Can We Geographically Validate a Natural Language Processing Algorithm for Automated Detection of Incidental Durotomy Across Three Independent Cohorts From Two Continents?

Karhade, Aditya; Oosterhoff, Jacobien H. F.; Groot, Olivier Q.; Agaronnik, Nicole; Ehresman, Jeffrey; Bongers, Michiel E. R.; Jaarsma, Ruurd L.; Poonnoose, Santosh; Sciubba, Daniel M.; Tobert, Daniel G.

Published in:
Clinical Orthopaedics and Related Research

DOI:
10.1097/CORR.0000000000002200

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2022

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Can We Geographically Validate a Natural Language Processing Algorithm for Automated Detection of Incidental Durotomy Across Three Independent Cohorts From Two Continents?

Aditya V. Karhade MD, MBA1, Jacobien H. F. Oosterhoff MD1,2, Olivier Q. Groot MD, PhD1, Nicole Agarunnik BS1, Jeffrey Ehresman MD3, Michiel E. R. Bongers MD, PhD1, Ruurd L. Jaarsma MD, PhD, FRACS, FAOOrthA4, Santosh I. Poonnoose MD5, Daniel M. Sciubba MD, MBA5, Daniel G. Tobert MD4, Job N. Doornberg MD, PhD4,6, Joseph H. Schwab MD, MS1

Received: 24 November 2021 / Accepted: 11 March 2022 / Published online: 12 April 2022
Copyright © 2022 by the Association of Bone and Joint Surgeons

Abstract
Background
Incidental durotomy is an intraoperative complication in spine surgery that can lead to postoperative complications, increased length of stay, and higher healthcare costs. Natural language processing (NLP) is an artificial intelligence method that assists in understanding free-text notes that may be useful in the automated surveillance of adverse events in orthopaedic surgery. A previously developed NLP algorithm is highly accurate in the detection of incidental durotomy on internal validation and external validation in an independent cohort from the...
same country. External validation in a cohort with linguistic differences is required to assess the transportability of the developed algorithm, referred to geographical validation. Ideally, the performance of a prediction model, the NLP algorithm, is constant across geographic regions to ensure reproducibility and model validity.

**Question/purpose** Can we geographically validate an NLP algorithm for the automated detection of incidental durotomy across three independent cohorts from two continents?

**Methods** Patients 18 years or older undergoing a primary procedure of (thoraco)lumbar spine surgery were included. In Massachusetts, between January 2000 and June 2018, 1000 patients were included from two academic and three community medical centers. In Maryland, between July 2016 and November 2018, 1279 patients were included from one academic center, and in Australia, between January 2010 and December 2019, 944 patients were included from one academic center. The authors retrospectively studied the free-text operative notes of included patients for the primary outcome that was defined as intraoperative durotomy. Incidental durotomy occurred in 9% (93 of 1000), 8% (108 of 1279), and 6% (58 of 944) of the patients, respectively, in the Massachusetts, Maryland, and Australia cohorts. No missing reports were observed. Three datasets (Massachusetts, Australian, and combined Massachusetts and Australian) were divided into training and holdout test sets in an 80:20 ratio. An extreme gradient boosting (an efficient and flexible tree-based algorithm) NLP algorithm was individually trained on each training set, and the performance of the three NLP algorithms (respectively American, Australian, and combined) was assessed by discrimination via area under the receiver operating characteristic curves (AUC-ROC; this measures the current procedural terminology (CPT) and ICD codes for durotomy, which otherwise could have led to underreporting. NLP is an artificial intelligence method that has the ability to understand human language by reading text (or hearing speech) and to interpret the human language (which is the meaning of the phrase “natural language”). NLP algorithms are able to provide automating text tasks and are also used to classify text.

Despite promising results on internal validation, the transferability of this algorithm to other clinical settings in the United States and abroad depends on repeated external validation in multiple contexts [27]. In particular, a systematic review of NLP algorithms showed that 88% of studies did not perform external validation and 78% claimed generalizability of an NLP algorithm without performing the necessary testing [16]. Given the recent surge in artificial intelligence studies in orthopaedic surgery, we believed it was important to examine the utility and transportability of the previously developed NLP algorithm in an international context, referred to geographical validation. Ideally, the performance of a prediction model, the NLP algorithm, is constant across geographic regions to ensure reproducibility and model validity [1].

**Results** The combined NLP algorithm (the combined Massachusetts and Australian data) achieved excellent performance on independent testing data from Australia (AUC-ROC 0.97 [95% confidence interval 0.87 to 0.99]), Massachusetts (AUC-ROC 0.99 [95% CI 0.80 to 0.99]) and Maryland (AUC-ROC 0.95 [95% CI 0.93 to 0.97]). The NLP developed based on the Massachusetts cohort had excellent performance in the Maryland cohort (AUC-ROC 0.97 [95% CI 0.95 to 0.99]) but worse performance in the Australian cohort (AUC-ROC 0.74 [95% CI 0.70 to 0.77]).

**Conclusion** We demonstrated the clinical utility and reproducibility of an NLP algorithm with combined datasets retaining excellent performance in individual countries relative to algorithms developed in the same country alone for detection of incidental durotomy. Further multi-institutional, international collaborations can facilitate the creation of universal NLP algorithms that improve the quality and safety of orthopaedic surgery globally. The combined NLP algorithm has been incorporated into a freely accessible web application that can be found at https://sorg-apps.shinyapps.io/nlp_incidental_durotomy/. Clinicians and researchers can use the tool to help incorporate the model in evaluating spine registries or quality and safety departments to automate detection of incidental durotomy and optimize prevention efforts.

**Introduction**

Incidental durotomy is an intraoperative complication of spine surgery that is associated with increased in-hospital complications [21], increased length of stay [21], and higher costs [18, 21, 24], with an incidence ranging from 0.3 to 11% in published studies [8, 11]. Sequelae of incidental durotomies may include postoperative headache, meningitis, a new-onset neurological deficit or pseudo-meningocele, which may lead to revision surgery. Early identification and automated detection may have several implications for monitoring this complication and improving quality and safety tracking. We developed a natural language processing (NLP) algorithm for automated detection of incidental durotomy in a single-center cohort of patients undergoing lumbar spine surgery [14]. In that study, we showed that the NLP algorithm outperforms current procedural terminology (CPT) and ICD codes for durotomy, which otherwise could have led to underreporting. NLP is an artificial intelligence method that has the ability to understand human language by reading text (or hearing speech) and to interpret the human language (which is the meaning of the phrase “natural language”). NLP algorithms are able to provide automating text tasks and are also used to classify text.

Volume 480, Number 9 Automated Detection of Incidental Durotomy 1767
We therefore asked: Can we geographically validate an NLP algorithm for the automated detection of incidental durotomy across three independent cohorts from two continents?

Patients and Methods

Guidelines

The Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis guidelines were followed in this study [4].

Study Design and Setting

Institutional review board (IRB) approval was granted for a retrospective study of electronic health records (EHR) at each of the three institutions [5, 14]. The three datasets originated from three different geographical locations: Massachusetts, Maryland, and Australia. In Massachusetts, patients from two academic and three community medical centers within one healthcare entity (Mass General Brigham, Boston, MA, USA) were included between January 2000 and June 2018 [14]. In Maryland, patients were included from one academic center (Johns Hopkins Medicine, Baltimore, MD, USA) between July 2016 and November 2018 [5]. In Australia, patients were included from one academic center (Flinders Medical Centre, Adelaide, SA, Australia) between January 2019 and December 2019. No missing reports were observed.

Study Participants

The selection criteria applied in the developmental study were applied [14]. The inclusion criteria for all three cohorts were age 18 years or older, primary procedure of posterior lumbar decompression and/or fusion for lumbar disc herniation, spondylolisthesis, or spinal stenosis, and free-text operative notes available for review in the EHR of the included institutions as described above. Patients undergoing surgery for trauma, malignancy, infection, pseudoarthrosis, and deformity were excluded. In Massachusetts, between January 2000 and June 2018, 1000 patients were included from two academic and three community medical centers. In Maryland, between July 2016 and November 2018, 1279 patients were included from one academic center, and in Australia, between January 2010 and December 2019, 944 patients were included from one academic center.

Patients’ Baseline Data

The authors retrospectively studied the free-text operative notes of included patients for the primary outcome which was defined as intraoperative durotomy. The proportions of incidental durotomies were 9% (93 of 1000) in Massachusetts, 6% (58 of 944) in Australia, and 8% (108 of 1279) in Maryland [5, 14].

Gold Standard for Diagnosis of Incidental Durotomy

The primary outcome was defined as an intraoperative incidental durotomy recorded in free-text operative notes. We considered dural rent, dural tear, and cerebrospinal fluid leak to be synonymous with incidental durotomy. Operative notes were manually reviewed using a binary classification (yes or no) by independent local researchers who were not involved in the care of these patients (AVK, OQG, MERB). No disagreements during study selection occurred between the three reviewers.

Training of the Three NLP Models

Three datasets (Massachusetts, Australian, and Combined Massachusetts and Australian) were divided into training and holdout test sets in an 80:20 ratio for training and internal validation (Fig. 1).

Subsequently, the free-text operative notes were processed to be suitable for model development, such as scrubbing redundant whitespace, line breaks, removing stop words (such as “to”, “the”) and stemming words (including “repairing” and “repaired” were converted to “repair”). Feature extraction was applied using the bag-of-word (BOW) approach, to identify which words in the operative notes were most important for detecting incidental durotomy. Relative variable importance plots were created for global explanation of the algorithm. These plots examine the most important words used by the algorithm for detecting incidental durotomy [22].

In total, three eXtreme Gradient Boosting (XGBoost) NLP machine-learning algorithms were developed to detect incidental durotomy: the American NLP model using 80% of the Massachusetts dataset, Australian NLP model using 80% of the Australian dataset, and a combined NLP model using the same 80% of the Massachusetts and Australian datasets [3]. The combined NLP model was therefore trained with both American and Australian operative notes. XGBoost is a flexible and efficient tree-based algorithm, using a gradient boosting framework. XGBoost is recommended for text classification tasks in small-to-medium datasets, it learns fast (efficient memory usage), and it applies regularization for avoiding overfitting of the model. Based on the meaningful words derived from the BOW approach and training of the XGBoost algorithm, NLP classified each operative note as describing an incidental durotomy or not.
Evaluation of the Three NLP Models

The American NLP was internally validated on 20% (200 of 1000) of the Massachusetts dataset, and externally validated on 100% of the Australian (944 of 944) and Maryland (1279 of 1279) datasets. Further details of the original American NLP that was built on Massachusetts data can be found in the developmental study [14]. The Australian NLP was internally validated on 20% (188 of 806) of the Australian dataset, and externally validated on 100% of the Massachusetts (1000) and Maryland (1279) datasets. The combined NLP was internally validated on 20% of the Massachusetts (200 of 1000) and Australian (188 of 994) datasets, and externally validated on 100% (1279 of 1279) of the Maryland dataset (Supplementary Table 1; http://links.lww.com/CORR/A777).

Performance Metrics

Each NLP model was assessed using the following performance metrics: area under the receiver operating characteristic curve (AUC-ROC), area under the precision-recall curve (AUC-PRC), Brier score (overall performance), calibration (intercept and slope), sensitivity (recall), specificity, positive predictive value (also known as precision), negative predictive value, F1-score, positive likelihood ratio, and negative likelihood ratio. The primary outcome of interest was the AUC-ROC. The secondary outcomes of interest were the other model performance metrics as mentioned above.

The ROC curve visualizes the performance of sensitivity and 1-specificity for binary classification across various threshold settings. The AUC-ROC score ranges from 0.50 to 1.0, with 1.0 indicating the highest discrimination score and 0.50 indicating the lowest. The higher the discrimination score, the better the model’s ability to distinguish between classes (that is, differentiating between patients who had the outcome of incidental durotomy from those who had not) [28].

The PRC shows the tradeoff between precision and recall for different thresholds. The AUC-PRC score ranges from 0.50 to 1.0, with 1.0 indicating perfect performance.
The null-model Brier score, which equals the probability of incidental durotomy in the dataset, was used to benchmark the algorithm’s Brier score. A Brier score lower than the null-model Brier score indicates superior performance of the NLP model to this null benchmark. Perfect NLP models would have a Brier score of 0 [2].

A calibration plot charts the estimated versus the observed probabilities for the primary outcome. A perfect calibration plot has an intercept of 0 (< 0 reflects overestimation, and > 0 reflects underestimation of the probability of the outcome) and a slope of 1 (the model is performing similarly in training and test sets) [29, 32]. In a small data set, the slope is often less than 1, reflecting model overfitting; probabilities are too extreme (low probabilities that are too low, and high probabilities that are too high).

Sensitivity, also known as the true positive rate or recall, corresponds to the proportion of positive observations that are correctly classified as positive compared with all predictions (TP / (FN+TP)). Sensitivity is 100% if all positive observations are classified as positive. Specificity, also known as the true negative rate, corresponds to the proportion of negative observations that are correctly classified as negative compared with all prediction (TN / (TN+FP)). Specificity is 100% if all negative observations are classified as negative.

Positive predictive value (PPV), also known as precision, refers to the proportion of patients with a positive prediction who actually had the outcome (TP / (TP+FP)). A perfect PPV is 100%, but is highly dependable on the prevalence of the outcome: If the prevalence decreases, then the PPV decreases. Negative predictive value (NPV) refers to the proportion of patients with a negative prediction who did not have the outcome (TN / (TN+FN)). A perfect NPV is also 100%, but also highly dependable on the prevalence: If the prevalence decreases, then the NPS will increase.

F1-score tries to find a balance between precision and recall, and ranges from 0 (lowest) to 1 (highest).

Positive likelihood ratio (LR+) gives insight into the odds of obtaining a positive predicted outcome in patients (sensitivity / (1-specificity). If a score is greater than 1, then you are more likely to have the outcome. Negative likelihood ratio (LR-) gives the change in the odds of having a negative predicted outcome in patients.

All performance metrics, discrimination curves, and global explanation plots were provided for each internal and external validation to allow accurate and reliable comparison of the NLP models. Anaconda Distribution (Anaconda Inc), Python (Python Software Foundation), R version (The R Foundation), and RStudio (RStudio) were used for data analysis.

**Internet Application**

The best performing NLP model was deployed as a free open-access web application that is accessible on desktops, smartphones, and tablets. Clinicians and researchers can use the application to enter their own operative notes.

**Ethical Approval**

Ethical approval for this study was waived by the IRBs at each of the three institutions: Massachusetts General Brigham, Boston, MA, USA; Flinders University, Adelaide, Australia; and Johns Hopkins University, Baltimore, MD, USA.

**Results**

On internal and external validation, the combined NLP model proved to be highly accurate across all three cohorts (derived from two academic US institutions and one academic English-speaking non-US institution), and thereby accurate in identifying an incidental durotomy on the two continents. The combined NLP model has been incorporated into a freely accessible web application that can be found at https://sorg-apps.shinyapps.io/nlp_incidental_durotomy/.

**American NLP Model**

The American NLP model had an AUC-ROC of 0.99 (95% confidence interval [CI] 0.96 to 1.00) on internal validation in Massachusetts (Table 1). The most important variables for automated detection of incidental durotomy were “repair,” “leak,” “csf,” and “tissue” (Fig. 2). On external validation in Maryland, the AUC-ROC was 0.97 (95% CI 0.95 to 0.99) (Supplementary Fig. 1; http://links.lww.com/CORR/A778). However, on external validation in Australia, the AUC-ROC was 0.74 (95% CI 0.70 to 0.77) (Supplementary Fig. 2; http://links.lww.com/CORR/A779).

**Australian NLP Model**

The Australian NLP model had an AUC-ROC of 0.96 (95% CI 0.80 to 0.99) on internal validation in Australia (Table 1). The most important variables for automated detection of incidental durotomy were “lesion,” “debulk,” “durotomi,” and “dura” (Fig. 3). On external validation, the AUC-ROC was 0.77 (95% CI 0.71 to 0.83) in Massachusetts (Supplementary Fig. 3; http://links.lww.com/CORR/A780) and 0.77 (95% CI 0.71 to 0.81) in Maryland (Supplementary Fig. 4; http://links.lww.com/CORR/A781).
Combined NLP Model

The combined NLP model achieved excellent performance across all cohorts (Massachusetts, Maryland, and Australian). On internal validation in the Australian and Massachusetts cohorts, the AUC-ROC was 0.97 (95% CI 0.87 to 0.99) and 0.99 (95% CI 0.80 to 0.99), respectively (Table 1). On external validation in the Maryland cohort, the AUC-ROC was 0.95 (95% CI 0.93 to 0.97) (Supplementary Fig. 5; http://links.lww.com/CORR/A782). The Brier score on external validation was 0.06 relative to the null-model Brier score of 0.08. At a threshold of 0.05 (discriminate significant from nonsignificant results derived by the NLP model), the combined NLP in the Maryland cohort had a specificity of 0.98 (95% CI 0.97 to 0.99), sensitivity of 0.61 (95% CI 0.51 to 0.70), positive predictive value of 0.74 (95% CI 0.64 to 0.83), and negative predictive value of 0.96 (95% CI 0.95 to 0.97) on external validation (Table 2). A global explanation of words used by the combined NLP algorithm showed that it included the important words from the American NLP algorithm (for example, “repair”) and the Australian NLP algorithm (for example, “durotomi”) for detecting incidental durotomy (Fig. 4).

Discussion

Early and accurate identification of an incidental durotomy may improve patient outcomes and improve quality and safety reporting. With the rise of data availability from the EHRs and automated data analytics using NLP, we are able to improve early and accurate identification of an incidental durotomy, which can aid clinicians and researchers in treating and preventing this complication [5, 14]. In this study, we aimed to geographically validate an NLP algorithm for automated identification of an incidental durotomy in three large healthcare institutions from two different continents. A combined algorithm developed using Massachusetts and Australian EHRs had excellent performance across all three independent cohorts, achieving an AUC-ROC of 0.99 in Massachusetts, AUC-ROC of 0.95 in Maryland, and AUC-ROC of 0.97 in Australia. The combined NLP algorithm has been incorporated into a freely accessible web application that can be found at https://sorg-apps.shinyapps.io/nlp_incidental_durotomy/. We demonstrated that developing a combined algorithm trained on US and non-US cohorts is an appropriate approach for improving the performance of a NLP algorithm in non-US settings to ensure geographical validation.
Limitations

We acknowledge that this study has several limitations. First, the algorithm, which was developed at Flinders University using only Australian EHRs, was not externally validated at another Australian institution. This means that the model cannot be automatically generalized to other Australian institutions. External validation at another Australian institution with different practice characteristics may provide a more accurate assessment of the model’s performance in the Australian population, especially if there are concerns around potential overfitting in the original training dataset [15, 26]. Second, in a prior study describing the development of the NLP algorithm [14], the performance of the algorithm was described using CPT and ICD codes. Although the algorithm outperformed standard codes, the performance of such codes may be different at other institutions with different billing practices. A comparison of the algorithm with standard codes may be appropriate, especially at institutions outside the US. Third, no baseline characteristics (such as age or sex) or a subanalysis of the three independent cohorts were provided because relevant words that are indicative of an incidental durotomy in the operative notes like “repair”, “leak”, “csf”, and “tis–seel” are—in our view—unrelated to the population.

In addition, the combined NLP algorithm described in our study may not be transferable to institutions in countries that primarily use a language other than English. However, our study has highlighted the challenge of applying NLP algorithms in non-US cohorts with different linguistic characteristics. Several approaches have been used for addressing this well-recognized challenge [19]. Our method used one of the most reliable yet tedious methods, which required training a model on a new dataset every time the model must be applied to a cohort with different linguistic characteristics. However, the use of a multilingual model with cross-lingual embeddings may be an alternative approach to developing an algorithm that can be used in a wider variety of international settings [33]. For example, the Multilingual Bidirectional Encoder Representations from Transformers model released by Google has been pretrained on 104 languages and has been popularized for its performance with named-entity recognition tasks (identification of entities such as “durotomy” from unstructured free text) [20]. Exploring such models for developing an NLP algorithm for identifying incidental durotomy may facilitate the transferability of the algorithm to a broader variety of clinical settings.

Discussion of Key Findings

Our study highlighted the importance of external geographic validation of an NLP algorithm for automated detection of incidental durotomy in patients undergoing (thoraco)lumbar spine surgery. We developed a highly accurate combined NLP algorithm using US and non-US patient cohorts, derived from Massachusetts and Australian operative notes. The combined NLP showed better results than the NLP algorithms in the relative countries alone, indicating that universal generalized models can be an advantage when developing text classification algorithms for detecting incidental durotomy. The deployed web application with the combined NLP can be used by clinicians...
and researchers to help incorporate the model when evaluating spine registries or by quality and safety departments to automate detection of incidental durotomy and optimize prevention efforts.

External validation of NLP algorithms across different clinical settings is essential for facilitating large-scale quality improvement efforts. Unstructured, free-text notes contain approximately 70% to 80% of clinically relevant information [17], but this information is often inaccessible for clinical or research purposes because of the inherent limitations of a burdensome manual record review. This has prompted the development of automated methods, such as NLP, for extracting such data. In orthopaedic surgery, NLP methods have been developed for extracting disease characteristics, including identifying and classifying periprosthetic femur fractures [31], interpreting findings on lumbar spine images [30], and identifying degenerative spine changes on MRI reports [10]. Some studies have also designed NLP methods for automated extraction of data elements to create retrospective research databases for quality review [23, 25, 34].

Most relevant to the context of this study, NLP algorithms have been designed to predict adverse outcomes, including wound infections resulting in reoperation after lumbar discectomy [12], periprosthetic joint infections [7], and intraoperative vascular injury [13]. Although these methods can identify documentation indicating outcomes of interest (such as incidental durotomy), a further manual record review is still required to exclude false positives. Similarly, when deploying NLP algorithms in large patient cohorts, it may not be feasible to identify all false negatives. Furthermore, the utility of NLP algorithms is only as good as the quality of the documentation [6, 9]. For example, an algorithm may be used to

<p>| Table 2. Performance of the three NLP models on internal and external validation in three cohorts |
|---------------------------------|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th>Performance metrics</th>
<th>American NLP</th>
<th>Australian NLP</th>
<th>Combined NLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australian cohort (n = 944)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity (recall)</td>
<td>0.47 (0.33 to 0.60) a</td>
<td>0.91 (0.59 to 1.00) b</td>
<td>0.82 (0.48 to 0.98) f</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.75 (0.72 to 0.78) a</td>
<td>0.97 (0.93 to 0.99) b</td>
<td>0.97 (0.93 to 0.99) f</td>
</tr>
<tr>
<td>Positive predictive value</td>
<td>0.11 (0.07 to 0.15) a</td>
<td>0.62 (0.35 to 0.85) b</td>
<td>0.60 (0.32 to 0.84) f</td>
</tr>
<tr>
<td>Negative predictive value</td>
<td>0.96 (0.94 to 0.97) a</td>
<td>0.99 (0.97 to 1.00) b</td>
<td>0.99 (0.96 to 1.00) f</td>
</tr>
<tr>
<td>F1-score</td>
<td>0.18 (0.12 to 0.24) a</td>
<td>0.74 (0.44 to 0.92) b</td>
<td>0.69 (0.39 to 0.90) f</td>
</tr>
<tr>
<td>LR (+)</td>
<td>1.85 (1.37 to 2.49) a</td>
<td>26.8 (12.0 to 60.2) b</td>
<td>24.1 (10.5 to 55.6) f</td>
</tr>
<tr>
<td>LR (-)</td>
<td>0.71 (0.56 to 0.91) a</td>
<td>0.09 (0.01 to 0.61) b</td>
<td>0.19 (0.05 to 0.66) f</td>
</tr>
<tr>
<td>Massachusetts cohort (n = 1000)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity (recall)</td>
<td>0.89 (0.65 to 0.99) d</td>
<td>0.04 (0.01 to 0.11) e</td>
<td>0.83 (0.59 to 0.96) f</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.99 (0.96 to 1.00) d</td>
<td>1.00 (0.99 to 1.00) e</td>
<td>0.98 (0.94 to 0.99) f</td>
</tr>
<tr>
<td>Positive predictive value</td>
<td>0.99 (0.96 to 1.00) d</td>
<td>0.80 (0.28 to 0.99) e</td>
<td>0.79 (0.54 to 0.94) f</td>
</tr>
<tr>
<td>Negative predictive value</td>
<td>0.89 (0.65 to 0.99) d</td>
<td>0.91 (0.89 to 0.93) e</td>
<td>0.98 (0.95 to 1.00) f</td>
</tr>
<tr>
<td>F1-score</td>
<td>0.89 (0.65 to 0.99) d</td>
<td>0.08 (0.02 to 0.19) e</td>
<td>0.81 (0.56 to 0.95) f</td>
</tr>
<tr>
<td>LR (+)</td>
<td>80.9 (20.2 to 324.1) d</td>
<td>39.0 (4.41 to 345.4) e</td>
<td>37.9 (14.1 to 102.1) f</td>
</tr>
<tr>
<td>LR (-)</td>
<td>0.11 (0.03 to 0.42) d</td>
<td>0.96 (0.92 to 1.00) e</td>
<td>0.17 (0.06 to 0.48) f</td>
</tr>
<tr>
<td>Maryland cohort (n = 1279)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity (recall)</td>
<td>0.84 (0.76 to 0.91) d</td>
<td>0.07 (0.03 to 0.14) e</td>
<td>0.61 (0.51 to 0.70) f</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.96 (0.94 to 0.97) d</td>
<td>1.00 (0.99 to 1.00) e</td>
<td>0.98 (0.97 to 0.99) f</td>
</tr>
<tr>
<td>Positive predictive value</td>
<td>0.65 (0.56 to 0.72) d</td>
<td>0.80 (0.44 to 0.97) e</td>
<td>0.74 (0.64 to 0.83) f</td>
</tr>
<tr>
<td>Negative predictive value</td>
<td>0.99 (0.98 to 0.99) d</td>
<td>0.92 (0.90 to 0.94) e</td>
<td>0.96 (0.95 to 0.97) f</td>
</tr>
<tr>
<td>F1-score</td>
<td>0.73 (0.64 to 0.80) d</td>
<td>0.14 (0.06 to 0.25) e</td>
<td>0.67 (0.57 to 0.76) f</td>
</tr>
<tr>
<td>LR (+)</td>
<td>19.7 (14.9 to 26.2) d</td>
<td>43.4 (9.33 to 202.7) e</td>
<td>31.1 (20.2 to 47.9) f</td>
</tr>
<tr>
<td>LR (-)</td>
<td>0.16 (0.11 to 0.25) d</td>
<td>0.93 (0.88 to 0.98) e</td>
<td>0.40 (0.31 to 0.50) f</td>
</tr>
</tbody>
</table>

The numbers in parentheses represent 95% CIs.

aExternal validation (n = 944).

bInternal validation (n = 188).

cInternal validation (n = 188).

dInternal validation (n = 200).

eInternal validation (n = 200); LR + = positive likelihood ratio; LR- = negative likelihood ratio.
identify patients in whom incidental durotomy was reported, but the algorithm may not be able to interpret the reason for the durotomy if descriptive information is sparse regarding the reason for the adverse outcome. Nonetheless, automated surveillance of adverse outcomes can substantially reduce the time required for identifying relevant procedures. A manual record review can then be used to investigate factors associated with certain outcomes (such as an iatrogenic injury described in an operative report). Although it would be ideal for an NLP algorithm to independently identify these factors, and thus entirely bypass a manual record review, sufficient performance metrics may not be feasible for this task, given the low prevalence of patients with the outcome of interest. However, continued efforts with cross-institutional external validation may facilitate the creation of a sufficiently large database of patients with incidental durotomy, allowing subsequent training of models for interpreting documented reasons for this adverse event.

**Conclusion**

We demonstrated the clinical utility of a geographically validated NLP algorithm for identifying incidental durotomy at two US and one non-US (Australian) institution, retaining excellent performance in individual countries relative to an NLP algorithm in the same country alone. This finding suggests transferability of an NLP algorithm across EHRs using different language rules. Further multi-institutional and international collaborations can facilitate the creation of universal NLP algorithms to improve patient outcomes as well as the quality and safety of orthopaedic surgery. The combined NLP tool is deployed as a freely accessible web application that can be found at [https://sorg-apps.shinyapps.io/nlp_incidental_durotomy/](https://sorg-apps.shinyapps.io/nlp_incidental_durotomy/). Clinicians and researchers can use the tool to help incorporate the model in evaluating spine registries or quality and safety departments to automate detection of incidental durotomy and optimize prevention efforts.

**References**


