



# Accuracy of claims-based algorithms for epilepsy research: Revealing the unseen performance of claims-based studies

\*†Lidia M. V. R. Moura, ‡Maggie Price, \*†Andrew J. Cole , \*†Daniel B. Hoch, and ‡§John Hsu

*Epilepsia*, 58(4):683–691, 2017  
doi: 10.1111/epi.13691

## SUMMARY

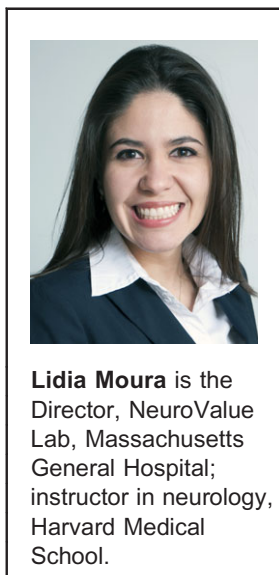
**Objective:** To evaluate published algorithms for the identification of epilepsy cases in medical claims data using a unique linked dataset with both clinical and claims data.

**Methods:** Using data from a large, regional health delivery system, we identified all patients contributing biologic samples to the health system's Biobank ( $n = 36K$ ). We identified all subjects with at least one diagnosis potentially consistent with epilepsy, for example, epilepsy, convulsions, syncope, or collapse, between 2014 and 2015, or who were seen at the epilepsy clinic ( $n = 1,217$ ), plus a random sample of subjects with neither claims nor clinic visits ( $n = 435$ ); we then performed a medical chart review in a random subsample of 1,377 to assess the epilepsy diagnosis status. Using the chart review as the reference standard, we evaluated the test characteristics of six published algorithms.

**Results:** The best-performing algorithm used diagnostic and prescription drug data (sensitivity = 70%, 95% confidence interval [CI] 66–73%; specificity = 77%, 95% CI 73–81%; and area under the curve [AUC] = 0.73, 95%CI 0.71–0.76) when applied to patients age 18 years or older. Restricting the sample to adults aged 18–64 years resulted in a mild improvement in accuracy (AUC = 0.75, 95%CI 0.73–0.78). Adding information about current antiepileptic drug use to the algorithm increased test performance (AUC = 0.78, 95%CI 0.76–0.80). Other algorithms varied in their included data types and performed worse.

**Significance:** Current approaches for identifying patients with epilepsy in insurance claims have important limitations when applied to the general population. Approaches incorporating a range of information, for example, diagnoses, treatments, and site of care/specialty of physician, improve the performance of identification and could be useful in epilepsy studies using large datasets.

**KEY WORDS:** Epidemiology, Epilepsy, Accuracy, Codes.



Lidia Moura is the Director, NeuroValue Lab, Massachusetts General Hospital; instructor in neurology, Harvard Medical School.

Epilepsy is an incurable, life-threatening neurologic disorder that affects an estimated 65 million people worldwide.<sup>1,2</sup> Despite the prevalence and seriousness of the

condition, there is limited information on the quality of epilepsy care. This dearth results in large part from the limited availability of population-wide databases and questions about the reliability of claims-based algorithms.<sup>3–6</sup>

There are significant challenges to creating an algorithm for identifying patients with epilepsy in insurance claims databases. Epilepsy is characterized by recurrent, spontaneous seizures, that is, clinical diagnoses based on symptoms and the exclusion of precipitating causes of seizures. The manifestation of seizures can vary substantially across patients, and the recorded diagnoses can vary across physicians, for instance, with the amount of specialized training of physicians.<sup>7,8</sup> In fact, many patients receive an epilepsy diagnosis without use of standardized

Accepted December 29, 2016; Early View publication February 15, 2017.

\*Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts, U.S.A.; †Department of Neurology, Harvard Medical School, Boston, Massachusetts, U.S.A.; ‡Mongan Institute, Massachusetts General Hospital, Boston, Massachusetts, U.S.A.; and §Departments of Health Care Policy and of Medicine, Harvard Medical School, Boston, Massachusetts, U.S.A.

Address correspondence to Lidia M. V. R. Moura, Wang 739 Neurology, Massachusetts General Hospital, 55 Fruit St., Boston, MA 02114, U.S.A. E-mail: lidia.moura@mgh.harvard.edu

Wiley Periodicals, Inc.

© 2017 International League Against Epilepsy

### KEY POINTS

- Accuracy of existing claims-based definitions of epilepsy is modest
- Combining diagnosis codes with medication use increases accuracy
- External validation is an important step in evaluating the performance of claims-based definitions of epilepsy

evaluation protocols, that is, receipt of or documentation of electroencephalography (EEG) or brain magnetic resonance imaging (MRI).<sup>9</sup>

As a result, there is uncertainty about even basic information such as the true population incidence and prevalence of epilepsy. Previous efforts to identify epilepsy patients have used multiple data types and medical services to identify epilepsy cases in both the general population and in higher risk subgroups (as the risk increases, so does the likely prevalence, which in turn affects test characteristics such as the positive and negative predictive values).

Unfortunately, the diversity of the health care systems being studied represents a serious limitation to the generalizability of the existing algorithms.<sup>4,5,10–12</sup> Validation using a small dataset and cross-validation using the same dataset used to develop the algorithm may not adequately reflect performance in the broader application context. In epilepsy research, the ultimate goal of epidemiologists is to provide accurate predictions for independent samples obtained in different settings. The problem with internal cross-validation is that it may produce inflated discrimination accuracy, when compared to cross-study, cross-population validation.<sup>13</sup>

In this study, we compare the published, validated algorithms for identification of epilepsy using a large, single dataset, and clearly stated sampling and evaluation criteria.

## METHODS

### Study design

This study involved three stages: (1) creation of a validation cohort seen at the Partners HealthCare System (PHS); (2) assessment of epilepsy status using a review of medical records; and (3) evaluation of test characteristics for the published algorithms.<sup>4,5,11,12</sup> The work was part of a larger study on the examination of the interactions of genes, lifestyle, and other factors in the development of epilepsy and other diseases, which focused on patients contributing biologic samples to the health system's Biobank (Fig. 1). As discussed in the legend of Figure 1, these patients enrolled in the Partners HealthCare Biobank project were implemented in Partners an affiliated ambulatory clinical practice, consisting of 17 centers and departments including

neurology, psychiatry, primary care, emergency, and internal medicine. Enrolled patients consented to the following: (1) a dedicated blood draw for preparation of DNA and blood derivatives for storage in the Biobank; (2) means to collect future discarded clinical specimens; (3) linkage of banked samples with their electronic medical record (EMR) and with health information collected through a secure survey; (4) specimen storage/distribution for broad use by institutional review board (IRB)-approved Partners' investigators; and (5) willingness to be re-contacted as part of collaborating studies.

### Patient population and data sources

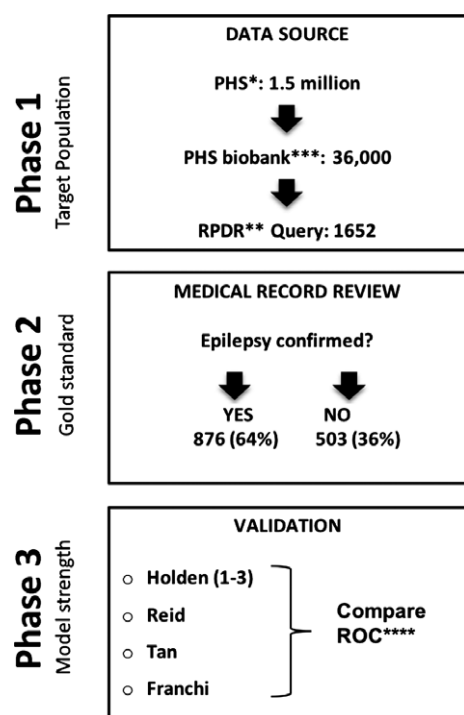
This study conducted the use of data from the PHS and the Partners HealthCare Research Patient Data Registry (RPDR), a data warehouse populated with data from several source systems, including the hospital and physician billing systems, as well as data from Partners' Clinical Data Repository (CDR), Epic, and the Enterprise Patient Master Index (EMPI). The EMPI assigns a 9-digit reference number to a patient and serves as a mechanism to assign multiple medical record numbers to one reference number, eliminating duplicate patients in the RPDR. This results in a comprehensive database that includes patient demographics, diagnoses (e.g., billing codes in claims), procedures (e.g., reports of brain MRI), inpatient pharmacy data, laboratories (e.g., antiepileptic drug level orders and results), transfusions, microbiology, inpatient and outpatient encounter information (e.g., physician's clinical notes), and provider data.

To compare basic test characteristics (e.g., sensitivity, specificity, positive predictive value, negative predictive value, receiver operating characteristic [ROC] curves, or C-statistics), we selected four articles that had reported key methodologic information in their methods, as detailed in Table 1.<sup>4,5,11,12</sup>

We identified 1,906 medical records of consented patients. Employees of the health care system were automatically excluded from the query as per our IRB protocol, yielding a final sample of 1,652 eligible patients. We recognized all subjects with no <1 diagnosis potentially consistent with epilepsy, for example, epilepsy, convulsions, syncope, or collapse, between 2014 and 2015, or who were seen at the epilepsy center (n = 1,217), in addition to an irregular sample of subjects with neither the diagnosis nor facility visits (n = 435). Then we conducted a medical chart review of a random subsample of 1,377 to examine the epilepsy diagnosis status. (Please find details of the medical records abstraction process in Appendix S12. A comprehensive description of the sample query, medical records review process and the algorithm selection can be found in Appendices S1–S3.)

### Statistical analysis

We calculated sensitivity, specificity, positive predictive values, and negative predictive values for each algorithm



against the reference standard derived from the chart review. Using the validation of the algorithms for epilepsy case definition, sensitivity represented the percentage of total epilepsy cases correctly identified as epilepsy cases by each algorithm. Specificity was defined as the percentage of total nonepilepsy cases correctly identified by the algorithm. Positive predictive values were determined by percentage of algorithm-identified epilepsy cases qualifying as “true” epilepsy cases. Negative predictive values were determined according to the percentage of algorithm-identified nonepilepsy cases qualifying as “true” nonepilepsy cases. The latter two predictive values, however, are influenced by the prevalence of true epilepsy cases in the sample, whereas sensitivity and specificity are not; therefore, we focus our presentation on sensitivity and specificity (Table 2).

We adapted the study design to allow for comparison among the different algorithms across different patient populations. For instance, Holden grouped age as 0–19 and 65+ years, whereas Reid used only adults. We performed the comparative analysis using the entire cohort of adult patients and repeated the same analysis stratified by age groups: 18–64 and >64 years old. In addition, we selected a 2-year time frame (2013–2014) to allow for comparison among the different algorithms. We also obtained a list of false positives and false negatives from the algorithm of highest performance to describe their demographic and clinic characteristics.

Appendix S4 shows a table of the generic and brand names of the antiepileptic drugs used in this study. Appendix S5 compares the diagnosis and procedure codes obtained using administrative claims. Appendix S6 shows

**Figure 1.**

Study design. \*PHS: Partners HealthCare System: We used the Partners HealthCare system (PHS)—integrated Epic-based Electronic Health Record or HER. In addition to the two founding academic medical centers in Boston, the PHS includes medical sites in Rhode Island, Connecticut, Massachusetts, and Maine, employing both primary care and specialty physicians at community hospitals, managed care organizations, specialty facilities, community health centers, and other health-related entities. Inpatient and outpatient records are collected on every patient in the PHS, which includes over 1.5 million covered lives. \*\*RPDR: The Research Patient Data Registry is a clinical data registry that aggregates all records throughout PHS, including those from the visit narrative (e.g., physician’s clinical notes), test reports (e.g., reports of brain MRI), laboratory results (e.g., antiepileptic drug level orders and results), or administrative systems (e.g., billing codes in claims). In this manuscript, we used the ICD-9 billing codes in professional claims, which represent the diagnostic judgment of a highly trained health care professional (as opposed to professional claims revised by coders or facility claims, which represents the judgment of professional coders who are often focused on resource utilization). \*\*\*Partners HealthCare Biobank: As a cohort in the validation work, we targeted the patients enrolled in the Partners HealthCare Biobank project implemented in Partners-affiliated ambulatory clinical practices. The clinical practices were part of 17 centers and departments including neurology, psychiatry, primary care, emergency, and internal medicine. As of May 2015, about 36,000 patients consented to be part of the Partners HealthCare Biobank registry. The Partners HealthCare Biobank is a large research program designed to help researchers understand how people’s health is affected by their genes, lifestyle, and environment. Enrolled patients consent to the following: (1) a dedicated blood draw for preparation of DNA and blood derivatives for storage in the Biobank; (2) means to collect future discarded clinical specimens; (3) linkage of banked samples with their electronic medical record (EMR) and with health information collected through a secure survey; (4) specimen storage/distribution for broad use by IRB-approved Partners’ investigators; and (5) willingness to be recontacted as part of collaborating studies. We obtained the necessary institutional ethics review board (IRB) approval as a collaborating study to perform this validation work. Patients were dichotomized into either confirmed epilepsy diagnosis or unconfirmed epilepsy diagnosis. Epilepsy diagnosis was defined according to the 2014 criteria adopted by the International League Against Epilepsy (ILAE), including at least one of the following: (1) at least two unprovoked (or reflex) seizures occurring more than 24 h apart; (2) one unprovoked (or reflex) seizure and a probability of further seizures similar to the general recurrence risk (at least 60%) after two unprovoked seizures; (3) clinical diagnosis of an epilepsy syndrome. \*\*\*\*Compare AUC, The area under the curve (AUC) is a measure of accuracy. It is created by plotting the true-positive rate (i.e., proportion of cases identified as epilepsy cases that were confirmed epilepsy cases or “sensitivity”) against the false-positive rate (i.e., proportion of cases identified as epilepsy cases that were not confirmed epilepsy cases or “1 - specificity”). Comparatively, the accuracy is greater in the algorithm with greater AUC value.

Epilepsia © ILAE

**Table 1. Validated claims-based algorithms for identification of incidence or prevalent case of epilepsy**

Group	Holden et al. 2005	Reid et al. 2012	Tan et al. 2015	Franchi et al. 2013
Database	Lovelace Health Plan, Lovelace Health System in New Mexico	The Alberta Health Care Insurance Plan Registry (AHCIP)	Health Information Services (HIS) Department at St. Vincent's Hospital, Melbourne, AU	Drug administrative database of the Lombardy region (Northern Italy)
Validation method	Chart review of sample	Chart review of sample	Chart review of sample	Physician survey (correlate for chart review)
Epilepsy definition <sup>a</sup>	Epilepsy	Epilepsy or convulsions	Epilepsy, status epilepticus, or convulsions	Epilepsy
Age <sup>b</sup>	0–19, 65+	18+	All	All
Time-frame <sup>c</sup>	1996–2001	2002–2007	2012–2013	2000–2008
Emergency room	–	ICD-9-CM (345) or ICD-10-CA (G40 or 41)	–	–
Outpatient claims	1 claim: ICD-9-CM, 345.xx, 333.2, 779, 780.3, 780.39, 780.2, 780.31	ICD-9-CM (345) or ICD-10-CA (G40 or 41)	ICD-10 AM (G40 or G41) <sup>d</sup>	–
Inpatient claims	–	1) ICD-9-CM (345) 2) ICD-10-CA (G40 or 41)	–	–
EEG (CPT)	And I CPT: 95812–95958	–	–	At least one
AED	At least one	–	At least one	At least one

ICD, International Classification of Diseases; EEG, electroencephalography; MRI, magnetic resonance imaging; CPT, Current Procedural Terminology; AED, antiepileptic drug.

We reviewed the literature to identify validated computational algorithms with sensitivity and specificity of at least 60%.<sup>2,5,6,10–12,14–20</sup> Four of the six algorithms included ICD-10 codes, which were not phased into U.S. billing until October 2015. As a result, we converted all instances to ICD-9 equivalents, as determined by current guidelines.<sup>11</sup>

<sup>a</sup>For consistency in case ascertainment, we used the updated 2014 criteria for epilepsy diagnosis adopted by the International League Against Epilepsy, including documentation of at least one of the following: (1) at least two unprovoked (or reflex) seizures occurring >24 h apart; (2) one unprovoked (or reflex) seizure and a probability of further seizures similar to the general recurrence risk (at least 60%) after two unprovoked seizures, occurring over the next 10 years; or (3) clinical diagnosis of an epilepsy syndrome documented in the medical record.

<sup>b</sup>We performed stratified analysis by age groups to allow for comparison among the different algorithms.

<sup>c</sup>We selected a 2-year time frame (2013–2014) to allow for comparison among the different algorithms.

<sup>d</sup>We replaced ICD-10-CA (G40 or 41) by ICD-9-CM (345). This substitution was performed because our database contained only ICD-9 codes.

the distribution of prescribed medications abstracted using administrative claims.

## RESULTS

The records of 875 patients (64%) contained information leading to a diagnosis of epilepsy and 502 records (36%) contained data supporting an alternative diagnosis (e.g., syncope and provoked seizures). Appendix S7 describes the demographic and clinical characteristics of the patients with ( $n = 875$ , 64%) and without confirmed diagnosis of epilepsy ( $n = 502$ , 36%) based on medical records review. For all analyses we defined the statistical significance level as  $p < 0.05$ . We used the SAS Studio software package (SAS Institute Inc. Cary, NC, U.S.A.) to perform statistical analysis.

The algorithm described by Holden had the highest accuracy (area under the curve [AUC] = 0.73, 95% confidence interval [CI] 0.71–0.76) when applied to this study's dataset (cohort of patients older than 18 years old). This model used diagnosis and antiepileptic drug as predictors and had a positive predictive value of 84.1% (the same as published in 2005). This model had a sensitivity of 69.6% (95% CI 66.4–72.6%) and specificity of 77.1% [(95% CI 73.2–80.7%)]. This model gained accuracy when applied only to the cohort of patients older than

18 years but younger than 65 years (AUC area = 0.75, 95% CI 0.73–0.78) and lost accuracy when applied to the population 65 years or older (AUC area = 0.66, 95% CI 0.61–0.71). Figure 2A,B, Table 1, and Appendixes S8 and S9 detail the indicators of model strength between the six algorithms.

In Holden 1 applied to the cohort of patients older than 18 years old, 609 (sensitivity 69.6%) of 875 were true positives and 266 (30.4%) of 875 were false-positive (FP) cases. FP indicates that the algorithm incorrectly identified the patient as having epilepsy but the patient did not have epilepsy based on medical records review. This same algorithm produced 387 of 502 (specificity 77.1%) true-negative and 115/502 (22.9%) false-negative cases (FN). FN indicates that the algorithm incorrectly indicated that the patient did not have epilepsy but the patient actually had epilepsy. We obtained the list of false positives and false negatives to describe their clinical characteristics (Appendices S10 and S11).

Restricting the sample to require at least two diagnoses suggestive of epilepsy and one prescription of antiepileptic drug (Modified Holden 1.1) decreased the test performance slightly (AUC area = 0.69, 95% CI 0.67–0.72). Finally, including antiepileptic drug use data (Modified Holden 1.2) increased the test performance (AUC = 0.78, 95% CI 0.76–0.80).

Table 2. Epilepsy validation results for algorithms applied to the overall cohort of all patients older than 18 years

	Sensitivity % (95% CI)	Specificity % (95% CI)	PPV % (95% CI)	NPV % (95% CI)	DLR+ (95% CI)	DLR- (95% CI)	AUC (95% CI)	OR (95% CI)
Holden 1	<b>69.6 (66.4, 72.6)</b>	<b>77.1 (73.2, 80.7)</b>	<b>84.1 (81.2, 86.7)</b>	<b>59.3 (55.4, 63.1)</b>	<b>3.04 (2.57, 3.59)</b>	<b>0.39 (0.35, 0.44)</b>	<b>0.73 (0.71, 0.76)</b>	<b>7.70 (5.98, 9.92)</b>
Holden 1.1	52.5 (49.1, 55.8)	86.3 (82.9, 89.1)	86.9 (83.8, 89.7)	51.0 (47.6, 54.4)	3.82 (3.04, 4.79)	0.55 (0.51, 0.60)	0.69 (0.67, 0.72)	6.92 (5.20, 9.22)
Holden 1.2	<b>66.5 (63.3, 69.6)</b>	<b>89 (86, 91.6)</b>	<b>91.4 (88.9, 93.4)</b>	<b>60.4 (56.8, 63.9)</b>	<b>6.07 (4.71, 7.82)</b>	<b>0.38 (0.34, 0.41)</b>	<b>0.78 (0.76, 0.80)</b>	<b>6.07 (4.71, 7.82)</b>
Holden 2	33.4 (30.3, 36.6)	86.1 (82.7, 89.0)	80.7 (76.2, 84.6)	42.6 (39.5, 45.7)	2.39 (1.89, 3.03)	0.77 (0.73, 0.82)	0.60 (0.58, 0.62)	3.09 (2.32, 4.12)
Holden 3	12.2 (10.1, 14.6)	94.4 (92.0, 96.3)	79.3 (71.4, 85.8)	38.2 (35.5, 40.9)	2.19 (1.47, 3.27)	0.93 (0.90, 0.96)	0.53 (0.52, 0.55)	2.36 (1.54, 3.62)
Reid	49.8 (46.5, 53.2)	83.3 (79.7, 86.4)	83.8 (80.4, 86.9)	48.8 (45.4, 52.2)	2.98 (2.42, 3.66)	0.60 (0.56, 0.65)	0.67 (0.64, 0.69)	4.94 (3.78, 6.47)
Tan	<b>58.5 (55.2, 61.8)</b>	<b>84.5 (81.0, 87.5)</b>	<b>86.8 (83.8, 89.4)</b>	<b>53.9 (50.3, 57.4)</b>	<b>3.77 (3.05, 4.65)</b>	<b>0.49 (0.45, 0.54)</b>	<b>0.71 (0.69, 0.74)</b>	<b>7.67 (5.82, 10.10)</b>
Franchi	34.7 (31.6, 38.0)	83.5 (79.9, 86.6)	78.6 (74.1, 82.5)	42.3 (39.2, 45.5)	2.10 (1.69, 2.61)	0.78 (0.73, 0.83)	0.59 (0.57, 0.61)	2.69 (2.05, 3.53)

PPV, positive predictive value; NPV, negative predictive value; DLR+, diagnostic likelihood ratio for a positive test (DLR+ N I indicates that a positive test is more likely in a subject with epilepsy than in a subject without epilepsy); DLR-, diagnostic likelihood ratio for a negative test; AUC, area under the curve. The area under the curve is a measure of accuracy. It is created by plotting the true positive rate (i.e., proportion of cases identified as epilepsy cases that were confirmed epilepsy cases or "sensitivity") against the false-positive rate (i.e., proportion of cases identified as epilepsy cases that were not confirmed epilepsy cases or "1-specificity"). Comparatively, the accuracy is greater in the algorithm with greater receiver operating characteristic (ROC) curve value. The algorithms with the highest accuracy, defined by the highest ROC area is shown in bold.

Time frame: 2013–2014. The gold standard diagnosis of epilepsy was based on medical records abstraction (binary, epilepsy confirmed vs. not). Holden 1 included one claim coded with epilepsy, convulsions, syncope, or collapse (ICD-9 codes 345.xx, 333.2, 779, 780.3, 780.39, 780.2, 780.31); and at least one code for antiepileptic drug prescription; Holden 2 included the criteria defined in Holden 1 plus one code for EEG performed; Holden 3 included the criteria defined in Holden 3 plus one code for diagnosis of Depression (ICD-9: 311, 300.4) or Anxiety (ICD-9: 300, 300.02); Holden 1.1 included two claims coded with epilepsy, convulsions, syncope, or collapse (ICD-9 codes 345.xx, 333.2, 779, 780.3, 780.39, 780.2, 780.31); and at least one code for antiepileptic drug prescription; Holden 1.2 included one claim coded with epilepsy, convulsions, syncope, or collapse (ICD-9 codes 345.xx, 333.2, 779, 780.3, 780.39, 780.2, 780.31); and at least one code for antiepileptic drug prescription or at least one code for antiepileptic drug used (medical records review data); Reid included two outpatient claims (ICD-9 codes 345.xx) and/or one hospitalization coded as epilepsy or convulsions (ICD-9 codes 345.xx); Tan included one outpatient claim coded with epilepsy or convulsions (ICD-9 codes 345.xx) and one code for antiepileptic drug prescription; Franchi included one code for EEG performed and one code for antiepileptic drug prescription.

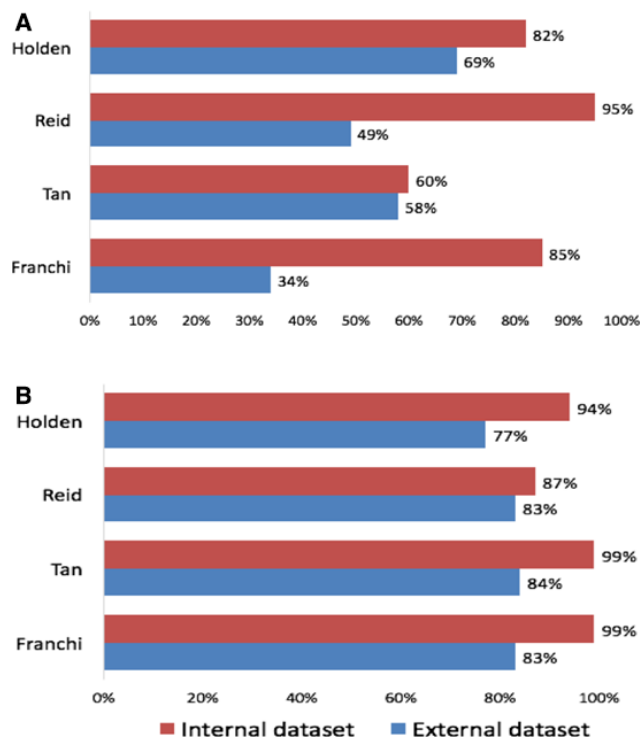


Figure 2.

(A) Sensitivity. Internal dataset (red) represents the sensitivity of the algorithms applied to the dataset in which they were developed (i.e., as published in the respective manuscripts). External dataset (blue) represents the sensitivity of the algorithms applied to our validation cohort. (B) Specificity. Internal dataset (red) represents the specificity of the algorithms applied to the dataset in which they were developed (i.e., as published in the respective manuscripts). External dataset (blue) represents the specificity of the algorithms applied to our validation cohort.

Epilepsia © ILAE

## DISCUSSION

Understanding the impact of epilepsy on patients broadly and assessing the quality of their care requires accurate identification of patients who have epilepsy using large datasets, such as those created by insurance claims. In the first comparison of existing algorithms for claims-based epilepsy identification, we find that the current algorithms perform modestly well at best, and all have important limitations. The most accurate algorithm for identifying patients with epilepsy includes one claim coded with the International Classification of Diseases, Ninth Revision (ICD-9) diagnosis of epilepsy, convulsions, syncope, or collapse, plus at least one code for an antiepileptic drug prescription or use. Incorporating more types of information (e.g., diagnosis, treatment, provider) could improve accuracy.

The scientific community often uses the area under the curve (AUC) for algorithm performance comparison. However, with some analyses, one might want to maximize sensitivity (e.g., screening of patients with possible epilepsy). The sensitivity of algorithms for identification of patients

who have epilepsy is also known to vary across different datasets and different algorithms. The sensitivity was highest in the algorithm by Reid et al., perhaps because this algorithm included inpatient and emergency rooms claims, which increases the number of care settings in which a patient may receive the diagnosis code. However, the sensitivity of the algorithm by Reid was substantially reduced when applied to our dataset, perhaps because there has been a high degree of referral leakage in our system (i.e., patients often travel long distances for outpatient epilepsy care but are encouraged to seek emergency care in local hospitals, which are often outside of our system). Of interest, the sensitivity of the algorithm by Tan did not substantially change when applied to our dataset, likely because our population characteristics were similar to the cohort used by Tan.

A common challenge facing clinical researchers is the timely acquisition of a statistically powerful and representative study sample.<sup>21</sup> Medical claims databases have emerged as a method for capitalizing on existing, codified data on regional and population-wide scales.<sup>16</sup> Research of this scale has the potential to strengthen epidemiologic surveillance as well as to monitor the impact of major health policy reform; however, its utility rests in the reliability of sample selection models.<sup>22,23</sup> According to the most recent epilepsy quality guidelines, quality care processes are recommended at least annually (e.g., personalized safety counseling, antiepileptic drug side effects query), which would require a minimum of two office visits within a time frame of 2 years.<sup>24</sup> Consequently, we kept the 2-year time frame and performed a second exploratory analysis, testing a modification of the best performing algorithm (modified Holden 1.1) to require at least two diagnoses suggestive of epilepsy and one prescription of antiepileptic drug.

Our study is consistent with prior studies suggesting that current drug therapy and EEG records provide only moderate sensitivity in identifying prevalent cases of epilepsy.<sup>5</sup> Of note, we compared algorithms based on populations in well-defined geographical regions of Italy, Canada, Australia, and the United States. Not surprisingly, Holden's, which was originally validated in a dataset from New Mexico (state located in the southwestern region of the United States), was the best performing algorithm when applied to our dataset.<sup>4,5,11,12</sup>

However, this cross-study comparison was limited by the general lack of congruence in the way the algorithms were developed and validated. In 2012, Reid used only diagnostic codes applied to The Alberta Health Care Insurance Plan Registry (AHCIP) in Canada. They explored 18 algorithms and suggested that the coding algorithm with the best diagnostic accuracy to identify epilepsy cases was two physician claims or one hospitalization over a 2-year time frame (Sn 88.9%, Sp 92.4%, PPV 89.2%, NPV 92.2%).<sup>5</sup> These results were quite different when the same algorithm was applied to our database. A possible explanation is that there was overfitting of the model. In overfitting, a statistical model describes

random error or noise instead of the underlying true predictive power. In particular, a model is typically trained by maximizing its performance on some set of training data (i.e., their own dataset, internal dataset). However, its efficacy is determined not by its performance on the training data but by its ability to perform well on unseen data (i.e., external dataset). As an extreme example, a simple model or learning process can predict the training data simply by memorizing the training data in its entirety, but such a model will typically fail drastically when making predictions about new or unseen data, since the simple model has not learned to generalize at all. To avoid overfitting, it is necessary to use additional techniques such as cross-validation or early stopping. Another possible explanation is the expected national and regional variation in coding patterns (e.g., secondary to different reimbursement incentives across different health care systems). Taking into consideration the national and regional variation, our study provides valuable information about each model's ability to generalize by evaluating their performance on a set of data not used for algorithm development, which is assumed to approximate the range of accuracy of the subsequent studies that used one of the six selected algorithms without prior validation.<sup>13,25</sup>

Later in 2013, Franchi et al. used a retrospective physician survey as the reference standard. This differs from chart review because it adds the physician recall bias. In fact, the Franchi et al. study included a small sample of epilepsy patients ( $n = 71$ ). They reported high accuracy (Sn 85%, Sp 99%, PPV 64%, NPV 99%), which was not replicated when applied to our dataset.<sup>4</sup> We believe that the robustness of our accuracy results is supported by our validation sample size of >700 epilepsy patients. Our results are also intuitively supported by the typical demographic and clinical characteristics of our cohort (e.g.: epilepsy patients often had abnormal brain imaging and neurophysiology studies compared to patients without epilepsy). Unfortunately, Franchi et al. did not report the characteristics of their dataset.

Most recently in 2015, Tan used data from the health information services (HIS) department at a hospital in Melbourne, Australia, to validate the algorithm utilizing ICD-10 codes for epilepsy and  $\geq 1$  antiepileptic drug (AED), which is essentially the same as in Holden but with ICD-10 codes. They reported a good accuracy (Sn 60, Sp 99.9% PPV 81.4%, NPV 99%). The comparison between this algorithm and the application to our dataset was threatened by our conversion of ICD-10 codes to ICD-9 codes, because ICD-10 codes were not phased into U.S. billing until October 2015. This conversion increased our uncertainty about the validity of the epilepsy identification algorithms.<sup>11</sup>

Accordingly, a comparison between International League Against Epilepsy (ILAE) disease classifications and ICD codes across the most recent iterations, both ICD-9 and ICD-10, showed limited cross-validation strength and considerable variation across studies.<sup>18</sup> Nevertheless, our study

builds on the literature in which suggests that quality of estimations of epilepsy based on claims data depends on the case definition of epilepsy as well as on the demographic and clinical characteristics of the population and the health care system.<sup>4,5,10–12</sup> The discussion about the secondary findings of this study is provided in Appendix S13.

A series of limitations of the previously published studies and our study serve as potential directions for further accuracy optimization. First, our study sample may not be well representative of a general population. For instance, the sampling from Partners HealthCare Biobank list of enrollees may have selected patients with higher educational level (e.g., able to understand the consent form procedures) and with more severe disease (e.g., willing to provide blood samples for researchers). Well-educated patients were able to provide a more accurate description of their events. In addition, more severely ill patients may have yielded more clinical evidence of the diagnosis of epilepsy (and more accurate medical documentation as the reference standard). Admittedly, we had strategically selected the Biobank enrollees in this validation effort as part of a larger project that will examine genetic and biomarker factors in relationship to epilepsy diagnoses and care. Nevertheless, our Appendix S7 shows a table that suggests that the clinical and demographic characteristics of the validation sample were representative of a broad cohort of epilepsy patients with multiple types of seizures and etiologies.

Second, in our validation sample, we noted that the majority of patients had documented two or more unprovoked (or reflex) seizures occurring more than 24 h apart (ILAE criteria “a”). Less often patients had documentation of a clinical diagnosis of an epilepsy syndrome (ILAE criteria “c”) or one unprovoked seizure and a probability of further seizures similar to the general recurrence risk after two unprovoked seizures (ILAE criteria “b”). However, we failed to track how many patients met each criteria or combinations. We also failed to track how many of the nonconfirmed cases were due to a lack of evidence to justify a diagnosis of epilepsy (e.g., events of unclear etiology) versus lack of documentation in the chart to allow for a diagnosis of epilepsy (i.e., loss to follow-up). In addition, it would be natural to consider that future studies may benefit from a validation dataset that includes the likelihood of diagnosis of epilepsy based on more variables such as response to prophylactic therapy, types of EEG and brain MRI abnormalities, or the specialty of the physician making the diagnosis (e.g., non-neurologist vs. general neurologist vs. epilepsy specialist). However, adding information about the results of brain MRI and EEG might cause more confusion because abnormalities on these tests alone have low predictive value in many circumstances (e.g., sensitivity of routine EEGs in patients with epilepsy is <50%). Symptom resolution after initiation of prophylactic therapy is also known to bias toward error, as patients with nonepileptic events often have symptom resolution with the same prophylactic therapies

(i.e., placebo effects) and up to 30% of patients with epilepsy may continue to have seizures despite adequate treatment.<sup>26</sup> Overall, until a biomarker becomes available, we may rely on physician judgment for case ascertainment of epilepsy. With that in mind, our medical records abstraction was performed by a well-trained medical student and a neurologist, under the close supervision of an epilepsy specialist in an effort to produce the most accurate categorization.

Similar to previous studies, our study has given little attention to the examination of the accuracy of epilepsy classification (e.g., whether claims indicating generalized epilepsy accurately represent a patient with generalized epilepsy syndrome). This is particularly important in comparative epilepsy research as efficacy of AEDs often differs across seizure types and syndromes. There are some published categorization methods that also merit cross-validation.<sup>16,27</sup> In addition, the extent to which clinically scaled measures, such as provider specialty involvement, may also increase accuracy remains unverified.<sup>28</sup>

Future studies may examine the accuracy of algorithms using sub-decimal ICD-9 codes for epilepsy. For instance, 65% (570/875) of patients with epilepsy received the code 345.1, which is titled “epilepsy, unspecified, without mention of intractable epilepsy.” In contrast, only 22% (112/502) of patients with epilepsy received this code (345.1), as shown in Appendix S5.

Future studies may also refine their algorithms based on the prescription patterns among patients who do and do not have a confirmed diagnosis of epilepsy, which has been described in the supporting information of the present study. For instance, gabapentin is classified as antiepileptic medication but has been more often prescribed for patients without epilepsy. This is consistent with the existing literature that describes that many AEDs (e.g., gabapentin, carbamazepine, and topiramate) are widely used in the treatment of neuropathic pain or headache. When used in patients with epilepsy, gabapentin is often a third- or fourth-line agent. Based on that, a reasonable nested algorithm that applies different algorithms depending on whether the patient is using gabapentin as monotherapy (likely not an epilepsy case) versus polytherapy (a possible epilepsy case) may be valuable.<sup>29</sup>

Similarly, our analysis of false-positive cases (Appendix S10) highlights two common problems: miscoding (i.e., claims of possible epilepsy for a patient who never had a seizure) and mis-management (i.e., long-term use of AEDs for patients who never had a seizure). Of note, current guidelines for traumatic brain injury (TBI) and aneurysmal subarachnoid hemorrhage recommend AEDs for a few days to decrease posttraumatic or posthemorrhage seizure risk.<sup>30</sup> Only a few patients may develop seizures, and they would then meet the criteria for symptomatic epilepsy and require longer-term prophylaxis and establishment of care with a neurologist. However, our medical records review demonstrated that long-term antiseizure prophylaxis has been widely prescribed for patients with other

conditions such as intracerebral tumors, craniotomy, ischemic stroke, or other forms of intracerebral hemorrhage who never had a seizure. Unfortunately, there is no clear evidence to support the long-term use of AEDs even for patients with aneurysmal subarachnoid hemorrhage.

Finally, new-onset epilepsy in older patients is often associated with other neurologic conditions, including Alzheimer's disease-related dementia (ADRD), stroke, and brain tumors.<sup>31–33</sup> Therefore, future studies may test the use of cardiovascular diseases and tumor comorbidities captured in ICD-9 codes to see if this can further increase accuracy of the algorithms applied to older adults. Future studies should also focus on the comparison of our cross-validation findings with the cross-validation of algorithms for identification of other neurologic diseases and even other chronic diseases.

In conclusion, this study highlights the immediate need for refinement and cross-validation of algorithms for identification of epilepsy in individuals who may later develop recurrent unprovoked seizures using a combination of data sources.

## DISCLOSURE OF CONFLICT OF INTEREST

No potential conflicts of interest exist for all authors for the last 3 years. Lidia Moura is the recipient of a 2015 Clinical Research Fellowship sponsored by the American Brain Foundation and reports no disclosures. Maggie Price, Daniel B. Hoch, and Andrew J. Cole report no disclosures. John Hsu receives grant funding from the National Institutes of Health (NIH) and the Agency for Healthcare Research and Quality (AHRQ) (1R01 CA164023, 2P01AG032952, R01 HD075121, R01 MH104560, and R01HS023128) and reports no disclosures. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

## REFERENCES

- Stovner LJ, Hoff JM, Svalheim S, et al. Neurological disorders in the Global Burden of Disease 2010 study. *Acta Neurol Scand Suppl* 2014;129:1–6.
- Faught E, Richman J, Martin R, et al. Incidence and prevalence of epilepsy among older US Medicare beneficiaries. *Neurology* 2012;78:448–453.
- Tsiropoulos I, Gichangi A, Andersen M, et al. Trends in utilization of antiepileptic drugs in Denmark. *Acta Neurol Scand* 2006;113:405–411.
- Franchi C, Giussani G, Messina P, et al. Validation of healthcare administrative data for the diagnosis of epilepsy. *J Epidemiol Community Health* 2013;67:1019–1024.
- Reid AY, Germaine-smith CS, Liu M, et al. Development and validation of a case definition for epilepsy for use with administrative health data. *Epilepsy Res* 2012;102:173–179.
- Betjemann JP, Josephson SA, Lowenstein DH, et al. Trends in status epilepticus—related hospitalizations and mortality redefined in US practice over time. *JAMA Neurol* 2016;72:650–655.
- Choi H, Hamberger MJ, Munger Clary H, et al. Seizure frequency and patient-centered outcome assessment in epilepsy. *Epilepsia* 2014;55:1205–1212.
- Blank L, Baxter S, Baird W, et al. Understanding referral patterns to an epilepsy clinic: professional perceptions of factors influencing the referral of older adults. *Seizure* 2013;22:698–702.
- Kuehn BM. New adult first seizure guideline emphasizes an individualized approach. *JAMA* 2015;314:111–113.
- Reid AY, Quan H, Hill MD, et al. How accurate is ICD coding for epilepsy? *Epilepsia* 2010;51:62–69.
- Tan M, Wilson I, Braganza V, et al. Epilepsy & Behavior Development and validation of an epidemiologic case definition of epilepsy for use with routinely collected Australian health data. *Epilepsy Behav* 2015;51:65–72.
- Holden EW, Nguyen HT, Grossman E, et al. Estimating prevalence, incidence, and disease-related mortality for patients with epilepsy in managed care organizations. *Epilepsia* 2005;46:311–319.
- Collins GS, Le Manach Y. Flaws in the development of the CHES score. *Eur J Neurol* 2016;23:41.
- Tu K, Wang M, Jaakkimainen RL, et al. Assessing the validity of using administrative data to identify patients with epilepsy. *Epilepsia* 2014;55:335–343.
- Pugh MJV, Copeland LA, Zeber JE, et al. The impact of epilepsy on health status among younger and older adults. *Epilepsia* 2005;46:1820–1827.
- Thurman DJ, Beghi E, Begley CE, et al. Standards for epidemiologic studies and surveillance of epilepsy. *Epilepsia* 2011;52:2–26.
- Bergen DC, Beghi E, Medina MT. Revising the ICD-10 codes for epilepsy and seizures. *Epilepsia* 2012;53:3–5.
- Zuberi SM, Jette N, Beghi E, et al. Special report ICD coding for epilepsy: past, present, and future — A report by the International League Against Epilepsy Task Force on ICD codes in epilepsy Codes for Epilepsy, and the Challenges of Mapping Epilepsy Codes between These Two ICD Versions. *Epilepsia* 2015;10:348–355.
- Fillit H, Geldmacher DS, Welter RT. Optimizing coding and reimbursement to improve management of Alzheimer's disease and related dementias. *J Am Geriatr Soc* 2002;50:1871–1878.
- Walker AM. Pattern recognition in health insurance claims databases. *Pharmacoepidemiol Drug Saf* 2001;10:393–397.
- Vickrey BG, Hirtz D, Waddy S, et al. Comparative effectiveness and implementation research: directions for neurology. *Ann Neurol* 2012;71:732–742.
- Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol* 2016;183:758–764.
- Baca CB, Barkley GL. Can big data help us close an epilepsy care gap? *Neurology* 2015;85:1014–1015.
- Fountain NB, Van Ness PC, Bennett A, et al. Quality improvement in neurology: epilepsy update quality measurement set. *Neurology* 2015;84:1483–1487.
- Petersen ML, van der Laan MJ. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology* 2014;25:418–426.
- Berg AT, Scheffer IE. New concepts in classification of the epilepsies: entering the 21st century. *Epilepsia* 2011;52:1058–1062.
- Kurth T, Lewis BE, Walker AM. Health care resource utilization in patients with active epilepsy. *Epilepsia* 2010;51:874–882.
- Roberts K, Costelloe D, Hutchinson M, et al. What difference does a neurologist make in a general hospital? Estimating the impact of neurology consultations on in-patient care. *Ir J Med Sci* 2007;176:211–214.
- Leong C, Muhammad M, Gomes T, et al. Antiepileptic use for epilepsy and nonepilepsy disorders. *Neurology* 2016;86:939–946.
- Rowe AS, Goodwin H, Brophy GM, et al. Seizure prophylaxis in neurocritical care: a review of evidence-based support. *Pharmacotherapy* 2014;34:396–409.
- Leppik IE. Epilepsy in the elderly: scope of the problem. *Int Rev Neurobiol* 2007;81:1–14.
- Martínez-Juárez IE, Alonso ME, Medina MT, et al. Juvenile myoclonic epilepsy syndromes: family studies and long-term follow-up. *Brain* 2006;129:1269–1280.
- Berg AT, Berkovic SF, Brodie MJ, et al. Revised terminology and concepts for organization of seizures and epilepsies: report of the ILAE Commission on Classification and Terminology, 2005–2009. *Epilepsia* 2010;51:676–685.
- Christensen J, Vestergaard M, Pedersen MG, et al. Incidence and prevalence of epilepsy in Denmark. *Epilepsy Res* 2007;76:60–65.



## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

**Appendix S1.** Explanation of the sample query.

**Appendix S2.** Medical records review clarification.

**Appendix S3.** Algorithm selection justification.

**Appendix S4.** Table listing generic and brand antiepileptic drug names.

**Appendix S5.** Table clarifying claims-based diagnosis between patients with and without epilepsy.

**Appendix S6.** Table comparison of the proportion of prescription fills between patients with and without epilepsy.

**Appendix S7.** Table assessment of demographic and clinical characteristics of the validation cohort.

**Appendix S8.** Table illustration of epilepsy validation results for algorithms applied to the overall cohort of patients 18 to 64 years old.

**Appendix S9.** Table demonstration of epilepsy validation results for algorithms applied to the overall cohort of patients older than 64 years.

**Appendix S10.** Table depiction of false-positive cases from the algorithm of highest performance.

**Appendix S11.** Table depiction of false-negative cases from the algorithm of highest performance.

**Appendix S12.** Description of abstraction detail.

**Appendix S13.** Discussion of secondary findings.