Vaccine: X 15 (2023) 100408

Contents lists available at ScienceDirect

Vaccine: X

journal homepage: www.elsevier.com/locate/jvacx

Validating pertussis data measures using electronic medical record data in Ontario, Canada 1986–2016

Shilo H. McBurney [a,b,*], Jeffrey C. Kwong [a,c,d,e,f], Kevin A. Brown [a,c,e], Frank Rudzicz [g,h,i], Branson Chen [e], Elisa Candido [e], Natasha S. Crowcroft [a,j]

[a] Dalla Lana School of Public Health, University of Toronto, 155 College Street, 6th Floor, Toronto, ON M5T 3M7, Canada
[b] Department of Epidemiology, Brown University School of Public Health, 121 South Main Street, Box G-S121-2, Providence, RI 02912, United States of America
[c] Public Health Ontario, 661 University Avenue, Suite 1701, Toronto, ON M5G 1M1, Canada
[d] Department of Laboratory Medicine and Pathobiology, University of Toronto, 1 King's College Circle, 6th Floor, Toronto, ON M5S 1A8, Canada
[e] ICES, G1 06, 2075 Bayview Avenue, Toronto, ON M4N 3M5, Canada
[f] Department of Family and Community Medicine, University of Toronto, 500 University Avenue, 5th Floor, Toronto, ON M5G 1V7, Canada
[g] Department of Computer Science, University of Toronto, 40 St. George Street, Room 4283, Toronto, ON M5S 2E4, Canada
[h] Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada
[i] Vector Institute for Artificial Intelligence, 661 University Ave Suite 710, Toronto, ON M5G 1M1, Canada
[j] Immunization, Vaccines and Biologicals, World Health Organization, Avenue Appia 20, 1211 Geneva 27, Switzerland

ARTICLE INFO

Keywords:
Pertussis
Vaccine-preventable diseases
Data validation
Diagnostic accuracy
Electronic medical records
Immunization

ABSTRACT

Background: Pertussis is a reportable disease in many countries, but ascertainment bias has limited data accuracy. This study aims to validate pertussis data measures using a reference standard that incorporates different suspected case severities, allowing for the impact of case severity on accuracy and detection to be explored.
Methods: We evaluated 25 pertussis detection algorithms in a primary care electronic medical record database between January 1, 1986 and December 30, 2016. We estimated sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). We used sensitivity analyses to explore areas of uncertainty and evaluated reasons for lack of detection.
Results: The algorithm including all data measures achieved the highest sensitivity at 20.6%. Sensitivity increased to 100% after reclassifying symptom-only cases as non-cases, but the PPV remained low. Age at first episode was significantly associated with detection in half of the tested scenarios, and false negatives often had some history of immunization.
Conclusions: Sensitivity improved by reclassifying symptom-only cases but remained low unless multiple data sources were used. Results demonstrate a trade-off between PPV and sensitivity. EMRs can enhance detection through patient history and clinical note data. It is essential to improve case identification of older individuals with vaccination history to reduce ascertainment bias.

Introduction

Canada continues to experience pertussis outbreaks, with case underestimation impeding understanding of transmission despite being a reportable disease [1–6]. Underestimation may be attributed to the failure to consider pertussis diagnostically, atypical presentations, infrequent diagnostic testing, suboptimal test accuracy, and reporting issues [1,7]. Case severity is related to age, with younger, severe cases more likely to be detected, tested, and to have a positive result, producing testing bias [5,7]. Mild cases are commonly missed in older individuals, with up to one third presenting without the classic paroxysmal cough [1,8]. Many of these cases may be due to waning acellular vaccine-induced immunity and could contribute to community transmission [8].

These issues have hampered surveillance data accuracy, leading to the exploration of other data sources to improve case detection. With accurate detection the first step in disease control, validation of these data measures is consequently critical [9]. However, such studies are challenging for pertussis due to difficulties in case ascertainment. A study validating Ontario Health Insurance Plan (OHIP) diagnostic codes

* Corresponding author at: Department of Epidemiology, 121 South Main Street, 2nd Floor Box G-S121-2, Providence, RI 02912, United States of America.
E-mail address: shilo.mcburney@mail.utoronto.ca (S.H. McBurney).

https://doi.org/10.1016/j.jvacx.2023.100408
Received 20 May 2022; Received in revised form 14 September 2023; Accepted 13 November 2023
Available online 21 November 2023

for identifying childhood infections could not report the accuracy of pertussis codes due to an insufficient sample size [10]. While an Ontario study was able to report that up to 98% of pertussis cases are missed by surveillance and health administrative data, only sensitivity could be evaluated [11]. Ensuring adequate cases are identified generally requires oversampling those suspected to have pertussis for verification, but if this strategy is not accounted for during analyses it can affect accuracy measure estimates by introducing partial verification bias [9,12–15].

An approach that has been growing in popularity is linking health databases to electronic medical records (EMRs) which contain rich data that are often unavailable elsewhere [16,17]. EMR data can be used to both validate data measures and identify previously undetected cases. In Ontario, primary care EMR data has largely been used to study chronic conditions including epilepsy, ischemic heart disease, multiple sclerosis, and rheumatoid arthritis [18–22]. The aim of this study is to validate data measures routinely used for pertussis research in Ontario. To improve understanding of differential case ascertainment, we explored the impact of case severity on accuracy estimates as well as undetected case characteristics.

Methods

The University of Toronto's Health Sciences Research Ethics Board (# 37885) and the Public Health Ontario Ethics Review Board (# 2019-006.02) approved this study. Data were linked using unique encoded identifiers and analyzed at ICES (formerly the Institute for Clinical Evaluative Sciences). ICES is an independent, non-profit research institute whose legal status under Ontario's health information privacy law allows it to collect and analyze health care and demographic data, without consent, for health system evaluation and improvement.

Study cohort and reference standard

We used a cohort-selected cross-sectional design to evaluate the accuracy of pertussis detection algorithms in the Electronic Medical Record Primary Care (EMRPC) database [23,24]. EMRPC is one of the only Canadian sources of primary care EMR data available for secondary use to support research and contains patients from over 350 practicing primary care physicians who use PS Suite, the most commonly used EMR software in Ontario [25–27]. In addition to standard health data such as OHIP physician billing codes and laboratory tests, EMRPC includes clinical notes from patient interactions (progress notes) and medical histories in the cumulative patient profile (CPP). All 404,922 study subjects had a data entry in the EMRPC on or after January 1, 1986, with the study period ending on December 30, 2016 (Fig. 1). To create a reference standard, we classified 800 individuals sampled using a stratified strategy as definite pertussis, possible pertussis, ruled-out pertussis (record has a relevant term but does not meet case criteria), or no mention of pertussis (record has no relevant term) based on record review by two trained abstractors after incorporating laboratory and immunization data. We subdivided these classifications to align with surveillance criteria (Table 1) [28–31]. Epidemiological linkage refers to close contact or common exposure with another case.

Participant characteristics included a unique identifier, sex, age, start year on the EMR, and EMR data collection date. We described whether patients were formally registered to a contributing physician using roster status. We assessed active status on the EMR by whether a visit occurred in the year before the upload date, with a visit defined as a same-day progress note and billing. We only included billings for services provided in physician offices. We identified pertussis-containing vaccines by applying a validated algorithm to immunization data collected using codes and curated terms [32–34]. We used number of doses by patient age at last entry to categorize immunization status as up-to-date, complete or incomplete primary, or unvaccinated (Supplementary Table S1) [33,35]. Status was stratified by those who did or did not join the EMRPC at birth, as the latter may have incomplete dose ascertainment.

We used prevalent cases from the reference standard to minimize uncertainty in estimates since recurrences of pertussis are uncommon, which was supported by findings from the reference standard. When classifying definite cases, there were only six additional cases when

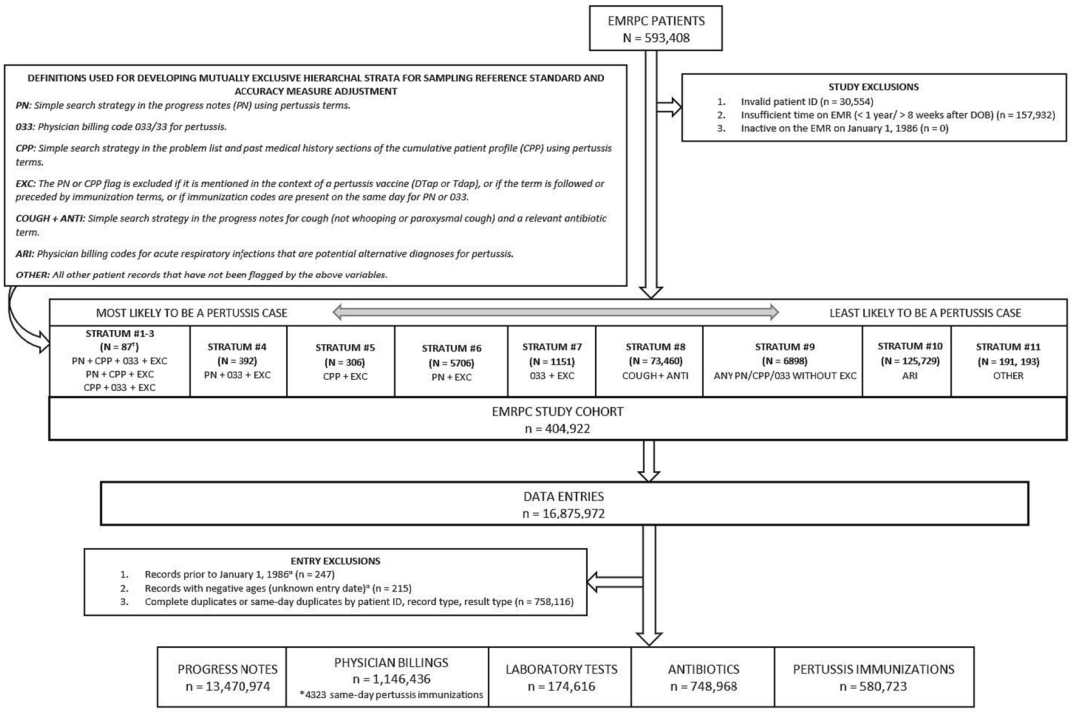

Fig. 1. Flow diagram of study cohort and data entries, 1986–2016. †combined for reporting due to low cell size (direct or by inference), αexcept for immunization data.

Table 1
Reference standard prevalent pertussis cases by classification and option number, 1986–2016.

Option #	Definite pertussis	# of cases	Possible pertussis	# of cases	Ruled-out pertussis	# of cases	No mention of pertussis	# of cases
1	A positive lab test result (culture or PCR) for *B. pertussis* AND clinical criteria (physician-diagnosed pertussis or physician-suspected pertussis or cough and key symptom).	24	A negative lab test result for *B. pertussis* within 0–28 days after initial entry in episode AND clinical criteria OR epidemiological link during episode.	110	Physician documents pertussis as ruled-out in progress notes.[α]	0	No pertussis-related data were present.	70
2	A positive lab test result for *B. pertussis* AND a cough.	0	Negative lab test result for *B. pertussis.*	7	Definite or possible pertussis[α] with subsequent confirmed alternative diagnosis within 90 days.	176	–	
3	A positive lab test result for *B. pertussis.*	1–5[*]	Indeterminate or unknown lab test result for *B. pertussis* within 0–28 days after initial entry in episode AND clinical criteria OR epidemiological link during episode.	20	Pertussis immunization given on same day as physician-diagnosed or -suspected pertussis.	1–5[*]	–	
4	Mention of epidemiological link AND clinical criteria.	25	Indeterminate PCR for *B. pertussis.*	0	Does not meet definite, possible, or other ruled-out criteria but a relevant keyword is present.	80–84 [*]	–	
5	Pertussis or whooping cough diagnosed by any physician in the CPP or progress notes.	90–94 [*]	Unknown lab test results for *B. pertussis.*	0	–		–	
6	Mention of epidemiological link AND history of cough in progress notes.	0	Physician-suspected pertussis with no additional information (not definite or ruled-out).	118–123 [*]	–		–	
7	Cough AND any key symptom: Paroxysms Inspiratory whoop Post-tussive vomitingApnea (if < 1 year)	64	Any key symptom[†] AND antibiotics prescribed.	1–5[*]	–		–	
8	–		Cough AND any symptom AND antibiotics prescribed.	0	–		–	
Total:		**208**		**261**		**261**		**70**

[*] suppressed for reporting due to low cell size (direct or by inference).
[†] includes cough lasting ≥ 2 weeks.
[α] excluding options pertaining to pertussis laboratory tests (definite options 1–3 and possible options.

comparing prevalent and incident case definitions. More uncertainty was evident when assessing possible cases, of which there were an additional 145 when using incident case definitions. Using prevalent cases also allowed CPP entries without specified dates to be incorporated to maximize case detection. Further details are available elsewhere [36].

EMR algorithms for identifying pertussis

We developed 25 distinct algorithms using the data from January 1, 1986 to December 30, 2016 in Fig. 1 to classify the EMRPC population into categories ("definite" pertussis, "possible" pertussis, "ruled-out" pertussis, and "no mention" of pertussis) that align with the reference standard (Supplementary Table S2). We used prevalent cases to minimize uncertainty in estimates since recurrences are uncommon, which allowed CPP entries without specified dates to be incorporated to maximize detection. We defined a prevalent case as ever being a case during the study period, with definite cases taking priority over possible cases when an individual was classified as both (Supplementary Figure S1). When we used incident cases for applying time-sensitive criteria, we defined them using 90-day episode rules (Supplementary Figure S1). Algorithms # 1–6 used pertussis data measures individually, including OHIP physician billing diagnostic codes, laboratory tests, and progress note and CPP mentions. We also assessed two codes or progress note mentions within 90 days. Algorithms # 7–10 used these data in combinations. Having or not having the measure of interest resulted in a "definite" or "no mention" classification respectively, with individuals who had the data measure but did not meet the necessary criteria considered "ruled-out." For algorithms including laboratory tests, we gave "possible" classifications if there was a negative or indeterminate test result within 28 days of the episode start, excepting participants with a positive pertussis laboratory test or another "definite" episode during the study period.

Algorithms # 11–22 had the same classifications except we used exclusion criteria to rule out additional entries from the previous algorithms. We identified entries with same-day immunization by expanding on the pertussis immunization strategy and excluded these as they are unlikely to be cases [33,35]. We used alternative diagnoses within 90 days to rule out suspected pertussis that was not the cause of illness. Diagnoses included acute respiratory infections that are most commonly responsible for misdiagnosed pertussis as well as other pertussis species [37]. Bronchitis or pneumonia only ruled out cases when a positive test for a pathogen other than pertussis was available since they can be pertussis complications [5]. The final three algorithms incorporated data measures with pertussis-associated antibiotics, including macrolides and trimethoprim-sulfamethoxazole [38,39], by applying complex rules that best aligned with the reference standard and full range of classifications. For example, for algorithm # 24 definite cases included positive laboratory results or physician diagnoses captured in the CPP. Alternatively, progress notes or OHIP codes were considered definite cases after attempting to decrease false positives by requiring two occurrences in 90 days. The OHIP codes also required same-day accompaniment with pertussis-associated antibiotics for either occurrence. Details on algorithm component development are available in Supplementary Table S3.

Accuracy measures and adjustment procedure

Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were selected as accuracy measures due to familiarity and relevance for clinical and public health audiences, with definite classifications considered to be cases and other classifications non-cases [12,40,41]. Begg and Greenes' method was used to avoid introducing partial verification bias through the reference standard sampling strategy [9,12–15]. Stratum-specific PPV and NPV were calculated using a binomial distribution and weighted using Bayes' theorem to obtain unbiased global estimates [9,13,14,42–44]. The average PPV or NPV was used for strata where a data measure was present but not selected for verification, with one person considered verified for calculating within-stratum variance. When normal approximation CI estimates extended beyond 0–100%, a logit transformation was used [45]. Cadieux et al.'s modified Begg and Greenes' approach was used to calculate sensitivity and specificity due to the large number of strata and resulting data missingness, with results compared to those from Obuchowski's method for verification [13,42]. With past studies reporting 1–68% sensitivity for pertussis health administrative data, 70% was used for the validity threshold [11,37]. Based on validation studies for other respiratory infections, the threshold for specificity and NPV was 90% and 50% for PPV [16]. An F2 score was used to complement the above accuracy measures by providing a composite measure that balances PPV and sensitivity as the harmonic mean between the two [46]. The F2 score is a version of the Fβ score which preferentially weights the importance of sensitivity using $\beta = 2$; this was selected to align with the study's focus on case detection [46]. F2 scores range between 0 and 1, with a higher score indicated better balance.

Sensitivity analyses

When building the reference standard, we noted possible cases had more uncertainty in true pertussis status [36]. Consequently, in a sensitivity analysis we assessed the inclusion of possible cases with definite cases as a secondary outcome. This allowed us to assess the impact of suspected case severity on accuracy, as we defined possible cases to generally be milder. To further explore the relationship between case definitions and accuracy, we used two analyses to reclassify physician-diagnosed and symptom-only pertussis as non-cases. These classifications were less specific but more sensitive, without laboratory confirmation or epidemiological linkage. We also conducted a sensitivity analysis that restricted results to those from the most reliable abstractor. To assess the impact of missing data on classification, we conducted two analyses limited to subjects most likely to have complete data (*i.e.*, rostered patients and patients with a visit in the 12 months prior to the EMR data collection date). Finally, we conducted an analysis that restricted progress notes to those related to visits, as they provide stronger evidence of pertussis-related healthcare interactions.

Discordance analysis

We examined classification subtypes of false negative pertussis cases from algorithms with the highest sensitivity or developed using commonly used data measures. We also explored age and immunization status at first episode as characteristics that may relate to ascertainment bias. Immunization status was described after removing doses in the 14 days prior to the first episode date due to insufficient time for an immunological response. We used a simple logistic regression model to evaluate the association between detection and age at first episode, removing CPP cases without dates for which age was unavailable. We could not directly evaluate the association between immunization status and detection as an insufficient number of participants joined the EMRPC at birth, meaning doses were likely incomplete.

Model fit was assessed after the addition of polynomials for age and natural cubic splines using the smallest Akaike information criterion (AIC) value [47]. Collinearity between age and the polynomial was considered to be present if the largest variance inflation factor (VIF) was greater than ten [48]. If present, age was centered by subtracting the mean from each observation and used in the model in place of the original. To further demonstrate the effect of age, we calculated the predicted probability of being undetected using the average age for those: less than one year of age, who experience the most severe pertussis infections; 1–27 years of age, who are eligible for publicly funded immunization with a lag of one year; and 28 or more years of age. We used sensitivity analyses that impacted estimates in addition to the primary analysis. When combining definite and possible cases, we used the date of first definite case for individuals with both classifications during the period. We considered results statistically significant at alpha ≤ 0.05 and used R for analyses [49].

Results

Algorithm counts

Algorithm counts by classification are available in Table 2. Over 96% of subjects had "no mention" of pertussis. The number of "definite" cases ranged between 55 for positive laboratory tests and 14,012 for any pertussis measure. More OHIP diagnostic code "definite" cases were ruled-out using alternative diagnoses than same-day immunizations, with the opposite true for progress notes. Of the 402 CPP mentions, 23% were accompanied by a pertussis-related progress note. Approximately two thirds of the CPP cases without dates were born before the study period. Sensitivity analyses with different counts are available in Supplementary Tables S4-6.

Validation of pertussis diagnostic algorithms

The algorithm with any pertussis measure (# 10) had the highest sensitivity of 20.6% (95% CI 8.25–42.8%), a PPV of 12.7% (95% CI 8.71–16.6%), and a F2 score of 0.183 (Table 3). The algorithm with complex rules applied to all measures (# 23) provided a similar balance between PPV (24.2%) and sensitivity (12.5%), although the higher PPV and lower sensitivity produced a lower F2 score of 0.138, as did algorithms based on progress notes. OHIP code (# 1) PPV and sensitivity were 13.6% (95% CI 9.28–17.9%) and 2.54% (95% CI 1.34–4.77%). Positive laboratory tests (# 3) had the lowest sensitivity for a single measure at 0.64% (95% CI 0.37–1.09%), but the PPV was 100%. Specificity and NPV remained above 95% for all algorithms. Sensitivity increased to 100% for the any measure algorithm (# 10) after reclassifying symptom-only cases (Table 4). However, the PPV and the F2 score were lower than other algorithms, at 4.87% and 0.204 respectively. For this analysis, the algorithm with complex rules applied to all measures with two of certain entries required within 90 days (# 24) provided a good balance between PPV and sensitivity, as did algorithms based on CPP mentions. The former had a PPV of 55.0% (95% CI 46.3–63.6%), 61.7% sensitivity (95% CI 6.78–97.3%), and an F2 score of 0.602.

Combining possible and definite cases generally decreased sensitivity and NPV and increased PPV, with little change to specificity (Supplementary Table S7). The algorithm for any measure (# 10) had a PPV of 45.7% (95% CI 38.6–52.8%), sensitivity of 14.7% (95% CI 9.44–22.0%), and an F2 score of 0.170. After shifting physician-diagnosed cases to non-cases, accuracy measures were largely unaffected (Supplementary Table S8). The exception to this was CPP mention algorithms, with PPVs dropping sharply. Estimates changed little after restricting to rostered patients and those with a visit in the last 12 months (Supplementary Tables S9-S10). With progress notes limited to visits, the PPV and sensitivity occasionally moved in either direction (Supplementary Table S11). Restricting results to those from the most reliable abstractor decreased sensitivity slightly, with the PPV moving marginally in either direction (Supplementary Table S12). Obuchowski's method produced similar accuracy point estimates across analyses

Table 2
Prevalent pertussis cases in the EMRPC study cohort (N = 404,922) by case classifications designated by rule-based algorithms.

Algorithm[σ]		CASES	NON-CASES		
Number	Description	"Definite"	"Possible"	"Ruled-out"	"No mention" (blank)
1	Physician billing pertussis diagnostic code (033/33)	1614	–	–	403,308
2	2x 033/33[†]	91	–	1523	403,308
3	Laboratory tests	55	766	–	404,101
4	Progress notes (PN)	13,095	–	–	391,827
5	2x PN[†]	1987	–	11,108	391,827
6	Cumulative patient profile (CPP)	402	–	–	404,520
7	PN + CPP	94	–	13,309	391,519
8	Same-day PN + 033/33	368	–	13,868	390,686
9	Ever PN + 033/33	473	–	13,763	390,686
10	Any flag with negative test exclusion (EXC)	14,012	660	–	390,250
11	033/33 with same-day immunization (SD immun) EXC	1557	–	57	403,308
12	033/33 with alternative diagnosis (alt dx) EXC	1390	–	224	403,308
13	033/33 with all EXC	1339	–	275	403,308
14	2x 033/33 with all EXC[†]	69	–	1545	403,308
15	PN with SD immun EXC	5899	–	7196	391,827
16	PN with alt dx EXC	11,162	–	1933	391,827
17	PN with all EXC	4343	–	8,752	391,827
18	2x PN with all EXC[†]	446	–	12,649	391,827
19	CPP with immun EXC terms	393	–	9	404,520
20	PN + CPP mention with all EXC	68	–	13,335	391,519
21	Same-day PN + 033/33 with all EXC	278	–	13,958	390,686
22	PN + 033/33 ever with all EXC	327	–	13,909	390,686
23	Complex rules (OR) with flags with all EXC[α]	4442	1603	8627	390,250
24	2x complex rules (OR) with flags with all EXC[†,α]	758	4409	9505	390,250
25	Complex rules (AND) with flags with all EXC[α]	108	5937	8627	390,250

[σ] detailed algorithm descriptions are available in Supplementary Table S2.
[†] 2x = two of the relevant data measure are required to be present within 90 days.
[α] OR = one of the listed data measures is sufficient (one OR the other), AND = both of certain data measures are required (one AND the other).

[42].

Discordance analysis

Most false negatives from the laboratory test (# 3) and OHIP code (# 1) algorithms were symptom-only or physician-diagnosed (Table 5). Of the laboratory test false negatives, 8–10% were missing positive laboratory results identified through abstraction. All false negatives for the algorithm with all data measures (# 10) were symptom-only. These participants were generally older, and despite most not having data from birth, approximately half of those for which immunization status was able to be determined had at least one pertussis-containing dose. Further details cannot be reported due to small cell counts. After combining definite and possible cases, most false negatives were physician-suspected (Table 5). When reclassifying symptom-only cases, we also assessed false negatives from the CPP algorithm with immunization exclusions (# 19). Missing positive tests accounted for 34–43% of these and epidemiological linkage with clinical criteria for 43–51%. Results were similar for the complex rule algorithm with two of certain entries required within 90 days (# 24).

When testing for differences in age at first episode between false negatives and true positives, the algorithm based on OHIP codes (# 1) had similar significant results under all three analyses (Table 6). In the primary analysis, for every year increase in age a patient had 1.02 (95% CI 1.01–1.04) times the odds of being a false negative (undetected) ($p = 0.005$). Patients who were five months of age had a probability of 0.32 (95% CI 0.21–0.45) of being undetected, compared to 0.58 (95% 0.46–0.68) for those 52 years of age, giving an odds ratio of 2.93. When combining possible and definite cases there was a significant association between age and detection for all other tested algorithms as well (Table 6). The odds ratio was 1.03 (95% CI 1.01–1.04, $p = 0.0003$) for all data measures (# 10) and 1.02 (95% CI 1.01–1.03, $p = 0.002$) for laboratory results (# 3). Model fit for the latter improved slightly after adding cubic splines. For each algorithm respectively, the probabilities of being undetected were 0.03 (95% CI 0.02–0.07) and 0.69 (95% CI 0.61–0.76) for those six months of age and 0.13 (95% CI 0.09–0.17) and 0.84 (95% CI 0.78–0.88) for those 53 years of age, producing odds ratios of 4.83 and 2.36.

Discussion

This study tested multiple pertussis data measures using a single reference standard and cohort-selected design. The results support the hypothesis that diagnosis and documentation issues have contributed to serious limitations in pertussis data accuracy, with the developed rule-based algorithms unable to adequately detect pertussis within primary care records. Sensitivity was particularly low, with findings supporting established theory that older or milder cases are less likely to be detected. While 100% sensitivity was achieved by an algorithm (# 10) when reclassifying symptom-only cases, the PPV was 4.87% and the F2 score was 0.204. This demonstrates the trade-off between these measures, which disallowed validity thresholds to be met. A better balance was achieved by algorithms applying complex rules to all measures or those based on progress notes or CPP mentions, evident through the higher F2 scores. This establishes the value of utilizing primary care EMR data with conventional sources for pertussis detection, in addition to identifying strategies that can be used to minimize the number of false positives while increasing sensitivity. NPVs only dropped below 90% when combining possible and definite cases increased prevalence. Specificity remained above 95%.

Laboratory results are in line with previously reported estimates of 1–50% sensitivity [11]. Accuracy estimates were lower than those from other EMRPC validation studies for acute or rare diseases, but the

Table 3
Accuracy measures from the primary evaluation of rule-based algorithms.

Algorithm[σ]	Cell Counts				Adjusted[§] Estimates % (95% CI)								
#	TP[†]	FP[†]	FN[†]	TN[†]	PPV[†]		NPV[†]		Sensitivity		Specificity		F2 Score
1	77	172	131	420	13.6	(9.28–17.9)	97.9	(97.9–98.0)	2.54	(1.34–4.77)	99.7	(99.6–99.7)	0.030
2	7	6	201	586	20.9	(14.2–27.6)	97.9	(97.9–97.9)	0.22	(0.07–0.71)	100	(100–100)	0.003
3	11	0	197	592	100	(100–100)	97.9	(97.9–97.9)	0.64	(0.37–1.09)	100	(100–100)[α]	0.008
4	140	378	68	214	11.8	(7.58–15.9)	98.2	(98.1–98.3)	17.8	(8.17–34.6)	97.1	(97.0–97.2)	0.162
5	53	81	155	511	12.1	(7.51–16.7)	97.9	(97.9–98.0)	2.79	(1.41–5.43)	99.6	(99.5–99.6)	0.033
6	98	28	110	564	79.1	(72.9–85.3)	98.0	(97.9–98.0)	3.69	(2.19–6.15)	100	(100–100)	0.046
7	35	12	173	580	76.8	(68.9–84.7)	97.9	(97.9–97.9)	0.84	(0.48–1.44)	100	(100–100)	0.010
8	61	88	147	504	38.5	(32.0–45.0)	97.9	(97.9–97.9)	1.64	(0.93–2.87)	99.9	(99.9–100)	0.020
9	73	113	135	479	34.8	(29.4–40.2)	97.9	(97.9–98.0)	1.91	(1.08–3.37)	99.9	(99.9–99.9)	0.024
10	199	401	9	191	12.7	(8.71–16.6)	98.3	(98.1–98.4)	20.6	(8.25–42.8)	96.9	(96.8–97.0)	0.183
11	77	169	131	423	12.8	(9.37–16.3)	97.9	(97.9–98.0)	2.31	(1.21–4.38)	99.7	(99.6–99.7)	0.028
12	68	142	140	450	13.7	(9.59–17.7)	97.9	(97.9–98.0)	2.20	(1.16–4.13)	99.7	(99.7–99.7)	0.026
13	*	*	*	*	12.7	(9.65–15.7)	97.9	(97.9–97.9)	1.97	(1.03–3.73)	99.7	(99.7–99.7)	0.024
14	*	*	*	*	20.1	(8.85–32.7)	97.9	(97.9–97.9)	0.17	(0.05–0.59)	100	(100–100)	0.002
15	135	317	73	275	18.4	(14.2–22.6)	98.1	(98.0–98.2)	12.6	(6.17–23.9)	98.8	(98.7–98.9)	0.134
16	117	307	91	285	11.5	(6.83–16.2)	98.1	(98.0–98.2)	14.9	(7.43–27.6)	97.5	(97.4–97.6)	0.141
17	110	246	98	346	19.1	(14.3–23.8)	98.1	(98.0–98.1)	9.61	(5.04–17.6)	99.1	(99.1–99.2)	0.107
18	34	32	174	560	29.1	(16.4–41.8)	97.9	(97.9–97.9)	1.50	(0.81–2.78)	99.9	(99.9–99.9)	0.019
19	98	28	110	564	79.3	(73.1–85.5)	98.0	(97.9–98.0)	3.61	(2.14–6.03)	100	(100–100)	0.045
20	28	10	180	582	75.8	(67.0–84.7)	97.9	(97.9–97.9)	0.60	(0.34–1.05)	100	(100–100)	0.007
21	50	64	158	528	43.6	(36.5–50.7)	97.9	(97.9–97.9)	1.41	(0.80–2.44)	100	(100–100)	0.017
22	58	80	150	512	41.3	(35.0–47.6)	97.9	(97.9–97.9)	1.57	(0.89–2.73)	100	(99.9–100)	0.019
23	178	231	30	361	24.2	(19.6–28.8)	98.1	(98.0–98.2)	12.5	(6.39–23.0)	99.2	(99.1–99.2)	0.138
24	125	48	83	544	56.3	(47.7–64.9)	98.0	(97.9–98.0)	4.95	(2.88–8.39)	99.9	(99.9–99.9)	0.061
25	35	9	173	583	82.5	(67.5–97.4)	97.9	(97.8–97.9)	1.03	(0.60–1.75)	100	(100–100)	0.013

[σ] detailed algorithm descriptions are available in Supplementary Table S2.
[§] adjusted to account for the stratified sampling strategy using a modified version of Begg and Greenes' procedure.
[†] TP = true positives, FP = false positives, FN = false negatives, TN = true negatives, PPV = positive predictive value, NPV = negative predictive value.
[*] suppressed due to residual small cell counts.
[α] variance unable to be directly calculated, reasonable prediction based on cell counts and estimates.

Table 4
Accuracy measures of rule-based algorithms, sensitivity analysis excluding those identified by symptoms alone from definite cases.

Algorithm[σ]	Cell Counts				Adjusted[§] Estimates % (95% CI)								
#	TP[†]	FP[†]	FN[†]	TN[†]	PPV[†]		NPV[†]		Sensitivity		Specificity		F2 Score
1	45	204	99	452	6.85	(3.96–9.74)	99.9	(99.8–99.9)	15.8	(1.98–63.7)	99.6	(99.6–99.7)	0.125
2	*	*	*	*	15.4	(7.21–23.6)	99.8	(99.8–99.9)	2.04	(0.24–15.2)	100	(100–100)	0.025
3	11	0	133	656	100	(100–100)	99.8	(99.8–99.9)	7.77	(1.43–32.8)	100	(100–100)[α]	0.095
4	*	*	*	*	3.36	(2.19–4.52)	99.9	(99.9–100)	63.7	(1.67–99.5)	96.9	(96.8–96.9)	0.139
5	41	93	103	563	8.06	(4.43–11.7)	99.9	(99.8–99.9)	23.7	(3.33–73.8)	99.6	(99.5–99.6)	0.171
6	*	*	*	*	78.8	(72.6–85.0)	99.9	(99.9–100)	46.1	(6.58–91.2)	100	(100–100)	0.503
7	*	*	*	*	75.4	(67.4–83.3)	99.9	(99.8–99.9)	10.4	(1.78–42.5)	100	(100–100)	0.126
8	41	108	103	548	25.3	(19.4–31.2)	99.9	(99.8–99.9)	13.6	(2.03–54.3)	99.9	(99.9–99.9)	0.150
9	*	*	*	*	20.0	(15.4–24.7)	99.9	(99.8–99.9)	13.8	(1.95–56.5)	99.9	(99.9–99.9)	0.147
10	144	456	0	200	4.87	(3.76–5.97)	100	(100–100)	100	(100–100)[α]	96.7	(96.6–96.8)	0.204
11	45	201	99	455	6.01	(4.84–7.18)	99.9	(99.8–99.9)	13.6	(1.67–59.4)	99.6	(99.6–99.7)	0.109
12	38	172	106	484	7.24	(4.15–10.3)	99.9	(99.8–99.9)	14.4	(1.92–59.1)	99.7	(99.7–99.7)	0.120
13	*	*	*	*	6.19	(4.88–7.50)	99.9	(99.8–99.9)	12.2	(1.58–54.4)	99.7	(99.7–99.7)	0.102
14	*	*	*	*	11.1	(0.02–98.7)	99.8	(99.8–99.9)	1.12	(0.10–11.1)	100	(100–100)	0.014
15	79	373	65	283	7.32	(4.74–9.90)	99.9	(99.9–99.9)	63.3	(2.50–99.2)	98.7	(98.6–98.7)	0.250
16	70	354	74	302	3.43	(2.20–4.66)	99.9	(99.9–99.9)	57.1	(2.97–98.3)	97.3	(97.3–97.4)	0.138
17	66	290	78	366	8.49	(5.37–11.6)	99.9	(99.9–99.9)	55.3	(3.87–97.4)	99.0	(99.0–99.1)	0.263
18	*	*	*	*	26.3	(13.7–39.0)	99.9	(99.8–99.9)	17.5	(2.80–60.8)	99.9	(99.9–99.9)	0.188
19	*	*	*	*	79.0	(72.8–85.1)	99.9	(99.9–100)	45.6	(6.45–91.0)	100	(100–100)	0.498
20	*	*	*	*	74.0	(65.1–82.9)	99.8	(99.8–99.9)	7.39	(1.27–33.1)	100	(100–100)	0.090
21	31	83	113	573	26.3	(20.0–32.5)	99.9	(99.8–99.9)	10.7	(1.67–45.7)	100	(99.9–100)	0.121
22	*	*	*	*	22.7	(17.4–28.1)	99.9	(99.8–99.9)	10.9	(1.64–47.2)	99.9	(99.9–100)	0.122
23	138	271	6	385	14.3	(11.2–17.4)	100	(100–100)	94.6	(0.05–100)	99.1	(99.0–99.1)	0.446
24	*	*	*	*	55.0	(46.3–63.6)	99.9	(99.9–100)	61.7	(6.78–97.3)	99.9	(99.9–99.9)	0.602
25	*	*	*	*	80.7	(65.0–96.5)	99.9	(99.8–99.9)	12.4	(2.20–47.4)	100	(100–100)	0.149

[σ] detailed algorithm descriptions are available in Supplementary Table S2.
[§] adjusted to account for the stratified sampling strategy using a modified version of Begg and Greenes' procedure.
[†] TP = true positives, FP = false positives, FN = false negatives, TN = true negatives, PPV = positive predictive value, NPV = negative predictive value.
[*] suppressed for reporting due to low cell size (direct or by inference).
[α] variance unable to be directly calculated, reasonable prediction based on cell counts and estimates.

Table 5
False negative pertussis cases from selected algorithms by classification option number and immunization status at first episode.

PRIMARY ANALYSIS	Algorithm # 1[†] (%)	Algorithm # 3[†] (%)	Algorithm # 10[†] (%)		
False negatives	**131**	**197**	**9**		
Definite classification (option number)[d]					
Positive test and clinical (# 1)	7 (5)	16–20* (8–10)	0 (0)		
Epidemiological link and clinical (# 4)	11 (8)	25 (13)	0 (0)		
Physician-diagnosed (# 5)	81 (62)	88–92* (45–47)	0 (0)		
Clinical symptoms (# 7)	32 (24)	64 (32)	9 (100)		
Pertussis immunization status at first episode					
Up-to-date for age	8 (6)	21 (11)	*		
At least one dose[γ]	23 (18)	43 (22)			
No doses documented or too young to be vaccinated	34 (26)	63–67* (32–34)			
Not applicable[α]	66 (50)	66–70* (34–36)			
SYMPTOM-ONLY AS NON-CASES	Algorithm # 1[†] (%)	Algorithm # 3[†] (%)	Algorithm # 10[†] (%)	Algorithm # 19[†] (%)	Algorithm # 24[†] (%)
False negatives	**99**	**133**	**0**	**47**	**24**
Definite classification (option number)[d]					
Positive test and clinical (# 1)	7 (7)	16–20* (12–15)	–	16–20* (34–43)	1–5* (4–21)
Positive test and cough (# 2)	0 (0)	–	–		0 (0)
Positive test alone (# 3)	0 (0)	–	–		0 (0)
Epidemiological link and clinical (# 4)	11 (11)	25 (19)	–	20–24* (43–51)	12–16* (50–67)
Physician-diagnosed (# 5)	81 (82)	88–92* (66–69)	–	7 (15)	7 (29)
Pertussis immunization status at first episode					
Up-to-date for age	16* (16)	7–11* (5–8)	–	10–14* (21–30)	16* (67)
At least one dose[γ]		22–26* (17–20)	–	17 (36)	
No doses documented or too young to be vaccinated	17 (17)	30–34* (23–26)	–	16–20* (34–43)	8 (33)
Not applicable[α]	66 (67)	66–70* (50–53)	–	–	–
DEFINITE AND POSSIBLE COMBINED	Algorithm # 1[†] (%)	Algorithm # 3[†] (%)	Algorithm # 10[†] (%)		
False negatives	**283**	**384**	**37**		
Definite classification (option number)[d]			1–5* (3–14)		
Positive test and clinical (# 1)	7 (2)	16–20* (4–5)			
Epidemiological link and clinical (# 4)	11 (4)	25 (7)			
Physician-diagnosed (# 5)	81 (29)	143–147* (37–38)			
Clinical symptoms (# 7)	32 (11)				
Possible classification (option number)[β]					
Negative lab and clinical/epidemiological link (# 1)	36 (13)	72* (19)			
Negative lab alone (# 2)	1–5* (0.4–2)				
Indeterminate lab and clinical/epidemiological link (# 3)	10 (4)				
Indeterminate lab alone (# 4)	0 (0)	0 (0)			
Symptom and antibiotics (# 7)	1–5* (0.4–2)	1–5* (0.3–1)			
Physician-suspected (# 6)	96–104* (34–37)	119–123* (31–32)	32–36* (86–97)		
Pertussis immunization status at first episode[β]					
Up-to-date for age	32 (11)	46 (12)	13–17* (35–46)		
At least one dose[γ]	73 (26)	100 (26)			
No doses documented or too young to be vaccinated	107–111* (38–39)	163–167* (42–43)	20–24* (54–65)		
Not applicable[α]	67–71* (24–25)	71–75* (18–20)	–		

[d] further description of the classifications and option numbers are available in Table 1.
[†] Algorithm # 1 = OHIP diagnostic codes, Algorithm # 3 = laboratory test results, Algorithm # 10 = all data measures, Algorithm # 19 = CPP mentions, Algorithm # 24 = all data measures with complex rules, two in 90 days.
[*] combined or suppressed due to small cell counts (direct or by inference).
[γ] patients in addition to those who are up-to-date with at least one dose.
[α] includes CPP cases without dates for which immunization status at time of episode could not be established.
[β] when combining definite and possible cases, first episode refers to first definite case for individuals with both a definite and positive case over the period.

accuracy of data requiring physician documentation may be inflated in those studies by basing the reference standard on physician diagnosis alone [10,22]. Cadieux et al. reported 100% PPV for pertussis ICD-9 codes, the basis for OHIP diagnostic codes, but only sampled nine codes for verification [13]. ICD-9 codes have also been demonstrated to have higher sensitivity and lower specificity (38.6% and 76.9%) [37]. However, these estimates were obtained using laboratory results as the reference standard and could be subject to testing bias. Requiring a positive laboratory result for confirmation, which is common for surveillance, emphasizes specificity at the risk of under-detection. This is exemplified through the exceptionally low sensitivity of laboratory results, with the high PPV an artefact of how we defined classifications. Contributing to this is that laboratory results were found to be incomplete after clinical note review, with all reports not entered in the EMRPC. Findings indicate that the low sensitivity of individual measures necessitates combining data sources. However, false positives increase with sensitivity, and efforts should be made to minimize and report misclassification using the findings from this study. For surveillance, this could include reporting a range of pertussis burden estimates with varying levels of certainty.

One strategy that could be further explored is using the algorithm including any measure (# 10) as a screening tool to minimize the number of cases in health administrative data that require additional review for case confirmation, as it has 100% sensitivity. After confirmation, newly identified pertussis cases could be incorporated into surveillance estimates to lessen the underestimation associated with reported cases. There were 14,012 definite cases flagged using this algorithm in the study cohort. While 14,012 records are substantially fewer to review than the entire cohort, this would still require significant resources, and was beyond the scope of this study. This highlights that to

Table 6
Effect of age at first episode[β] on the average odds and predicted probability of being undetected.

Primary analysis	Algorithm # 1[†]	95% CI	Algorithm # 3[†]	95% CI	Algorithm # 10[†]	95% CI
Intercept, OR (β)	0.46 (-0.77)	0.26–0.81	5.89 (1.77)	2.62–15.2	0.07 (-2.69)	0.02–0.18
Age at first episode, OR (β)	**1.02**[*] (0.02)	1.01–1.04	1.03 (0.03)	1.00–1.06	1.00 (0.001)	0.97–1.03
p-value for age	0.005[*]		0.09		0.95	
AIC	186.4		77.3		70.5	
Predicted probabilities (OR)						
< 1 year (avg. 5 months)[§]	0.32 (1.00)	0.21–0.45	0.86 (1.00)	0.72–0.93	0.06 (1.00)	0.02–0.17
1–27 years (avg. 9 years)[§]	0.36 (1.20)	0.26–0.47	0.88 (1.19)	0.79–0.94	0.06 (1.00)	0.03–0.15
28 + years (avg. 52 years)[§]	0.58 (2.93)	0.46–0.68	0.96 (3.91)	0.88–0.99	0.07 (1.18)	0.03–0.15

Symptom-only as non-cases	Algorithm # 1[†]	95% CI	Algorithm # 3[†]	95% CI	Algorithm # 19[†]	95% CI	Algorithm # 24[†]	95% CI	Algorithm # 19 with age^{2}[α]	95% CI
Intercept, OR (β)	0.38 (-0.96)	0.16–0.84	3.00 (1.10)	1.23–8.28	2.07 (0.73)	0.95–4.73	0.51 (-0.68)	0.22–1.10	0.88 (-0.13)	0.40–1.84
Age at first episode, OR (β)	**1.03**[*] (0.03)	1.00–1.05	1.03 (0.03)	1.00–1.07	0.99 (-0.01)	0.97–1.02	1.00 (-0.002)	0.98–1.02	1.00 (0.003)	0.98–1.03
Age at first episode^{2}, OR (β)									1.00 (0.002)	1.00–1.00
p-value for age (age, age^{2})	0.03[*]		0.12		0.59		0.87		0.85, 0.048[*]	
AIC	100.4		63.5		100.8		97.2		97.2	
Predicted probabilities (OR)										
< 1 year (avg. 4 months)[§]	0.28 (1.00)	0.15–0.46	0.75 (1.00)	0.54–0.88	0.67 (1.00)	0.48–0.82	0.36 (1.00)	0.19–0.53	0.81 (1.00)	0.59–0.93
1–27 years (avg. 9 years)[§]	0.33 (1.27)	0.20–0.48	0.79 (1.25)	0.64–0.89	0.66 (0.96)	0.51–0.79	0.33 (0.88)	0.21–0.49	0.67 (0.48)	0.50–0.81
28 + years (avg. 47 years)[§]	0.56 (3.27)	0.41–0.70	0.91 (3.37)	0.78–0.97	0.61 (0.77)	0.46–0.74	0.32 (0.84)	0.20–0.47	0.55 (0.29)	0.39–0.71

Definite and possible	Algorithm # 1[†]	95% CI	Algorithm # 3[†]	95% CI	Algorithm # 10[†]	95% CI	Algorithm # 3[†] with splines	95% CI	***Definite and possible***
Intercept, OR (β)	0.78 (-0.25)	0.56–1.09	2.22 (0.80)	1.53–3.25	0.04 (-3.34)	0.02–0.07	1.84 (0.61)	1.09–3.18	Intercept, OR (β)
Age at first episode, OR (β)	**1.01**[*] (0.01)	1.00–1.02	**1.02**[*] (0.02)	1.01–1.03	**1.03**[*] (0.03)	1.01–1.04	0.49 (-0.72)	0.12–1.86	Age at first episode, ns1[γ], OR (β)
							23.4[*] (3.15)	3.30–197.4	Age at first episode, ns2[γ], OR (β)
							14.6[*] (2.68)	2.59–133.2	Age at first episode, ns3[γ], OR (β)
p-value for age	0.004[*]		0.002[*]		0.0003[*]		0.30, 0.002[*], 0.007[*]		p-value for age (ns1, ns2, ns3[γ])
AIC	539.3		405.0		235.4		404.0		AIC
Predicted probabilities (OR)									
< 1 year (avg. 6 months)[§]	0.44 (1.00)	0.36–0.52	0.69 (1.00)	0.61–0.76	0.03 (1.00)	0.02–0.07	0.66 (1.00)	0.53–0.76	
1–27 years (avg. 10 years)[§]	0.47 (1.13)	0.40–0.54	0.72 (1.16)	0.66–0.78	0.04 (1.35)	0.02–0.08	0.77 (1.72)	0.69–0.83	
28 + years (avg. 53 years)[§]	0.60 (1.91)	0.53–0.66	0.84 (2.36)	0.78–0.88	0.13 (4.83)	0.09–0.17	0.79 (1.94)	0.81–0.85	

[β] when combining definite and possible cases first episode refers to the first definite episode for individuals with both a definite and positive case over the period, CPP cases without dates were excluded from analyses as age at time of episode could not be established.

[†] Algorithm # 1 = OHIP diagnostic codes, Algorithm # 3 = laboratory test results, Algorithm # 10 = all data measures, Algorithm # 19 = CPP mentions, Algorithm # 24 = all data measures with complex rules, two in 90 days.

[*] significant logistic regression model results at the alpha ≤ 0.05 level.

[§] for each age group, the average age was used to calculate the predicted probability of being undetected with significance determined and an odds ratio obtained by comparing to the youngest age group (<1 year).

[α] to address collinearity between age at first episode and age at first episode^{2} (squared), age was centered using the described procedure.

[γ] ns = natural spline.

be effectively used as a screening tool, it is essential to reduce the number of false positives while maintaining an adequate sensitivity. As a result, furthering understanding of why the positive predictive value of this algorithm was so poor would be a useful direction for future research. To use the algorithm as a screening tool, additional data or conditions may be required, or possibly a two-tiered testing protocol comprised of different algorithms could be used to produce a feasible number of records for additional review.

Reclassification of symptom-only cases produced the best accuracy estimates. If these are truly pertussis cases, one interpretation is that symptomatic, potentially milder cases are going undiagnosed, with misdiagnosis reported to be common even in the presence of a paroxysmal cough [7]. Alternatively, data may be missing, but this issue was mitigated by combining sources and using sensitivity analyses restricted to participants most likely to have complete data with little change in estimates. Providing further support, we found that older cases were significantly less likely to be detected using OHIP codes and for all tested algorithms with definite and possible cases combined. Although possible cases have greater uncertainty in true pertussis status, including milder case definitions demonstrated that older, less severe cases may be most frequently undetected. Failing to detect these cases impedes understanding of their relationship with waning immunity and transmission [1,8]. We could not directly test the relationship between immunization status and detection due to incomplete data, with immunizations possibly further underestimated using service codes from the EMRPC as doses received outside of primary care settings may be missed. However, age is somewhat a proxy measure as immunity wanes with age when immunizations are received in childhood. Descriptively, many undetected (*i.e.*, false negative) cases had at least one pertussis-containing vaccine dose. In addition to supporting established theory that older, milder cases are commonly undetected, this study describes how to use data to minimize this gap. Continued provider education on identification, documentation, and reporting could also improve detection. Furthermore, conducting additional testing among suspected mild cases using serological or oral fluid methods could assist with making a pertussis diagnosis [50,51]. Both strategies would have the added benefit of improving data accuracy and algorithm validity, meaning they could potentially be used as a component of case detection in the future.

While we used case episodes to apply exclusion criteria, prevalent cases were the focus. This provided the best chance for concordance between the algorithms and reference standard, meaning results can be conceptualized as the upper limit for episode-based accuracy measures. However, with only six additional occurrences of definite pertussis identified using incident cases, there is unlikely to be any meaningful change to accuracy estimates after incorporating these cases, excepting the sensitivity analysis with definite and possible cases combined. However, it could affect point and variance estimates by introducing correlation. Therefore, as most prevalent cases represent an individual's first and only occurrence, modeling prevalent cases increased model stability without a substantial cost to results [36].

In addition, while incident cases provide a good measure of risk, prevalent cases are particularly important for assessing population-level burden to assist in resource allocation such as vaccines and treatment. As a result, detecting prevalent cases is extremely useful for public health surveillance, clinicians, and public health practitioners. Finally, focusing on prevalent cases allowed us to incorporate CPP cases without specified dates to maximize detection and burden estimates, which were assumed to occur within the study period. Understanding the number of cases that have occurred over an extended period improves understanding of ongoing disease transmission, the degree of inaccuracy in past surveillance estimates, and possible reasons for misdiagnosis, even without knowing the exact diagnosis date.

If any of those born before 1986 were cases prior to the study period, disease misclassification would lead to underestimated sensitivity and NPV in algorithms that do not incorporate CPP mentions. However, we reclassified CPP cases as non-cases when reclassifying physician diagnoses with little change to estimates for non-CPP algorithms. In the future, this study's methods could be extended to analyze incident cases. Few validation studies have incorporated acute disease episodes, but it introduces new considerations. EMRPC studies focusing on chronic diseases have found requiring multiple entries to be a helpful classification tool by increasing diagnostic certainty [18–20,22]. In this study, it generally did not improve algorithm performance excepting for the complex algorithm (# 24) with symptom-only cases reclassified. Instead, we considered different approaches, such as using alternative diagnoses to rule out suspected cases. However, neither this nor same-day immunization greatly improved performance despite being used for reference standard development. In the future, the utility of using immunization history in combination with vaccine effectiveness as a pertussis diagnostic algorithm component could be explored.

As in most validation studies, an important limitation is that not all participants were verified [9,14]. The reference standard may also have further disease misclassification, such as that introduced by variation in how primary care physicians use EMRs [9,12,18]. In addition, EMRPC treatment data only captures about two thirds of the drug entries in the Ontario Drug Benefit program [26]; missing antibiotics could affect both the reference standard and algorithm classifications. The sampling strategy, adjustment procedure, and sensitivity analyses were undertaken to mitigate these potential sources of bias. Overall, the CIs for sensitivity were wider than predicted when optimizing the reference standard sample [36]. This can be partially explained by using a different method to calculate sensitivity. While producing the same point estimates, the selected approach provided more conservative variance estimates in situations when we suspected an insufficient number of false negatives were identified. As pertussis is rare and multiple data measures were tested, it is possible that we did not sample enough false negatives under all scenarios. In these cases, algorithm sensitivity cannot be adequately estimated because of small cell counts producing great variability [9,15]. One solution to further address this issue is to report findings with confidence intervals; although we used this strategy, it does not address the lack of precision. An additional limitation is that we could not completely stratify accuracy estimates by the reference standard subclassifications due to sample size limitations. While we evaluated the impact of severity on estimates through sensitivity analyses such as including possible with definite cases and shifting symptom-only cases to non-cases, this could be used to further assess severity in the future if enough cases are able to be verified.

Conclusions

This study simultaneously tested multiple pertussis data measures in the EMRPC. Sensitivity improved by reclassifying symptom-only cases but remained low unless multiple measures were used, suggesting that single measures frequently fail to identify cases. The algorithm with all data measures consistently produced the highest sensitivity, although a better trade-off between sensitivity and PPV was obtained after applying complex rules. However, the rule-based algorithms were unable to meet validity thresholds, indicating they are inadequate at detecting pertussis within primary care records. To maximize case detection, multiple sources should be used with efforts made to minimize and report ensuing false positives. EMRs can enhance case detection through patient history and clinical note data, however, identifying criteria that successfully enhance diagnostic certainty for acute diseases is an important area for future research. Additional gains in case identification and data accuracy can be achieved by implementing effective strategies for improving clinical diagnostic practice. It is essential to increase detection of older, milder cases with immunization history to reduce ascertainment bias in health data used for pertussis research. Failing to include these cases in vaccine effectiveness studies may artificially increase effectiveness estimates, particularly in terms of pertussis infection.

CRediT authorship contribution statement

Shilo H. McBurney: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration, Funding acquisition. **Jeffrey C. Kwong:** Conceptualization, Methodology, Investigation, Writing – review & editing, Supervision. **Kevin A. Brown:** Conceptualization, Methodology, Investigation, Writing – review & editing. **Frank Rudzicz:** Conceptualization, Methodology, Investigation, Writing – review & editing. **Branson Chen:** Software, Validation, Investigation, Data curation, Writing – review & editing. **Elisa Candido:** Methodology, Resources, Writing – review & editing. **Natasha S. Crowcroft:** Conceptualization, Methodology, Investigation, Resources, Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

The authors acknowledge the assistance of Andrew Wilton with data preparation and Arezou Saedi and Mohammad Ali Moinshaghaghi with record abstraction. The authors are grateful to the Ontario residents without whom this research would be impossible.

This study was supported by ICES, which is funded by an annual grant from the Ontario Ministry of Health (MOH) and the Ministry of Long-Term Care (MLTC). Parts of this material are based on data and/or information compiled and provided by ICES. The analyses, conclusions, opinions and statements expressed herein are solely those of the authors and do not reflect those of the funding or data sources; no endorsement is intended or should be inferred.

Funding

Shilo McBurney is supported by a Canadian Institutes of Health Research (CIHR) Doctoral Research Award (GSD-167037). Frank Rudzicz is supported by a CIFAR Chair in AI. JCK is supported by a Clinician-Scientist Award from the University of Toronto Department of Family and Community Medicine.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jvacx.2023.100408.

References

[1] Kilgore PE, Salim AM, Zervos MJ, Schmitt H-J. Pertussis: Microbiology, disease, treatment, and prevention. Clin Microbiol Rev 2016;29(3):449–86. https://doi.org/10.1128/CMR.00083-15.
[2] Doroshenko A, Qian W, Osgood ND. Evaluation of outbreak response immunization in the control of pertussis using agent-based modeling. PeerJ 2016;4:e2337.
[3] Barkoff AM, Grondahl-Yli-Hannuksela K, He Q. Seroprevalence studies of pertussis: What have we learned from different immunized populations. Pathog Dis 2015;73(7):ftv050. https://doi.org/10.1093/femspd/ftv050.
[4] Crowcroft NS, Stein C, Duclos P, Birmingham M. How best to estimate the global burden of pertussis? Lancet Infect Dis 2003;3(7):413–8. https://doi.org/10.1016/s1473-3099(03)00669-8.
[5] Crowcroft N, Miller E. Pertussis epidemiology. In: Rohani P, Scarpino S, editors. Pertussis: Epidemiology, Immunology, and Evolution. Oxford, UK: Oxford University Press; 2019. p. 66–86. https://doi.org/10.1093/oso/9780198811879.001.0001.
[6] Bolotin S, Quinn H, McIntyre P. Surveillance and diagnostics. In: Rohani P, Scarpino S, editors. Pertussis: Epidemiology, Immunology, and Evolution. Oxford, UK: Oxford University Press; 2019. p. 193–210. https://doi.org/10.1093/oso/9780198811879.001.0001.
[7] Deeks S, De Serres G, Boulianne N, Duval B, Rochette L, Dery P, et al. Failure of physicians to consider the diagnosis of pertussis in children. Clin Infect Dis 1999;28(4):840–6. https://doi.org/10.1086/515203.
[8] McGirr AA, Tuite AR, Fisman DN. Estimation of the underlying burden of pertussis in adolescents and adults in Southern Ontario, Canada. PLoS One 2013;8(12):e83850.
[9] Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. New York, NY: Oxford University Press; 2003.
[10] Hwee J, Sung L, Kwong JC, Sutradhar R, Tu K, Pole JD. Use of physician billing claims to identify infections in children. PLoS One 2018;13(11):e0207468.
[11] Crowcroft NS, Johnson C, Chen C, Li Y, Marchand-Austin A, Bolotin S, et al. Under-reporting of pertussis in Ontario: A Canadian Immunization Research Network (CIRN) study using capture-recapture. PLoS One 2018;13(5):e0195984.
[12] Umemneku Chikere CM, Wilson K, Graziadio S, Vale L, Allen AJ. Diagnostic test evaluation methodology: A systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard - An update. PLoS One 2019;14(10):e0223832.
[13] Cadieux G, Tamblyn R, Buckeridge DL, Dendukuri N. Validation of diagnostic groups based on health care utilization data should adjust for sampling strategy. Med Care 2017;55(8):e59–67. https://doi.org/10.1097/MLR.0000000000000324.
[14] Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983;39(1):207–15. https://doi.org/10.2307/2530820.
[15] Cronin AM, Vickers AJ. Statistical methods to correct for verification bias in diagnostic studies are inadequate when there are few false negatives: A simulation study. BMC Med Res Methodol 2008;8(75). https://doi.org/10.1186/1471-2288-8-75.
[16] Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: A systematic review. J Am Med Inform Assoc 2016;23(5):1007–15. https://doi.org/10.1093/jamia/ocv180.
[17] National Institute for Health Research. Clinical Practice Research Datalink. 2022. https://www.cprd.com/.
[18] Tu K, Wang M, Jaakkimainen RL, Butt D, Ivers NM, Young J, et al. Assessing the validity of using administrative data to identify patients with epilepsy. Epilepsia 2014;55(2):335–43. https://doi.org/10.1111/epi.12506.
[19] 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). Can J Cardiol 2010;26(7):e225–8. https://doi.org/10.1016/s0828-282x(10)70412-8.
[20] Widdifield J, Bombardier C, Bernatsky S, Paterson JM, Green D, Young J, et al. An administrative data validation study of the accuracy of algorithms for identifying rheumatoid arthritis: The influence of the reference standard on algorithm performance. BMC Musculoskelet Disord 2014;15:216. https://doi.org/10.1186/1471-2474-15-216.
[21] Widdifield J, Ivers NM, Young J, Green D, Jaakkimainen L, Butt DA, et al. Development and validation of an administrative data algorithm to estimate the disease burden and epidemiology of multiple sclerosis in Ontario. Canada *Mult Scler* 2015;21(8):1045–54. https://doi.org/10.1177/1352458514556303.
[22] Krysko KM, Ivers NM, Young J, O'Connor P, Tu K. Identifying individuals with multiple sclerosis in an electronic medical record. Mult Scler 2015;21(2):217–24. https://doi.org/10.1177/1352458514538334.
[23] Mathes T, Pieper D. An algorithm for the classification of study designs to assess diagnostic, prognostic and predictive test accuracy in systematic reviews. Syst Rev 2019;8(1):226. https://doi.org/10.1186/s13643-019-1131-4.
[24] Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(8):529–36. https://doi.org/10.7326/0003-4819-155-8-201110180-00009.
[25] Wilson SE, Chung H, Schwartz KL, Guttmann A, Deeks SL, Kwong JC, et al. Rotavirus vaccine coverage and factors associated with uptake using linked data: Ontario, Canada. PLoS One 2018;13(2):e0192809.
[26] Tu K, Mitiku TF, Ivers NM, Guo H, Lu H, Jaakkimainen L. Evaluation of Electronic Medical Record Administrative data Linked Database (EMRALD). Am J Manag Care 2014;20:e15–21.
[27] 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 Inform Decis Mak 2015:15. https://doi.org/10.1186/s12911-015-0195-x.
[28] Ministry of Health and Long-Term Care. Infectious Diseases Protocol - Appendix B: Provincial Case Definitions for Diseases of Public Health Signficiance, Pertussis (Whooping Cough). 2019.
[29] World Health Organization. Pertussis: Vaccine-Preventable Diseases Surveillance Standards. 2018. https://www.who.int/publications/m/item/vaccine-preventable-diseases-surveillance-standards-pertussis.
[30] Public Health Agency of Canada. National Case Definition: Pertussis. 2008. https://www.canada.ca/en/public-health/services/immunization/vaccine-preventable-diseases/pertussis-whooping-cough/health-professionals/national-case-definition.html.

[31] Centers for Disease Control and Prevention. Pertussis (Whooping Cough) - Surveillance and Reporting Case Definition. 2020. https://www.cdc.gov/pertussis/surv-reporting.html.
[32] Schwartz KL, Tu K, Wing L, Campitelli MA, Crowcroft NS, Deeks SL, et al. Validation of infant immunization billing codes in administrative data. Hum Vaccines Immunother 2015;11(7):1840–7. https://doi.org/10.1080/21645515.2015.1043499.
[33] Schwartz KL, Kwong JC, Deeks SL, Campitelli MA, Jamieson FB, Marchand-Austin A, et al. Effectiveness of pertussis vaccination and duration of immunity. CMAJ 2016;188(16):E399–406. https://doi.org/10.1503/cmaj.160193.
[34] Wilson SE, Wilton AS, Young J, Candido E, Bunko A, Buchan SA, et al. Assessing the completeness of infant and childhood immunizations within a provincial registry populated by parental reporting: A study using linked databases in Ontario. Canada *Vaccine* 2020;38(33):5223–30. https://doi.org/10.1016/j.vaccine.2020.06.003.
[35] Crowcroft NS, Schwartz KL, Chen C, Johnson C, Li Y, Marchand-Austin A, et al. Pertussis vaccine effectiveness in a frequency matched population-based case-control Canadian Immunization Research Network study in Ontario, Canada 2009–2015. Vaccine 2019;37(19):2617–23. https://doi.org/10.1016/j.vaccine.2019.02.047.
[36] McBurney SH, Kwong JC, Brown KA, Rudzicz F, Chen B, Candido E, et al. Developing a reference standard for pertussis by applying a stratified sampling strategy to electronic medical record data. Ann Epidemiol 2023;77:53–60. https://doi.org/10.1016/j.annepidem.2022.11.002.
[37] Fathima S, Simmonds KA, Drews SJ, Svenson LW, Kwong JC, Mahmud SM, et al. How well do ICD-9 physician claim diagnostic codes identify confirmed pertussis cases in Alberta, Canada? A Canadian Immunization Research Network (CIRN) study. BMC Health Serv Res 2017;17(1):479. https://doi.org/10.1186/s12913-017-2321-1.
[38] Schwartz KL, Wilton AS, Langford BJ, Brown KA, Daneman N, Garber G, et al. Comparing prescribing and dispensing databases to study antibiotic use: A validation study of the Electronic Medical Record Administrative data Linked Database (EMRALD). J Antimicrob Chemother 2019;74(7):2091–7. https://doi.org/10.1093/jac/dkz033.
[39] Schwartz KL, Langford BJ, Daneman N, Chen B, Brown KA, McIsaac W, et al. Unnecessary antibiotic prescribing in a Canadian primary care setting: A descriptive analysis using routinely collected electronic medical record data. CMAJ Open 2020;8(2):E360–9. https://doi.org/10.9778/cmajo.20190175.
[40] Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. J Am Med Inform Assoc 2017;24(1):198–208. https://doi.org/10.1093/jamia/ocw042.
[41] Zhou XH. Comparing accuracies of two screening tests in a two-phase study for dementia. J R Stat Soc Ser C Appl Stat 1998;47(1):135–47. https://doi.org/10.1111/1467-9876.00102.
[42] Obuchowski NA, Zhou XH. Prospective studies of diagnostic test accuracy when disease prevalence is low. Biostatistics 2002;3(4):477–92. https://doi.org/10.1093/biostatistics/3.4.477.
[43] Irwig L, Glasziou PP, Berry G, Chock C, Mock P, Simpson JM. Efficient study designs to assess the accuracy of screening tests. Am J Epidemiol 1994;140(8):759–69. https://doi.org/10.1093/oxfordjournals.aje.a117323.
[44] Holtman GA, Berger MY, Burger H, Deeks JJ, Donner-Banzhoff N, Fanshawe TR, et al. Development of practical recommendations for diagnostic accuracy studies in low-prevalence situations. J Clin Epidemiol 2019;114:38–48. https://doi.org/10.1016/j.jclinepi.2019.05.018.
[45] Ying GS, Maguire MG, Glynn RJ, Rosner B. Calculating sensitivity, specificity, and predictive values for correlated eye data. Investig Ophthalmol Vis Sci 2020;61(11):29. https://doi.org/10.1167/iovs.61.11.29.
[46] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 2015;15:29. https://doi.org/10.1186/s12880-015-0068-x.
[47] Akaike H. Akaike's Information Criterion. In Lovric M, (Ed). *International Encyclopedia of Statistical Science*. Berlin, Heidelberg: Springer Berlin Heidelberg. 2011:25-25. doi: 10.1007/978-3-642-04898-2_110.
[48] Marquardt DW. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics 1970;12(3):591–612. https://doi.org/10.2307/1267205.
[49] R-Core-Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.
[50] Litt DJ, Samuel D, Duncan J, Harnden A, George RC, Harrison TG. Detection of anti-pertussis toxin IgG in oral fluids for use in diagnosis and surveillance of Bordetella pertussis infection in children and young adults. J Med Microbiol 2006;55(Pt 9):1223–8. https://doi.org/10.1099/jmm.0.46543-0.
[51] Harnden A, Grant C, Harrison T, Perera R, Brueggemann AB, Mayon-White R, et al. Whooping cough in school age children with persistent cough: Prospective cohort study in primary care. BMJ 2006;333(7560):174–7. https://doi.org/10.1136/bmj.38870.655405.AE.