

Extracting autism spectrum disorder data from the electronic health record

Ruth A. Bush^{1,2}; Cynthia D. Connelly¹; Alexa Pérez¹; Halsey Barlow¹; George J. Chiang^{3,4}

¹Hahn School of Nursing and Health Science, Beyster Institute for Nursing Research, University of San Diego, San Diego, USA;

²Clinical Research Informatics, Rady Children's Hospital-San Diego, San Diego, USA;

³Rady Children's Institute for Genomic Medicine, Rady Children's Hospital San Diego, San Diego, CA, USA;

⁴Department of Surgery, University of California-San Diego, San Diego, USA

Keywords

Autism spectrum disorder, comparative effectiveness research, electronic health record, pediatrics

Summary

Background: Little is known about the health care utilization patterns of individuals with pediatric autism spectrum disorder (ASD).

Objectives: Electronic health record (EHR) data provide an opportunity to study medical utilization and track outcomes among children with ASD.

Methods: Using a pediatric, tertiary, academic hospital's Epic EHR, search queries were built to identify individuals aged 2–18 with International Classification of Diseases, Ninth Revision (ICD-9) codes, 299.00, 299.10, and 299.80 in their records. Codes were entered in the EHR using four different workflows: (1) during an ambulatory visit, (2) abstracted by Health Information Management (HIM) for an encounter, (3) recorded on the patient problem list, or (4) added as a chief complaint during an Emergency Department visit. Once individuals were identified, demographics, scheduling, procedures, and prescribed medications were extracted for all patient-related encounters for the period October 2010 through September 2012.

Results: There were 100,000 encounters for more than 4,800 unique individuals. Individuals were most frequently identified with an HIM abstracted code (82.6%) and least likely to be identified by a chief complaint (45.8%). Categorical frequency for reported race ($\chi^2 = 816.5$, $p < 0.001$); payor type ($\chi^2 = 354.1$, $p < 0.001$); encounter type ($\chi^2 = 1497.0$, $p < 0.001$); and department ($\chi^2 = 3722.8$, $p < 0.001$) differed by search query. Challenges encountered included, locating available discrete data elements and missing data.

Conclusions: This study identifies challenges inherent in designing inclusive algorithms for identifying individuals with ASD and demonstrates the utility of employing multiple extractions to improve the completeness and quality of EHR data when conducting research.

Correspondence to:

Ruth A. Bush PhD, MPH
Hahn School of Nursing and Health Science
Beyster Institute for Nursing Research
University of San Diego, San Diego, USA
Email: Rbush@sandiego.edu

Appl Clin Inform 2017; 8: 731–741

received: February 7, 2017

accepted: May 7, 2017

published: July 19, 2017

Citation: Bush RA, Connelly CD, Pérez A, Barlow H, Chiang GJ. Extracting autism spectrum disorder data from the electronic health record. Appl Clin Inform 2017; 8: 731–741

<https://doi.org/10.4338/ACI-2017-02-RA-0029>

Funding

This project was also supported in part by grant number K99/R00 HS022404 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.

1. Background and Significance

Autism spectrum disorder (ASD) is characterized by impairments in social interaction and communication along with restricted, repetitive, and stereotyped patterns of behaviour [1], affecting as many as 11.3 per 1,000 (one in 88) children [2]. The presentation of ASD can vary widely among affected individuals and within an individual over the lifespan [3, 4]. Among the conditions associated with ASD are intellectual disability, seizure disorders, hyperactivity, disorders of the gastrointestinal and immune system, and anxiety [5–11]. Medical treatments for children with ASD are primarily directed toward alleviating the co-morbid symptoms, rather than core symptoms [11]. The evidence for the utility of involving specialty treatment, particularly gastroenterology, dietitians/nutritionists, allergy/immunology, and prescribed medication is based on results from meta-analyses of heterogeneous methodology or pooled data. Limited research examines the comparative effectiveness in daily practice of the adopted techniques or of the effectiveness of adjunct medical or outpatient development treatments (e.g., neurology, speech therapy, dietary modification, etc.), which are being used daily.

Comparative effectiveness research (CER) is “designed to inform health care decisions by providing evidence on the effectiveness, benefits, and harms of different treatment options. The evidence is generated from research studies that compare drugs, medical devices, tests, surgeries, or ways to deliver health care” [12]. Since CER analyzes the health care delivery system as a whole and includes heterogeneous populations, therapy effectiveness is measured in a natural practice setting and results can be more easily generalized than the more controlled Randomized Controlled Trial (RCT) [13, 14] CER can identify interventions that are most effective under various circumstances and can take into account provider variability, institutional volume, and regional characteristics when providing information for patients, providers, and policy makers [15–19].

2. Objectives

Previously, given the reliance on administrative billing databases, data capture limitations have prevented the longitudinal tracking needed across health care delivery systems and across time to examine the impact of ASD treatment. The use of the Electronic Health Record (EHR) is an improvement; although EHR data elements vary from system to system; certified ambulatory systems contain the following data in a discrete format: age, gender, diagnosis, medical history, medications prescribed, lab and procedure orders and results, allergies, immunizations, and vital signs [20]. Studies in adult populations, demonstrate success in extracting EHR data for research with populations of several thousand individuals with conditions such as diabetes and cardiomyopathy [21, 22]. A recently conducted claims-based case identification of ASD compared against clinical review of medical charts demonstrates a positive predictive value of almost 90% [23]. It is important to determine if pediatric EHR data can provide the data needed to examine treatment effectiveness and to conduct CER with individuals with ASD.

3. Methods

The study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research involving Human Subjects, and was reviewed and approved by the University of California, San Diego Institutional Review Board. The study was conducted at an academic, pediatric hospital and its affiliated network, which draws from three counties in Southern California. The institution uses the Epic (Madison, WI) EHR system, which incorporates emergency department (ED), inpatient, outpatient (including satellite clinics), laboratory, and radiology input into an integrated system, which shares records within the organization. The EHR system has been fully operational at this location since 2010.

To develop and pilot test methods for using EHR data to capture and to measure medical treatment utilization patterns among patients with ASD, several data query techniques were employed to draw data from several potential sources including primary care, specialty, and urgent care/ED use.

Search queries were built to create a patient list based on the presence of the related International Classification of Diseases, Ninth Revision (ICD-9) codes in their record in four different ways: (1) Clinician assigned during ambulatory visit (Encounter Dx), (2) abstracted by health information management (HIM) for an encounter after review, research, and verification of patient information and clinical data, (3) recorded on the patient problem list (Prob List), or (4) added as a chief complaint during an ED visit (Chief Complaint).

Queries were executed for all patients who were part of the EHR system, aged 2–18 with ICD-9 codes 299.00, 299.10 and 299.80 as part of a record. Children younger than 2 years of age were excluded since children are infrequently diagnosed with ASD before age 2. Initially, the query techniques, using Business Objects Crystal Reports and EHR system's Clarity database, were applied to patients treated during the period 1 October through 31 December 2010 and validated before expanding to the period 1 October 2010 through 30 September 2012. Once a list of patients was created, data for all encounters related to the patient including demographics (age, race/ethnicity, gender and payor type); encounter date; type of care (e.g. outpatient, inpatient, medication refill, etc.); provider type (e.g., physician, nurse, occupational therapist, etc.); primary ICD9 code; secondary ICD-9 codes; ICD-9 procedure codes; and prescribed medications were extracted for the two year time period. A manual comparison of the data contained in the report of 10% of the individuals identified in the report were compared against the information in the electronic health record to ensure that 1) the query captured all of the patient encounters during the time period, 2) demographic information matched, and 3) the procedures and prescribed medications were complete. The comparison determined the data pull matched the EHR records.

Using SPSS[®] version 23 [24] descriptive frequencies were run for categorical variables and analytics such as mean, median, mode, standard deviation, skewness, and kurtosis of continuous variable to identify outliers, as well as to ascertain the type and impact of missing data. Analysis was conducted to identify the number of patients with short or minimal association with the healthcare system and to guide the approach to definitions of loss to follow-up and approaches for censored data. Analysis of variance and chi-square analyses, as appropriate for variable type, were used to examine group differences.

4. Results

The extraction identified nearly 100,000 encounters for more than 4,800 unique individuals. The demographic variable are presented in ► Table 1.

ASD patient encounters were most frequently identified; 82,450 encounters (82.6%) had an HIM abstracted code, of which 17,754 were identified solely by an HIM code. Encounters were least likely to be identified using a chief complaint applied during an ED visit; 45,741 or 45.8% captured using that methodology (► Figure 1). A total of 21,585 encounters (21.6%) were identified by all four methods and the majority were captured using at least two methods. Of note 32,201 encounters were identified through only one source. The sources of identification are enumerated in ► Table 2.

The most frequent encounter types were office visits (34.5%), development services (which includes speech, occupational, and physical therapy) (27.9%), and clinicians recording emails or telephone calls with patient/parents (14.7%). The departments with the most frequent encounters were pediatrics (23.6%), speech therapy (13.3%), occupational therapy (10.0%) and neurology (7.7%), and the most common provider types were physicians (44.7%), speech therapists (14.7%) and occupational therapists (10.7%).

Based on the noted differences by source type, *chi*-squared analysis of those encounters captured using HIM assigned codes versus codes assigned by the other three methodologies (Prob List, Chief Complaint, Encounter Dx) was used to determine if there were differences in patients captured by query type. Developmental services (Dev Services), and hospital encounters (Hospital), were over-represented by using only HIM coding ($X^2 = 3722.8$, $p < 0.001$); office visits and communication with patients/parents (Communication) were more likely to be identified through a query of non-HIM sources ($X^2 = 1497.0$, $p < 0.001$) (► Figure 2).

There were also noted differences in race depending on query type (► Figure 3). Black patients and patients who refused to identify their race were underrepresented by using only HIM coding

($X^2 = 816.5$, $p < 0.001$). Whether the payor was private insurance, government reimbursement, self-pay, and indigent also differed depending on the source of the ASD coding ($X^2 = 354.1$, $p < 0.001$). There were no significant differences in gender by source of the coding.

5. Discussion

Four different data queries extracting data from the same integrated pediatric EHR system yielded substantially different results. The differences demonstrated the workflow for the diagnosis to enter a patient's record varies notably. Although administrative or billing data provides the majority of information, the information gained from clinic documentation, such as records generated in the ED, are also important sources of patient identification, particularly for those individuals lacking private insurance and add to the diversity of patients captured. EHR-derived data may not be comprehensive enough for research unless multiple sources capturing several workflows are queried.

A significant strength of this project was the ability to employ different queries within a large heterogeneous healthcare delivery system that is the primary referral source for ASD in the geographic area and to have the statistical power to compare ASD capture within the EHR. This project demonstrated it is possible to identify patients with ASD and to capture needed data to identify and to quantify associated medical conditions. Such data is critical if medical intervention for ASD is to be studied and to have sufficient strength of evidence to evaluate either their potential benefit or adverse effects [11]. Data such as these will add to the growing body of clinical guidelines on the Agency for Healthcare Research and Quality (AHRQ) National Guideline Clearinghouse available to advise clinicians and administrators about the organization, financing and delivery of services to children with ASD, [25] as well as to provide the desired patient-centered approaches for treatment that recognize family dynamics and other social factors, demand an outcomes-based analysis.

The process of assigning ICD codes is complicated. There are numerous potential sources of error affecting ICD code accuracy including the amount and quality of information at admission, communication among patients and providers, the clinicians knowledge and experience with the illness, and the clinician's attention to detail [26]. Querying one source is not enough. For example, when using an algorithm designed to identify type 2 diabetes cases in the EHR, Pacheco et al. found just over half of patients were identified by searching the problem list and Kahn and Ranade found significantly different rates from safety source data from one hospital resulting from differences in workflow practice [22, 27]. A Canadian study found when analyzing administrative health data only 7% of obese children's condition were correctly identified with this information source, which relied primarily on inpatient hospital data. The child's weight was not noted during inpatient stay and outpatient visits were not included in the analysis, so the administrative data grossly underestimated the true population prevalence of obesity [28]. Similarly, it has been demonstrated among a small cohort of pediatric asthma patients there was a significant discrepancy between the presence in the EHR of clinical features compatible with a diagnosis of asthma in EHRs, but no ICD-9 reflecting the condition [29]. Under diagnosis of health conditions has tremendous implications for health planning.

This study demonstrated using a variety of data sources within the EHR may improve the accuracy and representativeness of the information capture. While the patient's medical history is generally captured in a narrative format, tools such as "smart notes" and history templates capture the information as discrete data elements. The EHR incorporates a computerized physician order entry (CPOE) system allowing providers to manage and communicate orders and results, which are recorded electronically. Among the benefits of using these data are the current data validation programs in place. Additionally, EHR clinical users undergo substantial training in order to have access to the system and to enter data. There are numerous, programmed validation checks of the data, which provide uniformity to the data captured in addition to detailed data dictionaries and documentation of the definitions applied to the captured data. Multiple source extraction illustrated overlap of data, greater inclusiveness of data capture than from a sole source, and the ability to cross-check when multiple sources are used. The findings support the capture of multiple workflows for greater patient and condition identification.

The overall utility of extracting such patient data for adult cohort studies are supported by the findings of Wells et al. who extracted echocardiographic data from the EHR and suggest it is possi-

ble to create EHR-based cohorts for use in the study of epidemiologic and genotype-phenotype associations in diverse populations [21]. The successful methodologic approach of building queries in this pediatric population had similar results to the work of Davis et al. who also used four algorithms based on ICD-9 codes and text keywords to identify adult individuals with Multiple Sclerosis [30]. Their approach, however, began with a training database of a smaller set of known individuals as well as using medications as a query approach. Lawrence et al. also noted the value of patient identification using a combination query approach of individuals with one or more outpatient diagnosis codes of diabetes *or* a prescription for insulin [31].

The health care system studied is the sole pediatric referral health care center for two large Southern California counties and part of a third, as well as being the highest volume autism services provider in the region. While it is estimated the EHR system used captures 80% of pediatric patients in the area, there are patients who do not seek treatment from the integrated delivery system provider or only part of their treatment is captured in the EHR. In conducting cross-verification of the encounters captured against a sample of 150 known ASD patients to verify the completeness of data, it became clear that many of the patients obtain clinical care outside the integrated delivery system because of limitations of insurance, additional school-based programs, and other reasons. The limitation in ability to access and to combine data across multiple platforms has also been noted in other studies in which the EHR data may be less comprehensive than claimed if derived from only one of the clinical practices from which a patient seeks care, unless the practice is part of an integrated delivery system that is providing the patient with all of his or her ambulatory and inpatient care [20, 31, 32].

This project demonstrates it was possible to leverage routine data entry by pediatric care providers via the EHR within a diverse, regional referral pediatric healthcare to construct a large clinical data set without the burden of manual data collection in the clinical setting. Additionally, the data collected contained extensive health care data to analyze utilization patterns and characterization of current medical treatment practices within an ethnically and social-economically diverse population. This particular study population will allow for better understanding of potential subgroups in what is known to be a heterogeneous condition. The volume and validity of data suggest it is possible to use the EHR data to address relevant CER questions in a timely manner, thereby avoiding the expense, extended follow-up period, and potential reluctance of patients and their families to be randomized, which are associated with a RCT.

6. Conclusion

This study examined the availability and utility of detailed clinical and administrative data contained in the EHR system. Its methodology recognizes the interrelatedness of child health domains, and, most importantly, address the paucity of available data sources related to the medical treatment of ASD patients. Extracted data from the EHR system is potentially rich resource for conducting comparative effectiveness research and epidemiologic surveillance, including longitudinal analyses, of medical utilization among children with ASD, as well as potential changes in clinical practices patterns among ASD patients. It is important to employ a variety of data extraction methods to capture patients who enter the EHR through different clinical workflows.

Multiple Choice Questions

Using which of the following data extractions was most likely to identify an autism spectrum disorder patient?

- A) Chief complaint during Emergency Department visit
- B) Clinician assigned during ambulatory visit
- C) Abstracted by health information management for an encounter
- D) Recorded on the patient problem list

The correct answer is C. Autism Spectrum Disorder patients were most frequently identified with an HIM abstracted code from an encounter (82.6%) and least likely to be identified by a chief complaint during an ED visit (45.8%).

Clinical Relevance Statement

There are few available data sources related to the medical treatment of autism spectrum disorder patients. Electronic health record offer detailed clinical and administrative data with the potential for use in comparative effectiveness research. This study evaluates the extracted EHR data and demonstrates that a variety of extraction methods are needed to capture a robust profile of ASD clinical data.

Conflicts of Interest

The authors have no conflicts of interest to disclose.

Protection of Human Subjects

The study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research involving Human Subjects, and was reviewed by University of California, San Diego Institutional Review Board.

Acknowledgements

The authors appreciate the assistance of Cynthia Sepulveda and Anouk Bellengi for their report writing assistance.

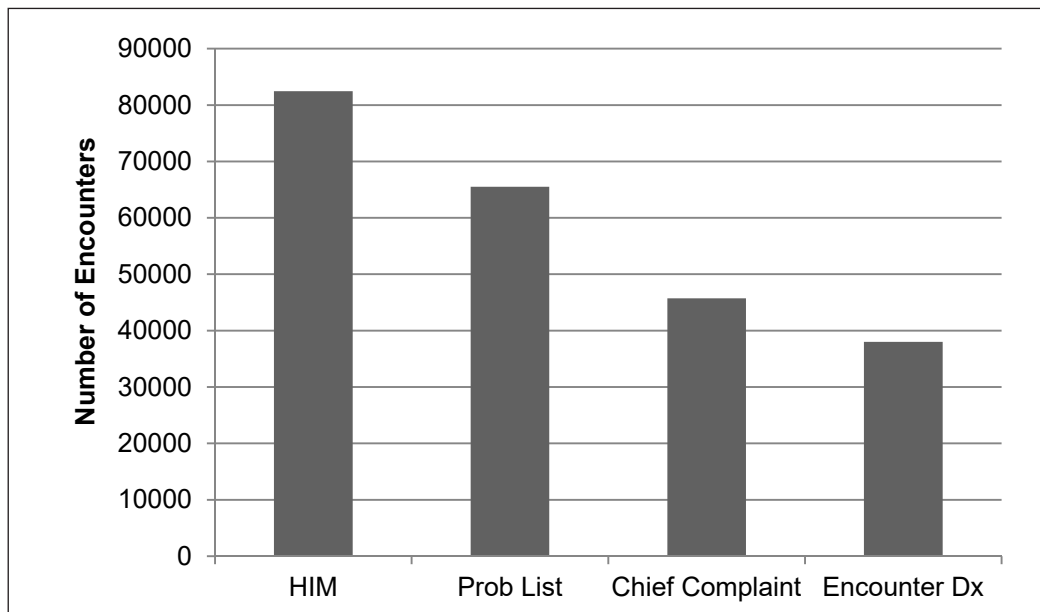


Fig. 1 Encounter Identification Source (Oct 2010 to Sep 2012; n = 99,847)

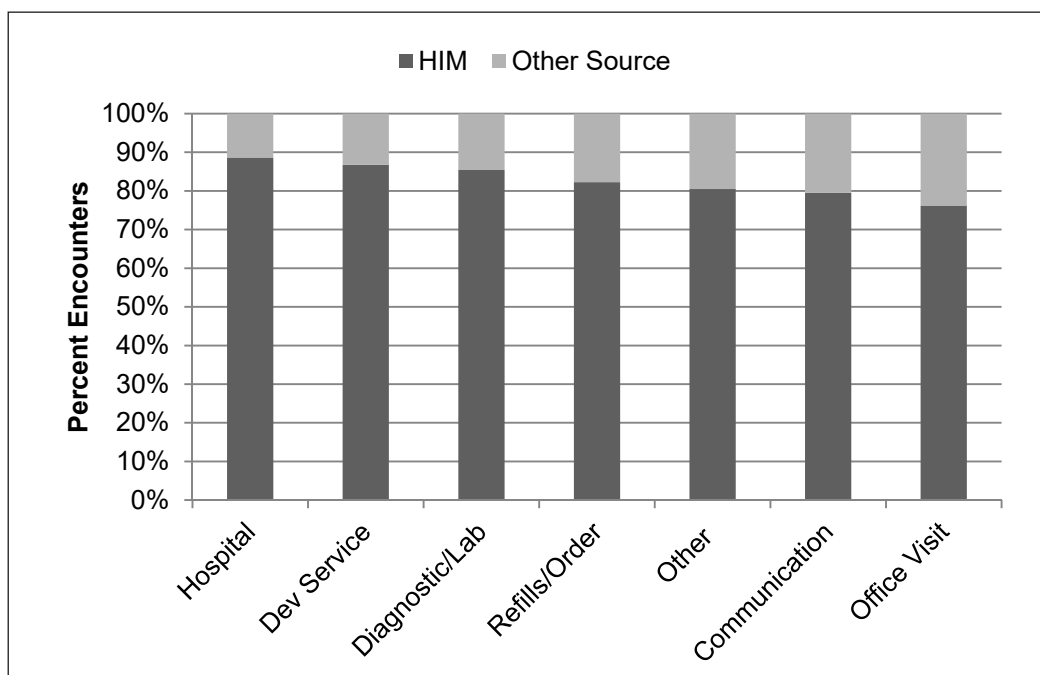


Fig. 2 Encounter Type by Source

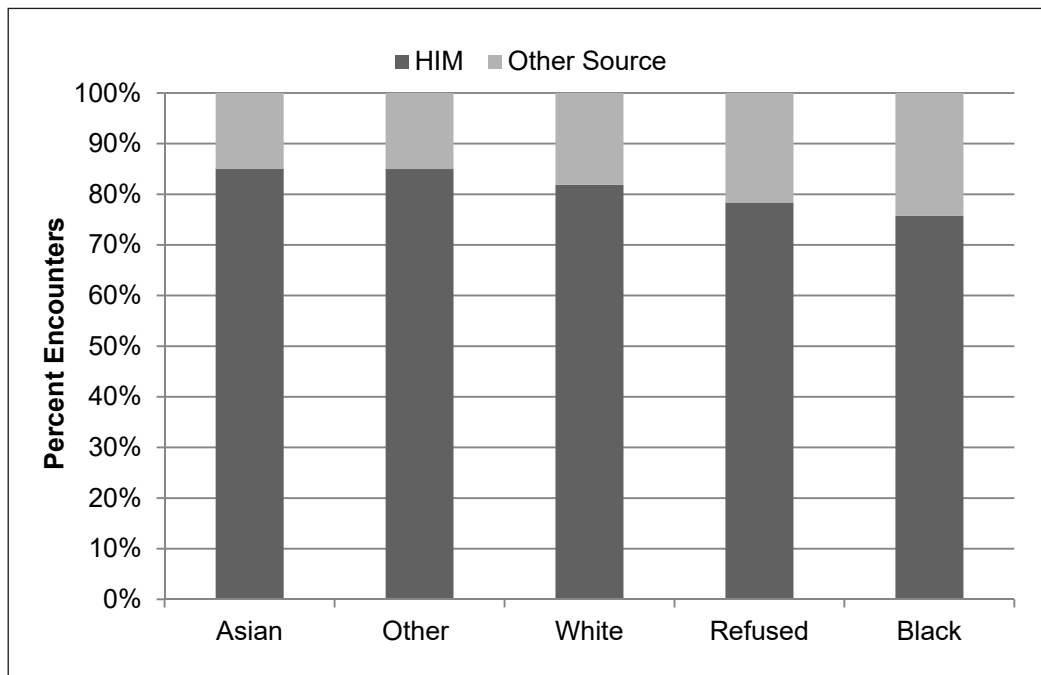


Fig. 3 Race by Source of Identification

	Mean	SD	
Age	7.22	4.2	
		Frequency	Percent
Gender	Male	77697	77.8
	Female	22150	22.2
Race	White	48538	48.6
	Other	34390	34.4
	Asian	4975	5.0
	African-American	4369	4.4
	Multi-racial	2352	2.4
	Native American/Eskimo	224	0.2
	Pac Islander/Hawaiian	132	0.1
	Missing	4867	4.9
Payor	MediCal/Medicare	42766	42.8
	Commercial	39536	39.6
	Tri-Care	12236	12.3
	Self-pay	5236	5.2
	Other	73	0.1

Table 1 Encounter Demographic Information

Table 2 Source of Encounter Identification

	Frequency	Percent	Cumulative Percent
All four sources	21585	21.6	21.6
HIM	17754	17.8	39.4
HIM, Problem List	14629	14.7	54.1
HIM, Problem List, Chief Complaint	10703	10.7	64.8
Enc Dx, HIM, Problem List	8069	8.1	72.9
Problem List	7772	7.8	80.6
Chief Complaint	5850	5.9	86.5
Enc Dx, HIM	4229	4.2	90.7
HIM, Chief Complaint	3433	3.4	94.2
Enc Dx, HIM, Chief Complaint	2269	2.3	96.4
Problem List, Chief Complaint	1707	1.7	98.2
Enc Dx, Problem List	828	0.8	99.0
Enc Dx	825	0.8	99.8
Enc Dx, Chief Complaint	194	0.2	100.0

HIM: Health Information Management, Enc Dx: Encounter Diagnosis

References

1. APA. Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR Fourth Edition (Text Revision). American Psychiatric Association; 2000.
2. Centers for Disease Control. Prevalence of autism spectrum disorders--Autism and Developmental Disabilities Monitoring Network, 14 sites, United States, 2008. *MMWR Surveill Summ* 2012; 61(3): 1–19.
3. Volkmar FR, Pauls D. Autism. *Lancet* 2003; 362(9390): 1133–1141.
4. Wing L. The autistic spectrum. *Lancet* 1997; 350(9093): 1761–1766.
5. Bauman ML. Medical comorbidities in autism: challenges to diagnosis and treatment. *Neurotherapeutics* 2010; 7(3): 320–327.
6. Buie T, Campbell DB, Fuchs GJ, Furuta GT, Levy J, Vandewater J, Whitaker AH, Atkins D, Bauman ML, Beaudet AL, Carr EG, Gershon MD, Hyman SL, Jirapinyo P, Jyonouchi H, Kooros K, Kushak R, Levitt P, Levy SE, Lewis JD, Murray KF, Natowicz MR, Sabra A, Wershil BK, Weston SC, Zeltzer L, Winter H. Evaluation, diagnosis, and treatment of gastrointestinal disorders in individuals with ASDs: a consensus report. *Pediatrics* 2010; 125 (Suppl.): S1–S18.
7. Coury D, Jones NE, Klatka K, Winklosky B, Perrin JM. Healthcare for children with autism: the Autism Treatment Network. *Curr Opin Pediatr* 2009; 21(6): 828–832.
8. Coury D. Medical treatment of autism spectrum disorders. *Curr Opin Neurol* 2010; 23(2): 131–136.
9. Gor RA, Fuhrer J, Schober JM. A retrospective observational study of enuresis, daytime voiding symptoms, and response to medical therapy in children with attention deficit hyperactivity disorder and autism spectrum disorder. *J Pediatr Urol* 2012; 8(3): 314–317.
10. Olivé H. The medical care of children with autism. *Eur J Pediatr* 2012; 171(5): 741–749.
11. McPheeters ML, Warren Z, Sathe N, Bruzek JL, Krishnaswami S, Jerome RN, Veenstra-Vanderweele J. A systematic review of medical treatments for children with autism spectrum disorders. *Pediatrics* 2011; 127(5): e1312–e1321.
12. Agency for Healthcare Research and Quality. What Is Comparative Effectiveness Research. Available from: <http://effectivehealthcare.ahrq.gov/index.cfm/what-is-comparative-effectiveness-research/1/>
13. Abdullah F, Ortega G, Islam S, Barnhart DC, St Peter SD, Lee SL, Glynn L, Teitelbaum DH, Arca MJ, Chang DC. Outcomes research in pediatric surgery. Part 1: overview and resources. *J Pediatr Surg* 2011; 46(1): 221–225.
14. Chang DC, Rhee DS, Papandria D, Aspelund G, Cowles RA, Huang EY, Chen C, Middlesworth W, Arca MJ, Abdullah F. Outcomes research in pediatric surgery. Part 2: how to structure a research question. *J Pediatr Surg* 2011; 46(1): 226–231.
15. Tunis SR, Benner J, McClellan M. Comparative effectiveness research: Policy context, methods development and research infrastructure. *Stat Med* 2010; 29(19): 1963–1976.
16. Birkmeyer JD, Stukel TA, Siewers AE, Goodney PP, Wennberg DE, Lucas FL. Surgeon volume and operative mortality in the United States. *N Engl J Med* 2003; 349(22): 2117–2127.
17. Birkmeyer JD, Siewers AE, Finlayson EVA, Stukel TA, Lucas FL, Batista I, Welch HG, Wennberg DE. Hospital volume and surgical mortality in the United States. *N Engl J Med* 2002; 346(15): 1128–1137.
18. Hall BL, Bilimoria KY, Ko CY. Investigations using clinical data registries: observational studies and risk adjustment. *Surgery* 2009; 145(6): 602–610.
19. Gabriel SE, Normand S-LT. Getting the methods right--the foundation of patient-centered outcomes research. *N Engl J Med* 2012; 367(9): 787–790.
20. West SL, Johnson W, Visscher W, Kluckman M, Qin Y, Larsen A. The challenges of linking health insurer claims with electronic medical records. *Health Informatics J* 2014; 20(1): 22–34.
21. Wells QS, Farber-Eger E, Crawford DC. Extraction of echocardiographic data from the electronic medical record is a rapid and efficient method for study of cardiac structure and function. *J Clin Bioinforma* 2014; 4: 12.
22. Pacheco JA, Thompson W, Kho A. Automatically detecting problem list omissions of type 2 diabetes cases using electronic medical records. *AMIA Annu Symp Proc* 2011; 2011: 1062–1069.
23. Burke JP, Jain A, Yang W, Kelly JP, Kaiser M, Becker L, Lawer L, Newschaffer CJ. Does a claims diagnosis of autism mean a true case? *Autism* 2014; 18(3): 321–330.
24. IBM Corp. Released. IBM SPSS Statistics for Windows, Version 23.0. 2015. 2015
25. AHRQ. National Guideline Clearing House 2012. Available from: <http://guideline.gov/index.aspx>
26. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. *Health Serv Res* 2005; 40(5 Pt 2): 1620–1639.
27. Kahn MG, Ranade D. The impact of electronic medical records data sources on an adverse drug event quality measure. *J Am Med Inform Assoc* 2010; 17(2): 185–191.

28. Kuhle S, Kirk SFL, Ohinmaa A, Veugelers PJ. Comparison of ICD code-based diagnosis of obesity with measured obesity in children and the implications for health care cost estimates. *BMC Med Res Methodol* 2011; 11: 173.
29. Juhn Y, Kung A, Voigt R, Johnson S. Characterisation of children's asthma status by ICD-9 code and criteria-based medical record review. *Prim Care Respir J* 2011; 20(1): 79–83.
30. Davis MF, Sriram S, Bush WS, Denny JC, Haines JL. Automated extraction of clinical traits of multiple sclerosis in electronic medical records. *J Am Med Inform Assoc* 2013; 20(e2): e334–e340.
31. Lawrence JM, Black MH, Zhang JL, Slezak JM, Takhar HS, Koebnick C, Mayer-Davis EJ, Zhong VW, Dabelea D, Hamman RF, Reynolds K. Validation of pediatric diabetes case identification approaches for diagnosed cases by using information in the electronic health records of a large integrated managed health care organization. *Am J Epidemiol* 2014; 179(1): 27–38.
32. Wasserman RC. Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research. *Acad Pediatr* 2011; 11(4): 280–287.