



## RESEARCH ARTICLE

# Assessing the validity of administrative health data for the identification of children and youth with autism spectrum disorder in Ontario

Jennifer D. Brooks<sup>1</sup>  | Jasleen Arneja<sup>1</sup>  | Longdi Fu<sup>2</sup> | Farah E. Saxena<sup>2</sup> | Karen Tu<sup>3,4</sup> | Virgiliu Bogdan Pinzaru<sup>2</sup> | Evdokia Anagnostou<sup>5,6</sup> | Kirk Nylen<sup>7,8</sup> | Natasha R. Saunders<sup>1,2,6,9</sup> | Hong Lu<sup>2</sup> | John McLaughlin<sup>1</sup> | Susan E. Bronskill<sup>1,2</sup>

<sup>1</sup>Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

<sup>2</sup>ICES, G1 06, Toronto, Ontario, Canada

<sup>3</sup>North York General Hospital, Toronto Western Hospital Family Health Team-University Health Network, Toronto, Ontario, Canada

<sup>4</sup>Department of Family and Community Medicine, University of Toronto, Toronto, Ontario, Canada

<sup>5</sup>Holland Bloorview Kids Rehabilitation Hospital, Toronto, Ontario, Canada

<sup>6</sup>Department of Pediatrics, University of Toronto, Toronto, Ontario, Canada

<sup>7</sup>Ontario Brain Institute, Toronto, Ontario, Canada

<sup>8</sup>Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada

<sup>9</sup>Department of Pediatric Medicine, The Hospital for Sick Children, Toronto, Ontario, Canada

## Correspondence

Jennifer Brooks, 155 College St. HSB 676 Toronto Ontario, M5T 3M7, Canada. Email: jennifer.brooks@utoronto.ca

## Funding information

McLaughlin Center, University of Toronto, Grant/Award Number: MC-2017-04; Ontario Brain Institute

## Abstract

Population-level identification of children and youth with ASD is essential for surveillance and planning for required services. The objective of this study was to develop and validate an algorithm for the identification of children and youth with ASD using administrative health data. In this retrospective validation study, we linked an electronic medical record (EMR)-based reference standard, consisting 10,000 individuals aged 1–24 years, including 112 confirmed ASD cases to Ontario administrative health data, for the testing of multiple case-finding algorithms. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and corresponding 95% confidence intervals (CI) were calculated for each algorithm. The optimal algorithm was validated in three external cohorts representing family practice, education, and specialized clinical settings. The optimal algorithm included an ASD diagnostic code for a single hospital discharge or emergency department visit or outpatient surgery, or three ASD physician billing codes in 3 years. This algorithm's sensitivity was 50.0% (95%CI 40.7–88.7%), specificity 99.6% (99.4–99.7), PPV 56.6% (46.8–66.3), and NPV 99.4% (99.3–99.6). The results of this study illustrate limitations and need for cautious interpretation when using administrative health data alone for the identification of children and youth with ASD.

**Lay Summary:** We tested algorithms (set of rules) to identify young people with ASD using routinely collected administrative health data. Even the best algorithm misses more than half of those in Ontario with ASD. To understand this better, we tested how well the algorithm worked in different settings (family practice, education, and specialized clinics). The identification of individuals with ASD at a population level is essential for planning for support services and the allocation of resources.

## KEYWORDS

administrative health data, algorithm, autism, Ontario

## INTRODUCTION

Autism spectrum disorders (ASD) are a group of lifelong neurodevelopmental disorders characterized by impaired communication and social interaction, repetitive

behaviors, and restricted interests. These core features of ASD are further complicated by the cooccurrence of other neurodevelopmental disorders (e.g., attention deficit hyperactivity disorder [ADHD]; Abdallah et al., 2011), many of which have overlapping traits, and can also

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. *Autism Research* published by International Society for Autism Research published by Wiley Periodicals LLC.

include intellectual disability (Taurines et al., 2012). Autistic individuals are also more likely to be diagnosed with a host of other medical conditions including: epilepsy, schizophrenia, gastrointestinal disturbances, sleep disorders, anxiety, and asthma, among others (Nazeen, Palmer, Berger, & Kohane, 2016). Together this means that individuals with ASD and their families, frequently require sustained support and increased resources, from both the public and private sectors, across the life-course (Liptak, Stuart, & Auinger, 2006; Weiss et al., 2018). These services are associated with a significant economic burden on publicly funded health, education, and social services, but are also associated with significant out-of-pocket expenses by the families of individuals with ASD (Buescher, Cidav, Knapp, & Mandell, 2014; Tsiplova et al., 2019).

In 2015, it was estimated that approximately one in 59 eight-year-olds in the United States had ASD (Baio et al., 2018). Recently, the Public Health Agency of Canada (PHAC) reported that one in 66 Canadian children between the ages of 5 and 17 years has the condition (Ofner et al., 2018). The diagnosis of autism is three- to four-times more common in males versus females, varies based on geographical region, and is increasing over time; factors that are reflected in both the American and Canadian estimates (Baio et al., 2018; Loomes, Hull, & Mandy, 2017; Ofner et al., 2018).

The identification of children and youth with ASD at a population level is essential for surveillance and planning for support services and resource allocation. Further, there is a need to provide evidence-based recommendations for programs and services for individuals with ASD and their families. To address this need, researchers are increasingly developing different approaches for the identification of individuals with ASD, both within electronic medical records (EMR; Brooks et al., 2021; Bush, Connelly, Perez, Barlow, & Chiang, 2017; Coleman et al., 2015; Lingren et al., 2016), and using administrative health data (Coo, Ouellette-Kuntz, Brownell, Shooshtari, & Hanlon-Dearman, 2018; Dodds et al., 2009).

Administrative data have the benefit of being population-wide and longitudinal in nature, allowing for the tracking of population trends over time. However, these data are not collected for research purposes, relying on routine practices which may not be consistent across health care providers or settings (Jutte, Roos, & Brownell, 2011). Further, the data used to generate population-based estimates of ASD vary across regions and can include information from social services, health services and educational data, alone or in combination. Researchers in some Canadian provinces have developed and validated algorithms for the identification of individuals with ASD using available administrative data. An algorithm developed in Nova Scotia using multiple administrative health databases achieved a sensitivity of 69.3% (Dodds et al., 2009). Researchers in Manitoba were able to do better by using a combination of administrative health and education data to generate a case-finding algorithm

with 88% sensitivity (Coo et al., 2018). In a recent study by Bickford et al. (2020), the investigators concluded that administrative data was insufficient for the identification of children with autism, in particular an inability to discriminate between children with ASD and children with other neurodevelopmental outcomes.

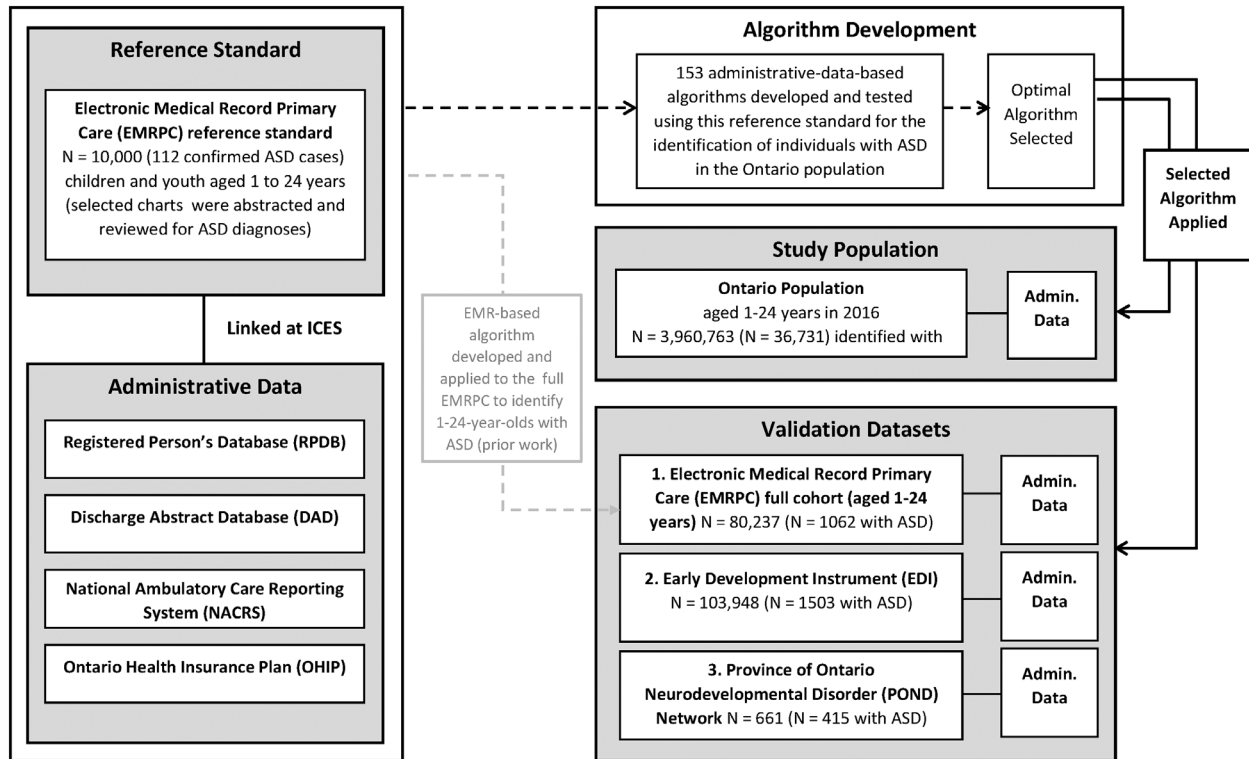
Given the importance of identifying individuals with ASD to support health system planning, resource allocation, and the monitoring of long-term outcomes in autistic individuals, the objective of this study was to develop and validate an algorithm for the identification of children and youth with ASD using administrative health data in Ontario Canada, where population-based estimates of ASD prevalence are not currently available.

## METHODS

We conducted a retrospective validation study of multiple algorithms to identify ASD in administrative health data, using a reference standard of individuals with known ASD status from family practice EMR (Figure 1). All research was carried out at ICES, an independent, nonprofit 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. ICES was formerly known as the Institute for Clinical Evaluative Sciences, and formally adopted the initialism ICES as its official name in 2018. This project was approved by the Research Ethics Board at the University of Toronto and Sunnybrook Health Sciences Center, Toronto, Canada.

### Reference standard

The Electronic Medical Record Primary Care (EMRPC; formerly known as EMRALD) data housed at ICES consist of the EMRs for over 400,000 patients, from over 350 family physicians using PS Suite EMR (formerly Practice Solutions). This includes data on a patient demographics, vital status, prescription medications, current diagnoses, progress notes, and billing information, as well as a free-text field for physician comments. Participation in EMRPC on the part of the physician is voluntary and only requires them to have used the EMR for at least 2 years (Tu et al., 2015). A reference standard for ASD was created using a population-based cohort of 10,000 children and youth aged 1 to 24 years within the EMRPC, details have been published elsewhere (Hauck, Lau, Wing, Kurdyak, & Tu, 2017). Briefly, trained nurse chart abstractors manually reviewed the EMR histories of 10,000 children and youth for diagnoses of ASD (including Asperger's syndrome and pervasive developmental delay [PDD] per DSM-V criteria), and other neurodevelopmental (e.g., ADHD) and mental health (e.g., depression) conditions. Abstractor-identified cases



**FIGURE 1** Flow diagram for the linkage of various data sources for the validation and development of an administrative data-based algorithm including application to the province of Ontario, and three external validation cohorts

were confirmed by a family physician (KT), and if confirmed, classified as “definitely” having ASD. In total, 112 individuals were identified from the cohort of 10,000 as having a confirmed diagnosis of ASD (mean age 12.7 years, 1:3 ratio females to males) and 9888 were determined to not have an autism diagnosis. These 112 individuals with ASD and 9888 without, were linked with Ontario administrative health data and used as the reference standard for the development and testing of algorithms for the identification of individuals with ASD in the Ontario population.

### Administrative health data

People living in Ontario are insured under a single-payer system that covers physician and hospital services and procedures. Administrative health data are generated through patient contact with the health care system and maintained in multiple databases that can be linked using a unique encoded identifier (Table S1). These include: the Registered Persons Database (RPDB) which provides information on demographic variables (e.g., age, sex, neighborhood-level income quintile, and geographic location [the latter two are derived from Statistics Canada Census data]), the Canadian Institute for Health Information Discharge Abstract Database (DAD) and National Ambulatory Care Reporting System (NACRS)

which provide information on diagnoses, procedures and other characteristics for hospital visits, community-based ambulatory care, outpatient clinics, outpatient surgery, and emergency department visits. The Ontario Health Insurance Plan (OHIP) physician billing database provides information on diagnoses and medical service billings by Ontario physicians (both general practitioners and specialists). These datasets were linked using unique encoded identifiers and analyzed at ICES (Figure 1).

### Algorithm development, testing, and selection

The EMR reference standard was first linked to administrative health data sources (RPDB, DAD, NACRS, and OHIP). Established methods for algorithm development, testing and selection, routinely employed at ICES (Tu et al., 2014), were used to generate and evaluate case-finding algorithms for ASD. Tested algorithms ( $N = 153$ ) included data from multiple sources: physician billing, hospital visits, outpatient surgeries, and emergency department visits. The OHIP billing code 299 “childhood psychoses e.g., autism” was used. While this billing code is not specific to autism, it is the code most frequently used by physicians to indicate the use of health services related to autism. Codes from all other sources included the ASD ICD-9 diagnostic codes (299.x) and ICD-10 codes (F84.x). OHIP billing codes from specific specialists (pediatricians,

neurologists, and psychiatrists) were also considered. The tested algorithms varied the type of data source(s) included (physician billing codes, hospital discharge, emergency department visits, and outpatient surgery), and the number and frequency of codes required (e.g., 1–4 codes in 1–4 years). A final, optimal algorithm was selected to maximize both the positive predictive value (PPV) and sensitivity of the algorithm (Figure 1).

## External validation

The optimal algorithm was validated using three external cohorts. These cohorts included individuals diagnosed or identified as having ASD in a range of contexts including primary care, education, and specialized clinical settings. The first validation was carried out in the primary care setting, among children and youth aged 1–24 years old in the full EMRPC cohort (as of March 1, 2016,  $N = 80,237$ ). Individuals were classified as having ASD or not using a previously developed EMR-based algorithm, which identified 1062 individuals with, and 79,175 without (sensitivity 82.1%, PPV 98.9%; Brooks et al., 2021). Notably this algorithm had good predictive ability and consisted of a key word search of the Cumulative Patient Profile, a free text field within the EMR that summarizes key patient information.

The second validation was conducted using a cohort of 5- to 6-year-olds in Ontario who completed the Early Development Instrument (EDI) between January and June 2015, at the end of their senior kindergarten year ( $N = 103,948$ ). The Early Development Instrument (EDI) is a questionnaire developed by Dr. Dan Offord and Dr. Magdalena Janus at the Offord Center for Child Studies at McMaster University, Hamilton, Ontario, Canada. The EDI is a teacher-completed checklist that measures children's developmental health in five domains: physical health and well-being, social competence, emotional maturity, language and cognitive development, and communication skills and general knowledge (Janus & Offord, 2007). Using this tool, 1503 individuals with, and 102,445 without ASD were identified.

Finally, the third validation was carried out using the Province of Ontario Neurodevelopmental Disorder (POND) Network. POND is an ongoing cross-sectional study of children and youth with neurodevelopmental disorders (e.g., ASD, ADHD, obsessive compulsive disorder [OCD]) for which recruitment began in 2012. Participants are identified through four recruitment centers in Ontario: Holland Bloorview Kids Rehabilitation Hospital and The Hospital for Sick Children in Toronto, McMaster University in Hamilton, and the Lawson Health Research Institute-Western University in London Ontario. This study population is referred to as the "clinical cohort" for the purpose of algorithm validation as these individuals were diagnosed within these specialized clinical settings. For the purpose of this validation, we

used 661 POND participants, 415 of whom have a diagnosis of ASD.

Each external validation cohort was linked to administrative health data at ICES using unique encoded identifiers to allow for testing of algorithm discrimination (Figure 1).

## Statistical analyses

The 153 tested algorithms developed using the reference standard ( $N = 10,000$  including  $N = 112$  with ASD) were applied to the population of Ontario residents aged 1–24 years in 2016 ( $N = 3,960,763$ ). Sensitivity, specificity, PPV, negative predictive value (NPV), and corresponding 95% confidence intervals (CI) were calculated for each algorithm. For further examination, the optimal algorithm was applied to the Ontario population, using the three cohorts: the broader EMRPC cohort using the EMR algorithm ( $N = 80,237$ ,  $N = 1062$  with ASD), those identified using EDI data ( $N = 103,948$ ,  $N = 1503$  with ASD) and those with a primary diagnosis of ASD included in the POND Study ( $N = 661$ ,  $N = 415$  with ASD), as the reference standard.

## RESULTS

A selection of the tested algorithms is presented in Table 1 (all algorithms are presented in Table S2). Overall, the algorithms tended to be highly specific (i.e., low false negative rate) while the sensitivity varied substantially. The most sensitive algorithm was one physician billing claim or a single hospital discharge or emergency department visit or outpatient surgery with a diagnosis of ASD (75.9%), however, this algorithm had a very low PPV (probability that individuals identified by the algorithm as having ASD actually had ASD) of only 42.9%. The optimal algorithm was selected to maximize both sensitivity and PPV. This algorithm included (three ASD physician billing codes in 3 years or a single hospital discharge or emergency department visit or outpatient surgery with a diagnosis of ASD). This algorithm had a sensitivity of 50.0% (95% CI 40.7–88.7%), specificity of 99.6% (95% CI 99.4–99.7), PPV of 56.6% (95% CI 46.8–66.3), and NPV of 99.4% (95% CI 99.3–99.6), maximizing both sensitivity and PPV.

Secondary review of the EMR data for all misclassified individuals in the reference standard revealed complex medical histories. False-positives (i.e., individuals in the reference standard classified as having ASD by the algorithm, but not by chart review) tended to have a high burden of other neurodevelopmental or mental health conditions including diagnoses or features of ADHD, anxiety and depression, or other possible NDDs. Conversely, many of the false-negatives (about half) were identified in the medical record review as having Asperger's Syndrome.

**TABLE 1** Selected algorithms tested for the identification of children and youth (ages 1–24 years) with ASD in Ontario

Description <sup>a</sup>	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
1 physician billing claim ever	74.1 (66.0–82.2)	98.9 (98.7–99.1)	42.6 (35.6–49.5)	99.7 (99.6–99.8)
1 physician billing claim code by any specialist	66.1 (57.3–74.8)	99.1 (98.9–99.2)	44.3 (36.8–51.8)	99.6 (99.5–99.7)
Hospital discharge or emergency department visit or outpatient surgery or 1 physician billing claim	75.9 (68.0–83.8)	98.9 (98.6–99.1)	42.9 (36.0–49.8)	99.7 (99.6–99.8)
Hospital discharge or emergency department visit or outpatient surgery or 1 physician billing claim by any specialist	67.9 (59.2–76.5)	99.0 (98.9–99.2)	44.7 (37.2–52.2)	99.6 (99.5–99.8)
Hospital discharge or emergency department visit or outpatient surgery or (2 physician billing claims in 2 years)	57.1 (48.0–66.3)	99.3 (99.1–99.5)	48.5 (40.0–57.0)	99.5 (99.4–99.7)
Hospital discharge or emergency department visit or outpatient surgery or (2 physician billing claims in 2 years at least 1 physician billing claim by any specialist)	52.7 (43.4–61.9)	99.4 (99.2–99.5)	48.4 (39.5–57.2)	99.5 (99.3–99.6)
Hospital discharge or emergency department visit or outpatient surgery or (2 physician billing claims in 3 years)	59.8 (50.7–68.9)	99.3 (99.1–99.5)	49.3 (40.9–57.7)	99.5 (99.4–99.7)
Hospital discharge or emergency department visit or outpatient surgery or (2 physician billing claims in 3 years at least 1 physician billing claim by any specialist)	53.6 (44.3–62.8)	99.4 (99.2–99.5)	48.4 (39.6–57.2)	99.5 (99.3–99.6)
Hospital discharge or emergency department visit or outpatient surgery or (3 physician billing claims in 2 years)	45.5 (36.3–54.8)	99.6 (99.4–99.7)	54.3 (44.2–64.3)	99.4 (99.2–99.5)
Hospital discharge or emergency department visit or outpatient surgery or (3 physician billing claims in 2 years at least 1 physician billing claims by any specialist)	45.5 (36.3–54.8)	99.6 (99.5–99.7)	56.0 (45.8–66.2)	99.4 (99.2–99.5)
Hospital discharge or emergency department visit or outpatient surgery or (3 physician billing claims in 3 years)	50.0 (40.7–59.3)	99.6 (99.4–99.7)	56.6 (46.8–66.3)	99.4 (99.3–99.6)
Hospital discharge or emergency department visit or outpatient surgery or (3 physician billing claims in 3 years at least 1 physician billing claim by any specialist)	49.1 (39.8–58.4)	99.6 (99.5–99.7)	57.9 (48.0–67.8)	99.4 (99.3–99.6)

Abbreviations: ASD, autism spectrum disorder; NPV, negative predictive value; PPV, positive predictive value.

<sup>a</sup>Physician billing claims included 299 “childhood psychoses e.g., autism.” Codes from all other sources (e.g., hospital discharge, emergency department visit) included ICD-9 codes (299.x) and ICD-10 codes (F84.x). Specialists include pediatricians, psychiatrists, and neurologists.

Asperger’s Syndrome used to be considered a clinically distinct higher functioning form of ASD, but since 2013 with the introduction of DSM-V is captured under the diagnosis of ASD.

When applied to the 2016 Ontario population (ages 1–24 years), the optimal algorithm resulted in 36,713 out of 3,960,763 children and youth being identified as having ASD. This corresponds to a prevalence of 0.93% or one in 108. The characteristics of individuals in Ontario identified with and without ASD are shown in Table 2. Individuals identified as having ASD were slightly older with a mean age of 12.7 years, as compared to those without ASD (mean age 12.5 years). The sex distribution also differed between the two groups with approximately 80% of those identified as having ASD being male.

Results from the external validation are presented in Table 3. Application of the algorithm to the EMRPC population (family practice), found discrimination of the algorithm was similar to that seen for the reference standard used in algorithm development (sensitivity = 51.8%, specificity = 99.7%, PPV = 62.5%, and NPV = 99.4%). In the EDI cohort (education), while the algorithm had a similar sensitivity to that seen using the reference standard or full EMRPC cohort (55.9%), the PPV was higher

at 70.1%. Finally, the algorithm performed well in the clinical cohort from the POND Study, with 72.8% sensitivity, 94.3% specificity, 95.6% PPV, and 67.2% NPV.

## DISCUSSION

We tested and validated over 150 algorithms for the identification of children and youth with ASD using population-based administrative health data in Ontario. The optimal algorithm used information from multiple sources and included: three ASD physician billing codes in 3 years or a single hospital discharge or emergency department visit or outpatient surgery, with an associated ASD diagnostic code. Notably, while this algorithm was considered “optimal” among all those that were tested, it is far from ideal with a sensitivity of only 50% and a PPV of 57%. To better understand who this algorithm is identifying as having autism, and conversely who it is not capturing, we applied this algorithm to the Ontario population and conducted several external validations.

When applied to the province of Ontario, this algorithm identified over 36,000 individuals between the ages 1–24 years as having ASD, corresponding to a prevalence

**TABLE 2** Characteristics of children and youth (ages 1–24 years) with and without ASD in Ontario (2016)

	No ASD <i>N</i> = 3,924,032	ASD <i>N</i> = 36,731
Mean age ( $\pm$ SD) <sup>a</sup>	12.5 $\pm$ 7.3	12.7 $\pm$ 5.8
Age group (years), <i>N</i> (%)		
1–4	726,767 (18.5)	2,631 (7.2)
5–9	761,472 (19.4)	9,847 (26.8)
10–14	751,718 (19.2)	10,124 (27.6)
15–19	793,060 (20.2)	8,369 (22.8)
$\geq$ 20	891,015 (22.7)	5,760 (15.7)
Sex		
Female	1,924,366 (49.0)	7,201 (19.6)
Male	1,999,666 (51.0)	29,530 (80.4)
Geographic Location		
Rural	412,632 (10.5)	3,304 (9.0)
Suburban	282,152 (7.2)	2,812 (7.7)
Urban	3,229,248 (82.3)	30,615 (83.3)
Neighborhood Income Quintile		
1 (Lowest)	769,673 (19.6)	7,815 (21.3)
2	728,467 (18.6)	7,110 (19.4)
3	779,880 (19.9)	7,205 (19.6)
4	808,081 (20.6)	7,258 (19.8)
5 (Highest)	829,172 (21.1)	7,216 (19.6)

Abbreviation: ASD, Autism spectrum disorder.

<sup>a</sup>Age as of March 1, 2016.

of 0.93%. This estimate is lower than the estimated prevalence of ASD in Canada and the United States of approximately 1.5% (Baio et al., 2018; Ofner et al., 2018). Sex differences in the prevalence of ASD are well documented (Baio et al., 2018; Ofner et al., 2018). Here, we found that of the individuals determined by the algorithm to have ASD, about 80% were male. This is slightly higher than the expected ratio of about 1:3 or 1:4 (Baio et al., 2018; Loomes et al., 2017; Ofner et al., 2018), suggesting that the algorithm preferentially identifies male over female individuals as having ASD.

Autism can be reliably diagnosed as early as 24 months of age (Corsello, Akshoomoff, & Stahmer, 2013; Steiner, Goldsmith, Snow, & Chawarska, 2012). In Ontario, and other Canadian provinces, the earliest opportunity to assess developmental concerns across multiple domains (e.g., language, social) is typically with a child's physician, as routine visits are frequent in the early years of life. Still, most children are not clinically diagnosed until after the age of four (Baio et al., 2018; Brett, Warnell, McConachie, & Parr, 2016; Coe et al., 2012; Ofner et al., 2018).

When examining the characteristics of the individuals identified by the algorithm as having ASD versus not, there is a slight difference in the age distribution. A lower proportion of children under 5 years of age are classified as having ASD (7.2% with ASD and 18.5% without ASD, Table 2). This is likely to be partly due to the timing of

**TABLE 3** Algorithm validation in external cohorts

Cohort details	EMRPC <sup>a</sup>	EDI <sup>b</sup>	POND <sup>c</sup>
Total <i>N</i>	80,237	103,948	661
Number with ASD	1062	1503	415
Estimated ASD Prevalence (%)	1.3	1.4	N/A <sup>d</sup>
Age range	1–24 years	5–6 years	1–21 years
Performance of selected algorithm			
Sensitivity (%)	51.8	55.9	72.8
Specificity (%)	99.7	99.6	94.3
PPV (%)	62.5	70.1	95.6
NPV (%)	99.4	99.4	67.2

Abbreviations: ASD, autism spectrum disorder; EDI: early development instrument; EMRPC: electronic medical record primary care; NPV, negative predictive value; POND: province of Ontario neurodevelopmental disorders; PPV, positive predictive value.

<sup>a</sup>EMRPC data, housed at ICES. Individuals with ASD were identified using an EMR-based case-finding algorithm.

<sup>b</sup>EDI is a teacher-completed measure of early development outcomes of children in kindergarten. Data is from a cohort of 5–6-year-old children in Ontario in 2015.

<sup>c</sup>POND Study is an ongoing cross-sectional study of children and youth with neurodevelopmental disorders.

<sup>d</sup>Prevalence of ASD is not estimated in POND as it is a cohort of children and youth with neurodevelopmental disorders.

clinical diagnosis, but may also be related to a bias in the algorithm towards the identification of older children, with one of the elements of the algorithm being the occurrence of three physician billing codes in 3 years. This requires that the individual be alive, and in Ontario for at least 3 years to be classified as having ASD based on this criterion. However, algorithms that used shorter time periods (e.g., 1 or 2 years) did not perform as well (Table 1).

Discrimination of the optimal algorithm was tested against multiple cohorts including children with ASD diagnosed/identified across a variety of settings including primary care, education, and specialized clinics. When we linked these cohorts to the administrative data and observed who was captured by the algorithm, we saw differences in the algorithm's ability to discriminate between individuals with and without ASD across the different settings. The algorithm performed somewhat better in the EDI data, where the sensitivity was still low (55.9%) but the PPV was higher at 70.1%. This could be because the EDI data include individuals that were 5- to 6-years old in 2015 and so are able to be captured by the three OHIP codes in 3 years criterion. Not surprisingly, the algorithm performed best in the specialized clinical cohort from the POND study with a sensitivity of 72.8% and a PPV of 95.6%, this is likely to be partly due to the high prevalence of ASD in this study cohort. However, it is also likely because these patients are being seen within a highly specialized clinical setting, suggesting that the algorithm is likely missing individuals who are not as intensely engaged in clinical care.

A particular challenge to the use of these data for the identification of individuals with ASD is that ASD is one

of a series of neurodevelopmental disorders (e.g., ADHD and OCD) that share common traits (Krakowski et al., 2020; Kushki et al., 2019; Taurines et al., 2012), and etiology (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013; Faraone & Larsson, 2019; Gonzalez-Mantilla, Moreno-De-Luca, Ledbetter, & Martin, 2016; Lowther et al., 2017; Zarrei et al., 2019), frequently co-occurring within individuals (Lai et al., 2019). A recent meta-analysis found that 28% of individuals with ASD also have ADHD and 9% have OCD (Lai et al., 2019). The combination of within disorder heterogeneity, co-occurrence, and cross-disorder homogeneity, makes the use of administrative data for case-identification particularly challenging.

Indeed, when we reviewed the EMRs of individuals classified as false positives, they were found to have complex medical histories with multiple on-going neurodevelopmental diagnoses. Conversely, approximately half of the individuals with ASD not captured by the algorithm were found to have Asperger's Syndrome, a higher functioning autism. Together this suggests that the algorithm, favors the identification of individuals with more complex needs, requiring multiple interactions with the health care system.

Prior efforts have examined the potential of using provincial administrative health data to identify individuals with ASD with variable degrees of success (Bickford et al., 2020; Coo et al., 2018; Dodds et al., 2009), and those able to incorporate data from multiple settings, outside of administrative health data having the highest sensitivity (Coo et al., 2018). Our results, like those of Bickford et al. (2020), suggest that the use of administrative health data alone is not sufficient for the identification of individuals with ASD. This is likely because autistic individuals receive services from other health professionals (e.g., psychologists, social workers, occupational, and physical therapists), and educational support staff, all of whom are known to play an important role in the diagnosis and care of individuals with ASD (Lai, Anagnostou, Wiznitzer, Allison, & Baron-Cohen, 2020), but whose activity is not captured through physician billing.

Numerous studies have used administrative health data to identify individuals with autism (Alexeeff et al., 2017; Angell et al., 2021; Burke et al., 2014; Coleman et al., 2015; Jain et al., 2015; Zerbo et al., 2018). While this approach has several obvious potential benefits including the ability to capture population-level estimates of ASD prevalence, prevalence of comorbidities, and health system utilization, both retrospectively and prospectively, the results of the current study illustrate some of the limitations of identifying children and youth with ASD in this manner.

Overall, we found that our optimal algorithm for the identification of children and youth with ASD within administrative health data favors the identification of individuals with ASD who are older (>age 5 years), male,

seen within a specialized clinical setting and/or with more complex presentation. This reinforces the need for validation studies across multiple jurisdictions (Nicholls, Langan, Sørensen, Petersen, & Benchimol, 2016) and the incorporation of data outside of administrative health data alone, to better capture the complex and varied pathways to an ASD diagnosis.

The current study has numerous strengths. In particular, the use of a reference standard developed through chart abstraction and confirmation by a family physician. This is further strengthened by our use of multiple cohorts for additional external validation, to better understand who is and is not being captured by the algorithm. There are of course several limitations that must also be considered. First, is our current inability to utilize education data or services administered by other health professionals into the case-finding algorithms. We were also not able to distinguish between individuals with autism with and without intellectual disability or language impairment (as described in DSM-V). It is possible that these individuals may have different patterns of health services use and may be differentially captured by the algorithm. Further, while the use of multiple validation cohorts is a strength of the study, each of these cohorts has their own limitations. First, autistic individuals in the EMPRC cohort were algorithm identified (Brooks et al., 2021). While having good accuracy, we cannot be 100% certain patients have been correctly classified. The same is true for the EDI instrument which is not a screening instrument but rather a teacher-recorded confirmation of a predetermined/preexisting diagnosis. Finally, this study did not address the issue of identifying adults with autism in the population, but rather focused on children and youth. Autistic individuals require sustained but changing supports across the life course (Lai et al., 2020) and an understanding of these needs is essential. Ultimately, it is hoped that individuals with autism can be identified in childhood and followed-up longitudinally to understand long-term outcomes.

The identification of children and youth with ASD is essential for ASD surveillance, identification of trends in incidence and prevalence, and planning for services, supports and resource allocation. The challenge of accurately identifying individuals with autism using administrative health data limits our ability to conduct population-level research on ASD and associated outcomes. This has widespread implications for the study of autism using administrative health data, requiring cautious interpretation of studies using administrative health data alone to identify individuals with autism, supporting the need for validation studies across multiple jurisdictions, as well as the integration of data across sectors (e.g., health, education, and social services).

## ACKNOWLEDGMENTS

This study was supported by ICES, which is funded by an annual grant from the Ontario Ministry of Health and

Long-Term Care (MOHLTC). This study also received funding from the McLaughlin Center at the University of Toronto and the Ontario Brain Institute (Accelerator Grant: MC-2017-04). Karen Tu is supported by a Research Scholar Award from the Department of Family and Community Medicine at the University of Toronto. Parts of this material are based on data and information compiled and provided by Canadian Institute for Health Information (CIHI). 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.

## CONFLICT OF INTEREST

The authors declare no conflicts of interest.

## ORCID

Jennifer D. Brooks  <https://orcid.org/0000-0001-7574-4256>

Jasleen Arneja  <https://orcid.org/0000-0003-1913-6699>

## REFERENCES

- Abdallah, M. W., Greaves-Lord, K., Grove, J., Norgaard-Pedersen, B., Hougaard, D. M., & Mortensen, E. L. (2011). Psychiatric comorbidities in autism spectrum disorders: Findings from a Danish Historic Birth Cohort. *European Child and Adolescent Psychiatry*, 20(11–12), 599–601. <https://doi.org/10.1007/s00787-011-0220-2>
- Alexeeff, S. E., Yau, V., Qian, Y., Davignon, M., Lynch, F., Crawford, P., Davis, R., & Croen, L. A. (2017). Medical conditions in the first years of life associated with future diagnosis of ASD in children. *Journal of Autism and Developmental Disorders*, 47(7), 2067–2079. <https://doi.org/10.1007/s10803-017-3130-4>
- Angell, A. M., Deavenport-Saman, A., Yin, L., Zou, B., Bai, C., Varma, D., & Solomon, O. (2021). Sex differences in co-occurring conditions among autistic children and youth in Florida: A retrospective cohort study (2012–2019). *Journal of Autism and Developmental Disorders*. <https://doi.org/10.1007/s10803-020-04841-5>
- Baio, J., Wiggins, L., Christensen, D. L., Maenner, M. J., Daniels, J., Warren, Z., ... White, T. (2018). Prevalence of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2014. *MMWR Surveillance Summaries*, 67(6), 1–23. <https://doi.org/10.15585/mmwr.ss6706a1>
- Bickford, C. D., Oberlander, T. F., Lanphear, N. E., Weikum, W. M., Janssen, P. A., Ouellette-Kuntz, H., & Hanley, G. E. (2020). Identification of pediatric autism spectrum disorder cases using health administrative data. *Autism Research*, 13(3), 456–463. <https://doi.org/10.1002/aur.2252>
- Brett, D., Warnell, F., McConachie, H., & Parr, J. R. (2016). Factors affecting age at ASD diagnosis in UK: No evidence that diagnosis age has decreased between 2004 and 2014. *Journal of Autism and Developmental Disorders*, 46(6), 1974–1984. <https://doi.org/10.1007/s10803-016-2716-6>
- Brooks, J. D., Bronskill, S. E., Fu, L., Saxena, F. E., Arneja, J., Pinzaru, V. B., ... Tu, K. (2021). Identifying children and youth with autism Spectrum disorder in electronic medical records: Examining health system utilization and comorbidities. *Autism Research*, 14(2), 400–410. <https://doi.org/10.1002/aur.2419>
- Buescher, A. V. S., Cidav, Z., Knapp, M., & Mandell, D. S. (2014). Costs of autism spectrum disorders in the United Kingdom and the United States. *JAMA Pediatrics*, 168(8), 721–728. <https://doi.org/10.1001/jamapediatrics.2014.210>
- Burke, J. P., Jain, A., Yang, W., Kelly, J. P., Kaiser, M., Becker, L., Lawer, L., & Newschaffer, C. J. (2014). Does a claims diagnosis of autism mean a true case? *Autism*, 18(3), 321–330. <https://doi.org/10.1177/1362361312467709>
- Bush, R. A., Connelly, C. D., Perez, A., Barlow, H., & Chiang, G. J. (2017). Extracting autism spectrum disorder data from the electronic health record. *Applied Clinical Informatics*, 8(3), 731–741. <https://doi.org/10.4338/aci-2017-02-ra-0029>
- Coleman, K. J., Lutsky, M. A., Yau, V., Qian, Y., Pomichowski, M. E., Crawford, P. M., Lynch, F. L., Madden, J. M., Owen-Smith, A., Pearson, J. A., Pearson, K. A., Rusinak, D., Quinn, V. P., & Croen, L. A. (2015). Validation of autism spectrum disorder diagnoses in large healthcare systems with electronic medical records. *Journal of Autism and Developmental Disorders*, 45(7), 1989–1996. <https://doi.org/10.1007/s10803-015-2358-0>
- Coo, H., Ouellette-Kuntz, H., Lam, M., Yu, C. T., Dewey, D., Bernier, F. P., Chudley, A. E., Hennessey, P. E., Breitenbach, M. M., Noonan, A. L., Lewis, M. E., & Holden, J. J. (2012). Correlates of age at diagnosis of autism spectrum disorders in six Canadian regions. *Chronic Diseases and Injuries in Canada*, 32(2), 90–100.
- Coo, H., Ouellette-Kuntz, H., Brownell, M., Shoostari, S., & Hanlon-Dearman, A. (2018). Validating an administrative data-based case definition for identifying children and youth with autism spectrum disorder for surveillance purposes. *Canadian Journal of Public Health. Revue Canadienne de Santé Publique*, 108(5-6), e530–e538. <https://doi.org/10.17269/cjph.108.5963>
- Corsello, C. M., Akshoomoff, N., & Stahmer, A. C. (2013). Diagnosis of autism spectrum disorders in 2-year-olds: A study of community practice. *Journal of Child Psychology and Psychiatry*, 54(2), 178–185. <https://doi.org/10.1111/j.1469-7610.2012.02607.x>
- Cross-Disorder Group of the Psychiatric Genomics Consortium. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: A genome-wide analysis. *The Lancet*, 381(9875), 1371–1379. [https://doi.org/10.1016/S0140-6736\(12\)62129-1](https://doi.org/10.1016/S0140-6736(12)62129-1)
- Dodds, L., Spencer, A., Shea, S., Fell, D., Armson, B. A., Allen, A. C., & Bryson, S. (2009). Validity of autism diagnoses using administrative health data. *Chronic Diseases in Canada*, 29(3), 102–107.
- Faraoane, S. V., & Larsson, H. (2019). Genetics of attention deficit hyperactivity disorder. *Molecular Psychiatry*, 24(4), 562–575. <https://doi.org/10.1038/s41380-018-0070-0>
- Gonzalez-Mantilla, A. J., Moreno-De-Luca, A., Ledbetter, D. H., & Martin, C. L. (2016). A cross-disorder method to identify novel candidate genes for developmental brain disorders. *JAMA Psychiatry*, 73(3), 275–283. <https://doi.org/10.1001/jamapsychiatry.2015.2692>
- Hauck, T. S., Lau, C., Wing, L. L. F., Kurdyak, P., & Tu, K. (2017). ADHD treatment in primary care: Demographic factors, medication trends, and treatment predictors. *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie*, 62(6), 393–402. <https://doi.org/10.1177/0706743716689055>
- Jain, A., Marshall, J., Buikema, A., Bancroft, T., Kelly, J. P., & Newschaffer, C. J. (2015). Autism occurrence by MMR vaccine status among US children with older siblings with and without autism. *JAMA*, 313(15), 1534–1540. <https://doi.org/10.1001/jama.2015.3077>
- Janus, M., & Offord, D. R. (2007). Development and psychometric properties of the early development instrument (EDI): A measure of Children's school readiness. *Canadian Journal of Behavioural Science: Revue Canadienne Des Sciences Du Comportement*, 39(1), 1–22. <https://doi.org/10.1037/cjbs2007001>
- Jutte, D. P., Roos, L. L., & Brownell, M. D. (2011). Administrative record linkage as a tool for public health research. *Annual Review of Public Health*, 32, 91–108. <https://doi.org/10.1146/annurev-publhealth-031210-100700>



- Krakowski, A. D., Cost, K. T., Anagnostou, E., Lai, M.-C., Crosbie, J., Schachar, R., Georgiades, S., Duku, E., & Szatmari, P. (2020). Inattention and hyperactive/impulsive component scores do not differentiate between autism spectrum disorder and attention-deficit/hyperactivity disorder in a clinical sample. *Molecular Autism*, *11*(1), 28–28. <https://doi.org/10.1186/s13229-020-00338-1>
- Kushki, A., Anagnostou, E., Hammill, C., Duez, P., Brian, J., Iaboni, A., Schachar, R., Crosbie, J., Arnold, P., & Lerch, J. P. (2019). Examining overlap and homogeneity in ASD, ADHD, and OCD: A data-driven, diagnosis-agnostic approach. *Translational Psychiatry*, *9*(1), 318. <https://doi.org/10.1038/s41398-019-0631-2>
- Lai, M.-C., Kasse, C., Besney, R., Bonato, S., Hull, L., Mandy, W., Szatmari, P., & Ameis, S. H. (2019). Prevalence of co-occurring mental health diagnoses in the autism population: A systematic review and meta-analysis. *The Lancet Psychiatry*, *6*(10), 819–829. [https://doi.org/10.1016/S2215-0366\(19\)30289-5](https://doi.org/10.1016/S2215-0366(19)30289-5)
- Lai, M.-C., Anagnostou, E., Wiznitzer, M., Allison, C., & Baron-Cohen, S. (2020). Evidence-based support for autistic people across the lifespan: Maximising potential, minimising barriers, and optimising the person–environment fit. *The Lancet Neurology*, *19*(5), 434–451. [https://doi.org/10.1016/S1474-4422\(20\)30034-X](https://doi.org/10.1016/S1474-4422(20)30034-X)
- Lingren, T., Chen, P., Bochenek, J., Doshi-Velez, F., Manning-Courtney, P., Bickel, J., Wildenger Welchons, L., Reinhold, J., Bing, N., Ni, Y., Barbaresi, W., Mentch, F., Basford, M., Denny, J., Vazquez, L., Perry, C., Namjou, B., Qiu, H., Connolly, J., ... Savova, G. (2016). Electronic health record based algorithm to identify patients with autism Spectrum disorder. *PLoS One*, *11*(7), e0159621. <https://doi.org/10.1371/journal.pone.0159621>
- Liptak, G. S., Stuart, T., & Auinger, P. (2006). Health care utilization and expenditures for children with autism: Data from U.S. national samples. *Journal of Autism and Developmental Disorders*, *36*(7), 871–879. <https://doi.org/10.1007/s10803-006-0119-9>
- Loomes, R., Hull, L., & Mandy, W. P. L. (2017). What is the male-to-female ratio in autism spectrum disorder? A systematic review and meta-analysis. *Journal of the American Academy of Child and Adolescent Psychiatry*, *56*(6), 466–474. <https://doi.org/10.1016/j.jaac.2017.03.013>
- Lowther, C., Speevak, M., Armour, C. M., Goh, E. S., Graham, G. E., Li, C., Zeeman, S., Nowaczyk, M. J. M., Schultz, L. A., Morra, A., Nicolson, R., Bikangaga, P., Samdup, D., Zaazou, M., Boyd, K., Jung, J. H., Siu, V., Rajguru, M., Goobie, S., ... Bassett, A. S. (2017). Molecular characterization of NRXN1 deletions from 19,263 clinical microarray cases identifies exons important for neurodevelopmental disease expression. *Genetics in Medicine*, *19*(1), 53–61. <https://doi.org/10.1038/gim.2016.54>
- Nazeen, S., Palmer, N. P., Berger, B., & Kohane, I. S. (2016). Integrative analysis of genetic data sets reveals a shared innate immune component in autism spectrum disorder and its co-morbidities. *Genome Biology*, *17*, 228. <https://doi.org/10.1186/s13059-016-1084-z>
- Nicholls, S. G., Langan, S. M., Sørensen, H. T., Petersen, I., & Benchimol, E. I. (2016). The RECORD reporting guidelines: Meeting the methodological and ethical demands of transparency in research using routinely-collected health data. *Clinical Epidemiology*, *8*, 389–392. <https://doi.org/10.2147/CLEP.S110528>
- Ofner, M., Coles, A., Decou, M., Do, M., Bienek, A., Snider, J., & Ugnat, A. (2018). *Autism spectrum disorder among children and youth in Canada 2018*. Ottawa, ON: Public Health Agency of Canada.
- Steiner, A. M., Goldsmith, T. R., Snow, A. V., & Chawarska, K. (2012). Practitioner's guide to assessment of autism Spectrum disorders in infants and toddlers. *Journal of Autism and Developmental Disorders*, *42*(6), 1183–1196. <https://doi.org/10.1007/s10803-011-1376-9>
- Taurines, R., Schwenck, C., Westerwald, E., Sachse, M., Siniatchkin, M., & Freitag, C. (2012). ADHD and autism: Differential diagnosis or overlapping traits? A selective review. *Attention Deficit and Hyperactivity Disorders*, *4*(3), 115–139. <https://doi.org/10.1007/s12402-012-0086-2>
- Tsiplova, K., Ungar, W. J., Flanagan, H. E., den Otter, J., Waddell, C., Murray, P., D'Entremont, B., Léger, N., Garon, N., Bryson, S., & Smith, I. M. (2019). Types of services and costs of programs for preschoolers with autism Spectrum disorder across sectors: A comparison of two Canadian provinces. *Journal of Autism and Developmental Disorders*, *49*(6), 2492–2508. <https://doi.org/10.1007/s10803-019-03993-3>
- Tu, K., Wang, M., Jaakkimainen, R. L., Butt, D., Ivers, N. M., Young, J., Green, D., & Jetté, N. (2014). Assessing the validity of using administrative data to identify patients with epilepsy. *Epilepsia*, *55*(2), 335–343. <https://doi.org/10.1111/epi.12506>
- Tu, K., Widdifield, J., Young, J., Oud, W., Ivers, N. M., Butt, D. A., Leaver, C. A., & Jaakkimainen, L. (2015). 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*, *15*, 67. <https://doi.org/10.1186/s12911-015-0195-x>
- Weiss, J. A., Isaacs, B., Diepstra, H., Wilton, A. S., Brown, H. K., McGarry, C., & Lunskey, Y. (2018). Health concerns and health service utilization in a population cohort of young adults with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *48*(1), 36–44. <https://doi.org/10.1007/s10803-017-3292-0>
- Zarrei, M., Burton, C. L., Engchuan, W., Young, E. J., Higginbotham, E. J., MacDonald, J. R., Trost, B., Chan, A. J. S., Walker, S., Lamoureux, S., Heung, T., Mojarad, B. A., Kellam, B., Paton, T., Faheem, M., Miron, K., Lu, C., Wang, T., Samler, K., ... Scherer, S. W. (2019). A large data resource of genomic copy number variation across neurodevelopmental disorders. *NPJ Genomic Medicine*, *4*, 26–26. <https://doi.org/10.1038/s41525-019-0098-3>
- Zerbo, O., Modaresi, S., Goddard, K., Lewis, E., Fireman, B. H., Daley, M. F., Irving, S. A., Jackson, L. A., Donahue, J. G., Qian, L., Getahun, D., DeStefano, F., McNeil, M. M., & Klein, N. P. (2018). Vaccination patterns in children after autism spectrum disorder diagnosis and in their younger siblings. *JAMA Pediatrics*, *172*(5), 469–475. <https://doi.org/10.1001/jamapediatrics.2018.0082>

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Brooks JD, Arneja J, Fu L, et al. Assessing the validity of administrative health data for the identification of children and youth with autism spectrum disorder in Ontario. *Autism Research*. 2021;14:1037–1045. <https://doi.org/10.1002/aur.2491>