

58. Liu R, Zhang L, McHale CM, Hammond SK. Paternal smoking and risk of childhood acute lymphoblastic leukemia: systematic review and meta-analysis. *Journal of oncology.* 2011;2011:854584.

59. Brauer M, Hoek G, Van Vliet P, et al. Air pollution from traffic and the development of respiratory infections and asthmatic and allergic symptoms in children. *Am J Respir Crit Care Med.* 2002;166(8):1092-1098.

60. Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, Guttmann A. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *Journal of clinical epidemiology.* 2011;64(8):821-829.

61. Gilham C, Peto J, Simpson J, et al. Day care in infancy and risk of childhood acute lymphoblastic leukaemia: findings from UK case-control study. *BMJ.* 2005;330(7503):1294.

62. Wiemels J. Perspectives on the causes of childhood leukemia. *Chem Biol Interact.* 2012;196(3):59-67.

63. Chang JS, Zhou M, Buffler PA, Chokkalingam AP, Metayer C, Wiemels JL. Profound deficit of IL10 at birth in children who develop childhood acute lymphoblastic leukemia. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology.* 2011;20(8):1736-1740.

64. Couper KN, Blount DG, Riley EM. IL-10: the master regulator of immunity to infection. *Journal of immunology (Baltimore, Md : 1950).* 2008;180(9):5771-5777.

65. Carey AJ, Tan CK, Ulett GC. Infection-induced IL-10 and JAK-STAT: A review of the molecular circuitry controlling immune hyperactivity in response to pathogenic microbes. *Jak-stat.* 2012;1(3):159-167.

66. JA W. Chapter 10 Principles of Diagnosis. In: S B, ed. *Medical Microbiology*. Galveston (TX): University of Texas Medical Branch at Galveston; 1996.

67. Steingart KR, Flores LL, Dendukuri N, et al. Commercial serological tests for the diagnosis of active pulmonary and extrapulmonary tuberculosis: an updated systematic review and meta-analysis. *PLoS Med.* 2011;8(8):e1001062.

68. Dowdy DW, Steingart KR, Pai M. Serological Testing Versus Other Strategies for Diagnosis of Active Tuberculosis in India: A Cost-Effectiveness Analysis. *PLOS Medicine.* 2011;8(8):e1001074.

69. Tugwell P, Dennis DT, Weinstein A, et al. LAboratory evaluation in the diagnosis of lyme disease. *Annals of internal medicine.* 1997;127(12):1109-1123.

70. Kwong JCC, N. S.; Campitelli, M. A.; Ratnasingham, S.; Daneman, N.; Deeks, S. L.; Manuel, D. G. *Ontario Burden of Infectious Disease Study (ONBOIDS): An OAHPP/ICES Report.* Toronto, Ontario: Ontario Burden of Infectious Disease Study Advisory Group and Institute for Clinical Evaluative Sciences;2010.

71. Hijiya N, Ness KK, Ribeiro RC, Hudson MM. Acute Leukemia as a Secondary Malignancy in Children and Adolescents: Current Findings and Issues. *Cancer.* 2009;115(1):23-35.

72. Roizen NJ, Patterson D. Down's syndrome. *The Lancet.*361(9365):1281-1289.

73. Zipursky A, Poon A, Doyle J. Leukemia in Down Syndrome: A Review. *Pediatric Hematology and Oncology.* 1992;9(2):139-149.

74. Ward E, DeSantis C, Robbins A, Kohler B, Jemal A. Childhood and adolescent cancer statistics, 2014. *CA: A Cancer Journal for Clinicians.* 2014;64(2):83-103.

75. Muenchhoff M, Goulder PJR. Sex Differences in Pediatric Infectious Diseases. *The Journal of Infectious Diseases.* 2014;209(Suppl 3):S120-S126.

76. Ray GT, Suaya JA, Baxter R. Incidence, microbiology, and patient characteristics of skin and soft-tissue infections in a U.S. population: a retrospective population-based study. *BMC Infectious Diseases.* 2013;13(1):252.

77. Bailey HD, Infante-Rivard C, Metayer C, et al. Home pesticide exposures and risk of childhood leukemia: Findings from the childhood leukemia international consortium. *International journal of cancer Journal international du cancer.* 2015;137(11):2644-2663.

78. Bailey HD, Fritschi L, Infante-Rivard C, et al. Parental occupational pesticide exposure and the risk of childhood leukemia in the offspring: findings from the childhood leukemia

international consortium. *International journal of cancer Journal international du cancer.* 2014;135(9):2157-2172.

79. Cook DG, Strachan DP. Summary of effects of parental smoking on the respiratory health of children and implications for research. *Thorax.* 1999;54(4):357-366.

80. Carlos-Wallace FM, Zhang L, Smith MT, Rader G, Steinmaus C. Parental, In Utero, and Early-Life Exposure to Benzene and the Risk of Childhood Leukemia: A Meta-Analysis. *American journal of epidemiology.* 2016;183(1):1-14.

81. Metayer C, Milne E, Clavel J, et al. The Childhood Leukemia International Consortium. *Cancer Epidemiology.* 2013;37(3):336-347.

82. Institute for Clinical Evaluative Sciences (ICES). Privacy at ICES 2017; https://www.ices.on.ca/Data-and-Privacy/Privacy-at-ICES. Accessed November 12, 2017, 2017.

83. Matheson FI, Dunn JR, Smith KL, Moineddin R, Glazier RH. Development of the Canadian Marginalization Index: a new tool for the study of inequality. *Canadian journal of public health = Revue canadienne de sante publique.* 2012;103(8 Suppl 2):S12-16.

84. Mustard CA, Derksen S, Berthelot JM, Wolfson M. Assessing ecologic proxies for household income: a comparison of household and neighbourhood level income measures in the study of population health status. *Health & place.* 1999;5(2):157-171.

85. Chan B, Anderson GM, Theriault ME. High-billing general practitioners and family physicians in Ontario: how do they do it? An analysis of practice patterns of GP/FPs with annual billings over $400,000. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne.* 1998;158(6):741-746.

86. Jaakkimainen L, Upshur, R.E.G., Klein-Geltink, J.E., Maaten, S., Schultz, S.E., Leong, A., Wang, L. *Primary Care in Ontario: ICES Atlas.* Toronto, Canada2006.

87. JPPC. *Funding hospital based ambulatory care. Final report.* Ontario Joint Policy and Planning Committee;2002.

88. CIHI. *Canadian Institute for Health Information. Data Quality Documentation, National Ambulatory Care Reporting System - Current-Year Information, 2016-2017.* Ottawa2017.

89. CIHI. *Canadian Institute for Health Information. Data Quality Documentation, Discharge Abstract Database, Current-Year Information 2016-2017.* Ottawa2017.

90. Tu K, Widdifield J, Young J, et al. Are family physicians comprehensively using electronic medical records such that the data can be used for secondary purposes? A Canadian perspective. *BMC medical informatics and decision making.* 2015;15:67.

91. Birken CS, Tu K, Oud W, et al. Determining rates of overweight and obese status in children using electronic medical records. *Cross-sectional study.* 2017;63(2):e114-e122.

92. Greenberg ML, Barr RD, DiMonte B, McLaughlin E, Greenberg C. Childhood cancer registries in Ontario, Canada: lessons learned from a comparison of two registries. *International journal of cancer Journal international du cancer.* 2003;105(1):88-91.

93. Kramarova E, Stiller CA. The international classification of childhood cancer. *International journal of cancer Journal international du cancer.* 1996;68(6):759-765.

94. DiMonte B, Pole, J.D., Agha, M., Greenberg, M.L.,. *Chapter 2: POGONIS: Methods and Data Sources.* Toronto, Canada2015.

95. Ontario POGo. Data Requests. 2018; http://www.pogo.ca/research-data/pogonis-childhood-cancer-database/data-requests/. Accessed April 24, 2018, 2018.

96. Rezai MR, Maclagan LC, Donovan LR, Tu JV. Classification of Canadian immigrants into visible minority groups using country of birth and mother tongue. *Open Medicine.* 2013;7(4):e85-e93.

97. da Conceicao Nunes J, de Araujo GV, Viana MT, Sarinho ES. Association of atopic diseases and parvovirus B19 with acute lymphoblastic leukemia in childhood and adolescence in the northeast of Brazil. *Int J Clin Oncol.* 2016.

98. Martin-Lorenzo A, Hauer J, Vicente-Duenas C, et al. Infection exposure is a causal factor in B-cell precursor acute lymphoblastic leukemia as a result of Pax5-inherited susceptibility. *Cancer Discovery.* 2015;5(12):1328-1343.

99. Simpson J, Smith A, Ansell P, Roman E. Childhood leukaemia and infectious exposure: A report from the United Kingdom Childhood Cancer Study (UKCCS). *European Journal of Cancer.* 2007;43(16):2396-2403.

100. McNally RJ, Eden TO. An infectious aetiology for childhood acute leukaemia: a review of the evidence. *Br J Haematol.* 2004;127(3):243-263.

101. Ma XM, Urayama K, Chang J, Wiemels JL, Buffler PA. Infection and pediatric acute lymphoblastic leukemia. *Blood Cells Mol Dis.* 2009;42(2):117-120.

102. Maia Rda R, Wunsch Filho V. Infection and childhood leukemia: review of evidence. *Rev Saude Publica.* 2013;47(6):1172-1185.

103. Buffler PA, Kwan ML, Reynolds P, Urayama KY. Environmental and genetic risk factors for childhood leukemia: Appraising the evidence. *Cancer Invest.* 2005;23(1):60-75.

104. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA.* 2000;283(15):2008-2012.

105. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159-174.

106. MacArthur AC, McBride ML, Spinelli JJ, Tamaro S, Gallagher RP, Theriault GP. Risk of childhood leukemia associated with vaccination, infection, and medication use in childhood: the Cross-Canada Childhood Leukemia Study. *American journal of epidemiology.* 2008;167(5):598-606.

107. Neglia JP, Linet MS, Shu XO, et al. Patterns of infection and day care utilization and risk of childhood acute lymphoblastic leukaemia. *British Journal of Cancer.* 2000;82(1):234-240.

108. Nishi M, Miyake H. A case-control study of non-T cell acute lymphoblastic leukaemia of children in Hokkaido, Japan. *Journal of Epidemiology and Community Health.* 1989;43(4):352-355.

109. Rosenbaum PF, Buck GM, Brecher ML. Allergy and infectious disease histories and the risk of childhood acute lymphoblastic leukaemia. *Paediatric and Perinatal Epidemiology.* 2005;19(2):152-164.

110. Schlehofer B, Blettner M, Geletneky K, et al. Sero-epidemiological analysis of the risk of virus infections for childhood leukaemia. *Int J Cancer.* 1996;65(5):584-590.

111. Rosella L, Bowman C, Pach B, Morgan S, Fitzpatrick T, Goel V. The development and validation of a meta-tool for quality appraisal of public health evidence: Meta Quality Appraisal Tool (MetaQAT). *Public Health.* 2016;136:57-65.

112. CASP Case Control Checklist CASP; 2014. http://media.wix.com/ugd/dded87_63fb65dd4e0548e2bfd0a982295f839e.pdf. Accessed December 12, 2015.

113. CASP Cohort Study Checklist. CASP; 2014. http://media.wix.com/ugd/dded87_e37a4ab637fe46a0869f9f977dacf134.pdf. Accessed December 12, 2015.

114. Greenland S. Quantitative methods in the review of epidemiologic literature. *Epidemiologic reviews.* 1987;9:1-30.

115. Bae J-M. Comparison of methods of extracting information for meta-analysis of observational studies in nutritional epidemiology. *Epidemiology and Health.* 2016;38:e2016003.

116. Gart JJ, Nam J. Approximate interval estimation of the ratio of binomial parameters: a review and corrections for skewness. *Biometrics.* 1988;44(2):323-338.

117. Greenland S, Longnecker MP. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *American journal of epidemiology.* 1992;135(11):1301-1309.

118. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled clinical trials.* 1986;7(3):177-188.

119. Lagakos SW. The challenge of subgroup analyses--reporting without distorting. *N Engl J Med.* 2006;354(16):1667-1669.

120. Altman DG, Bland JM. Interaction revisited: the difference between two estimates. *BMJ.* 2003;326(7382):219.

121. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of clinical epidemiology.* 2008;61(10):991-996.

122. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ.* 1997;315(7109):629-634.

123. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. *2010.* 2010;36(3):48.

124. Paltiel O, Laniado DE, Yanetz R, et al. The risk of cancer following hospitalization for infection in infancy: A population-based cohort study. *Cancer Epidemiol Biomarkers Prev.* 2006;15(10):1964-1968.

125. Till M, Rapson N, Smith PG. Family studies in acute leukaemia in childhood: a possible association with autoimmune disease. *British journal of cancer.* 1979;40(1):62-71.

126. Ibrahem WN, Hasony HJ, Hassan JG. Human parvovirus B19 in childhood acute lymphoblastic leukaemia in Basrah. *Jpma.* 2014;The Journal of the Pakistan Medical Association. 64(1):9-12.

127. Kerr JR, Barah F, Cunniffe VS, et al. Association of acute parvovirus B19 infection with new onset of acute lymphoblastic and myeloblastic leukaemia. *Journal of Clinical Pathology.* 2003;56(11):873-875.

128. Mackenzie J, Gallagher A, Clayton RA, et al. Screening for herpesvirus genomes in common acute lymphoblastic leukemia. *Leukemia.* 2001;15(3):415-421.

129. Salonen MJH, Siimes MA, Salonen EM, Vaheri A, Koskiniemi M. Antibody status to HHV-6 in children with leukaemia. *Leukemia.* 2002;16(4):716-719.

130. Tesse R, Santoro N, Giordano P, Cardinale F, De Mattia D, Armenio L. Association between DEFB1 gene haplotype and herpes viruses seroprevalence in children with acute lymphoblastic leukemia. *Pediatr Hematol Oncol.* 2009;26(8):573-582.

131. Zaki ME, Hassan SA, Seleim T, Lateef RA. Parvovirus B19 infection in children with a variety of hematological disorders. *Hematology.* 2006;11(4):261-266.

132. Zaki MES, Ashray RE. Clinical and hematological study for Parvovirus b19 infection in children with acute leukemia. *International Journal of Laboratory Hematology.* 2010;32(2):159-166.

133. Lin JN, Lin CL, Lin MC, et al. Risk of leukaemia in children infected with enterovirus: a nationwide, retrospective, population-based, Taiwanese-registry, cohort study. *The Lancet Oncology.* 2015;16(13):1335-1343.

134. Ahmed HG, Osman SI, Ashankyty IM. Incidence of Epstein-Barr virus in pediatric leukemia in the Sudan. *Clin Lymphoma Myeloma Leuk.* 2012;12(2):127-131.

135. Ateyah ME, Hashem ME, Abdelsalam M. Epstein-Barr virus and regulatory T cells in Egyptian paediatric patients with acute B lymphoblastic leukaemia. *J Clin Pathol.* 2017;70(2):120-125.

136. Loutfy SA, Alam El-Din HM, Ibrahim MF, Hafez MM. Seroprevalence of herpes simplex virus types 1 and 2, Epstein-Barr virus, and cytomegalovirus in children with acute lymphoblastic leukemia in Egypt. *Saudi Med J.* 2006;27(8):1139-1145.

137. Mahjour SB, Ghaffarpasand F, Fattahi MJ, Ghaderi A, Fotouhi Ghiam A, Karimi M. Seroprevalence of human herpes simplex, hepatitis B and epstein-barr viruses in children with acute lymphoblastic leukemia in southern iran. *Pathol Oncol Res.* 2010;16(4):579-582.

138. Petridou E, Dalamaga M, Mentis A, et al. Evidence on the infectious etiology of childhood leukemia: the role of low herd immunity (Greece). *Cancer Causes Control.* 2001;12(7):645-652.

139. Surico G, Muggeo P. Epstein-Barr Virus Infection at the Onset of Acute Lymphoblastic Leukaemia in Children. *The International Cancer Journal of Australia and Asia.* 2005;4(1):19-24.

140. Lim JYS, Bhatia S, Robison LL, Yang JJ. Genomics of Racial and Ethnic Disparities in Childhood Acute Lymphoblastic Leukemia. *Cancer.* 2014;120(7):955-962.

141. Whitehead TP, Metayer C, Wiemels JL, Singer AW, Miller MD. Childhood Leukemia and Primary Prevention. *Current problems in pediatric and adolescent health care.* 2016;46(10):317-352.

142. Swaminathan S, Klemm L, Park E, et al. Mechanisms of clonal evolution in childhood acute lymphoblastic leukemia. *Nat Immunol.* 2015;16(7):766-774.

143. Kerr JR, Mattey DL. The role of parvovirus B19 and the immune response in the pathogenesis of acute leukemia. *Reviews in medical virology.* 2015;25(3):133-155.

144. Young NS, Brown KE. Parvovirus B19. *New England Journal of Medicine.* 2004;350(6):586-597.

145. Riley AW, Duncan GJ. Completing a national birth cohort in the united states. *JAMA pediatrics.* 2016;170(9):829-830.

146. Brown RC, Dwyer T, Kasten C, et al. Cohort Profile: The International Childhood Cancer Cohort Consortium (I4C). *International journal of epidemiology.* 2007;36(4):724-730.

147. Larsen PS, Kamper-Jorgensen M, Adamson A, et al. Pregnancy and birth cohort resources in europe: a large opportunity for aetiological child health research. *Paediatr Perinat Epidemiol.* 2013;27(4):393-414.

148. Caliendo AM, Gilbert DN, Ginocchio CC, et al. Better Tests, Better Care: Improved Diagnostics for Infectious Diseases. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America.* 2013;57(Suppl 3):S139-S170.

149. Amitay EL, Keinan-Boker L. Breastfeeding and childhood leukemia incidence: A meta-analysis and systematic review. *JAMA pediatrics.* 2015;169(6):e151025.

150. Kwan ML, Buffler PA, Abrams B, Kiley VA. Breastfeeding and the risk of childhood leukemia: a meta-analysis. *Public Health Reports.* 2004;119(6):521-535.

151. Martin RM, Gunnell D, Owen CG, Smith GD. Breast-feeding and childhood cancer: A systematic review with metaanalysis. *International journal of cancer Journal international du cancer.* 2005;117(6):1020-1031.

152. CIHI. *Canadian Institute for Health Information. A Snapshot of Health Care in Canada as Demonstrated by Top 10 Lists, 2011.* Ottawa, Canada2012.

153. Niska R, Bhuiya F, Xu J. National Hospital Ambulatory Medical Care Survey: 2007 emergency department summary. *National health statistics reports.* 2010(26):1-31.

154. Monto AS, Sullivan KM. Acute respiratory illness in the community. Frequency of illness and the agents involved. *Epidemiology and infection.* 1993;110(1):145-160.

155. Toivonen L, Karppinen S, Schuez-Havupalo L, et al. Burden of Recurrent Respiratory Tract Infections in Children: A Prospective Cohort Study. *The Pediatric infectious disease journal.* 2016;35(12):e362-e369.

156. Villers MS, Ramsey AF, Mitchell DK, et al. Utilization of Health Care for Infectious Illnesses at a Pediatric Practice 549. *Pediatric Research.* 1998;43:96.

157. Shiff NJ, Jama S, Boden C, Lix LM. Validation of administrative health data for the pediatric population: a scoping review. *BMC health services research.* 2014;14:236.

158. Population by sex and age group, by province and territory (Number, both sexes). July 1, 2016. 2016. http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/demo31a-eng.htm. Accessed July 6, 2017.

159. Guttmann A, Nakhla M, Henderson M, et al. Validation of a health administrative data algorithm for assessing the epidemiology of diabetes in Canadian children. *Pediatric diabetes.* 2010;11(2):122-128.

160. Gershon AS, Wang C, Guan J, Vasilevska-Ristovska J, Cicutto L, To T. Identifying patients with physician-diagnosed asthma in health administrative databases. *Canadian Respiratory Journal : Journal of the Canadian Thoracic Society.* 2009;16(6):183-188.

161. Schwartz KL, Tu K, Wing L, et al. Validation of infant immunization billing codes in administrative data. *Human Vaccines & Immunotherapeutics.* 2015;11(7):1840-1847.

162. Thomas EM. Recent trends in upper respiratory infections, ear infections and asthma among young Canadian children. *Health reports.* 2010;21(4):47-52.

163. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *Journal of clinical epidemiology.* 2005;58(8):859-862.

164. Tu K, Mitiku T, Lee DS, Guo H, Tu JV. Validation of physician billing and hospitalization data to identify patients with ischemic heart disease using data from the Electronic Medical Record Administrative data Linked Database (EMRALD). *The Canadian journal of cardiology.* 2010;26(7):e225-228.

165. Tu K, Wang M, Young J, et al. Validity of administrative data for identifying patients who have had a stroke or transient ischemic attack using EMRALD as a reference standard. *The Canadian journal of cardiology.* 2013;29(11):1388-1394.

166. du Plessis V, Beshiri R, Bollman RD, Clemenson H. Definitions of "Rural". In: Division A, ed. Vol 61. Ottawa, Canada: Statistics Canada; 2002.

167. MHLTC. Ontario Ministry of Health and Long-Term Care. PHPDB - Medical Services User Guide. In: Database PHP, ed. Toronto: Ministry of Health and Long-Term Care; 2008.

168. Hwee J, Tait C, Sung L, Kwong JC, Sutradhar R, Pole JD. A systematic review and meta-analysis of the association between childhood infections and the risk of childhood acute lymphoblastic leukaemia. *British Journal Of Cancer.* 2017;118:127.

169. Feudtner C, Hays RM, Haynes G, Geyer JR, Neff JM, Koepsell TD. Deaths attributed to pediatric complex chronic conditions: national trends and implications for supportive care services. *Pediatrics.* 2001;107(6):E99.

170. Cohen E, Berry JG, Camacho X, Anderson G, Wodchis W, Guttmann A. Patterns and costs of health care use of children with medical complexity. *Pediatrics.* 2012;130(6):e1463-1470.

171. Austin PC. Using the Standardized Difference to Compare the Prevalence of a Binary Variable Between Two Groups in Observational Research. *Communications in Statistics - Simulation and Computation.* 2009;38(6):1228-1234.

172. Pisesky A, Benchimol EI, Wong CA, et al. Incidence of Hospitalization for Respiratory Syncytial Virus Infection amongst Children in Ontario, Canada: A Population-Based Study Using Validated Health Administrative Data. *PLoS One.* 2016;11(3):e0150416.

173. Hsu VP, Staat MA, Roberts N, et al. Use of active surveillance to validate international classification of diseases code estimates of rotavirus hospitalizations in children. *Pediatrics.* 2005;115(1):78-82.

174. Williams DJ, Shah SS, Myers A, et al. Identifying pediatric community-acquired pneumonia hospitalizations: Accuracy of administrative billing codes. *JAMA pediatrics.* 2013;167(9):851-858.

175. O'Sullivan CE, Baker MG. Proposed epidemiological case definition for serious skin infection in children. *Journal of paediatrics and child health.* 2010;46(4):176-183.

176. Beitel AJ, Olson KL, Reis BY, Mandl KD. Use of emergency department chief complaint and diagnostic codes for identifying respiratory illness in a pediatric population. *Pediatric emergency care.* 2004;20(6):355-360.

177. Shaklee J, Zerr DM, Elward A, et al. Improving surveillance for pediatric Clostridium difficile infection: derivation and validation of an accurate case-finding tool. *The Pediatric infectious disease journal.* 2011;30(3):e38-40.

178. Tieder JS, Hall M, Auger KA, et al. Accuracy of administrative billing codes to detect urinary tract infection hospitalizations. *Pediatrics.* 2011;128(2):323-330.

179. Bourgeois FT, Olson KL, Brownstein JS, McAdam AJ, Mandl KD. Validation of syndromic surveillance for respiratory infections. *Annals of emergency medicine.* 2006;47(3):265.e261.

180. Jaakkimainen RL, Shultz SE, Tu K. Effects of implementing electronic medical records on primary care billings and payments: a before–after study. *CMAJ Open.* 2013;1(3):E120-E126.

181. Glazier RZ, BM; Rayner, J. *Comparison of Primary Care Models in Ontario by Demographics, Case Mix and Emergency Department Use, 2008/09 to 2009/10.* Toronto: Institute for Clinical Evaluative Sciences;2012.

182. Lazarus R, Klompas M, Campion FX, et al. Electronic Support for Public Health: Validated Case Finding and Reporting for Notifiable Diseases Using Electronic Medical Data. *Journal of the American Medical Informatics Association.* 2009;16(1):18-24.

183. Hasegawa K, Tsugawa Y, Cohen A, Camargo CAJ. Infectious Disease-related Emergency Department Visits Among Children in the US. *The Pediatric Infectious Disease Journal.* 2015;34(7):681-685.

184. Pui C-H, Robison LL, Look AT. Acute lymphoblastic leukaemia. *The Lancet.* 2008;371(9617):1030-1043.

185. Johnston BL, Conly JM. The changing face of Canadian immigration: Implications for infectious diseases. *The Canadian Journal of Infectious Diseases & Medical Microbiology.* 2008;19(4):270-272.

186. Dores GM, Devesa SS, Curtis RE, Linet MS, Morton LM. Acute leukemia incidence and patient survival among children and adults in the United States, 2001-2007. *Blood.* 2012;119(1):34-43.

187. Albert RH. Diagnosis and treatment of acute bronchitis. *American family physician.* 2010;82(11):1345-1350.

188. Dan L, Anthony F, Dennis K, Stephen H, Jameson J, Joseph L. *Chapter 31: Pharyngitis, Sinusitis, Otitis, and Other Upper Respiratory Tract Infections* McGraw-Hill Professional; 2012.

189. Dan L, Anthony F, Dennis K, Stephen H, Jameson J, Joseph L. *Viral Gastroenteritis.* McGraw-Hill Professional; 2012.

190. Kamper-Jørgensen M, Wohlfahrt J, Simonsen J, Grønbæk M, Benn CS. Population-Based Study of the Impact of Childcare Attendance on Hospitalizations for Acute Respiratory Infections. *Pediatrics.* 2006;118(4):1439-1446.

191. Fendrick A, Monto AS, Nightengale B, Sarnes M. The economic burden of non–influenza-related viral respiratory tract infection in the united states. *Archives of internal medicine.* 2003;163(4):487-494.

192. Turner RB. The common cold. *Pediatric annals.* 1998;27(12):790-795.

193. Kleinbaum DG, and Klein, Mitchel. *Logistic Regression: A Self-Learning Text.* Vol 3. New York: Springer-Verlag 2010.

194. Kuh D, Ben-Shlomo Y, Lynch J, Hallqvist J, Power C. Life course epidemiology. *Journal of Epidemiology and Community Health.* 2003;57(10):778-783.

195. Wilson A, Chiu YM, Hsu HL, Wright RO, Wright RJ, Coull BA. Potential for Bias When Estimating Critical Windows for Air Pollution in Children's Health. *American journal of epidemiology.* 2017;186(11):1281-1289.

196. Nelson W. *Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications.*

197. Sinha M. *Spotlight on Canadians: Results from the General Social Survey. Child care in Canada.* Ottawa: Statistics Canada;2014.

198. Couper KN, Blount DG, Riley EM. IL-10: The Master Regulator of Immunity to Infection. *The Journal of Immunology.* 2008;180(9):5771-5777.

199. Oft M. IL-10: master switch from tumor-promoting inflammation to antitumor immunity. *Cancer immunology research.* 2014;2(3):194-199.

200. Dennis KL, Blatner NR, Gounari F, Khazaie K. Current status of IL-10 and regulatory T-cells in cancer. *Current opinion in oncology.* 2013;25(6):637-645.

201. LO W-J, CHANG W-S, HSU H-F, et al. Significant Association of Interleukin-10 Polymorphisms with Childhood Leukemia Susceptibility in Taiwan. *In Vivo.* 2016;30(3):265-269.

202. Zhang G, Rowe J, Kusel M, et al. Interleukin-10/interleukin-5 responses at birth predict risk for respiratory infections in children with atopic family history. *Am J Respir Crit Care Med.* 2009;179(3):205-211.

203. Ding S, Wang X, Chen W, et al. Decreased Interleukin-10 Responses in Children with Severe Mycoplasma pneumoniae Pneumonia. *PLOS ONE.* 2016;11(1):e0146397.

204. Hoebee B, Bont L, Rietveld E, et al. Influence of Promoter Variants of Interleukin-10, Interleukin-9, and Tumor Necrosis Factor–α Genes on Respiratory Syncytial Virus Bronchiolitis. *The Journal of Infectious Diseases.* 2004;189(2):239-247.

205. Wilson J, Rowlands K, Rockett K, et al. Genetic variation at the IL10 gene locus is associated with severity of respiratory syncytial virus bronchiolitis. *J Infect Dis.* 2005;191(10):1705-1709.

206. Mege J-L, Meghari S, Honstettre A, Capo C, Raoult D. The two faces of interleukin 10 in human infectious diseases. *The Lancet Infectious Diseases.* 2006;6(9):557-569.

207. Taber JM, Leyva B, Persoskie A. Why do People Avoid Medical Care? A Qualitative Study Using National Data. *Journal of General Internal Medicine.* 2015;30(3):290-297.

208. Riley AW, Finney JW, Mellits ED, et al. Determinants of children's health care use: an investigation of psychosocial factors. *Medical care.* 1993;31(9):767-783.

209. Guttmann A, Shipman SA, Lam K, Goodman DC, Stukel TA. Primary Care Physician Supply and Children's Health Care Use, Access, and Outcomes: Findings From Canada. *Pediatrics.* 2010.

210. Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *Journal of clinical epidemiology.* 2004;57(3):229-236.

211. Okonny-Myers I. The Interprovincial Mobility of Immigrants in Canada.: Citizenship and Immigration Canada. Government of Canada; June 2010.

212. Bruijnzeels MA, Foets M, van der Wouden JC, van den Heuvel WJ, Prins A. Everyday symptoms in childhood: occurrence and general practitioner consultation rates. *The British*

journal of general practice : the journal of the Royal College of General Practitioners. 1998;48(426):880-884.

213. Holme CO. Incidence and prevalence of non-specific symptoms and behavioural changes in infants under the age of two years. *The British journal of general practice : the journal of the Royal College of General Practitioners.* 1995;45(391):65-69.

214. Bouche G, Migeot V. Parental use of the Internet to seek health information and primary care utilisation for their child: a cross-sectional study. *BMC Public Health.* 2008;8:300.

215. Abat C, Chaudet H, Rolain J-M, Colson P, Raoult D. Traditional and syndromic surveillance of infectious diseases and pathogens. *International Journal of Infectious Diseases.* 2016;48:22-28.

216. Kwong JC, Ratnasingham S, Campitelli MA, et al. The Impact of Infection on Population Health: Results of the Ontario Burden of Infectious Diseases Study. *PLOS ONE.* 2012;7(9):e44103.

217. Kwong JC, Schwartz KL, Campitelli MA, et al. Acute Myocardial Infarction after Laboratory-Confirmed Influenza Infection. *New England Journal of Medicine.* 2018;378(4):345-353.

218. *Association of Public Health Epidemiologists in Ontario Core Indicators Work Group; Ontario Agency for Health Protection and Promotion (Public Health Ontario). Gaps in public health indicators and data in Ontario. Revised ed.* . Toronto: Association of Public Health Epidemiologists in Ontario;2016.

219. Gupta S, Gibson P, Pole JD, Sutradhar R, Sung L, Guttmann A. Predictors of diagnostic interval and associations with outcome in acute lymphoblastic leukemia. *Pediatric blood & cancer.* 2015;62(6):957-963.

220. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: Introducing the e-value. *Annals of internal medicine.* 2017;167(4):268-274.

221. Population by year, by province and territory (Number). Statistics Canada; 2017. http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/demo02a-eng.htm. Accessed April 26, 2018.

222. Canada S. Chapter 13 - Ethnic diversity and immigration. *Canada Year Book 2011 - Catalouge no. 11-402-X*. Ottawa, Canada: Statistics Canada; 2011.

223. Sorensen HT, Pedersen L, Olsen J, Rothman K. Seasonal variation in month of birth and diagnosis of early childhood acute lymphoblastic leukemia. *Jama.* 2001;285(2):168-169.

224. Nyari TA, Kajtar P, Bartyik K, Thurzo L, McNally R, Parker L. Seasonal variation of childhood acute lymphoblastic leukaemia is different between girls and boys. *Pathology oncology research : POR.* 2008;14(4):423-428.

225. Feltbower RG, Pearce MS, Dickinson HO, Parker L, McKinney PA. Seasonality of birth for cancer in Northern England, UK. *Paediatric and Perinatal Epidemiology.* 2001;15(4):338-345.

226. Marcotte EL, Ritz B, Cockburn M, Yu F, Heck JE. Exposure to Infections and Risk of Leukemia in Young Children. *Cancer Epidemiology Biomarkers & Prevention.* 2014;23(7):1195-1203.

227. Lipsitch M, Tchetgen ET, Cohen T. Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. *Epidemiology (Cambridge, Mass).* 2010;21(3):383-388.

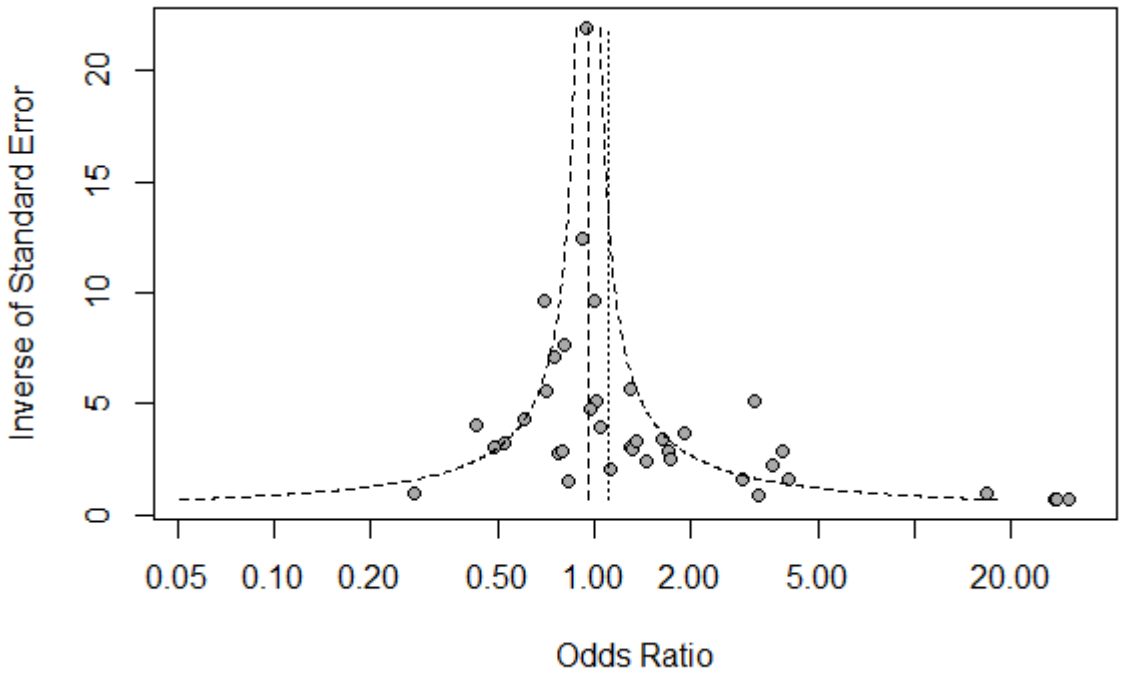
228. Canada Go. Equality + Growth. A Strong Middle Class. 2018 Budget Plan: Chapter 2 - Progress. In: Canada DoF, ed. Ottawa: Government of Canada; 2018:85-95.

229. VanderWeele TJ. Mediation Analysis: A Practitioner's Guide. *Annual Review of Public Health.* 2016;37(1):17-32.

Appendices

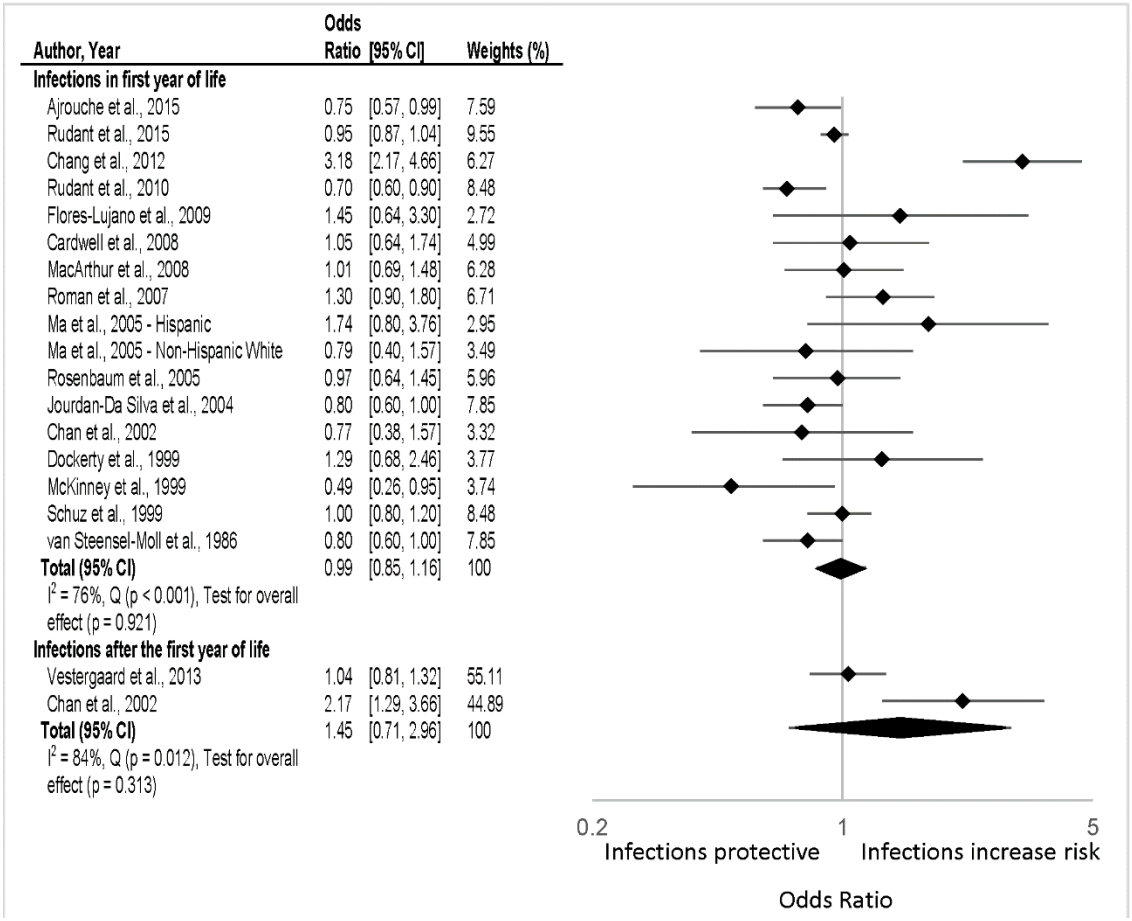
Appendices A Supplementary Information for Objective 1

Supplementary Figure 3.1 Egger's test and funnel plot for the presence of publication bias.



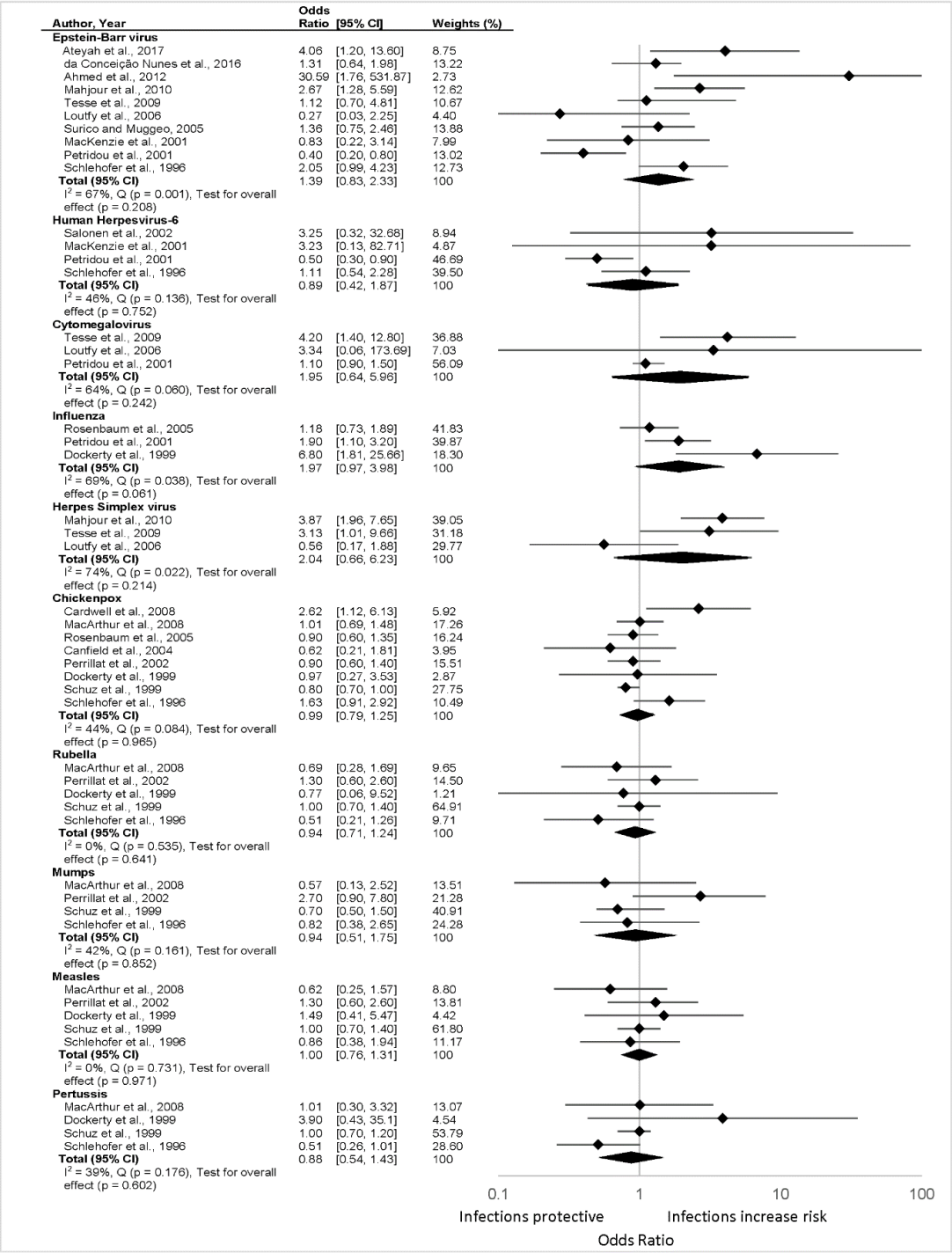
Egger's bias coefficient was 1.19 (95% CI: 0.30, 2.08).

Supplementary Figure 3.2 Random effects model examining the association between the timing of infections and odds of childhood acute lymphoblastic leukemia



CI represents confidence interval. Common infections are reported as a two-class variable, or highest vs lowest in more than 2 categories.

Supplementary Figure 3.3 Random effects model examining the association between infectious pathogens and odds of childhood acute lymphoblastic leukemia



CI represents confidence interval. Common infections are reported as a two-class variable, or highest vs lowest in more than 2 categories.

Supplementary Table 3.1 Search Strategy by Database MEDLINE(R) 1946 to February 21, 2017

MEDLINE(R) 1946 to February 21, 2017

	Searches	**Results**
1	leukemia, lymphoid/ or leukemia, b-cell/ or leukemia, prolymphocytic, b-cell/ or leukemia, biphenotypic, acute/ or leukemia, prolymphocytic/ or leukemia, prolymphocytic, t-cell/ or leukemia, t-cell/ or leukemia, large granular lymphocytic/ or precursor cell lymphoblastic leukemia-lymphoma/ or precursor b-cell lymphoblastic leukemia-lymphoma/ or precursor t-cell lymphoblastic leukemia-lymphoma/	50217
2	("acute biphenotypic leukaemia*" or "acute biphenotypic leukemia*" or "acute leukaemia* biphenotypic" or "acute leukaemia* hybrid" or "acute leukaemia* mixed-lineage" or "acute leukemia* biphenotypic" or "acute leukemia* hybrid" or "acute leukemia* mixed-lineage" or "acute t cell leukaemia*" or "acute t cell leukemia*" or "acute t lymphocytic leukaemia*" or "acute t lymphocytic leukemia*" or "all childhood" or "b and t cell leukaemia* acute" or "b and t cell leukemia* acute" or "b cell leukaemia*" or "b cell leukemia*" or "biphenotypic acute leukaemia*" or "biphenotypic acute leukemia*" or "biphenotypic leukaemia* acute" or "biphenotypic leukemia* acute" or "childhood all" or "hybrid acute leukaemia*" or "hybrid acute leukemia*" or "large granular lymphocyte leukemia*" or "leukaemia* acute biphenotypic" or "leukaemia* acute t cell" or "leukaemia* acute t lymphocytic" or "leukaemia* b cell" or "leukaemia* b lymphocytic" or "leukaemia* biphenotypic acute" or "leukaemia* hybrid acute" or "leukaemia* lgl" or "leukaemia* lymphocytic" or "leukaemia* lymphoid" or "leukaemia* mixed cell" or "leukaemia* mixed lineage acute" or "leukaemia* nk-lgl" or "leukaemia* pre-b cell" or "leukaemia* prolymphocytic" or "leukaemia* t lymphocytic" or "leukaemia* t cell" or "leukaemia* t lgl" or "leukaemia*, CALLA-positive" or "leukaemia*, lymphocytic, acute" or "leukaemia*, mixed cell" or "leukaemia*, null cell" or "leukaemia*,acute lymphatic" or "leukaemia*,acute lymphoblastic" or "leukemia* acute biphenotypic" or "leukemia* acute lymphoblastic" or "leukemia* acute lymphocytic" or "leukemia* acute lymphoid" or "leukemia* acute t cell" or "leukemia* acute t lymphocytic" or "leukemia* b cell" or "leukemia* biphenotypic acute" or "leukemia* b lymphocytic" or "leukemia* hybrid acute" or "leukemia* l1 lymphocytic" or "leukemia* l2 lymphocytic" or "leukemia* large granular lymphocytic" or "leukemia* lgl" or "leukemia* lymphoblastic" or "leukemia* lymphocytic" or "leukemia* lymphoid" or "leukemia* mixed b and t cell" or "leukemia* mixed cell" or "leukemia* mixed lineage acute" or "leukemia* natural killer cell large granular lymphocytic" or "leukemia* nk lgl" or "leukemia* pre b cell" or "leukemia* prolymphocytic" or "leukemia* t cell" or "leukemia* t lgl" or "leukemia* t lymphocytic" or "leukemia*, b cell" or "leukemia*, biphenotypic, acute" or "leukemia*, CALLA- positive" or "leukemia*, large granular lymphocytic" or "leukemia*, lymphocytic, acute" or "leukemia*, lymphoid" or "leukemia*, mixed cell" or "leukemia*, null cell" or "leukemia*, prolymphocytic" or "leukemia*, t cell" or "leukemia*,acute lymphatic" or "leukemia*,acute lymphoblastic" or "lgl leukaemia*" or "lgl leukemia*" or "lymphatic leukaemia*,acute" or "lymphatic leukemia*,acute" or "lymphoblastic leukaemia*" or "lymphoblastic leukemia*" or "lymphoblastic lymphoma" or "lymphocytic leukaemia*" or "lymphocytic leukemia*" or "lymphoid leukaemia*" or "lymphoid leukemia*" or "lymphoma lymphoblastic" or "lymphoproliferative disease of granular lymphocytes" or "lymphoproliferative disease of large granular lymphocytes" or "mixed lineage acute leukaemia*" or "mixed cell leukaemia*" or "mixed cell leukemia*" or "mixed-lineage acute leukaemia*" or "mixed-lineage acute leukemia*" or "nk lgl leukemia*" or "nk-lgl leukaemia*" or "nk-lgl leukemia*" or "pre b all" or "pre b cell leukemia*" or "pre-b cell leukaemia*" or "prolymphocytic leukaemia*" or "prolymphocytic leukemia*" or "t all" or "t cell leukaemia*" or "t cell leukemia*" or "t cell leukemia*" or "t lgl leukemia*" or "t-lgl leukaemia*" or "t-lgl leukemia*").tw.	60896
3	1 or 2	81461
4	Infection/	35780
5	infect*.mp.	1781046
6	4 or 5	1781046
7	3 and 6	10324

8	limit 7 to "all child (0 to 18 years)"	3390
9	(infan* or neonat* or child* or adolescen* or juvenile or teen* or girl* or boy* or youth* or toddler* or paediatric* or pediatric*).tw.	1779541
10	7 and 9	2355
11	**8 or 10**	**3620**

Supplementary Table 3.2 Study Definitions of Common Infections Variable

Study, Year (Reference)	Definition for common infections	Specific infections investigated and definitions
Ateyah et al. 2017	*	Latent infection: EBV IgG antibody titer to VCA and EBV nuclear antigen
Conceicao Nunes et al. 2016	*	Total immunoglobin type E; Immunoglobin for parvovirus B19 specific IgG antibodies; EBV anti-VCA IgG
Lin et al. 2015	• Enterovirus defined as ≥3 clinic visits with an associated ICD-9-CM diagnosis code: 008.67, 047, 048, 074, 079.1, or 079.2	*
Rudant et al. 2015	• Pooled analysis from 8 different studies, definitions range from a combination of the following: any infection, ear, nose, throat infection, gastroenteritis and any other, tonsillitis, otitis media, upper respiratory tract infections, bronchiolitis and other lower respiratory tract infections, gastroenteritis, urinary, pneumonia, cold, persistent cough, diarrhea	Ear, nose, throat infections; Otitis media; Lower respiratory tract infections; Gastroenteritis; Ear, nose, throat surgery
Ajrouche et al. 2015	• Common infections included: tonsillitis, otitis media, upper respiratory tract infections, gastroenteritis, bronchiolitis and other lower respiratory tract infections, and urinary tract infections	Ear, nose, throat surgery for repeated common infections (adenoidectomy, tonsillectomy, paracentesis) before age 4; Pediatric infections (measles, rubella, chickenpox, mumps, whooping cough, scarlet fever, hand, foot and mouth disease, meningitis, mononucleosis); History of hospitalization for infections and other causes; Tonsillitis; Otitis media; Rhinopharyngitis; Laryngitis; Conjunctivitis; Bronchiolitis; Pulmonary infection; Gastroenteritis; Urinary tract infections
Ibrahem et al. 2014	*	Parvovirus B19 IgG antibodies
Vestergaard et al. 2013	• Defined as hospitalizations using ICD-8 and ICD-10 codes • Severe infections: bacterial meningitis, viral central nervous system infections, septicaemia, pyelonephritis, osteomyelitis, ethmoiditis • Less severe infections: upper respiratory infections, pneumonia, bronchitis, lower urinary tract infections, gastroenteritis, conjunctivitis, influenza	Bacterial meningitis: ICD-8: 013-013.09, 027.01, 320, 036.09; ICD-10: A17, A32.1, G00, G01, G05.0; Viral central nervous system infections: ICD-8: 045, 052.01, 053.02, 054.03, 055.01, 056.01, 075.01, 075.02, 079.29, 323.00, 323.08, 323.09, 065; ICD-10: A85-A87, B00.3, B00.4, B01.0, B01.1, B02.0, B02.1, B05.0, B05.1, B06.0A, B06.0B, B06.0C, B26.1, B26.2, G02.0, G05.1, G05.2; Septicaemia: ICD-8: 036.10, 036.11, 038; ICD-10: A02.1, A32.7, A37.7, A39.3, A40, A41; Pyelonephritis: ICD-8: 590.10, 590.11, 590.12, 590.13; ICD-10: N10.9, N12; Osteomyelitis: ICD-8: 720.00–720.09, 015.09; ICD-10: M46.2, M46.5, M68.2, M86.0, M86.1; Ethmoiditis: ICD-8: 461.02; ICD-10: J01.2; Upper

		respiratory: ICD-8: 034.00, 034.01, 034.09, 381.01, 381.02, 382.09, 383.09, 460, 461.00, 461.01, 461.03, 461.04, 461.08, 461.09, 462, 463, 464, 465, 501.99, 508.00–508.09; ICD-10: B53, H66.0, H67.0, H67.1, H70.0, H73.0, J00, J01.3, J01.4, J01.8, J01.9, J02–J06, J36; Pneumonia: ICD-8: 011, 480–483, 485, 486; ICD-10: A15.0–A15.3, A16.0–A16.2, A48.1, B01.2, B05.2, B06.8A, J12–J17, J18.0, J18.1, J18.8, J18.9, J22; Bronchitis: ICD-8: 466; ICD-10: J20, J21; Lower urinary tract: ICD-8: 595.00, 595.01; ICD-10: N30.0; Gastroenteritis: ICD-8: 000-009; ICD-10: A00, A01, A02.0, A03–A05, A06.0, A07–A09, K93.0; Conjunctivitis: ICD-8: 053.00, 054.04, 078.00–078.09, 360.00; ICD-10: A74.0, B00.5, B02.3, B30, H10.0, H13.1, H19.1; Influenza: ICD-8: 470-474; ICD-10: J01.0, J01.1
Ahmed et al. 2012	*	PCR for EBV DNA
Chang et al. 2012	• Defined as ambulatory care visits, and hospitalizations using ICD-9 CM codes • Common infection included: otitis media, acute respiratory infections, pneumonia and influenza, unspecified bronchitis, intestinal infectious diseases, conjunctivitis and perinatal infections	Otitis media: 381 and 382; Acute respiratory infections: 460-466; Pneumonia and influenza: 480-488; Unspecified bronchitis: 490; Intestinal infectious diseases: 001-009; Conjunctivitis: 372.0-372.3, 771.6; Infections specific to the perinatal period: 771
Mahjour et al. 2010	*	EBV anti-VCA IgG; HSV IgG antibodies; and Hepatitis B Virus antibodies
Rudant et al. 2010	• Common infections included: tonsillitis, otitis, upper respiratory tract infections, gastroenteritis, bronchiolitis and other lower respiratory tract infections, and urinary tract infections • Repeated common infections defined as 4 or more episodes of infection of at least 1 given site or 1–3 episodes of infection of at least 4 sites	Tonsillitis; Otitis media; Upper respiratory tract infections; Bronchiolitis and other lower respiratory tract infections; Gastroenteritis; Urinary tract infections
Zaki and Ashray 2010	*	Parvovirus B19 IgG antibodies
Flores-Lujano et al. 2009	• Common infections included: upper respiratory tract infections, bronchopneumonia, pneumonias, gastrointestinal infections, and others (not defined)	Hospitalizations for infections included: gastrointestinal, respiratory tract infection, and others (not defined)
Tesse et al. 2009	*	HSV 1 and 2 IgG antibodies; EBV IgG antibodies; and CMV IgG antibodies
Cardwell et al. 2008	• Medical records were abstracted using OXMIS and READ codes • Definition included infections and symptoms for: diarrhoea, fever, pyrexia, sore throat, earache, snuffles, vomiting and diarrhoea, dysuria, otorrhoea and chesty cough	Upper respiratory tract; Lower respiratory tract; Otitis media; Conjunctivitis; Gastrointestinal; Urinary tract; Non-invasive fungal disease; Chickenpox

MacArthur et al. 2008	*	Mumps; Measles; Rubella; Multiple ear infection; Chickenpox; Pertussis; Other illness (not defined)
Roman et al. 2007	• Medical records were abstracted and coded using ICD-10 codes • Common infections: A00–B99, H10, H66, J00–J11, J18–J22, L00–L03, L08, and P35–P39	Upper respiratory tract: ICD-10: J-J00-J06, J11.1; Lower respiratory tract: ICD-10: J18-J22; Otitis media: ICD-10: H66; Conjunctivitis: ICD-10: H10, P39.1; Gastrointestinal: ICD-10: A02-A09; Non-invasive fungal disease: ICD-10: B35, B37, P37.5
Loutfy et al. 2006	*	EBV anti-VCA IgG; HSV IgG antibodies; Cytomegalovirus IgG antibodies
Paltiel et al. 2006	• Hospitalizations for infections defined using ICD-7 as at least one admission for: 2-138.9, 300, 309, 310, 340, 390-394.9, 400-402.9, 430-432.9, 468-468.2, 470-475.9, 480-483.9, 490-493.9, 500-502.9, 510-513.9, 516-519.9, 521, 523-527, 530-532, 536-540, 543, 550-553.9, 571, 572, 575-576.9, 580-582, 585, 587, 590-592.9, 600, 601, 607, 609, 611, 614, 626, 630, 690-698.9, 700, 701, 720, 730, 743	*
Zaki et al. 2006	*	Parvovirus B19 IgG antibodies
Ma, et al. 2005	• Common infections included: severe diarrhea and vomiting, ear infections, persistent cough, mouth infection, eye infection, influenza, other infections	Severe diarrhea and vomiting; Ear infection; Persistent cough
Rosenbaum et al. 2005	*	Colds; Otitis media; Streptococcal and sinus infections; Vomiting; Diarrhoea; Influenza; Croup; Bronchiolitis; Pneumonia; Chickenpox; Results not reported: meningitis, septicaemia, skin infection, Coxsackie viral infections, other (not defined), measles, mumps, rubella, fifth disease
Surico and Muggeo 2005	*	EBV anti-VCA IgG and EBV nuclear antigen IgG
Jourdan-Da Silva et al. 2004	• Common infections included: ear, nose and throat, gastrointestinal, and other infections (not defined) • Infantile diseases included: chickenpox, measles, mumps, rubella	Ear, nose and throat infections; Gastrointestinal
Canfield et al. 2004	• Common infections included: chickenpox, ear infections, measles, colds, and bronchial infections	Chickenpox; Ear infections; Colds and bronchial infections; Other (not defined); Results not reported: mumps, rubella
Kerr et al. 2003	•	
Chan et al. 2002	• Common infections included: roseola, measles, mumps, rubella, varicella, pertussis, herpes simplex, pneumonia, ear infections, eye infection, and other fever with rash	Tonsillitis; Roseola and/or fever and rash
Perrillat et al. 2002	• Common infections not clearly defined	Measles; Rubella; Chickenpox; Mumps; Glandular fever

		Viral hepatitis; Surgical procedures as a measure of repeated ear, nose and throat infections
Salonen et al. 2002	*	HHV-6 IgG antibodies
MacKenzie et al. 2001	*	Quantitative PCR of HHV-6 DNA
Petridou et al. 2001	*	Adenovirus IgG; Epstein-Barr virus anti-VCA IgG; Human herpes virus-6 IgG antibodies; Influenza A IgG antibodies; Influenza B IgG antibodies; Parainfluenza 1, 2, 3 IgG antibodies; Parvovirus B19 IgG antibodies; Respiratory syncytial virus IgG antibodies; Cytomegalovirus IgG antibodies; Mycoplasma antibodies
Neglia et al. 2000	*	Ear infection; Lung infection; Gastroenteritis (vomiting and diarrhoea)
Schuz et al. 1999	• Common infections included: chickenpox, measles, mumps, rubella, Scarlet fever, pneumonia, bronchitis, pertussis, inflammation of the middle ear, diphtheria, tetanus, poliomyelitis, croup, herpes labialis, rheumatic fever, hepatitis, Pfeiffer's disease and Sticker's disease	Chickenpox; Measles; Mumps; Rubella; Scarlet fever; Pneumonia; Bronchitis; Pertussis; Inflammation of the middle ear; Other infections (not defined)
McKinney et al. 1999	• Infections coded using ICD-10 • Common infections included: respiratory tract, gastrointestinal tract, fungal, conjunctivitis, skin infections, other	Respiratory tract; Gastrointestinal tract; Fungal; Conjunctivitis; Skin infections; Other: not completed defined, but includes ICD-10: P36 (bacterial sepsis of newborn), P39.9 infections in perinatal period
Dockerty et al. 1999	• Common infections in 1st year of life included: whooping cough, measles, rubella, chickenpox, mouth infection, eye infection, ear infection, influenza, colds, persistent cough, diarrhoea and vomiting, other infection (not defined) • Infections any time prior to diagnosis date: glandular fever, cold sores, giardiasis, hepatitis B, poliomyelitis, cytomegalovirus infection	Whooping cough; Measles; Rubella; Chickenpox; Mouth infection; Eye infection; Ear infection; Influenza; Colds; Persistent cough; Diarrhoea and vomiting; Other infection; Results not reported: hepatitis B, poliomyelitis, cytomegalovirus infection, mumps
Schlehofer et al. 1996	*	Whooping cough; Rubella; Mumps; Measles; Chickenpox; Herpes labialis; Unspecific exanthema Parvovirus B19 IgG antibodies; HHV-6 antibodies, EBV anti-VCA IgG, adeno-associated parvovirus IgG antibodies
Nishi et al. 1989	*	Measles and measles vaccination combined, however there are percentages of cases and controls with measles infections that can allow for back-calculating an effect estimate; Results not reported: chickenpox, rubella, mumps, others (not defined)
McKinney et al. 1987	• Common infections coded as ICD-9 and included: viral diseases (chickenpox, rubella, measles, mumps, viral meningitis, and viral influenza) • Viral diseases ICD-9: 045-079, 408	*

van Steensel-Moll et al. 1986	• Since there was no common infection measure, we used the common colds variable to represent common infections	Bronchitis; Primary infections (measles, chickenpox, mumps, or rubella); Otitis media; Common colds Periods of fever (temperature >38 C for 2 days or longer); Hospitalization/consultation for infections (most common were pneumonia, bronchitis, meningitis, otitis, tonsillectomy, skin infections, urinary tract infections, diarrhea, and unspecified fever or viral infections); Infections
Till et al. 1979	• Common infections (counts): infantile gastroenteritis, pneumonia, pyogenic infections, upper respiratory infection, urinary tract, infectious mononucleosis, infective hepatitis, viral meningitis	Infantile gastroenteritis; Pneumonia; Pyogenic infections; Upper respiratory infection; Urinary tract; Infectious mononucleosis; Infective hepatitis; Viral meningitis

*Represent studies that did not have either a common infection definition or did not examine specific infection types. VCA is viral capsid antigen, PCR is polymerase chain reaction, EBV is Epstein-Barr virus, HSV is herpes simplex viruses, HHV-6 is human herpesvirus-6

Supplementary Table 3.3a Risk of Bias Assessment Using the CASP Tool of the Included Case-Control Studies

Case-Control Study	Did the study address a clearly focused issue?	Did the authors use an appropriate method to answer their question	Were the cases recruited in an acceptable way?	Were the controls selected in an acceptable way?	Was the exposure accurately measured to minimize bias?	Have authors taken account of potential confounding factors in design and analysis?	Do you believe in the results?	Can the results be applied to local population?	Do the results of this study fit with other available evidence?	TOTAL
Ahmed et al. 2012	Yes	Yes	Can't tell	Can't tell	No	No	No	No	Can't tell	2
Ajrouche et al. 2015	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	9
Ateyah et al. 2017	Yes	Yes	Can't tell	Can't tell	No	No	No	No	Can't tell	2
Canfield et al. 2004	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	9
Cardwell et al. 2008	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	7
Chan et al. 2002	Yes	No	Yes	No	No	No	No	No	Yes	3
Chang et al. 2012	Yes	Yes	Yes	Yes	Yes	No	No	No	No	5
Conceicao Nunes et al. 2016	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	6
Dockerty et al. 1999	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	7
Flores-Lujano et al. 2009	Yes	Yes	Yes	No	No	No	No	No	No	3
Ibrahem et al. 2014	Yes	Yes	Can't tell	Can't tell	No	No	No	No	Can't tell	2
Jourdan-Da Silva et al. 2004	Yes	Yes	Yes	Yes	No	No	No	No	Yes	5
Kerr et al. 2003	Yes	Yes	Can't tell	No	Yes	No	No	No	Can't tell	3

Loutfy et al. 2006	Yes	No	Can’t tell	No	No	No	No	No	Can’t tell	1
Ma et al. 2005	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	9
MacArthur et al. 2008	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	9
MacKenzie et al. 2001	Yes	Yes	Can’t tell	Can’t tell	Yes	No	No	No	Can’t tell	3
Mahjour et al. 2009	Yes	Yes	Can’t tell	Can’t tell	Yes	No	No	No	Can’t tell	3
McKinney et al. 1987	Yes	Yes	Can't tell	Can't tell	Can't tell	No	No	No	No	2
McKinney et al. 1999	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	8
Neglia et al. 2000	Yes	Yes	Yes	Yes	Can't tell	No	No	No	No	4
Nishi et al. 1989	No	No	Yes	No	No	No	No	No	No	1
Paltiel et al. 2006	Yes	Yes	Can't tell	Yes	Can't tell	Can't tell	No	No	No	3
Perrillat et al. 2002	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	7
Petridou et al. 2001	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	6
Roman et al. 2007	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	8
Rosenbaum et al. 2005	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	7
Rudant et al. 2010	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	7
Rudant et al. 2015	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	7
Salonen et al. 2002	No	Yes	Can’t tell	Can’t tell	Yes	No	No	No	Can’t tell	2

Schlehofer et al. 1996	Yes	Yes	Yes	No	No	No	No	No	No	3
Schuz et al. 1999	Yes	Yes	Yes	No	No	No	Yes	No	Yes	5
Surico and Muggeo 2005	Yes	Yes	Can't tell	Can't tell	No	No	No	No	Can't tell	2
Tesse et al. 2009	Yes	Yes	Can't tell	Can't tell	Yes	No	No	No	Can't tell	3
Till et al. 1979	No	No	Yes	No	No	No	No	Yes	Can't tell	2
van Steensel-Moll et al. 1986	Yes	Yes	Yes	Yes	No	No	Can't tell	No	Yes	5
Zaki and Ashray 2010	No	No	Yes	No	Yes	No	Can't tell	No	Can't tell	2
Zaki et al. 2006	Yes	Yes	Can't tell	No	Yes	No	Can't tell	No	Can't tell	3

Supplementary Table 3.3b Risk of Bias Assessment Using the CASP Tool of the Included Cohort Studies

Study	**Did the study address a clearly focused issue?**	**Were the cohort recruited in an acceptable way?**	**Was the exposure accurately measured to minimize bias?**	**Was the outcome accurately measured to minimize bias?**	**Have the authors identified all important confounding factors?**	**Have authors taken account of the confounding factors in design and/or analysis?**	**Was the follow up of subjects complete enough?**	**Was the follow up of subjects long enough?**	**Do you believe the results?**	**Can the results be applied to the local population?**	**Do the results of the study fit with other available evidence?**	**TOTAL**
Lin et al. 2015	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	9
Vestergaard et al. 2013	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	10

Supplementary Table 3.4 Meta-regression analysis to explore the heterogeneity among the included studies

Study Characteristic		Model 1 Univariate: Overall effect			Model 2 with Data Source: Overall effect		
		OR [95% CI]	I^2 (%)	R^2 (%)	OR [95% CI]	I^2 (%)	R^2 (%)
Risk of Bias	Low-risk	Ref	84	14	Ref	77	35
	Moderate-risk	1.15 (0.71-1.87)			1.01 (0.62-1.65)		
	High-risk	**1.94 (1.29-2.95)**			1.48 (0.92-2.39)		
Region	North America	Ref	77	32	Ref	71	47
	Europe	1.03 (0.68-1.56)			0.90 (0.60-1.35)		
	Asia	0.81 (0.41-1.59)			0.78 (0.41-1.49)		
	Other	**2.38 (1.41-3.98)**			1.63 (0.93-2.83)		
Publication Era	≥2010	Ref	87	0	Ref	76	41
	2000-2009	0.84 (0.53-1.33)			0.99 (0.68-1.44)		
	≤1999	1.03 (0.59-1.84)			1.49 (0.93-2.38)		
Source of Controls	General population	Ref	86	10	Ref	82	28
	General practitioner list	1.34 (0.75-2.40)			1.11 (0.63-1.97)		
	Hospital control	**1.68 (1.21-2.52)**			1.07 (0.66-1.75)		
Data Source	Self-reported	Ref	80	35			
	Administrative/medical records data	1.09 (0.72-1.65)					
	Laboratory investigation	**2.37 (1.55-3.62)**					

The univariate models included only 1 covariate as indicated, and model 2 included the indicated covariate and data source. Laboratory investigation remained an important factor in all bivariate models (model 2).

Appendices B Supplementary Information for Objective 2

Supplementary Table 4.1 Patient characteristics of those excluded from the analysis due to misalignment of the visit date on the electronic medical record and the billing date in Ontario Health Insurance Plan

Characteristic	EMRALD patients, n (%)	Standardized difference or p-value for comparison to cohort of included patients
Number of patients	264	
Female	118 (44.7)	p=0.23
Age, average (SD)	7.1 (5.3)	p=0.08
0 to < 2	45 (17.0)	**0.19**
2 to 5	75 (28.4)	**0.11**
6 to 9	48 (18.2)	0.01
10 to 14	65 (24.6)	0.04
15 to 18	31 (11.7)	0.03
Rural residence	73 (27.7)	**p<0.01**
Residential instability		
1 least	48 (18.2)	0.06
2	58 (22.0)	0.00
3	52 (19.7)	0.04
4	53 (20.1)	0.01
5 most	44 (16.7)	0.07
Material deprivation		
1 least	69 (26.1)	0.06
2	52 (19.7)	0.04
3	44 (16.7)	0.06
4	53 (20.1)	0.08
5 most	37 (14.0)	0.01
Dependency		
1 least	79 (29.9)	0.03
2	60 (22.7)	0.07
3	42 (15.9)	0.05
4	37 (14.0)	0.02
5 most	37 (14.0)	0.03
Ethnic concentration		
1 least	51 (19.3)	0.08
2	50 (18.9)	0.04
3	54 (20.5)	0.03
4	60 (22.7)	0.08
5 most	40 (15.2)	0.01
Chronic conditions or illnesses*		
Complex Chronic Conditions	7 (2.7)	p=0.43
Allergies	≤5	p=0.73
Asthma or reactive airways	14 (7.0)	p=0.16

Behavioral and emotional disorders with onset usually occurring in childhood and adolescence	10 (5.0)	p=0.52
Mood disorders	≤5	p=0.72
Pervasive and specific developmental disorders	≤5	p=0.72

There are 9 missing individuals in the residential instability, material deprivation, dependency, and ethnic concentration variables. Standardized difference >0.10 indicates an imbalance in the prevalence of the covariate between the included and excluded patients. A p-value >0.05 in the χ^2 test indicates a difference between included and excluded patients. One-way ANOVA test was used for mean age comparison. Some cells (≤5) suppressed because of small cell size (direct or by inference), which cannot be reported as per privacy regulations.

Supplementary Table 4.2 Performance measures of the Ontario Health Insurance Plan physician billing claims for identifying infectious syndromes compared to electronic medical records, by age group, sex, rural and urban residence, presence of asthma or reactive airways, and presence of chronic complex conditions

	Classification of infection	% infection in EMR	% infection in AD	Sensitivity [95% CI]	Specificity [95% CI]	PPV [95% CI]	NPV [95% CI]
Age 0-2, n=546	Any infection	22.5	20.1	76 (67-83)	96 (94-98)	85 (76-91)	93 (90-95)
	Respiratory infection	19.0	17.0	78 (69-85)	97 (95-99)	87 (79-93)	95 (92-97)
	Skin and soft tissue infection	2.0	1.1	27 (6-61)	99 (98-100)	50 (12-88)	99 (97-99)
	Gastrointestinal infection	2.0	1.6	64 (31-89)	100 (99-100)	78 (40-97)	99 (98-100)
	Urinary tract infections	0.0	0.0				
	Otitis externa (ear) infection	0.0	≤1.0*				
Age 2-5, n=519	Any infection	44.9	39.5	78 (72-83)	92 (88-95)	88 (83-92)	83 (79-87)
	Respiratory infection	33.7	31.0	78 (71-84)	93 (89-95)	84 (78-90)	89 (85-92)
	Skin and soft tissue infection	7.7	4.8	55 (38-71)	99 (98-100)	88 (69-97)	96 (94-98)
	Gastrointestinal infection	2.7	1.9	57 (29-82)	100 (99-100)	80 (44-97)	99 (97-100)
	Urinary tract infections	1.9	1.3	50 (19-81)	100 (99-100)	71 (29-96)	99 (99-100)
	Otitis externa (ear) infection	≤1.0*	≤1.0*				
Age 6-9, n=390	Any infection	41.3	36.4	78 (70-84)	93 (88-96)	88 (82-93)	85 (80-90)
	Respiratory infection	24.1	23.3	79 (69-86)	94 (91-97)	81 (72-89)	93 (90-96)
	Skin and soft tissue infection	13.1	9.5	67 (52-79)	99 (97-100)	92 (78-98)	95 (92-97)

	Gastrointestinal infection	2.6	≤1.4*				
	Urinary tract infections	2.8	2.1	55 (23-83)	99 (98-100)	75 (35-97)	99 (97-100)
	Otitis externa (ear) infection	1.5	≤1.4*				
Age 10-14, n=497	Any infection	31.0	24.1	66 (58-74)	95 (92-97)	85 (77-91)	86 (82-90)
	Respiratory infection	17.9	16.5	76 (66-85)	97 (94-98)	83 (73-90)	95 (92-97)
	Skin and soft tissue infection	12.5	5.2	35 (24-49)	99 (98-100)	85 (65-96)	92 (89-94)
	Gastrointestinal infection	≤1.0*	≤1.0*				
	Urinary tract infections	≤1.0*	≤1.0*				
	Otitis externa (ear) infection	1.6	≤1.0*				
Age 15+, n=233	Any infection	24.5	15.9	61 (48-74)	99 (96-100)	95 (82-99)	89 (84-93)
	Respiratory infection	12.9	9.0	70 (51-85)	100 (98-100)	100 (84-100)	96 (92-98)
	Skin and soft tissue infection	7.7	4.3	44 (22-69)	99 (97-100)	80 (44-97)	96 (92-98)
	Gastrointestinal infection	≤2.15*	≤2.15*				
	Urinary tract infections	≤2.15*	≤2.15*				
	Otitis externa (ear) infection	≤2.15*	≤2.15*				
Female, n=1066	Any infection	33.5	26.9	71 (66-76)	95 (94-97)	89 (84-92)	87 (84-89)
	Respiratory infection	21.7	19.3	77 (71-82)	97 (95-98)	86 (81-91)	94 (92-95)
	Skin and soft tissue infection	8.3	4.1	42 (31-53)	99 (99-100)	84 (70-93)	95 (93-96)
	Gastrointestinal infection	2.7	1.4	48 (29-67)	100 (99-100)	93 (68-100)	99 (98-99)

	Urinary tract infections	2.1	1.3	45 (24-68)	100 (99-100)	71 (42-92)	99 (98-100)
	Otitis externa (ear) infection	1.0	0.8	45 (17-77)	100 (99-100)	63 (24-91)	99 (99-100)
Male, n=1119	Any infection	33.2	29.2	76 (71-80)	94 (92-96)	86 (82-90)	89 (86-91)
	Respiratory infection	23.3	21.6	77 (72-82)	95 (94-97)	83 (78-88)	93 (91-95)
	Skin and soft tissue infection	8.3	5.4	56 (45-66)	99 (98-100)	87 (75-94)	96 (95-97)
	Gastrointestinal infection	1.3	1.2	64 (35-87)	100 (99-100)	69 (39-91)	100 (99-100)
	Urinary tract infections	≤0.5*	0.7				
	Otitis externa (ear) infection	0.7	≤0.5*				
Rural, n=416	Any infection	41.1	34.4	75 (68-81)	94 (90-97)	90 (83-94)	84 (79-88)
	Respiratory infection	27.4	24.0	75 (66-82)	95 (92-97)	85 (76-91)	91 (87-94)
	Skin and soft tissue infection	11.5	7.9	60 (45-74)	99 (97-100)	88 (72-97)	95 (92-97)
	Gastrointestinal infection	2.4	1.4	60 (28-88)	100 (99-100)	100 (54-100)	99 (98-100)
	Urinary tract infections	≤1.2*	≤1.2*				
	Otitis externa (ear) infection	≤1.2*	≤1.2*				
Urban, n=1767	Any infection	31.5	26.7	73 (69-77)	95 (93-96)	87 (83-90)	89 (87-90)
	Respiratory infection	21.4	19.7	78 (74-82)	96 (95-97)	85 (81-88)	94 (93-95)
	Skin and soft tissue infection	7.6	4.0	45 (36-54)	99 (99-100)	85 (74-92)	96 (95-97)
	Gastrointestinal infection	1.9	1.2	52 (34-69)	100 (99-100)	77 (55-92)	99 (99-99)
	Urinary tract infections	1.4	1.1	50 (29-71)	100 (99-100)	60 (36-81)	99 (99-100)

	Otitis externa (ear) infection	1.0	0.6	41 (18-67)	100 (100-100)	70 (35-93)	99 (99-100)
Asthma or reactive airways, n=210	Any infection	34.2	31.2	74 (62-84)	91 (86-96)	81 (69-90)	88 (81-93)
	Respiratory infection	22.4	21.9	74 (60-86)	93 (88-97)	76 (61-87)	93 (88-96)
	Skin and soft tissue infection	9.0	4.8	47 (24-71)	99 (97-100)	90 (56-100)	95 (91-98)
	Gastrointestinal infection	≤2.4*	≤2.4*				
	Urinary tract infections	≤2.4*	≤2.4*				
	Otitis externa (ear) infection	≤2.4*	≤2.4*				
No asthma or reactive airways, n=1975	Any infection	33.4	27.9	74 (70-77)	95 (94-96)	88 (85-91)	88 (86-89)
	Respiratory infection	22.5	20.4	78 (73-81)	96 (95-97)	86 (82-89)	94 (92-95)
	Skin and soft tissue infection	8.3	4.8	49 (41-57)	99 (99-100)	85 (76-92)	96 (95-96)
	Gastrointestinal infection	2.1	1.3	54 (37-69)	100 (99-100)	85 (65-96)	99 (98-99)
	Urinary tract infections	1.3	1.0	50 (30-70)	100 (99-100)	65 (41-85)	99 (99-100)
	Otitis externa (ear) infection	0.8	0.5	38 (15-65)	100 (100-100)	67 (30-93)	99 (99-100)
Complex Chronic Conditions, n=78	Any infection	24.4	21.8	79 (54-94)	97 (88-100)	88 (64-99)	93 (84-98)
	Respiratory infection	20.5	17.9	75 (48-93)	97 (89-100)	86 (57-98)	94 (85-98)
	Skin and soft tissue infection	≤6.4*	≤6.4*				
	Gastrointestinal infection	≤6.4*	≤6.4*				
	Urinary tract infections	≤6.4*	≤6.4*				
	Otitis externa (ear) infection	0.0	0.0				

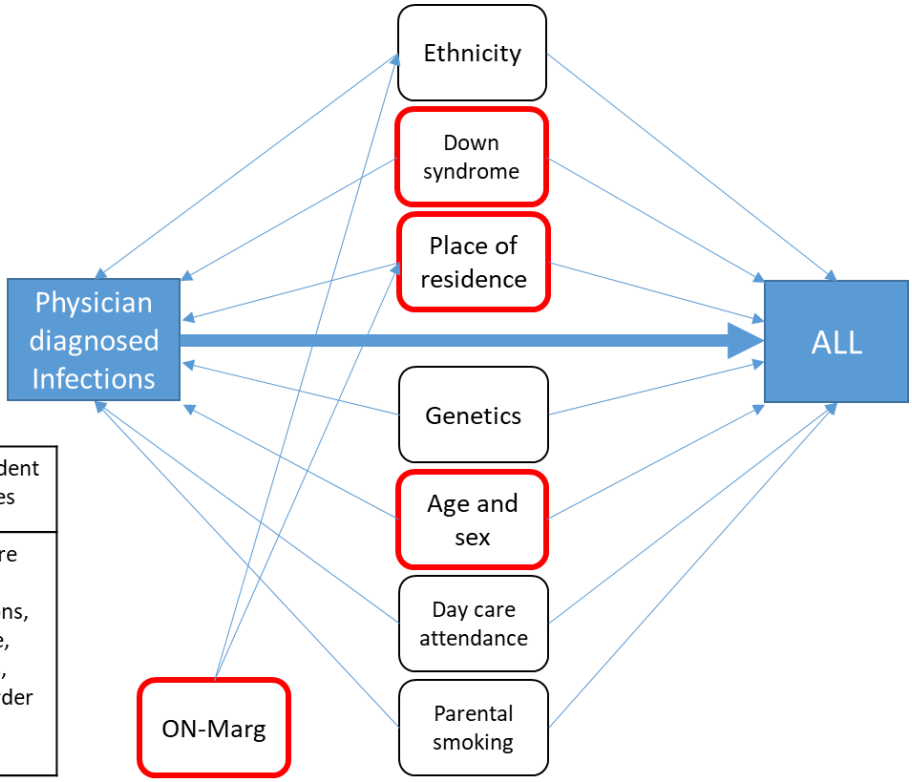
*Cells suppressed because of small cell size (direct or by inference), which cannot be reported as per privacy regulations, and performance characteristics have deliberately not been reported due to the potential to back-calculate the small cell sizes. Cells with ≤5 persons have been suppressed. EMR=electronic medical records, AD=administrative data, PPV=positive predictive value, NPV=negative predictive value.

Appendices C Supplementary Information for Objective 3

Supplementary Table 5.1 Included syndromes in the definitions of the types of infections

Type of infection	Included syndromes
Respiratory Infection	Acute bronchitis Acute conjunctivitis Acute laryngitis Acute mastoiditis Acute nasopharyngitis / common cold / upper respiratory infection Acute sinusitis Acute tonsillitis Infectious mononucleosis Influenza Otitis media Pertussis Pneumonia Streptococcal sore throat
Gastrointestinal infection	Diarrhea, gastro-enteritis, viral gastro-enteritis Food poisoning Pinworm infestation
Otitis externa infection	Otitis externa
Skin and soft tissue infection	Blepharitis, chalazion, stye Chickenpox Candidiasis, thrush Cellulitis Dental caries Herpes simplex, cold sore Impetigo Other mycoses Pilonidal cyst Pyoderma, pyogenic granuloma, other local infections Warts
Urinary tract infections	Cystitis Other disorders of urinary tract
Invasive infections	Bacterial meningitis Encephalitis, encephalomyelitis Meningitis due to enterovirus Meningitis due to other organisms Meningococcal infection or meningitis Other viral diseases of central nervous system Septicemia, blood poisoning

Supplementary Figure 5.1 Causal Diagram of Prior Infections and Childhood ALL



Other control variables	Mediators	Antecedent variables
Birth weight, access to health services, social interaction, maternal age, ionizing radiation, air pollution	Immune modulation, asthma, allergy, atopic disease	Exposure to infections, hygiene, siblings, birth order

The diagram represents confounders to the relationship, and the Ontario Marginalization Index as an antecedent variable to confounders ethnicity and place of residence. The table of other control variables, mediators, and antecedent variables were not considered for the modeling of the relationship between physician diagnosed infections and childhood acute lymphoblastic leukemia (ALL). Red boxes indicate confounders that were available in the data.

Supplementary Table 5.2 Subgroup analyses of the association between rate of infections and ALL in children aged 2-14 years from Ontario, Canada between 1993-2014, among non-immigrants and those without down syndrome

	Adjusted model estimates	
Parameters	**OR**	**95% CI**
N=15,180		
Rate of any infection		
≤0.25 infection per year	Ref	
>0.25 to 0.50 infection per year	**1.44**	**(1.02-2.04)**
>0.50 to 1 infection per year	**1.40**	**(1.04-1.87)**
>1 to 2 infections per year	**1.31**	**(1.00-1.73)**
>2 infections per year	**1.51**	**(1.16-1.97)**
Dependency		
1: Least marginalized	Ref	
2	0.92	(0.79-1.09)
3	0.95	(0.80-1.14)
4	0.93	(0.77-1.13)
5: Most marginalized	0.96	(0.78-1.19)
Missing	0.81	(0.26-2.53)
Material deprivation		
1: Least marginalized	Ref	
2	1.13	(0.95-1.35)
3	0.97	(0.80-1.17)
4	0.93	(0.77-1.14)
5: Most marginalized	*	
Missing	0.85	(0.38-1.89)
Ethnic concentration		
1: Least marginalized	Ref	
2	1.16	(0.95-1.43)
3	1.16	(0.94-1.44)
4	0.98	(0.78-1.23)
5: Most marginalized	*	
Missing	1.21	(0.55-2.63)
Residential instability		
1: Least marginalized	Ref	
2	1.04	(0.88-1.25)
3	1.02	(0.84-1.23)
4	1.03	(0.85-1.25)
5: Most marginalized	*	
Missing	0.95	(0.43-2.11)

ALL represents acute lymphoblastic leukemia. There were 1,380 cases and 13,800 controls. These are adjusted conditional logistic regression models of complete matched sets of cases and controls were matched on date of birth, sex, rural residence at start of observation, and covariates dependency, material deprivation, ethnic concentration, and residential instability. OR represents odds ratio. CI represents confidence interval.

Supplementary Table 5.3 Sensitivity analyses of the association between rate of infections and ALL children aged 2-14 years from Ontario, Canada between 1993-2014, restricted to visits to primary care physician offices

Physician diagnosed infections in primary care settings	**Cases**		**Controls**		**Crude model estimates**		**Adjusted model estimates**	
	n	**%**	**n**	**%**	**OR**	**95% CI**	**OR**	**95% CI**
N	1,600		16,000					
Rate of any infection								
≤0.25 infection per year	103	6.4	1,329	8.3	Ref		Ref	
>0.25 to 0.50 infection per year	93	5.8	844	5.3	**1.46**	**(1.08-1.97)**	**1.45**	**(1.06-1.97)**
>0.50 to 1 infection per year	202	12.6	2,096	13.1	1.28	(0.99-1.65)	**1.30**	**(1.00-1.70)**
>1 to 2 infections per year	396	24.8	4,248	26.6	1.24	(0.98-1.57)	**1.28**	**(1.00-1.63)**
>2 infections per year	806	50.4	7,483	46.8	**1.45**	**(1.16-1.82)**	**1.44**	**(1.14-1.82)**

ALL represents acute lymphoblastic leukemia. These are adjusted conditional logistic regression models of complete matched sets of cases and controls were matched on date of birth, sex, rural residence at start of observation, and includes confounders immigrant status and down syndrome, and covariates dependency, material deprivation, ethnic concentration, and residential instability. OR represents odds ratio. CI represents confidence interval.