RESEARCH ARTICLE

# Use of physician billing claims to identify infections in children

**Jeremiah Hwee**[1,2]*, **Lillian Sung**[3,4,5], **Jeffrey C. Kwong**[1,2,6,7,8], **Rinku Sutradhar**[1,2,5], **Karen Tu**[6,8], **Jason D. Pole**[1,2,9]

**1** Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada, **2** Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada, **3** Department of Pediatrics, Division of Haematology/Oncology, The Hospital for Sick Children, Toronto, Ontario, Canada, **4** Program in Child Health Evaluative Sciences, The Hospital for Sick Children, Peter Gilgan Centre for Research and Learning, Toronto, Ontario, Canada, **5** Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada, **6** Department of Family and Community Medicine, University of Toronto, Toronto, Ontario, Canada, **7** Public Health Ontario, Toronto, Ontario, Canada, **8** Toronto Western Family Health Team, University Health Network, Toronto, Ontario, Canada, **9** Pediatric Oncology Group of Ontario, Toronto, Ontario, Canada

* jeremiah.hwee@mail.utoronto.ca

## Abstract

While medical records have detailed information, they are limited in reach to the availability and accessibility of those records. On the other hand, administrative data while limited in scope, have a much further reach in coverage of an entire population. However, few studies have validated the use of administrative data for identifying infections in pediatric populations. Pediatric patients from Ontario, Canada 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. 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. We found when using linked EMR data as the reference standard, administrative billing codes are reasonably accurate in identifying infections in a pediatric population.

OPEN ACCESS

**Citation:** Hwee J, Sung L, Kwong JC, Sutradhar R, Tu K, Pole JD (2018) Use of physician billing claims to identify infections in children. PLoS ONE 13(11): e0207468. https://doi.org/10.1371/journal.pone.0207468

**Editor:** Dafna Yahav, Rabin Medical Center, Beilinson Hospital, ISRAEL

**Received:** August 13, 2018

**Accepted:** October 31, 2018

**Published:** November 12, 2018

**Data Availability Statement:** Data cannot be made publicly available due to the ethical and legal requirements implemented by the Institute for Clinical Evaluative Sciences (ICES) and Ontario's Information and Privacy Commissioner (IPC). The data sets used in this study are held securely in coded form at ICES; data sharing agreements with the various data providers prohibit making these data publicly available due to privacy concerns and risk of re-identification. Access may be granted to those who meet pre-specified criteria for confidential access by submitting an ICES DAS Public or Private Sector Request Form, available at www.ices.on.ca/DAS or by e-mailing das@ices.on.ca.

**Funding:** Funding was provided by the Canadian Institutes of Health Research - Institute of Cancer Research Frederick Banting and Charles Best Canada Graduate Scholarship to Honour Nelson Mandela Doctoral award (JH), the Department of Family and Community Medicine, University of Toronto (JCK and KT), and the Canadian Institutes of Health Research New Investigator Award (JCK). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This study was supported by the Institute for Clinical Evaluative Sciences (ICES), which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred.

**Competing interests:** The authors have declared that no competing interests exist.

## 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.[1–4] 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 administrative data for identifying infections in pediatric populations.

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.[5] 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 receipt of immunizations.[6–8] 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.

## Methods

The study was approved by University of Toronto's Health Sciences Research Ethics Board and Sunnybrook Health Sciences Centre's Research Ethics Board. The Institute of Clinical Evaluative Sciences (ICES) is named as a prescribed entity under provincial 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 conducted by ICES, independently or on behalf of policy-makers or other stakeholders.

### 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. The globally unique identifier approach was used to generate a random sample using Microsoft SQL Server Management Studio 2012 (Microsoft Corporation). We limited patients to only one visit 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.[9] 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 (CI) of 80%.[10]

### 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.[11] 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 (Practice Solutions)

Suite EMR. EMRALD contains clinical information such as a cumulative patient profile, progress notes, laboratory results, and prescriptions. Physicians participating in EMRALD 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), and quintile of neighbourhood material deprivation from the Ontario Marginalization Index, with 1 being the least deprived and 5 being the most deprived.[12, 13] 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 (ICD) 8 and 9.[14] 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.

## 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 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.[15] 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 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,[16] and other chronic conditions from the cumulative patient profile. Since the abstractor did not have clinical experience, 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 to verify their validity. If multiple diagnoses were made, both were kept and compared to the corresponding billing code.

## Analysis

Duplicate abstraction of a random sample of 200 patient visits was performed to assess intrarater 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.[17] We compared the demographic characteristics of the included and excluded patients using standardized differences and $\chi^2$ test for categorical variables, and one-way ANOVA test for mean age.[18] A standardized difference >0.10 indicates a potential imbalance in the prevalence of a variable between included and excluded patients. Diagnoses of

**Table 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 |

https://doi.org/10.1371/journal.pone.0207468.t001

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. These measures are recommended for studies describing the diagnostic performance of administrative data for identifying diseases.[19] A binomial distribution was used for the performance measures to calculate 95% 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. All datasets were linked using unique, encoded identifiers, and were analyzed at ICES.

## 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 study cohort are summarized in Table 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 (S1 Table). There were 2,185 unique OHIP billing

**Table 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.

https://doi.org/10.1371/journal.pone.0207468.t002

**Table 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.

https://doi.org/10.1371/journal.pone.0207468.t003

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 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 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). 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 (S2 Table).

## 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

Table 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 | <0.03 | 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 | 0.04 | 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. Cells suppressed because of small cell size (direct or by inference).

https://doi.org/10.1371/journal.pone.0207468.t004

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.[20–26] 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 (45% to 99%), moderate-to-high specificity (69% to 100%), poor-to-high PPV (55% to 100%), and had to trade-off higher sensitivity for lower specificity or vice versa by expanding the number of ICD diagnosis codes, the number of data fields, or the diagnosis types. Our estimates for any infection, respiratory infection, and specific infectious syndromes such as otitis media and conjunctivitis performed well compared to these studies.

We found infections accounted for 33.3% of all visits to a physician, respiratory infections accounted for 67.6% of those infections. Followed by skin and soft tissue infections represented a quarter of the visits for an infection, gastrointestinal infections represented 6.0%, urinary tract infections represented 3.9%, and otitis externa represented 2.7%. Infections continue to represent one of the most frequent reasons to seek healthcare in children and adolescents aged <18 years.[1–4, 27]

This study utilizes a population-based primary care cohort and is the largest to date that gives evidence on the diagnostic performance of administrative data in identifying infections in children within a primary care setting. The contributions of the study are important for a field with limited evidence and demonstrates the validity of administrative data in identifying infections in children for clinicians, researchers, and decision makers. This will allow for future studies in this area to examine larger populations and changes over time. However, our study had several limitations. First, only one abstractor without clinical experience was used and this could have implications on the validity of the study. 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).[28] Another limitation is that there were differences between those included and excluded in this study and this may have implications for generalizability of the study.

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 one of the primary care reform models; therefore the results of this study might not be generalizable to other physicians. EMRALD physicians were found 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.[29] 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.[30] 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.[31] 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 generalizable to Ontario specialists or family physicians not participating in EMRALD, or to other jurisdictions where physician billing practices or disease classification systems may differ. An important limitation is that this study was conducted in Canada and our results may not be generalizable to other countries. However, they are more likely to be applicable to countries with similar healthcare systems and more specifically, studies conducted in the United States and New Zealand showed consistent findings in the performance of the administrative data to identify infections.[21–26]

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.[32] 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.[33] 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.[27, 33] 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.

## Supporting information

**S1 File. STROBE Statement—Checklist of items that should be included in reports of observational studies.** *Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.
(DOC)

**S1 Table. 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.** 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 $\chi^2$ 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.
(DOCX)

**S2 Table. 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.** *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.
(DOCX)

## Acknowledgments

We would like to thank Shirin Jabbari and Atul Sivaswamy for preparing the data.

## Author Contributions

**Conceptualization:** Jeremiah Hwee, Lillian Sung, Jeffrey C. Kwong, Karen Tu, Jason D. Pole.

**Data curation:** Jeremiah Hwee.

**Formal analysis:** Jeremiah Hwee.

**Funding acquisition:** Jeremiah Hwee.

**Investigation:** Jeremiah Hwee, Lillian Sung, Jeffrey C. Kwong, Karen Tu, Jason D. Pole.

**Methodology:** Jeremiah Hwee, Lillian Sung, Jeffrey C. Kwong, Rinku Sutradhar, Karen Tu, Jason D. Pole.

**Project administration:** Jeremiah Hwee.

**Resources:** Jeremiah Hwee, Karen Tu, Jason D. Pole.

**Supervision:** Lillian Sung, Jason D. Pole.

**Validation:** Jeremiah Hwee, Lillian Sung, Jeffrey C. Kwong, Jason D. Pole.

**Visualization:** Jeremiah Hwee.

**Writing – original draft:** Jeremiah Hwee.

**Writing – review & editing:** Lillian Sung, Jeffrey C. Kwong, Rinku Sutradhar, Karen Tu, Jason D. Pole.

## References

1. Niska R, Bhuiya F, Xu J. National Hospital Ambulatory Medical Care Survey: 2007 emergency department summary. National health statistics reports. 2010;( 26):1–31. Epub 2010/08/24. PMID: 20726217.
2. Monto AS, Sullivan KM. Acute respiratory illness in the community. Frequency of illness and the agents involved. Epidemiology and infection. 1993; 110(1):145–60. Epub 1993/02/01. PMID: 8432318; PubMed Central PMCID: PMCPMC2271959.
3. Toivonen L, Karppinen S, Schuez-Havupalo L, Teros-Jaakkola T, Vuononvirta J, Mertsola J, et al. Burden of Recurrent Respiratory Tract Infections in Children: A Prospective Cohort Study. The Pediatric infectious disease journal. 2016; 35(12):e362–e9. Epub 2016/07/28. https://doi.org/10.1097/INF.0000000000001304 PMID: 27455443.
4. CIHI. Canadian Institute for Health Information. A Snapshot of Health Care in Canada as Demonstrated by Top 10 Lists, 2011. Ottawa, Canada: 2012.
5. Population by sex and age group, by province and territory (Number, both sexes). July 1, 2016 [Internet]. 2016 [cited July 6, 2017]. Available from: http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/demo31a-eng.htm.
6. Guttmann A, Nakhla M, Henderson M, To T, Daneman D, Cauch-Dudek K, et al. Validation of a health administrative data algorithm for assessing the epidemiology of diabetes in Canadian children. Pediatric diabetes. 2010; 11(2):122–8. Epub 2009/06/09. https://doi.org/10.1111/j.1399-5448.2009.00539.x PMID: 19500278.
7. 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):183–8. PMC2807792.
8. Schwartz KL, Tu K, Wing L, Campitelli MA, Crowcroft NS, Deeks SL, et al. Validation of infant immunization billing codes in administrative data. Human Vaccines & Immunotherapeutics. 2015; 11(7):1840–7. https://doi.org/10.1080/21645515.2015.1043499 PMC4514409. PMID: 26075651
9. Thomas EM. Recent trends in upper respiratory infections, ear infections and asthma among young Canadian children. Health reports. 2010; 21(4):47–52. Epub 2011/01/29. PMID: 21269011.
10. 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):859–62. Epub 2005/07/16. https://doi.org/10.1016/j.jclinepi.2004.12.009 PMID: 16018921.

11. Tu K, Wang M, Young J, Green D, Ivers NM, Butt D, 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):1388–94. Epub 2013/10/01. https://doi.org/10.1016/j.cjca.2013.07.676 PMID: 24075778.

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

13. 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):S12–6. Epub 2012/01/01. PMID: 23618065.

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

15. 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. https://doi.org/10.1038/bjc.2017.360 https://www.nature.com/articles/bjc2017360#supplementary-information. PMID: 29065105

16. 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):e1463–70. Epub 2012/11/28. https://doi.org/10.1542/peds.2012-0175 PMID: 23184117; PubMed Central PMCID: PMCPMC4528341.

17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159–74. Epub 1977/03/01. PMID: 843571.

18. 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):1228–34. https://doi.org/10.1080/03610910902859574

19. 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):821–9. Epub 2011/01/05. https://doi.org/10.1016/j.jclinepi.2010.10.006 PMID: 21194889.

20. Pisesky A, Benchimol EI, Wong CA, Hui C, Crowe M, Belair MA, 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. Epub 2016/03/10. https://doi.org/10.1371/journal.pone.0150416 PMID: 26958849; PubMed Central PMCID: PMCPMC4784925.

21. Hsu VP, Staat MA, Roberts N, Thieman C, Bernstein DI, Bresee J, 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. Epub 2005/01/05. https://doi.org/10.1542/peds.2004-0860 PMID: 15629984.

22. Williams DJ, Shah SS, Myers A, Hall M, Auger K, Queen MA, et al. Identifying pediatric community-acquired pneumonia hospitalizations: Accuracy of administrative billing codes. JAMA pediatrics. 2013; 167(9):851–8. Epub 2013/07/31. https://doi.org/10.1001/jamapediatrics.2013.186 PMID: 23896966; PubMed Central PMCID: PMCPMC3907952.

23. O'Sullivan CE, Baker MG. Proposed epidemiological case definition for serious skin infection in children. Journal of paediatrics and child health. 2010; 46(4):176–83. Epub 2010/01/29. https://doi.org/10.1111/j.1440-1754.2009.01658.x PMID: 20105249.

24. 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):355–60. Epub 2004/06/05. PMID: 15179142.

25. Shaklee J, Zerr DM, Elward A, Newland J, Leckerman K, Asti L, 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. Epub 2010/11/17. https://doi.org/10.1097/INF.0b013e3182027c22 PMID: 21079527.

26. Tieder JS, Hall M, Auger KA, Hain PD, Jerardi KE, Myers AL, et al. Accuracy of administrative billing codes to detect urinary tract infection hospitalizations. Pediatrics. 2011; 128(2):323–30. Epub 2011/07/20. https://doi.org/10.1542/peds.2010-2064 PMID: 21768320; PubMed Central PMCID: PMCPMC3146355.

27. Villers MS, Ramsey AF, Mitchell DK, Holmes SJ, McCraw AS, Hannon PM, et al. Utilization of Health Care for Infectious Illnesses at a Pediatric Practice 549. Pediatric Research. 1998; 43:96. https://doi.org/10.1203/00006450-199804001-00570

28. 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.e1. Epub 2006/02/24. https://doi.org/10.1016/j.annemergmed.2005.11.022 PMID: 16492494.

29. 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–E6. https://doi.org/10.9778/cmajo.20120039 PMC3985899. PMID: 25077111

30. 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.

31. Tu K, Widdifield J, Young J, Oud W, Ivers NM, Butt DA, 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. Epub 2015/08/14. https://doi.org/10.1186/s12911-015-0195-x PMID: 26268511; PubMed Central PMCID: PMCPMC4535372.

32. Lazarus R, Klompas M, Campion FX, McNabb SJN, Hou X, Daniel J, 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. https://doi.org/10.1197/jamia.M2848 PMID: 18952940

33. 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):681–5. https://doi.org/10.1097/INF.0000000000000704 00006454-201507000-00001. PMID: 25860534