

## Special Article

# Identifying cases of spinal cord injury or disease in a primary care electronic medical record database

John Shepherd <sup>1</sup>, Karen Tu <sup>2,3,4,5</sup>, Jacqueline Young <sup>6</sup>, Jawad Chishtie <sup>1</sup>,  
B. Catharine Craven <sup>3,7,8</sup>, Rahim Moineddin <sup>2,6,9</sup>, Susan Jaglal <sup>3,6,7,10</sup>

<sup>1</sup>Rehabilitation Sciences Institute, University of Toronto, Toronto, Ontario, Canada, <sup>2</sup>Department of Family and Community Medicine, University of Toronto, Toronto, Ontario, Canada, <sup>3</sup>Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada, <sup>4</sup>North York General Hospital, Toronto, Ontario, Canada, <sup>5</sup>Toronto Western Hospital Family Health Team, University of Toronto, Toronto, Ontario, Canada, <sup>6</sup>Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada, <sup>7</sup>KITE, Toronto Rehab – University Health Network, Toronto, Ontario, Canada, <sup>8</sup>Division of Physical Medicine and Rehabilitation, Department of Medicine, University of Toronto, Toronto, Ontario, Canada, <sup>9</sup>Dalla Lana School of Public Health, Toronto, Ontario, Canada, <sup>10</sup>Department of Physical Therapy, University of Toronto, Toronto, Ontario, Canada

**Objective:** To identify cases of spinal cord injury or disease (SCI/D) in an Ontario database of primary care electronic medical records (EMR).

**Design:** A reference standard of cases of chronic SCI/D was established via manual review of EMRs; this reference standard was used to evaluate potential case identification algorithms for use in the same database.

**Setting:** Electronic Medical Records Primary Care (EMRPC) Database, Ontario, Canada.

**Participants:** A sample of 48,000 adult patients was randomly selected from 213,887 eligible patients in the EMRPC database.

**Interventions:** N/A.

**Main Outcome Measure(s):** Candidate algorithms were evaluated using sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F-score.

**Results:** 126 cases of chronic SCI/D were identified, forming the reference standard. Of these, 57 were cases of traumatic spinal cord injury (TSCI), and 67 were cases of non-traumatic spinal cord injury (NTSCI). The optimal case identification algorithm used free-text keyword searches and a physician billing code, and had 70.6% sensitivity (61.9–78.4), 98.5% specificity (97.3–99.3), 89.9% PPV (82.2–95.0), 94.7% NPV (92.8–96.3), and an F-score of 79.1.

**Conclusions:** Identifying cases of chronic SCI/D from a database of primary care EMRs using free-text entries is feasible, relying on a comprehensive case definition. Identifying a cohort of patients with SCI/D will allow for future study of the epidemiology and health service utilization of these patients.

**Keywords:** Spinal cord injury, Primary health care, Case identification, Electronic medical records

Spinal cord injury or disease (SCI/D) is damage to the spinal cord resulting in neurological impairment; it can

be caused either by external injury, known as traumatic spinal cord injury (TSCI), or by other causes such as congenital conditions or disease processes, known as non-traumatic spinal cord injury (NTSCI).<sup>1</sup> Historically, health services research in the field of “spinal cord injury” has focused only on TSCI, but

Correspondence to: John Shepherd, Rehabilitation Sciences Institute, University of Toronto, 500 University Ave, Toronto, Ontario, Canada. Email: john.shepherd@mail.utoronto.ca

Supplemental data for this article can be accessed on the publisher's website: <https://doi.org/10.1080/10790268.2021.1971357>.

recent efforts to classify the potential causes of NTSCI have made it feasible to identify and include non-traumatic cases.<sup>2</sup> This is important as the incidence of NTSCI is expected to increase with an aging population.<sup>3</sup>

Although prevalence estimates are sparse and inconsistent, SCI/D is typically considered a rare disorder;<sup>4</sup> however, it has a disproportionately large impact on healthcare utilization<sup>5</sup> due to the many associated impairments and co-morbidities<sup>6</sup> experienced by patients during the decades of life post-injury.<sup>7</sup> Most research in SCI/D has focused on the early experience of patients, during acute in-hospital care and initial specialized rehabilitation, while their experience of the years of life post-discharge is not well described.<sup>8–10</sup> In recent years, administrative datasets have been used to address this gap;<sup>11–14</sup> however, these data sources typically capture only incident cases over a particular time period and cannot be used to identify prevalent cases of chronic SCI/D. Lack of a clear case definition for SCI/D adds complexity; this is particularly so for NTSCI, which comprises a wide range of rare conditions.<sup>2</sup>

Following the widespread adoption of electronic medical records (EMR), databases of EMRs aggregated from primary care practices have become available for research.<sup>15,16</sup> Methods for identifying various conditions of interest within these databases have been described, employing evidence such as disease classification codes, specific medications, and key diagnostic test or lab values.<sup>17–21</sup> However, when implementing a similar case identification method in SCI/D, the types of evidence (such as disease codes) used in other conditions may be unreliable,<sup>22,23</sup> so that the use of additional material from free-text fields in the EMR may be necessary.

An advantage of using primary care rather than hospital records is the possibility of identifying prevalent cases, creating a cohort that is more representative of the SCI/D population overall. Additionally, the capture of mild cases of NTSCI can be improved, since these cases may not result in hospitalization.<sup>24</sup> Once a cohort is reliably identified, it can be linked to other datasets and used to study the long-term experience of people living with SCI/D.

Our objectives were to (i) develop a method for identifying cases of chronic SCI/D in a primary care EMR database using a comprehensive case definition; (ii) conduct a detailed manual chart review in order to establish a reference standard cohort of cases of chronic SCI/D; and (iii) use the reference standard cohort to validate case finding algorithms that can be used to identify all cases of SCI/D in the EMR database.

## Methods

EMR data were extracted from the Electronic Medical Record Primary Care (EMRPC) database (formerly known as EMRALD). EMRPC aggregates data from the EMR systems of participating family physicians across Ontario, Canada, a province with a population of approximately 14.5 million and a single-payer health-care system. Data were collected semi-annually from primary care clinics that volunteer to participate. At the time of data extraction for this study (December 31, 2016), EMRPC comprised 376 family physicians (approximately 3% of the family physicians in Ontario) in 43 clinics, and 443,038 patients. EMRPC is among the largest EMR databases in the world (those representing more than 1% of their reference population)<sup>25</sup> and is unusual in that it contains large amounts of free-text data.<sup>16</sup> EMRPC contains all clinically relevant information aggregated from EMRs (see [Table 1](#)).

EMRPC is maintained by ICES, an independent, non-profit research institute funded by an annual grant from the Ontario Ministry of Health (MOH) and the Ministry of Long-Term Care (MLTC).<sup>26</sup> As a prescribed entity under Ontario's privacy legislation, ICES is authorized to collect and use health care data for the purposes of health system analysis, evaluation and decision support. Secure access to these data is governed by policies and procedures that are approved by the Information and Privacy Commissioner of Ontario. The use of the data in this project is authorized under section 45 of Ontario's Personal Health Information Protection Act (PHIPA) and does not require review by a Research Ethics Board.

The study cohort included patients who had a valid date of birth and a valid health insurance number, were assigned to an active primary care physician, and were more than 14 years old as of the date of the data extraction. We excluded the pediatric population to facilitate chart review, with the age cut-off of 14 years chosen to align with other major studies of this population.<sup>3,4</sup> Patients were excluded if they had less than one year of entries in the EMR, or if their physician had been using the EMR for less than 2 years before data extraction, in order to ensure accuracy and completeness of data. After application of the inclusion and exclusion criteria, 213,887 patients comprised the study cohort. A random sample of 48,000 patients was selected from the study cohort to generate a feasible number of charts for manual review. The study was conducted and reported according to recommended guidelines for reporting of case definition validation studies.<sup>17</sup>

**Table 1 Components of EMRPC database.**

Component	Author
<ul style="list-style-type: none"> <li>• <b>Cumulative Patient Profile (CPP)</b> <ul style="list-style-type: none"> <li>○ <b>Problem list</b></li> <li>○ <b>Medical history</b></li> <li>○ Risk factors</li> <li>○ Allergies</li> <li>○ Medications</li> </ul> </li> <li>• <b>Progress notes (PN)</b></li> </ul>	Primary care physician
<ul style="list-style-type: none"> <li>• <b>Consultation letters (CONS)</b></li> <li>• Discharge summaries</li> <li>• Diagnostic procedures</li> <li>• Prescriptions</li> </ul>	

**Table 2 Initial search terms.**

spinal cord injury  
 paraplegia/-paresis,  
 tetraplegia/-paresis,  
 quadriplegia/-paresis  
 quadraplegia/-paresis (*common misspellings*)  
 hemiparaplegia  
 anterior cord syndrome  
 central cord syndrome  
 posterior cord syndrome  
 conus medullaris syndrome  
 brown sequard syndrome  
 neurogenic bladder  
 neurogenic bowel

**Keyword search strategy**

A preliminary keyword search was used to identify potential cases of SCI/D for manual review from the 48,000 patient study sample. Care was taken to ensure that the keyword search strategy was inclusive of all potential cases. Previous studies employing the same database to identify other conditions of interest used a similar keyword search approach.<sup>18,27,28</sup>

A list of 17 search terms related to SCI/D including associated impairments and syndromes (Table 2) was developed in consultation with a primary care physician (with expertise in treating SCI/D patients) and used to conduct an initial keyword search in certain free-text components of patient records: the cumulative patient profile (CPP), progress notes (PN), and consultation letters (CONS). Within the CPP, the problem list and medical history sections were considered.

It was hypothesized that these terms would risk missing some NTSCI cases because primary care physicians may identify them by the etiology rather than the impairment; this was confirmed when a preliminary review of 100 patients identified using a list of NTSCI-related etiologies revealed 5 possible cases. Therefore, in order to maximize capture of potential NTSCI cases, an exhaustive list of NTSCI-related

keywords including etiologies was compiled using authoritative sources of NTSCI-related terms.<sup>2,29</sup> The inclusion of an additional 126 search terms in the CPP problem list was considered warranted as problem lists maintained by primary care physicians have been shown to be well-maintained and accurate.<sup>30-32</sup>

For the list of terms in the keyword search strategy, see Supplemental Table 1.

**Case definition**

A comprehensive case definition for identifying chronic SCI/D for chart review was developed (Table 3), based on prior work to characterize TSCI and the diagnostic criteria for NTSCI.<sup>22,24</sup>

This definition was designed to be maximally comprehensive, and includes spina bifida, which is sometimes treated as a separate entity.<sup>4</sup> All cases of multiple sclerosis were excluded.

**Chart review**

During the preliminary chart review process, the reliability of different chart components was noted. A mention of a keyword in the CPP problem list was more likely to identify a case accurately than a mention in progress notes (PN), which contained varied information and were therefore more likely to generate spurious keyword matches resulting from, for example, queries, differential diagnoses, and family histories. The use of diagnostic terms in consultation letters (CONS) tended to be clear and unambiguous, and consultation letters from certain specialties such as neurosurgery, neurology, and psychiatry often contained additional evidence to validate the diagnosis, including reports of neurological examinations. Consultation letters from urology often clearly documented neurogenic bladder. The chart review process was designed to take these observations into account

**Table 3 Case definition: criteria for chronic TSCI and NTSCI case identification.**

Criterion	TSCI	NTSCI	NTSCI (due to degeneration)
Type of neurological impairment	Any of motor, sensory, bowel or bladder impairment	Motor or sensory impairment AND bowel or bladder impairment	Only motor impairment (subject to review)
Duration of impairment		Any duration	
Injury site		Spinal cord and cauda equina (but not nerve roots)	
Cause/etiology	External physical force	Etiologies as identified by New <sup>2</sup> : congenital, genetic, and acquired (non-degenerative)	Etiologies as identified by New <sup>2</sup> : acquired degenerative
Point of maximal impairment	Typically immediately following surgery/hospitalization	Variable	Often immediately preceding surgery/hospitalization

in the evidentiary value accorded to different EMR components.

EMRPC chart data was reviewed via a custom online abstraction platform which has previously been used in similar studies.<sup>27,33</sup> The chart review process is shown in Fig. 1.

All charts were reviewed by a principal reviewer (JS), while a secondary reviewer (JC) validated a random sample of 10% of the charts, and reviewed all cases rated “possible” by the first reviewer, reclassifying them where possible. Cohen’s kappa for the 10% validation sample indicated strong agreement at 87.37%. Complex cases were referred to a senior clinical expert (CC) for adjudication.

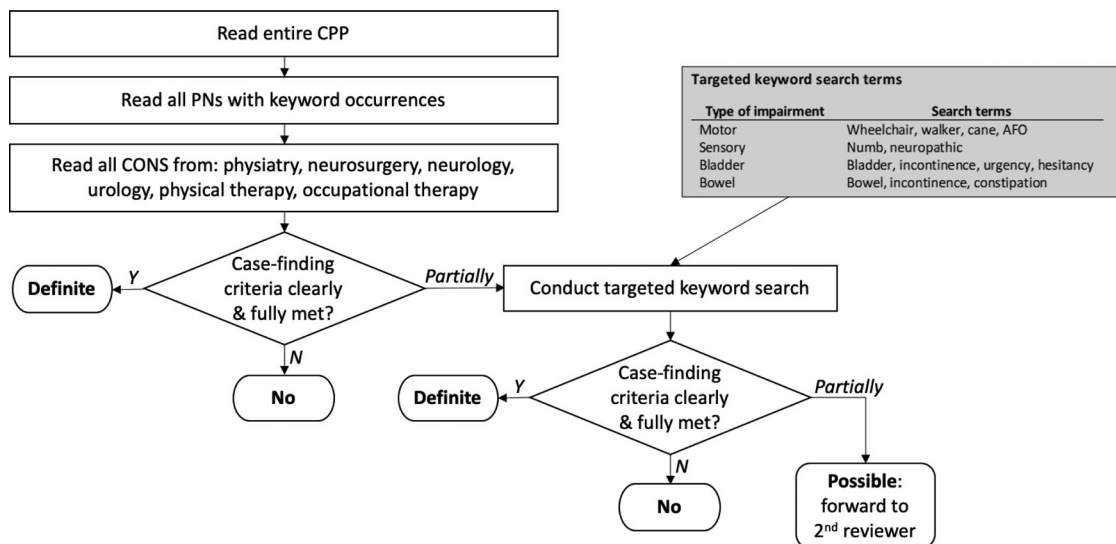
**Algorithm development**

The reference standard cohort contained all definite cases of chronic SCI/D in the 48,000-patient study sample. In order to identify cases of chronic SCI/D in the larger study cohort, algorithms were developed,

and their performance was evaluated using the reference standard. These algorithms were composed of text strings (keywords, names of medications) and physician billing codes that, if found when searching the EMRPC database, could accurately identify cases of SCI/D.

Given the different significance and evidentiary value of keywords depending on where they appeared in the chart, these were considered separately in the CPP and in the PN and CONS. In both cases, to determine the value of each keyword for case identification, occurrences in the reference standard charts were analyzed. Based on this analysis, keyword lists were constructed for use in algorithm development.

In order to identify medications commonly used in the SCI/D population, we consulted the available literature on pharmacological use in SCI/D.<sup>34-38</sup> However, this was of limited value due to the non-specific nature of most medications commonly used in this population, such as bowel agents, analgesics,



**Figure 1 Chart review process.**

antidepressants, and antibiotics. As a second step, a database of indication-medication pairs<sup>39</sup> was consulted to determine which drugs were most commonly prescribed for SCI/D. Of these, the reverse pairs (medication-indication) were then examined to determine which drugs were prescribed more commonly for SCI/D than for other indications. From this analysis, three medications were included as an algorithm component: oxybutynin, baclofen, and dantrolene (the brand names Ditropan, Lioresal, and Dantrium were also included).

ICD-9 disease classification codes related to SCI/D were identified in the literature<sup>22,23</sup> and were used as another algorithm component. The three ICD-9 codes (344, 806, and 952) that corresponded to physician billing codes in EMRPC were used.

Using the reference standard cohort as the criterion, confusion matrix values (true positives, true negatives, false positives, false negatives) were tabulated for each algorithm component, and the sensitivity, specificity, PPV, NPV, and F-scores were calculated using Microsoft SQL and Excel. Individual algorithm components were then combined in various permutations to create case identification algorithms, and the performance of these combination algorithms was similarly evaluated. For this study, sensitivity and PPV were considered to be primary outcomes (to select

algorithms able to identify all cases of SCI/D, i.e. sensitivity, and only cases of SCI/D, i.e. PPV). F-score, the harmonic mean of sensitivity and PPV, was included as a single summary metric of performance. Performance metrics for all algorithms (including specificity and NPV) are reported in Supplemental Table 2.

## Results

### Chart review and case identification

Through the manual review of electronic medical records, this study identified 126 cases of chronic SCI/D. Using this cohort as a reference standard, case identification algorithms were developed and evaluated. The overall process is depicted, and key results summarized, in Fig. 2. The 126 validated SCI/D cases in the shaded box became the reference standard cohort (RSC).

The demographic characteristics of the RSC compared with the study cohort are presented in Table 4; results are shown for the RSC as a whole, and for TSCI and NTSCI cases separately (two cases could not be classified as either TSCI or NTSCI).

### Algorithm development

#### Keyword analysis

The case identification algorithms developed made use of different types of information in the EMR database.

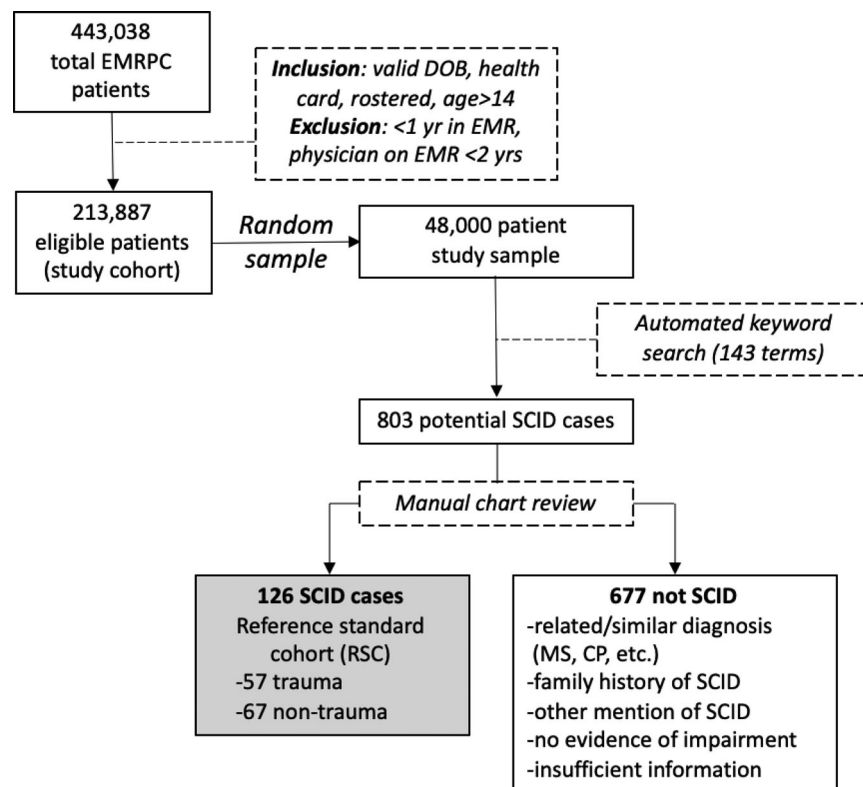
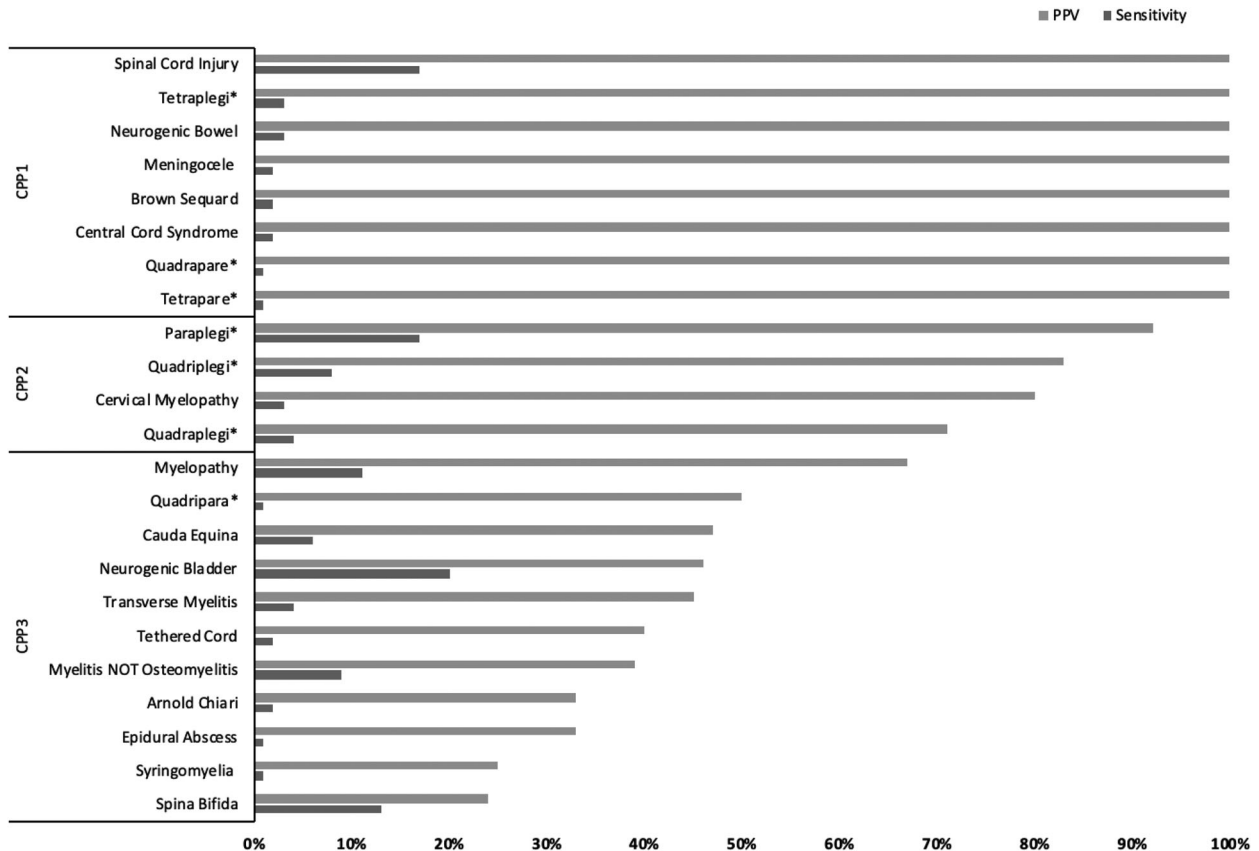


Figure 2 Flowchart for SCI/D case identification.

**Table 4 Demographic characteristics.**

	RSC-Total (n = 126)	RSC-TSCI (n = 57)	RSC-NTSCI (n = 67)	Study cohort (n = 213,887)
Sex				
Female, n (%)	50 (39.7)	16 (28.1)	34 (50.7)	124,776 (58.3)
Male, n (%)	76 (60.3)	41 (71.9)	33 (49.3)	89,111 (41.7)
Age				
Mean age, years (sd)	55.1 (17.0)	55.6 (13.8)	54.3 (19.2)	50.8 (19.1)
Age 14–34, n (%)	16 (12.7)	5 (8.8)	11 (16.4)	50,357 (23.5)
Age 35–64, n (%)	74 (58.7)	39 (68.4)	34 (50.7)	108,946 (50.9)
Age ≥65, n (%)	36 (28.6)	13 (22.8)	22 (32.8)	54,584 (25.5)



**Figure 3 CPP keyword analysis.**

The CPP component identified charts with a keyword match in the CPP (specifically in the problem list and medical history portions). Similarly, the P/C component identified charts with a keyword match in the progress notes (PN) or consultation letters (CONS).

In order to determine which keywords to use in these algorithm components, the occurrence of keywords in the charts of the RSC was analyzed. Because a given keyword (or keyword phrase) can occur at most once in the CPP, it was possible to calculate the sensitivity and PPV of each individual CPP keyword. This information, shown in Fig. 3, was used to construct keyword lists for use in case identification algorithms. The first list of CPP keyword terms, called CPP1, contains eight

keywords all of which have a PPV of 100%, meaning that these keywords appeared only in the CPP of true positive cases. The next two lists added further terms in descending order of PPV, using 70% as a breakpoint; grouping the terms in this way facilitated the evaluation of different algorithm permutations with respect to the tradeoff between sensitivity and predictive value.

In the case of keywords occurring in the P/C components, it was possible for a given keyword to occur multiple times in the same record, making it infeasible to use the same analysis as with CPP keywords. Rather, the total numbers of keyword occurrences were tabulated, and the occurrences in cases were compared with those in non-cases to determine the value of each

**Table 5 Terms included in algorithm components.**

Component	Description	Terms included	Terms excluded
CPP1	Keywords in CPP (8)	brown sequard central cord syndrome meningocele neurogenic bowel quadrapare* spinal cord injury tetrapare* tetraplegi*	cerebral palsy CP guillain barre MS (MS cases identified using EMRPC MS algorithm)
CPP2	Keywords in CPP (12)	CPP1+ cervical myelopathy paraplegi* quadraplegi* quadriplegi*	cerebral palsy CP guillain barre MS (MS cases identified using EMRPC MS algorithm)
CPP3	Keywords in CPP (23)	CPP2+ arnold chiari cauda equina cervical myelopathy epidural abscess myelitis myelopathy neurogenic bladder quadripare* spina bifida syringomyelia tethered cord transverse myelitis	cerebral palsy CP guillain barre MS (MS cases identified using EMRPC MS algorithm) occulta osteomyelitis
P/C	Keywords in progress notes and consultation letters (13)	brown sequard central cord syndrome hemiparaplegia neurogenic bowel parapare* paraplegi* quadrapare* quadraplegi* quadripare* quadriplegi* spinal cord injury tetrapare* tetraplegi*	n/a

Note: "\*" indicates a wildcard character.

keyword for case identification. This algorithm component could specify either a certain number of unique terms from the list of keywords (e.g. three different terms) or a certain number of occurrences of any of the terms from the list (e.g. three occurrences of any term, including multiple occurrences of the same term).

The complete lists of inclusion and exclusion terms for CPP and P/C are provided in Table 5.

**Algorithm component evaluation**

Algorithm components using the three CPP keyword lists illustrate the tradeoff between sensitivity and PPV. The best overall CPP algorithm component is CPP2 with an F-score of 67.4%. This component has

an excellent PPV of 94.3% but the sensitivity of 52.4% is inadequate.

Algorithm components using keywords found in progress notes and consultation letters demonstrated a similar tradeoff between sensitivity and PPV. Overall, the best-performing of these achieved F-scores comparable to the best-performing CPP algorithm components, which is a significant result since this method (using keywords from progress notes and consultation letters in case identification algorithms) has not been documented in previous studies.

The performance of algorithm components using medication names and physician billing codes was also evaluated. The medication name algorithm

component performed poorly, with a sensitivity of only 29.4% and PPV of 47.4%. Occurrences of physician billing codes were analyzed using the same method as in other studies: looking for single or multiple (“x2,” “x3,” etc.) occurrences, and also looking for multiple occurrences in a single year (“x2in1”). The only ICD-9 code that identified a significant number of cases was 806 (“fracture of vertebral column with spinal cord injury”), with 20 true positives and 0 false positives. Since there were no false positives, there was no improvement in performance from requiring multiple occurrences.

The other code that identified SCI/D cases was 344 (“other paralytic syndromes”), with three true positives and zero false positives; however, all three of the true positive cases also used the 806 code, so there was no incremental value in using both rather than using 806 alone. For the sake of parsimony ICD-9 code 344 was dismissed in further analysis. ICD-9 code 952 (“spinal cord injury without evidence of spinal bone injury”) did not identify any true positive cases and was dismissed in further analysis.

### Algorithm performance evaluation

Using various permutations of the algorithm components, 125 candidate algorithms were created, and their performance evaluated. By combining algorithm components, a synergistic improvement in performance was obtained, increasing true positives and hence sensitivity without commensurately increasing false negatives and diminishing PPV. While the highest F-score of any single component is 68.1%, an optimal combination of components can yield an F-score as high as 79.1%.

The optimal algorithm selects cases that have any of three components: (i) a CPP keyword on the CPP2 list; or (ii) three or more occurrences of any of the keywords in the P/C list; or (iii) an occurrence of ICD-9 code 806. This algorithm has a sensitivity of 70.6% and a PPV of 89.9%.

Table 6 presents the performance of selected algorithm components and algorithms. For a list of all algorithms considered, see Supplemental Table 2.

### Discussion

This study has demonstrated that it is possible to reliably identify cases of chronic SCI/D in a primary care EMR database using a detailed case definition, a comprehensive keyword search strategy, and a rigorous manual chart review process. With respect to the development of algorithms for use in the automated identification of cases of SCI/D in EMRPC (and potentially in

other similar EMR databases), the precise methods demonstrated in previous studies (using disease codes and prescriptions) proved infeasible in SCI/D. However, it was possible to construct satisfactory case identification algorithms by making more extensive use of free-text keyword searching in the unstructured portions of the EMRPC database.

Unlike many other chronic conditions where diagnosis or billing codes (using a standard classification like ICD-9 or ICD-10) can be used, the relevant codes for SCI/D were found to be unreliable, with only 15.9% of the RSC showing an SCI/D-related physician billing code. This is perhaps because primary care visits by patients with SCI/D do not typically result in an SCI/D-related billing code. When working with a different sample, it would be prudent to re-examine the validity of disease classification codes.

Furthermore, prescriptions and lab test values were similarly unhelpful, although they have been shown to be reliable in identifying cases of other conditions such as MS and diabetes.<sup>18,20</sup> With respect to the three medications identified as being both relatively widely used within and specific to SCI/D, only 29.4% of the RSC cases included a prescription for at least one of these; furthermore, 89 false positives were identified, for a PPV of 47.4% for the medication algorithm component.

In order to improve algorithm performance, additional unstructured free-text elements of the EMRPC database such as progress notes and consultation letters were used. Occurrences in these fields of highly specific terms such as “tetraplegia,” “neurogenic bowel,” and “spinal cord injury” proved to be a valuable addition to the contents of the CPP for the purpose of case identification. This approach leveraged the unique inclusion in EMRPC of these substantial free-text elements.

With respect to previous case identification studies in SCI/D,<sup>22,24,40,41</sup> this is the first to use EMRs as a data source, and the first to use the information other than diagnostic codes (such as ICD-9 or ICD-10). The algorithms developed in this study performed similarly to those developed to identify other disease populations using the EMRPC database.<sup>18,19,27,28</sup> The sensitivity of the optimal algorithm (70.6%) was lower than in some of these studies, which may be attributed to the difficulty of capturing every instance of a diagnosis (SCI/D) that can be designated in many different ways, particularly for cases of NTSCI. The PPV of the optimal algorithm (89.9%), however, is compared favorably. When compared with a broader sample of 40 case-finding studies covering 47 different conditions, using a variety of



**Table 6 Performance of selected algorithms.**

Description	TP	TN	FN	FP	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	F-score
<b>A. Algorithm Components</b>									
Keywords from 1 of 3 lists of terms found in Cumulative Patient Profile (CPP)									
CPP1 (8 keywords included, 4 excluded)	40	677	86	0	31.7 (23.7-40.6)	100 (99.5-100)	100 (91.2-100)	88.7 (86.3-90.9)	48.2
CPP2 (12 included, 4 excluded)	66	673	60	4	52.4 (43.3-61.3)	99.4 (98.5-99.8)	94.3 (86.0-98.4)	91.8 (89.6-93.7)	67.3
CPP3 (23 included, 6 excluded)	95	598	31	79	75.4 (66.9-82.6)	88.3 (85.7-90.7)	54.6 (46.9-62.1)	95.1 (93.1-96.6)	63.3
Keywords from single list of 13 terms found in Progress Notes and Consultation Letters (P/C)									
≥ 5 unique keywords	6	677	120	0	4.8 (1.8-10.1)	100 (99.5-100)	100 (54.1-100)	84.9 (82.3-87.4)	9.1
≥ 4 unique keywords	14	677	112	0	11.1 (6.2-17.9)	100 (99.5-100)	100 (76.8-100)	85.8 (83.2-88.2)	20.0
≥ 3 unique keywords	28	675	98	2	22.2 (15.3-30.5)	99.7 (98.9-100)	93.3 (77.9-99.2)	87.3 (84.8-89.6)	35.9
≥ 5 keyword occurrences (any)	48	674	78	3	38.1 (29.6-47.2)	99.6 (98.7-99.9)	94.1 (83.8-98.8)	89.6 (87.2-91.7)	54.2
≥ 4 keyword occurrences (any)	58	672	68	5	46.0 (37.1-55.1)	99.3 (98.3-99.8)	92.1 (82.4-97.4)	90.8 (88.5-92.8)	61.4
≥ 2 unique keywords	63	668	63	9	50.0 (41.0-59.0)	98.7 (97.5-99.4)	87.5 (77.6-94.1)	91.4 (89.1-93.3)	63.6
≥ 3 keyword occurrences (any)	68	669	58	8	54.0 (44.9-62.9)	98.8 (97.7-99.5)	89.5 (80.3-95.3)	92.0 (89.8-93.9)	67.3
≥ 2 keyword occurrences (any)	78	652	48	25	61.9 (52.8-70.4)	96.3 (94.6-97.6)	75.7 (66.3-83.6)	93.1 (91.0-94.9)	68.1
<b>Medications prescribed</b>									
Medication name (oxybutynin, baclofen, dantrolene sodium)	37	636	89	41	29.4 (21.6-38.1)	93.9 (91.9-95.6)	47.4 (36.0-59.1)	87.7 (85.1-90.0)	36.3
<b>ICD-9 Billing codes</b>									
344	3	677	123	0	2.4 (0.5-6.8)	100 (99.5-100)	100 (29.2-100)	84.6 (81.9-87.1)	4.7
806	20	677	106	0	15.9 (10.0-23.4)	100 (99.5-100)	100 (83.2-100)	86.5 (83.9-88.8)	27.4
806 x 2	16	677	110	0	12.7 (7.4-19.8)	100 (99.5-100)	100 (79.4-100)	86.0 (83.4-88.4)	22.5
806 x 2 in 1 year	11	677	115	0	8.7 (4.4-15.1)	100 (99.5-100)	100 (71.5-100)	85.5 (82.8-87.9)	16.1
952	0	677	126	0	0.0 (0.0-2.9)	100 (99.5-100)	0.0 (0.0-100)	84.3 (81.6-86.8)	0.0
<b>B. Top Performing Algorithms</b>									
CPP2 <i>or</i> P/C ≥ 3 keywords (any) <i>or</i> Billing code 806	89	667	37	10	70.6 (61.9-78.4)	98.5 (97.3-99.3)	89.9 (82.2-95.0)	94.7 (92.8-96.3)	79.1
CPP2 <i>or</i> P/C ≥ 4 keywords (any) <i>or</i> Billing code 806	85	670	41	7	67.5 (58.5-75.5)	99.0 (97.9-99.6)	92.4 (84.9-96.9)	94.2 (92.3-95.8)	78.0
CPP2 <i>or</i> P/C ≥ 3 keywords (any)	87	667	39	10	69.0 (60.2-77.0)	98.5 (97.3-99.3)	89.7 (81.9-94.9)	94.5 (92.5-96.0)	78.0
CPP2 <i>or</i> P/C ≥ 2 unique keywords <i>or</i> Billing code 806	86	666	40	11	68.3 (59.4-76.3)	98.4 (97.1-99.2)	88.7 (80.6-94.2)	94.3 (92.4-95.9)	77.1
CPP2 <i>or</i> P/C ≥ 4 keywords (any)	83	670	43	7	65.9 (56.9-74.1)	99.0 (97.9-99.6)	92.2 (84.6-96.8)	94.0 (92.0-95.6)	76.9

Abbreviations: TP, true positive; FP, false positive; TN, true negative; FN, false negative; PPV, positive predictive value; NPV, negative predictive value.

primary care databases,<sup>17</sup> the sensitivity was well within the range observed, with the PPV at the higher end of the range. Although there is no definitive threshold for the performance of classification algorithms, a summary of reviews of health algorithm studies proposed classifying PPVs over 70% as high, 50% to 70% as moderate, and below 50% as low.<sup>42</sup>

The prevalence of chronic SCI/D in the study sample (calculated using the number of cases in the RSC) is: 262.5 per 100,000 overall; 118.8 per 100,000 for TSCI; and 139.6 per 100,000 for NTSCI. These results agree with the most recent estimate of SCI/D prevalence in Canada:<sup>3</sup> 252.5 per 100,000 overall; 129.8 per 100,000 for TSCI; and 122.7 per 100,000 for NTSCI.

An important strength of this study is the use of an EMR database (EMRPC) that has been shown to be fairly representative of the overall Ontario SCI/D population.<sup>15</sup> Furthermore, EMRPC uniquely includes a large amount of unstructured free-text information which, in this study, proved important in case identification. Another strength is the inclusive and evidence-based case definition of SCI/D, which includes both TSCI and NTSCI. Finally, the rigorous and detailed chart review method helped ensure that cases were identified comprehensively and reliably.

There are several important limitations to this study. First, the emphasis on comprehensiveness in the case definition may have resulted in the inclusion of some patients (for example those with a temporary impairment lasting >48 h) that would not be considered cases of chronic SCI/D according to more restrictive definitions. Second, the use of only two reviewers imposed a limitation on the number of charts that could feasibly be manually reviewed (each chart review took on average approximately 15 min). Third, there is insufficient evidence in EMRPC to reliably determine important information such as level, severity, and completeness of SCI/D. Fourth, the date of onset of SCI/D is not clearly recorded in EMR, making identification of incident cases infeasible. Fifth, it is possible that greater precision may be achieved by considering progress notes and consultation letters separately, rather than combining them into a single component. Finally, it is important to note that the results of this study may not be translatable to other EMR databases that do not contain the same free-text entries.

The novel work using free text keywords documented in this study may be useful to other researchers looking to identify cases of SCI/D in free text or unstructured data sources including EMRs. For such researchers, lessons learned from this study include the importance

of including the many potential etiologies of NTSC and the necessity of a precise, detailed case definition. Going forward, data from EMR databases can help address longstanding research gaps with respect to post-rehabilitation health system surveillance for this population.<sup>8</sup> The SCI/D cohort identified in EMRPC can be linked with other databases to develop an understanding of the health care utilization, outcomes, cost and care patterns across the entire healthcare system of the SCI/D population, creating a fuller picture of patient journeys throughout their life course, and helping to improve the quality of care.

### Abbreviations

CONS: consultation letters; CPP: cumulative patient profile; EMR: electronic medical record; EMRPC: Electronic Medical Records Primary Care database; NPV: negative predictive value; NTSCI: non-traumatic spinal cord injury; P/C: progress notes and consultation letters; PN: progress notes; PPV: positive predictive value; RSC: reference standard cohort; SCI/D: spinal cord injury and disease; TSCI: traumatic spinal cord injury

### Acknowledgements

Dr Craven wishes to acknowledge the support of the Toronto Rehab Foundation for the Chair in SCI Rehabilitation. 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 nor should be inferred.

### Disclaimer statements

**Contributors** None.


**Funding** This study was supported by the ICES, which is funded by an annual grant from the Ontario Ministry of Health (MOH) and the Ministry of Long-Term Care (MLTC).

**Conflicts of interest** The authors have no conflicts of interest to report.

### ORCID

John Shepherd  <http://orcid.org/0000-0001-8072-3775>

Karen Tu  <http://orcid.org/0000-0003-0883-4934>

Jacqueline Young  <http://orcid.org/0000-0003-0994-1934>

Jawad Chishtie  <http://orcid.org/0000-0001-8650-4469>

B. Catharine Craven  <http://orcid.org/0000-0001-8234-6803>

Rahim Moineddin  <http://orcid.org/0000-0002-5506-084X>

Susan Jaglal  <http://orcid.org/0000-0002-2930-1443>

## References

- Chhabra HS. ISCoS Textbook on Comprehensive Management of Spinal Cord Injuries. Chhabra HS, (ed.). New Delhi: Wolters Kluwer (India); 2015. 1209 p.
- New PW, Marshall R. International spinal cord injury data sets for non-traumatic spinal cord injury. *Spinal Cord* 2014;52(2): 123–32.
- Noonan VK, Fingas M, Farry A, Baxter D, Singh A, Fehlings MG, et al. Incidence and prevalence of spinal cord injury in Canada: a national perspective. *Neuroepidemiology* 2012;38(4):219–26.
- Bickenbach J, Boldt I, Brinkhof M, Chamberlain J, Cripps R, Fitzharris M, et al. A global picture of spinal cord injury. In: Bickenbach JE, (ed.) *International Perspectives on Spinal Cord Injury*. Geneva: WHO-ISCOS; 2013. p. 11–41.
- Dryden DM, Saunders LD, Rowe BH, May LA, Yiannakoulis N, Svenson LW, et al. Utilization of health services following spinal cord injury: a 6-year follow-up study. *Spinal Cord* 2004; 42(9):513–25.
- Jensen MP, Truitt AR, Schomer KG, Yorkston KM, Baylor C, Molton IR. Frequency and age effects of secondary health conditions in individuals with spinal cord injury: a scoping review. *Spinal Cord* 2013;51(12):882–92.
- DeVivo MJ, Maetz HM, Fine PR, Stover SL. Prevalence of spinal cord injury: a reestimation employing life Table techniques. *Arch Neurol*. 1980;37(11):707–8.
- Dvorak MF, Cheng CL, Fallah N, Santos A, Atkins D, Humphreys S, et al. Spinal cord injury clinical registries: improving care across the SCI care continuum by identifying knowledge gaps. *J Neurotrauma* 2017;34(20):2924–33.
- Gerber LH, Deshpande R, Prabhakar S, Cai C, Garfinkel S, Morse L, et al. Narrative review of clinical practice guidelines for rehabilitation of people with spinal cord injury. *Am J Phys Med Rehabil*. 2020;100(5):501–12.
- Rowan CP, Chan BCF, Jaglal SB, Catharine Craven B. Describing the current state of post-rehabilitation health system surveillance in Ontario – an invited review. *J Spinal Cord Med*. 2019;42 (sup1):21–33.
- Jaglal SB, Munce SEP, Guilcher SJ, Couris CM, Fung K, Craven BC, et al. Health system factors associated with rehospitalizations after traumatic spinal cord injury: a population-based study. *Spinal Cord* 2009;47(8):604–9.
- Guilcher SJT. Physician utilization among adults with traumatic spinal cord injury in Ontario. *Spinal Cord* 2009;47:470–476.
- Munce SEP, Wodchis WP, Guilcher SJT, Couris CM, Verrier M, Fung K, et al. Direct costs of adult traumatic spinal cord injury in Ontario. *Spinal Cord* 2013;51(1):64–9.
- Guilcher SJT, Munce SEP, Couris CM, Fung K, Craven BC, Verrier M, et al. Health care utilization in non-traumatic and traumatic spinal cord injury: a population-based study. *Spinal Cord* 2010;48(December 2008):45–50.
- 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 Med Inform Decis Mak*. 2015;15 (1):67.
- Tu K, Mitiku TF, Ivers NM, Guo H, Lu H, Jaakkimainen L, et al. Evaluation of electronic medical record administrative data linked database (EMRALD). *Am J Manag Care*. 2014;20(1):15–21.
- McBrien KA, Souri S, Symonds NE, Rouhi A, Lethebe BC, Williamson TS, et al. Identification of validated case definitions for medical conditions used in primary care electronic medical record databases: a systematic review. *J Am Med Inform Assoc*. 2018;25:1567–78.
- Krysko KM, Ivers NM, Young J, O'Connor P, Tu K. Identifying individuals with multiple sclerosis in an electronic medical record. *Mult Scler J*. 2015;21(2):217–24.
- Ivers N, Pylypenko B, Tu K. Identifying patients with ischemic heart disease in an electronic medical record. *J Prim Care Community Health* 2011;2(1):49–53.
- Tu K, Manuel D, Lam K, Kavanagh D, Mitiku TF, Guo H. Diabetics can be identified in an electronic medical record using laboratory tests and prescriptions. *J Clin Epidemiol*. 2011;64(4):431–5.
- Lee TM, Tu K, Wing LL, Gershon AS. Identifying individuals with physician-diagnosed chronic obstructive pulmonary disease in primary care electronic medical records: a retrospective chart abstraction study. *npj Prim Care Respir Med*. 2017;27(1):34. <http://doi.org/10.1038/s41533-017-0035-9>.
- Hagen E, Rekan T, Gilhus N, Gronning M. Diagnostic coding accuracy for traumatic spinal cord injuries. *Spinal Cord* 2008; 47:367–71.
- St. Germaine-Smith C, Metcalfe A, Pringsheim T, Roberts JJ, Beck CA, Hemmelgarn BR, et al. Recommendations for optimal ICD codes to study neurologic conditions a systematic review. *Neurology* 2012;79:1049–55.
- Ho C, Guilcher SJT, McKenzie N, Mouneimne M, Williams A, Voth J, et al. Validation of algorithm to identify persons with non-traumatic spinal cord dysfunction in Canada using administrative health data. *Top Spinal Cord Inj Rehabil*. 2017;23(4):333–42.
- Gentil ML, Cuggia M, Fiquet L, Hagenbourger C, Le Berre T, Banâtre A, et al. Factors influencing the development of primary care data collection projects from electronic health records: a systematic review of the literature. *BMC Med Inform Decis Mak*. 2017;17(1):139. <http://doi.org/10.1186/s12911-017-0538-x>.
- Schull MJ, Azimae M, Marra M, Cartagena R, Vermulen M, Ho M, et al. ICES: data, discovery. *Better Health. Int J Popul Data Sci*. 2020;4(2):1–9.
- Butt DA, Tu K, Young J, Green D, Wang M, Ivers N, et al. A validation study of administrative data algorithms to identify patients with Parkinsonism with prevalence and incidence trends. *Neuroepidemiology* 2014;43(1):28–37.
- Breiner A, Young J, Green D, Katzberg HD, Barnett C, Bril V, et al. Canadian administrative health data can identify patients with myasthenia gravis. *Neuroepidemiology* 2015;44:108–13.
- New PW, Delafosse V. What to call spinal cord damage not due to trauma? Implications for literature searching. *J Spinal Cord Med*. 2012;35(2):89–95.
- Wright A, Febowitz J, Maloney FL, Henkin S, Bates DW. Use of an electronic problem list by primary care providers and specialists. *J Gen Intern Med*. 2012;27(8):968–73.
- Luna D, Franco M, Plaza C, Otero C, Wassermann S, Gambarte ML, et al. Accuracy of an electronic problem list from primary care providers and specialists. *Stud Health Technol Inform*. 2013;192(1–2):417–21.
- Zhou X, Zheng K, Ackerman MS, Hanauer DA. Cooperative documentation: the patient problem list as a nexus in electronic health records. In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW '12)*. 2012. p. 911–20.
- Weisman A, Tu K, Young J, Kumar M, Austin PC, Jaakkimainen L, et al. Validation of a type 1 diabetes algorithm using electronic medical records and administrative healthcare data to study the population incidence and prevalence of type 1 diabetes in Ontario, Canada. *BMJ Open Diabetes Res Care* 2020;8(1):1–11.
- Patel T, Milligan J, Lee J. Medication-related problems in individuals with spinal cord injury in a primary care-based clinic. *J Spinal Cord Med*. 2017;40(1):54–61.
- Hwang M, Zbracki K, Vogel LC. Medication profile and polypharmacy in adults with pediatric-onset spinal cord injury. *Spinal Cord* 2015;53(9):673–8.
- Høgholen H, Storhaug A, Kvernørød K, Kostovski E, Viktil KK, Mathiesen L. Use of medicines, adherence and attitudes to medicines among persons with chronic spinal cord injury. *Spinal Cord* 2017;56:35–40.
- Guilcher SJT, Hogan ME, Calzavara A, Hitzig SL, Patel T, Packer T, et al. Prescription drug claims following a traumatic spinal cord injury for older adults: a retrospective population-based study in Ontario, Canada. *Spinal Cord* 2018;56(11): 1059–68.

- 38 Rouleau P, Guertin PA. Traumatic and nontraumatic spinal-cord-injured patients in Quebec, Canada. Part 3: pharmacological characteristics. *Spinal Cord* 2011;49(2):186–95.
- 39 Wei W-Q, Cronin RM, Xu H, Lasko TA, Bastarache L, Denny JC. Development and evaluation of an ensemble resource linking medications to their indications. *J Am Med Inform Assoc*. 2013;20(5):954–61.
- 40 Noonan V, Thorogood NP, Fingas M, Batke J, Belanger LM, Kwon BK, *et al*. The validity of administrative data to classify patients with spinal column and cord injuries. *J Neurotrauma* 2012;30:173–80.
- 41 Welk B, Loh E, Shariff SZ, Liu K, Siddiqi F. An administrative data algorithm to identify traumatic spinal cord injured patients: a validation study. *Spinal Cord* 2014;52(1):34–8.
- 42 Carnahan RM. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative data: summary of findings and suggestions for future research. *Pharmacoepidemiol Drug Saf*. 2012;21(S1):90–9.