# The accuracy of using administrative healthcare data to identify epilepsy cases: A systematic review of validation studies

Received: 30 July 2019 | Revised: 28 April 2020 | Accepted: 29 April 2020

DOI: 10.1111/epi.16547



Epilepsia®

CRITICAL REVIEW AND INVITED COMMENTARY

# The accuracy of using administrative healthcare data to identify epilepsy cases: A systematic review of validation studies

Gashirai K. Mbizvo[1] | Kyle H. Bennett[1] | Christian Schnier[2] | Colin R. Simpson[2,3] | Susan E. Duncan[1,4] | Richard F.M. Chin[1,5]

[1]Muir Maxwell Epilepsy Centre, Centre for Clinical Brain Sciences, The University of Edinburgh, Edinburgh, UK

[2]Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, UK

[3]School of Health, Faculty of Health, Victoria University of Wellington, Wellington, NZ

[4]Department of Clinical Neurosciences, Western General Hospital, Edinburgh, UK

[5]Royal Hospital for Sick Children, Edinburgh, UK

**Correspondence**
Gashirai K. Mbizvo, Muir Maxwell Epilepsy Centre, Child Life and Health, 20 Sylvan Place, EH9 1UW, Edinburgh, UK.
Email: gashiraimbizvo@hotmail.com

**Funding information**
Epilepsy Research UK, Grant/Award Number: R44007; The Juliet Bergqvist Memorial Fund, Grant/Award Number: N/A

## Abstract

Our objective was to undertake a systematic review ascertaining the accuracy of using administrative healthcare data to identify epilepsy cases. We searched MEDLINE and Embase from 01/01/1975 to 03/07/2018 for studies evaluating the diagnostic accuracy of routinely collected healthcare data in identifying epilepsy cases. Any disease coding system in use since the International Classification of Diseases, Ninth Revision (ICD-9) was permissible. Two authors independently screened studies, extracted data, and quality-assessed studies. We assessed positive predictive value (PPV), sensitivity, negative predictive value (NPV), and specificity. The primary analysis was a narrative synthesis of review findings. Thirty studies were included, published between 1989 and 2018. Risks of bias were low, high, and unclear in 4, 14, and 12 studies, respectively. Coding systems included ICD-9, ICD-10, and Read Codes, with or without antiepileptic drugs (AEDs). PPVs included ranges of 5.2%–100% (Canada), 32.7%–96.0% (USA), 47.0%–100% (UK), and 37.0%–88.0% (Norway). Sensitivities included ranges of 22.2%–99.7% (Canada), 12.2%–97.3% (USA), and 79.0%–94.0% (UK). Nineteen studies contained at least one algorithm with a PPV >80%. Sixteen studies contained at least one algorithm with a sensitivity >80%. PPV was highest in algorithms consisting of disease codes (ICD-10 G40-41, ICD-9 345) in combination with one or more AEDs. The addition of symptom codes to this (ICD-10 R56; ICD-9 780.3, 780.39) lowered PPV. Sensitivity was highest in algorithms consisting of symptom codes with one or more AEDs. Although using AEDs alone achieved high sensitivities, the associated PPVs were low. Most NPVs and specificities were >90%. We conclude that it is reasonable to use administrative data to identify people with epilepsy (PWE) in epidemiological research. Studies prioritizing high PPVs should focus on combining disease codes with AEDs. Studies prioritizing high sensitivities should focus on combining symptom codes with AEDs. We caution against the use of AEDs alone to identify PWE.

[Correction added on June 09, 2020, after first online publication: "USA" was removed from affiliation 3.]

**KEYWORDS**
diagnostic test accuracy, international classification of diseases, positive predictive value, routine data, seizures, sensitivity

## 1 | INTRODUCTION

Administrative healthcare data consist of routine diagnostic and procedural information collected about patients when they use healthcare services.[1] These national data sets are widely available and less-intrusive potential resources for medical research.[1] However, their diagnostic accuracy requires validation because the data were collected originally for routine nonscientific purposes. Poor or incomplete hospital discharge letters and clinical coding errors are examples of potential sources for administrative data inaccuracies.[2] Systematic reviews of administrative data validation studies increase confidence in case-ascertainment accuracy estimates for a particular condition and scrutinize the quality of available evidence.[3-11] To date, there has been only one systematic review of studies validating the accuracy of administrative epilepsy data sets.[12] Although this was helpful, the 11 studies reviewed were from the United States and Canada alone and published up to the year 2010 only. The International Classification of Diseases, Ninth Revision (ICD-9) system was used in all but one study, and the review lacked a risk of bias assessment. We now provide an updated systematic review of epilepsy validation studies worldwide. We include risk of bias assessments and evaluate ICD-9, ICD-10, and other coding systems routinely used.[13,14] We focus on positive predictive value (PPV) and sensitivity, as these are the most commonly reported validation outcomes in the literature.[11,12] However, we also report negative predictive value (NPV) and specificity, where available.

**Key Points**

- Administrative healthcare data can be used confidently to identify people with epilepsy in epidemiological research.
- Studies prioritizing high positive predictive values should focus on combining disease codes with antiepileptic drugs.
- Studies prioritizing high sensitivities should focus on combining symptom codes with antiepileptic drugs.
- Antiepileptic drugs alone are unlikely to accurately capture persons with epilepsy from administrative data.

## 2 | METHODS

The aims and inclusion criteria were established before conduct of the review and the a priori study protocol was registered on PROSPERO (CRD42017081212-https://bit.ly/2VOdoNj),[15] and published.[16] The review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) system of reporting.[17] Ethical approval was not required.

### 2.1 | Eligibility criteria

We included studies according to the following criteria:

- *Study country:* No restrictions;
- *Language:* No restrictions: translations were sought for any non-English texts;
- *Study design:* No study type was excluded;
- *Participants:* People with epilepsy of all ages;
- *Validated database:* Studies evaluating the diagnostic accuracy of routinely collected (administrative) healthcare data using ICD-9, ICD-10, or any other diagnostic coding system in use since the advent of ICD-9 in 1975[18];
- *Reference standard:* There were no minimum requirements for the types of gold standard used. The risk of study bias was assessed using the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool[19];
- *Observations and outcomes:* Studies had to report at least a PPV, sensitivity, NPV, or specificity estimate, or provide data from which these could be calculated;
- *Timeframe:* Studies conducted from 01/01/1975 onward.

### 2.2 | Information sources, literature searches, and study selection

MEDLINE and Embase (including Embase gray literature)[15,20,21] were searched from 01/01/1975 to 03/07/2018 (date last searched: 04/07/2018) for potentially eligible studies (see Appendix S1 for search strategy). Reference lists of manuscripts screened were also searched to identify potentially eligible studies. Any studies made aware to us by colleagues were also screened. Two authors (GKM and KB) independently screened titles, abstracts, or full-length articles for eligibility and inclusion, with disagreements resolved by

consensus or, where necessary, third reviewer adjunction (RFMC).

### 2.3 | Data abstraction

GKM and KB independently abstracted data pertaining to study location, participant age, administrative data setting, coding system, sample size, algorithm(s), reference standard details, PPV, sensitivity, NPV, specificity, and confidence intervals (CIs). We used a pre-piloted data abstraction form (Appendix S2). We made attempts to collect any relevant missing or unpublished study data by contacting the corresponding study author by email. We settled any differences in abstracted data by mutual review of the relevant study article.

### 2.4 | Quality assessment

We assessed the risk of bias and applicability concern for included studies using the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2)[19] tool adapted by Wilkinson et al[11] for administrative validation studies (Table 1). GKM and KB independently graded each study's risk of bias and applicability concern as low, unclear, or high across the four categories: patient selection, administrative database ("index test"), reference standard, and study participant flow.[19] We used the QUADAS-2 abstraction form in Appendix S3 (16 quality decisions per study), with disagreements resolved by consensus or, where necessary, third reviewer adjunction (RFMC).

### 2.5 | Summary measures and synthesis of results

Cohen's kappa statistic was used estimate interrater agreement (above chance) between GKM and KB for study eligibility selection and QUADAS-2 quality assessment.[22,23] We reported all of the algorithms and associated PPVs, sensitivities, NPVs, and specificities for epilepsy provided by each study, but we excluded any algorithms attempting to further identify epilepsy by seizure type or epilepsy syndrome to reduce clinical heterogeneity of the condition being assessed. The primary analysis was a narrative synthesis of results for PPV, sensitivity, NPV, and specificity.[10-12,24] This included results summarized as ranges across studies for different countries to assist in correcting for potential differences in epilepsy prevalence between different countries and also to facilitate reporting and interpreting results from a wide range of countries (each with potentially varying coding practices). Results were also summarized as ranges across studies exclusively using medical records as their diagnostic gold standard to allow interpretation of results when applied across a consistent optimal reference standard. We created forest plots consisting of the best and worst PPV, sensitivity, NPV, and specificity from each study, allowing us to discuss the study characteristics and diagnostic algorithms associated with the best cases and worst cases of PPV, sensitivity, NPV, and specificity. We did not pool PPV and NPV estimates into a formal meta-analysis, as this approach is cautioned against owing to the expected heterogeneity from variation in disease prevalence and different disease positivity thresholds.[11,12,24] We used a Reitsma random-effects bivariate meta-analysis model to estimate a summary sensitivity and specificity because these two outcomes are relatively resilient to changes in disease prevalence.[25,26] This analysis is only possible across studies providing all of the true positives, false positives, true negatives, and false negatives.[26] All estimates were reported with 95% CIs where possible. We considered participants aged ≥18 years as adults and <18 years children. Data were processed and analyzed using RSTUDIO Version 1.2.1335, Boston, MA: RStudio, Inc.

## 3 | RESULTS

### 3.1 | Search and study characteristics

Thirty studies[14,27-55] were included from 197 potentially eligible records (Figure 1). The level of study selection agreement between authors was graded as moderate,[23] detailed in Appendix S4. Twenty-seven studies were identified from searching MEDLINE and Embase, and three[36,50,55] were identified from hand-searching the reference lists of potentially eligible studies. One study had unpublished validation outcome data for epilepsy, which were kindly provided to us separately by the corresponding author upon request.[40] The included studies were all published between 1989 and 2018 (only two before 2000).[31,49] All were carried out in high-income countries. Excluded studies are listed in Appendix S5.[21] Participants were adults in eight studies,[29,33,40,42,46,48,52,53] children in two studies,[27,41] both adults and children in 14 studies,[14,28,34-39,43,49-51,54,55] and unclear age in six studies.[30-32,44,45,47] Epilepsy was the target diagnostic condition in all included studies. The majority (19 studies) used hand-searched medical records as the diagnostic gold standard.[14,30,32,37-43,45-48,50-54] In the remaining 11 studies, the other gold standards used were the general practitioner (GP: 3 studies),[35,36,49] being on antiepileptic drugs (AEDs: three studies, one of which used rufinamide only),[31,33,44] medical records and/or parent telephone interview,[27] an epilepsy center patient list,[28] a previous epilepsy prevalence study,[29] another administrative data set,[34] and a specialist epilepsy database.[55] Administrative data sets validated included inpatients, outpatients, pharmacy, accident and emergency, physician claims, and primary care. Algorithms included ICD-9, ICD-10 (including two studies

**TABLE 1** Risk of bias and applicability concern judgements in QUADAS-2[19] adapted for administrative validation[11]

| Domain | Patient selection | Administrative database | Reference standard | Flow and timing |
|---|---|---|---|---|
| Description | Describe methods of patient selection: | Describe the administrative database and how it was used and interpreted: | Describe the reference standard and how it was conducted and interpreted: | Describe any patients in the validation cohort who were not found within the reference standard or who were excluded from cross-tabulation of the administrative data diagnoses results against the results of the reference standard diagnoses: |
| | What is the study question? | Where available, include comment on how coding was done, by whom, and whether or not there was reimbursement for coding. | Where available, include comment on quality of the reference standard, including the level of experience of clinicians making the diagnosis, access to diagnostic tests such as electroencephalography and telemetry, and the thresholds/criteria used to make a diagnosis of epilepsy. | Describe the time interval and any interventions between administrative database diagnosis and reference standard diagnosis: |
| Signalling questions (yes/no/unclear) | Was a consecutive or random sample of patients enrolled? | Were the administrative database diagnosis results interpreted without knowledge of the results of the reference standard diagnosis? | Is the reference standard likely to correctly classify the epilepsy? | Was there an appropriate interval between administrative database diagnosis and reference standard diagnosis? |
| | Did the study avoid inappropriate exclusions? | If a diagnostic threshold was used, was it pre-specified? | Were the reference standard results interpreted without knowledge of the results of the administrative database diagnosis? | Did all patients receive a reference standard?<br>Did all patients receive the same reference standard?<br>Were all patients included in the analysis? |
| Risk of bias: High/low/unclear | Could the selection of patients have introduced bias? | Could the conduct or interpretation of the administrative database have introduced bias? | Could the reference standard, its conduct, or its interpretation have introduced bias? | Could the patient flow have introduced bias? |
| Concerns regarding applicability: High/low/unclear | Are there concerns that the included patients do not match the study question? | Are there concerns that the administrative database, its conduct, or interpretation differ from the study question? | Are there concerns that epilepsy, as defined by the reference standard, does not match the study question? | |

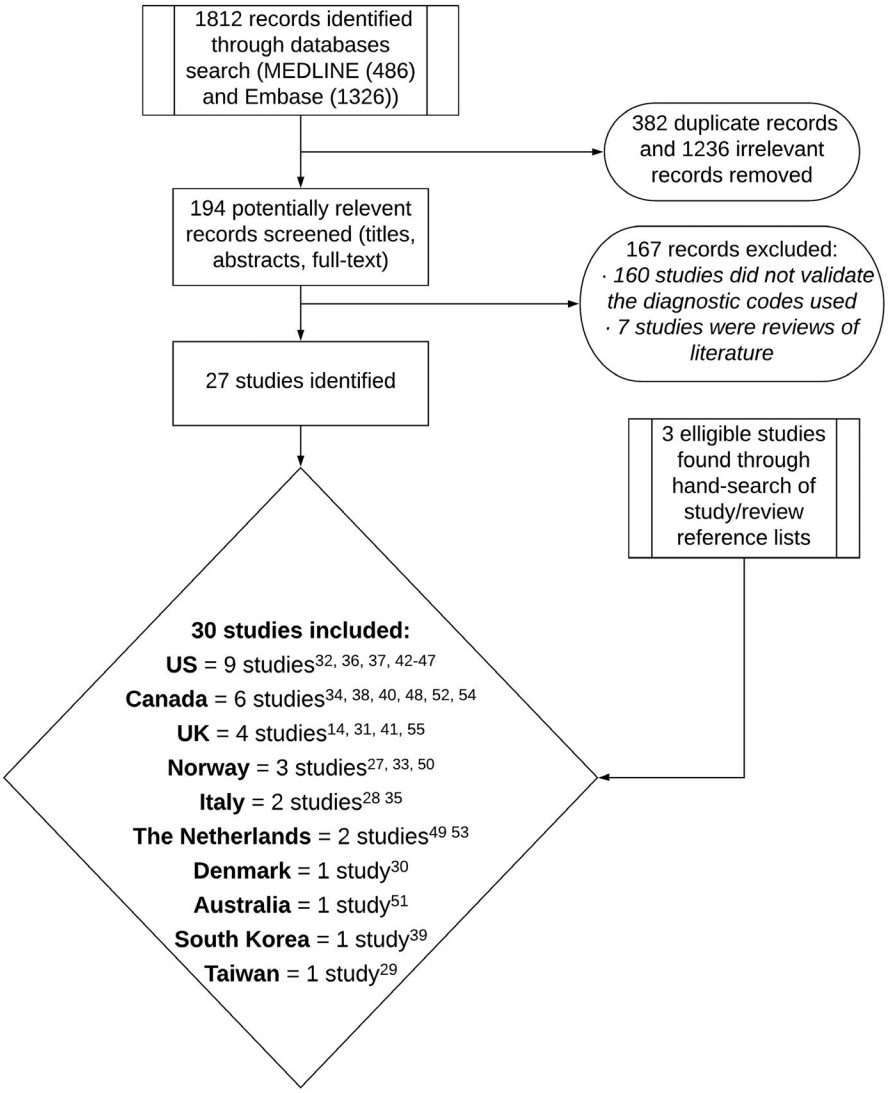


**FIGURE 1** Flow diagram of study selection process. The three studies identified from hand-searching reference lists of screened literature were Pickrell 2015[55] (found in the included study Fonferko-Shadrach 2017[14]), Syvertsen 2015[50] (found in the excluded study Aaberg 2016[59]), and Frost 2000[36] (found in the systematic review Kee 2012[12])

that combined ICD-8 and ICD-10 data sets together),[30,52] Read Codes (Version 2), AEDs, hospitalizations, and procedures such as electroencephalography (EEG) or vagus nerve stimulation (VNS). Twenty-eight studies assessed PPV,[14,27-43,45-54] of which 14 also assessed sensitivity.[14,28,29,33-35,37,38,40,42,48,51,52,54] Two studies assessed sensitivity alone.[44,55] Fourteen studies assessed NPV and specificity,[14,28,29,33-35,37,38,40,42,48,51,52,54] and one study assessed specificity without NPV.[44]

## 3.2 | Quality assessment

The mean level of quality assessment agreement between authors was graded as fair,[23] detailed in Appendix S6A. Four studies had low risks of bias and low applicability concerns across all categories (Appendix S6B). Fourteen studies had one or more high-risk ratings. High risks of bias from reference standard conduct were common. This was largely due to inadequate blinding of the reference standard reader,[35,36,49] use of unvalidated administrative data sets as the reference standard,[34] and use of AEDs as the reference standard.[31,33,44] High risks of bias from study flow and timing were also common, largely due to all study participants not receiving the same reference standard,[27,34,36,40] or due to the diagnostic alternatives of "epilepsy" or "not epilepsy" for missing participants not being explored in a "best-case" or "worst-case" sensitivity analysis, respectively.[14,49,54] The remaining 12 studies had ratings of both low and unclear risks across different categories, resulting in overall risks being graded unclear.[28,32,37,41-43,45-47,50,52,55] The most common reason for this was lack of clarity about whether or not those reading the reference standard were blinded to the participants' administrative codes.

## 3.3 | Narrative analysis

### 3.3.1 | Overall PPV and sensitivity

Overall, 28 studies generated 172 algorithms estimating PPV and 121 of these algorithms (generated by 13 studies)

also estimated sensitivity. Figure 2A,B illustrates the range of these PPV and sensitivity results (arranged by country). There was no clear visual pattern[11] to suggest that any particular country's PPVs or sensitivities performed better than another's. Figure 2C,D illustrates the range of PPVs and sensitivities arranged by overall risks of bias or applicability concern. A larger proportion of PPVs (81% of 70 algorithms) and sensitivities (83% of 46 algorithms) from low-risk studies were of a high magnitude (>80%). By contrast, a smaller proportion of PPVs (12% of 57 algorithms) and sensitivities (52% of 42 algorithms) from high-risk studies were of a high magnitude (>80%). For unclear-risk studies, a small proportion of PPVs (33% of 45 algorithms) and sensitivities (29% of 35 algorithms) were of a high magnitude (>80%). Figure 3A,B illustrates the range of these PPV and sensitivity results grouped by the 19 studies exclusively using medical records as the diagnostic gold standard against the 11 studies that used the other methods. PPV ranged 5.2%–100% and sensitivity ranged 61.0%–100% across the studies using medical records as the diagnostic gold standard. PPV ranged 5.0%–93.0% and sensitivity ranged 36.6%–97.3% across the studies using other methods as the diagnostic gold standard. The lookup table in Appendix S7 ranks all the algorithm results from highest to lowest PPV alongside their sensitivity results (where available), other associated study characteristics, and overall risks of study bias or applicability concern.

Figure 4A,B demonstrates the optimal diagnostic PPV and sensitivity algorithms, respectively, taken from each study and arranged in a forest plot. At their best, PPV or sensitivity were >80% in the majority of studies (19 of 28 studies for PPV and 12 of 16 studies for sensitivity). From visual inspection[11] of these figures, it was difficult to appreciate any potential effect of varying diagnostic gold standards used on optimal PPV and sensitivity estimates because hand-searched medical records predominated the gold standards used across studies. Inspecting the figures[11] otherwise revealed that the high PPV and sensitivity estimates >80% appeared to cross multiple different coding systems and algorithms used (including ICD-8, ICD-9, ICD-10, or Read Codes with and without AEDs or procedures), across various population ages and settings including accident and emergency, inpatients, outpatients, and primary care. Nearly all of the studies with an optimal PPV ≤80% had a high risk of bias or applicability concern (six of nine studies, with unclear risks in the remaining three studies, Figure 4A). The quality ratings were more variable for studies generating optimal sensitivity algorithms (Figure 4B). Studies providing multiple within-study PPVs and sensitivities were used to generate forest plots illustrating the worst-performing diagnostic PPV and sensitivity algorithms, respectively, in Figure 5A,B. These figures illustrate that PPV was low, <80%, in 19 of the 20 worst-performing PPV algorithms, and sensitivity was <80% for nine of the 10 worst-performing sensitivity algorithms. Once again, hand-searched medical records predominated the diagnostic gold standard across this comparison.

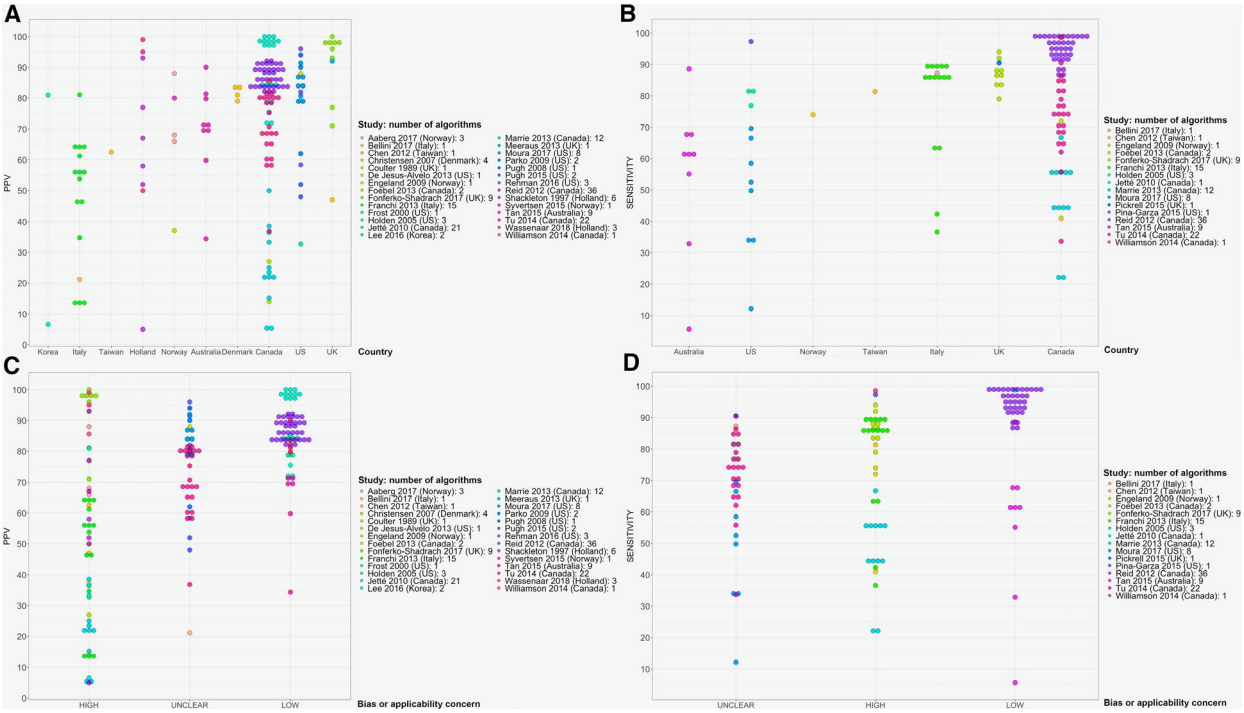


**FIGURE 2** Positive predictive value (PPV, %) and sensitivity (%) dot plots: Dots represent the PPV or sensitivity for each algorithm provided by studies, organized by country (2A = PPV, 2B = sensitivity), and by Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2)[19] score for overall risk of study bias or applicability concern (2C = PPV, 2D = sensitivity)

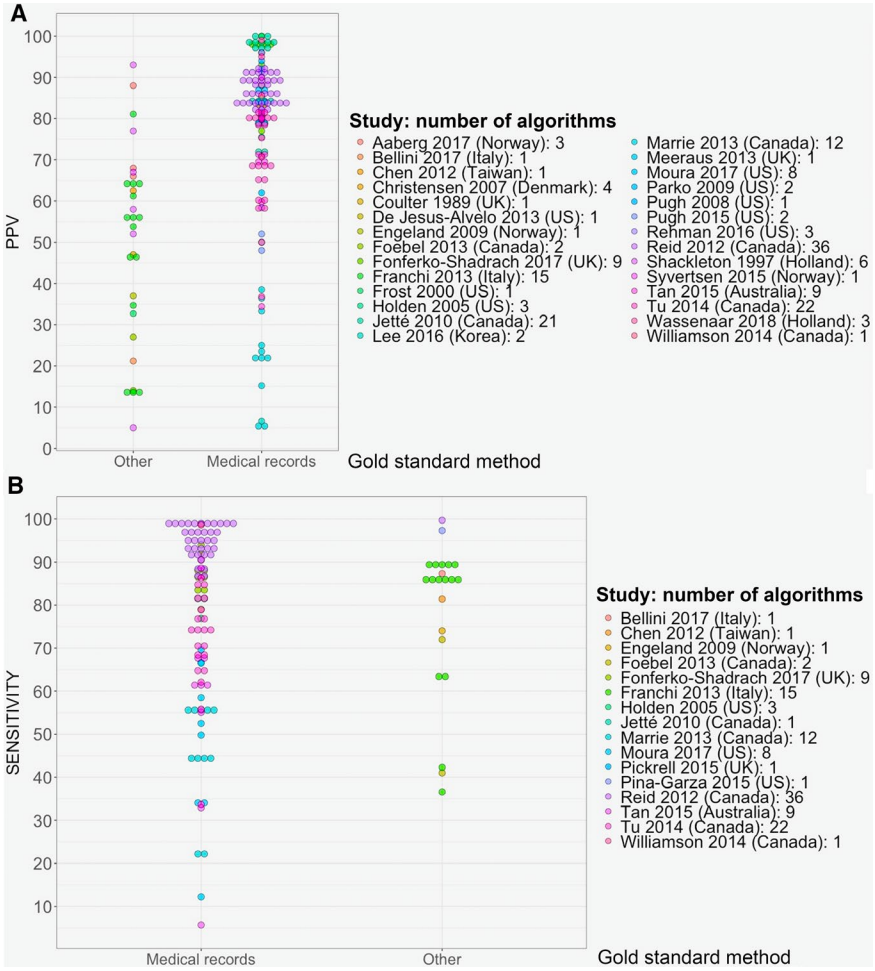


**FIGURE 3** Gold standard method dot plots: Dots represent the PPV (3A, %) or sensitivity (3B, %) for each algorithm, grouped by whether the reference standard used was exclusively medical records or other methods (ie, the general practitioner,[35,36,49] antiepileptic drugs,[31,33,44] medical records and/or parent telephone interview,[27] an epilepsy center patient list,[28] a previous epilepsy prevalence study,[29] another administrative data set,[34] or a specialist epilepsy database)[55]

## 3.4 | PPVs and sensitivities by country

### 3.4.1 | Canada

Positive predictive value ranged from 5.2%–100% and sensitivity ranged from 22.2%–99.7% across six Canadian studies (Appendix S7).[34,38,40,48,52,54] Jetté 2010[38] and Reid 2012[48] had low risks of bias and applicability concerns (Appendix S6B). All but one study (Foebel 2013[34]) used medical records as their diagnostic gold standard, and PPV and sensitivity range remained 5.2%–100% and 22.2%–99.7%, respectively, across these five studies. Jetté 2010[38] estimated an optimal PPV for epilepsy of 100% (95% CI 93.2%–100%) for ICD-10 codes G40 (epilepsy) and G41 (status epilepticus) combined (Figure 4A). The addition of ICD-10 code R56 (convulsions) lowered this PPV to a worst estimate of 71.6% (CI 60.5%–80.6%, Figure 5A). The sensitivity for G40-41 was 98.8% (CI 93.3%–99.8%) and was not tested with R56. The study also tested the PPVs of ICD-9 coding and found that performance between ICD-9 and ICD-10 was similar, and that there was also little difference between results for children and adults. Reid 2012[48] tested algorithms consisting of ICD-9 code 345 (epilepsy) or ICD-10 G40-41 when used as various combinations of physician claims, hospitalizations, or accident and emergency visits over 1-2 years. Spreading these codes over one physician claim, one hospitalization, or one accident and emergency visit over 2 years had the highest sensitivity (99.7%, CI 99.3%–100%, Figure 4B), demonstrating an ability to capture almost all persons with epilepsy in the population, albeit with the trade-off of also giving the worst PPV of 81.6% (CI 79.0%–84.2%, Figure 5A),[56] indicating that many false positives were also captured by this use of multiple sites. By limiting to those algorithms with two physician claims over 1 year, PPV was raised to an optimal of 92.1% (CI 90.1%–94.1%, Figure 4A) at a trade-off of lowering sensitivity to, at worst, 86.2% (CI 83.3%–90.2%, Figure 5B); suggesting that a good proportion of the epilepsy population was still captured. Williamson 2014[54] validated primary care using ICD-9 codes and had high risks of bias or applicability concerns (Appendix S6B). The algorithm was epilepsy ICD-9 345 or convulsions 780.3 (with a requirement for AEDs in "petit mal" 345.2, "grand mal" 345.3, or in 780.3). This generated a very high sensitivity of 98.6% (CI 96.6%–100%, Figure 4B) with a trade-off of PPV 85.6% (CI 80.2%–91.1%, Figure 4A), perhaps suggesting

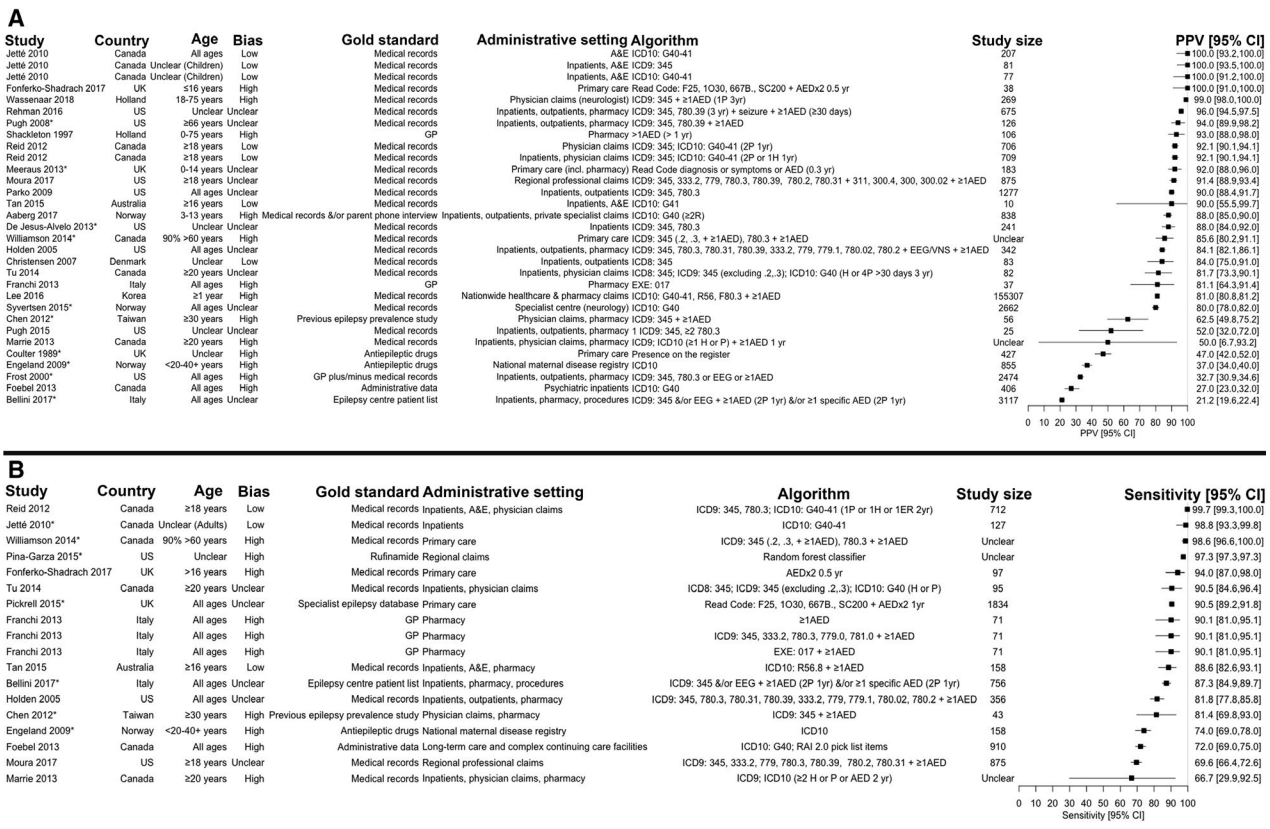


**FIGURE 4** Forest plot showing the optimal diagnostic algorithms (4A = highest positive predictive value [PPV, %], 4B = highest sensitivity [%]) in each included study that estimated PPV or sensitivity, alongside various study characteristics. * = Studies that provided one PPV or sensitivity algorithm only. Bias = Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2)[19] score for overall risk of study bias or applicability concern for that study. A and E: Accident and emergency; AED: Antiepileptic drug; CI: Confidence interval (%); EEG: Electroencephalography; EXE: Exception code; F25, 1O30, 667B., SC200: Epilepsy Read Codes; H: Hospitalization; ICD: International Classification of Diseases; ICD10 F80.3: Acquired aphasia with epilepsy; ICD10 G40: Epilepsy; ICD10 G41; Status epilepticus; ICD8 345: Epilepsy; ICD9 311, 300.4, 300, 300.02: Anxiety and depression codes; ICD9 333.2: Myoclonus; ICD9 345: Epilepsy; ICD9 779: Perinatal illness; ICD9 780.2: Syncope and collapse; ICD9 780.3: Convulsions; ICD9 780.31: Febrile convulsions; ICD9 780.39: convulsions; P: Physician claim; RAI: Resident Assessment Instrument; VNS: Vagus nerve stimulation; Yr: year

that the convulsions code in ICD-9 lowers PPV similarly to the ICD-10 convulsions code R56, as was seen in Jetté 2010.[38] Other studies included the Tu 2014[52] study, which had an unclear risk of bias and tested an algorithm consisting of ICD-8 345 (epilepsy), ICD-9 345, or ICD-10 G40 as hospitalization or physician claims over varying times. The optimal PPV of 81.7% (CI 73.3%–90.1%, Figure 4A) was generated by one hospitalization or four physician claims separated by ≥30 days within a 3-year period and greatly contrasted with the worst PPV of 36.8% (CI 30.6%–42.9%, Figure 5A) generated by only one hospitalization or physician claim. However, this poor PPV was associated with a trade-up in sensitivity to 90.5% (CI 84.6%–96.4%, Figure 4B). The worst sensitivity consisted of hospitalization alone (33.7%, CI 24.2%–43.2%) indicating that hospitalization alone is not a reliable way of capturing all persons with epilepsy. Another study, by Marrie 2013,[40] validated epilepsy codes as a comorbidity of multiple sclerosis and had high risks of bias and low PPV and sensitivity estimates (with wide CIs, Figures 4 and 5, Appendix S7). These results were difficult to interpret in the context of the different algorithms provided. Foebel 2013[34] had high risks of bias and validated ICD-10 G40 or Resident Assessment Instrument (RAI) 2.0 pick list items within psychiatric inpatients, long-term care, and complex continuing care, generating low PPV and sensitivity estimates, which suggests that these environments and/or pick list items brought inaccuracy to estimates (Figures 4 and 5). It is possible that the study's use of an administrative data set as the diagnostic gold standard compounded the poorer accuracy estimates.

### 3.4.2 | United States

Positive predictive value and sensitivity ranged from 32.7%–96.0% and 12.2%–97.3%, respectively, across nine US studies (Appendix S7).[32,36,37,42-47] Risks of bias and applicability concern were unclear in all but two studies,[36,44] where risks were

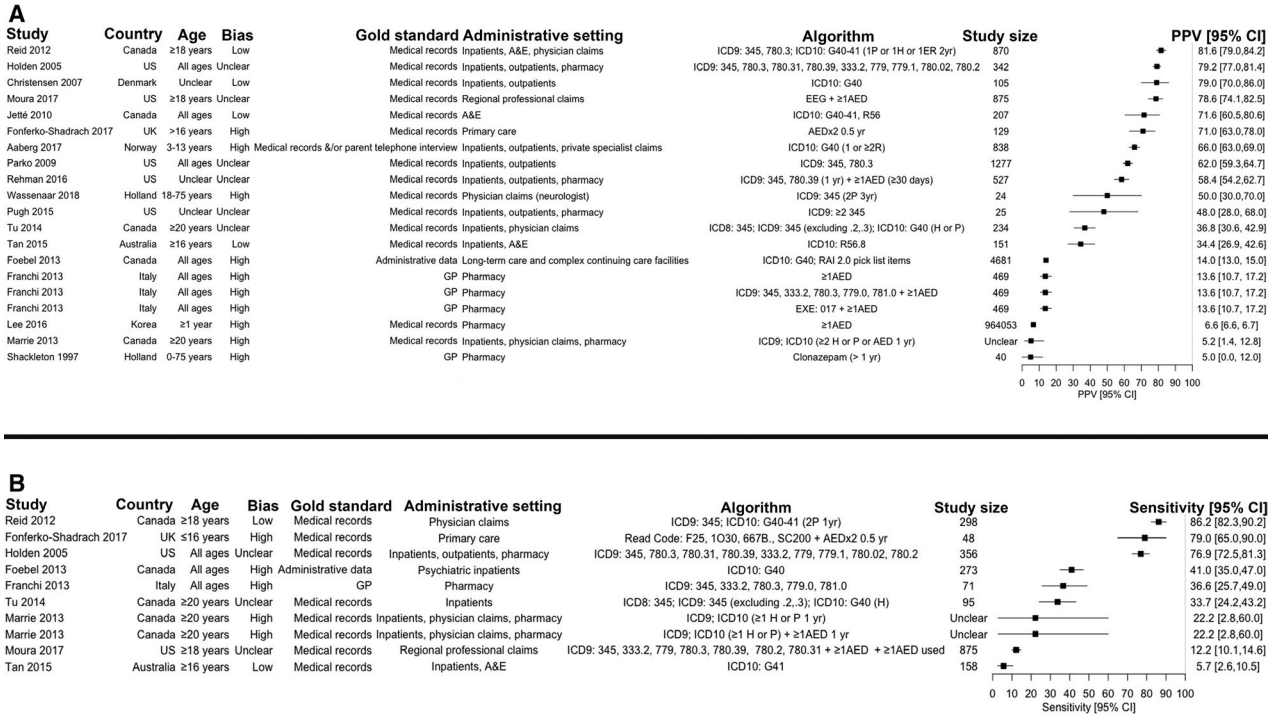


**FIGURE 5** Forest plot showing the worst-performing diagnostic algorithms (5A = lowest positive predictive value [PPV, %], 5B = lowest sensitivity [%]) in each included study that estimated PPV or sensitivity, alongside various study characteristics. Bias = Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2)[19] score for overall risk of study bias or applicability concern for that study. A and E: Accident and emergency; AED: Antiepileptic drug; CI: Confidence interval (%); EEG: Electroencephalography; EXE: Exception code; F25, 1O30, 667B., SC200: Epilepsy Read Codes; H: Hospitalization; ICD: International Classification of Diseases; ICD10 F80.3: Acquired aphasia with epilepsy; ICD10 G40: Epilepsy; ICD10 G41; Status epilepticus; ICD8 345: Epilepsy; ICD9 311, 300.4, 300, 300.02: Anxiety and depression codes; ICD9 333.2: Myoclonus; ICD9 345: Epilepsy; ICD9 779: Perinatal illness; ICD9 780.2: Syncope and collapse; ICD9 780.3: Convulsions; ICD9 780.31: Febrile convulsions; ICD9 780.39: Convulsions; P: Physician claim; RAI: Resident Assessment Instrument; VNS: vagus nerve stimulation; Yr: year

high (Appendix S6B). Six US studies used medical records as their diagnostic gold standard (PPV range 48.0%–96.0%, sensitivity range 12.2%–81.8%).[37,42,43,45-47] The US studies all tested ICD-9 codes, including at least ICD-9 345 (epilepsy) and either 780.3 or 780.39 (seizures). When these diagnostic codes were looked at without a requirement for AEDs, PPV ranged between 48.0% and 90.0% (seven algorithms from five studies),[32,37,42,43,45] with PPV <80% in four of the seven algorithms. When diagnostic codes were looked at with a requirement for AEDs, PPV ranged between 58.4% and 94.0% (nine algorithms from four studies).[37,42,46,47] PPV was <80% in only two of nine algorithms, indicating that an AED requirement increased the likelihood of higher PPV estimates. When diagnostic codes were looked at with a requirement for AEDs and procedures (EEG or VNS), PPV ranged between 32.7% and 86.9% (four algorithms from three studies,[36,37,42] with two algorithm PPVs <80%). The best and worst algorithm combinations from the United States are shown in Figures 4 and 5. In the Parko 2009[43] study, the algorithm ICD-9 345 or 780.3 had a PPV of 90.0% (CI 88.4%–91.7%) when capturing epilepsy or seizures (Figure 4A), and fell to 62.0% (CI 59.3-64.7) when capturing epilepsy. ICD-9 345 without 780.3 was not tested.

### 3.4.3 | United Kingdom

Positive predictive value and sensitivity ranged from 47.0%–100% and 79.0%–94.0%, respectively, across four UK studies (Appendix S7).[14,31,41,55] Two studies (Fonferko-Shadrach 2017[14] and Meeraus 2013[41]) used medical records as their diagnostic gold standard (PPV range 71.0%–100%, sensitivity range 79.0%–94.0%). The Fonferko-Shadrach 2017[14] study had a high risk of bias or applicability concern (Appendix S6B). This study demonstrated that within Read Codes, two prescriptions of the same AED within 6 months provided the optimal sensitivity (94.0%, CI 87.0%–98%, Figure 4B), capturing nearly all persons with epilepsy. However, many false positives were also captured, generating the lowest PPV (71.0%, CI 63.0%–78.0%, Figure 5A). This was in participants >16 years of age and rose to 98% (CI 94%–100%) in participants ≤16 years of age. AEDs are seldom prescribed for indications other than epilepsy in children in the UK.[14] This was the only included study to allow direct comparison of AEDs alone between adults and children. Diagnostic epilepsy Read Codes combined with two AED prescriptions in 6 months generated

the highest PPV (100%, CI 91.0%–100%, Figure 4A), at a trade-off of slightly poorer sensitivity (79.0%, CI 65.0%–90.0%, Figure 5B). PPVs and sensitivities for diagnostic Read Codes alone ranged between 93.0% and 98.0% and 83.0% and 88.0%, respectively. Pickrell 2015[55] had unclear risks of bias (Appendix S6B) and demonstrated a similar sensitivity result (90.5%, CI 89.2%–91.8%) to Fonferko-Shadrach 2017[14] by also using diagnostic epilepsy Read Codes combined with two AED prescriptions, although over 12 months. Pickrell 2015[55] used a specialist epilepsy database as the diagnostic gold standard. Meeraus 2013[41] had unclear risks of bias and demonstrated high PPVs for Read Code epilepsy diagnosis, symptoms, or AED prescription repeated within 4 months (92.0%, CI 88.0%–96.0%). Coulter 1989,[31] the oldest included study, had high risks of bias. The chronic primary care disease register validated was found to have a poor PPV (47.0%, CI 42.0%–52.0%). It is possible that this was compounded by the study's use of AEDs as a diagnostic gold standard.

### 3.4.4 | Norway

Positive predictive value ranged from 37.0%–88.0% across three Norwegian studies (Appendix S7).[27,33,50] Risks of bias were high in all but one study, which had unclear risks (Appendix S6B).[50] This same study (Syvertsen 2015[50]) was the only one to exclusively use medical records as the diagnostic gold standard (PPV 80%). Both Syvertsen 2015[50] and Aaberg 2017[27] validated ICD-10 code G40 alone (two or more registrations in Aaberg 2017)[27] and achieved good optimal PPVs ranging from 80%–88% (Figure 4A). Reducing to one or more registrations lowered PPV to 66.0% (CI 63.0%–69.0%, Figure 5A).[27] Aaberg 2017[27] used medical records and/or parent telephone interview as the diagnostic gold standard. Engeland 2009[33] validated a maternal disease registry to capture epilepsy in pregnant women using ICD-10 (codes were not further specified), identifying globally poor estimates of PPV 37.0% (CI 34.0%–40.0%, Figure 4A) and sensitivity 74.0% (CI 69.0%–78.0%, Figure 4B). It is possible that this was compounded by the study's use of AEDs as the diagnostic gold standard.

### 3.4.5 | Italy

Positive predictive value ranged from 13.6%–81.1% and sensitivity ranged from 36.6%–90.1% across two Italian studies (Appendix S7), one using the GP[35] and the other an epilepsy center patient list[28] as the diagnostic gold standard. Risks of bias were high in Franchi 2013[35] and unclear in Bellini 2017[28] (Appendix S6B). The highest PPV came from the former study's use of lone exemption codes (EXE) for epilepsy 017 (based on ICD-9): 81.1% (CI 64.3%–91.4%, Figure 4A).[35] These are codes that qualify people with epilepsy for free-of-charge healthcare. However, sensitivity was low (42.3%, CI 30.8%–54.5%) indicating that these were not able to capture all persons with epilepsy. The addition of AEDs to exemption codes or to diagnostic ICD-9 codes (or use of AEDs alone) dropped PPV to its worst (13.6%, CI 10.7%–17.2%, Figure 5A). Correspondingly, sensitivity was highest from use of the diagnostic ICD-9 or exemption codes with an additional requirement for AEDs or from AEDs alone (90.1%, CI 81.0%–95.1%, Figure 4B).[35] Bellini 2017[28] used a complex case-ascertainment algorithm arranged as follows: [≥1 ICD-9 345 code] and/or [≥1 EEG recorded in a person prescribed ≥2 AEDs over 12 months] and/or [≥2 prescriptions of one epilepsy-specific AED over 12 months]. This algorithm had a good sensitivity (87.3%, CI 84.9%–89.7%, Figure 4B) at a trade-off of generating a particularly low PPV (21.2%, CI 19.6%–22.4%, Figure 4A), suggesting that despite capturing most epilepsy cases there were many additional false positives. The PPV and sensitivity ranges across nine algorithms that included codes for procedures (EEG) were 21.2%–64.2% and 85.9%–88.7%, respectively.[28,35] PPV and sensitivity ranges across seven algorithms without procedures were 13.6%–81.1% and 36.6%–90.1%, respectively, suggesting that there was a more stable tendency for good sensitivity when procedures were included but PPVs were variable (Appendix S7).

### 3.4.6 | The Netherlands

Positive predictive value ranged from 5.0%–99.0% across two Dutch studies (Appendix S7), one using the GP[49] and the other medical records[53] as the diagnostic gold standard. Risks of bias or applicability concern were high for both studies (Appendix S6B). AED monotherapy or ICD-9 coding without AEDs tended to perform poorly, with PPV range 5.0%–77.0% for the former notwithstanding requirement for continuous prescription over more than 1 year, and PPV 50.0 (CI 30.0%–70.0%) for the latter notwithstanding requirement for two attendances over 3 years. Polytherapy (>1 AED) improved PPV to 93.0% (CI 88.0%–98.0%), and adding in ICD-9 codes to the AED algorithm improved PPV to 95.0%–99.0%.

### 3.4.7 | Other countries

Little additional information was gained from the findings of countries within which only one study was conducted. These studies are summarized in Appendix S8 (Denmark,[30] Australia,[51] South Korea,[39] Taiwan).[29] In the South Korean study (gold standard medical records), one or more AEDs

alone generated a worst PPV of 6.6% (CI 6.6%–6.7%, Figure 5A), reinforcing the potential problems of using AEDs alone.[39]

## 3.5 | Overall NPV and specificity

Less than half of the studies provided data on NPV and specificity: 14 studies, generating 131 algorithms estimating NPV (range 13.2%–100%), with 121 of these algorithms also estimating specificity (range 61.0%–100%).[14,28,29,33-35,37,38,40,42,48,51,52,54] One study provided specificity (95.6%) without NPV.[44] Figures 6 and 7 illustrate the range of these NPV and specificity results, arranged by country (Figure 6A,B), study quality score (Figure 6C,D), and gold standard method (Figure 7A,B). Nearly all of the NPV estimates (93% of 131 algorithms) and specificity estimates (95% of 122 algorithms) were high, >80%, with little visible influence seen from grouping by country or study quality (Figure 6). Grouping by gold standard method suggested that the highest NPV and specificity estimates were found in the studies using methods other than medical records as the gold standard: All 20 algorithms from five studies generated NPVs >97% and all 21 algorithms from six studies generated specificities >94% (Figure 7). Figure 8A,B demonstrates the optimal NPV and specificity algorithms taken from each study and arranged in a forest plot. In each figure, all but one algorithm generated a very high NPV or specificity (>90%). This holds true despite multiple different coding systems and algorithms used, across various population ages and settings, and between different countries. Most studies used medical records as their diagnostic gold standard. The worst-performing NPV and specificity algorithms (Figure 9A,B) demonstrated similar trends, with nearly all achieving stable estimates >80%–90% despite varying study characteristics. NPV and specificity estimates are also included in the Appendix S7 look-up table.

## 3.6 | Meta-analysis of sensitivity and specificity

Only 10 of the 30 included studies provided true positives, false positives, true negatives, and false negatives for the algorithms being tested (91 algorithms, Appendix S9A), allowing limited opportunity for comprehensive meta-analysis. The bivariate diagnostic random-effects meta-analysis revealed an overall sensitivity estimate of 88.6% (CI 85.7%–91.0%), with an overall specificity estimate of 97.9% (CI 98.5%–96.9%) false positive rate 0.021 (CI 0.015–0.031, Appendix S9B). This matches closely with our narrative assessment that in general, sensitivity, and specificity estimates were high across studies. We were unable to proceed to a

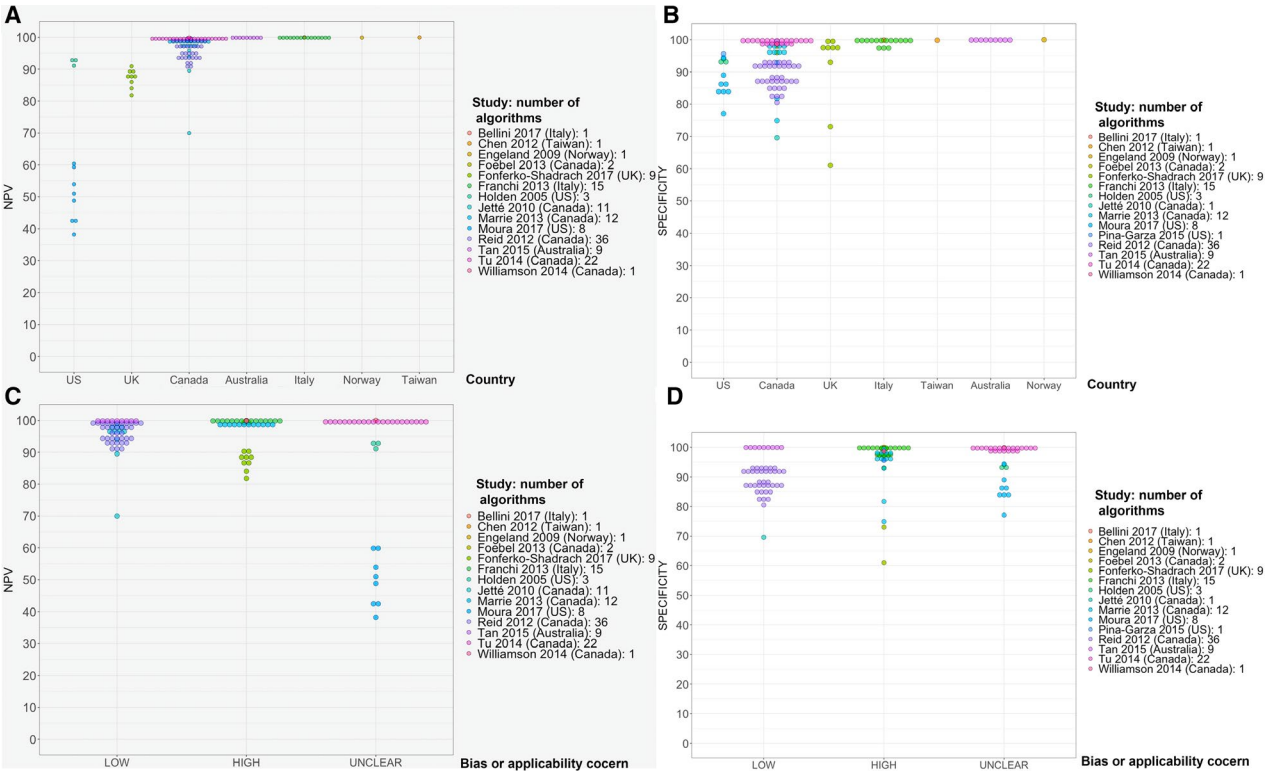


**FIGURE 6** Negative predictive value (NPV, %) and specificity (%) dot plots: Dots represent the NPV or specificity for each algorithm provided by studies, organized by country (6A = NPV, 6B = specificity), and by Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2)[19] score for overall risk of study bias or applicability concern (6C = NPV, 6D = specificity)

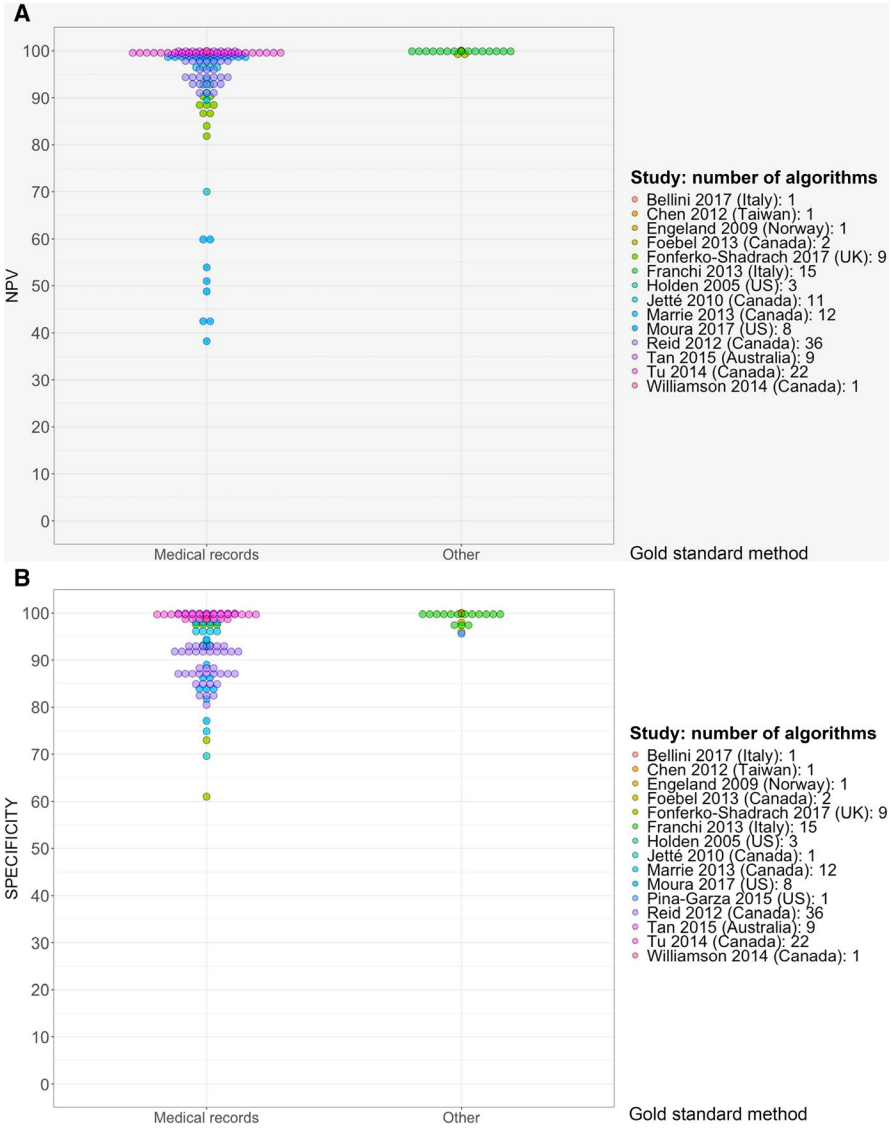


**FIGURE 7** Gold standard method dot plots: Dots represent the NPV (7A, %) or specificity (7B, %) for each algorithm, grouped by whether the reference standard used was exclusively medical records or other methods

valid bivariate meta-regression using study characteristics as covariates because the full spectrum of available study characteristics from the 30 included studies were not represented in a small sample of only 10 studies. Furthermore, such a meta-regression would assume each of the within-study algorithms providing estimates were independent, but they are not.

## 4 | DISCUSSION

### 4.1 | Summary of findings

Overall, this systematic review illustrates that it is reasonable to use administrative healthcare data to identify people with epilepsy in epidemiological research, with studies tending to achieve high estimates (>80%) of PPV, sensitivity, NPV, and specificity. Because the large majority of validation studies in epilepsy have used hand-searched medical records as the diagnostic gold standard, limiting analysis to only such studies makes little difference to the overall results. Little difference is also seen in optimal epilepsy case-identification accuracy between the different ICD coding versions and between the ICD system and others. However, for a particular disease-coding system to perform well in and of itself, the most important elements seem to be algorithm structure and composition. There is a known trade-off relationship between PPV and sensitivity.[56] In situations where researchers wish to prioritize achieving a high PPV, our findings suggest that the algorithm should consist of disease codes (ICD-10 G40-41, ICD-9 345) without symptom codes (ICD-10 R56, ICD-9 780.3, 780.39), and one or more AEDs should be included. Most optimal case-identification algorithms capturing PPVs >80% (and even above 90%) used this basic arrangement, with or without the need to have the disease code registered more than once (Figure 4A). Where a balance needs to be struck between PPV and sensitivity, researchers may wish to choose disease codes alone (ie, without AEDs). This managed

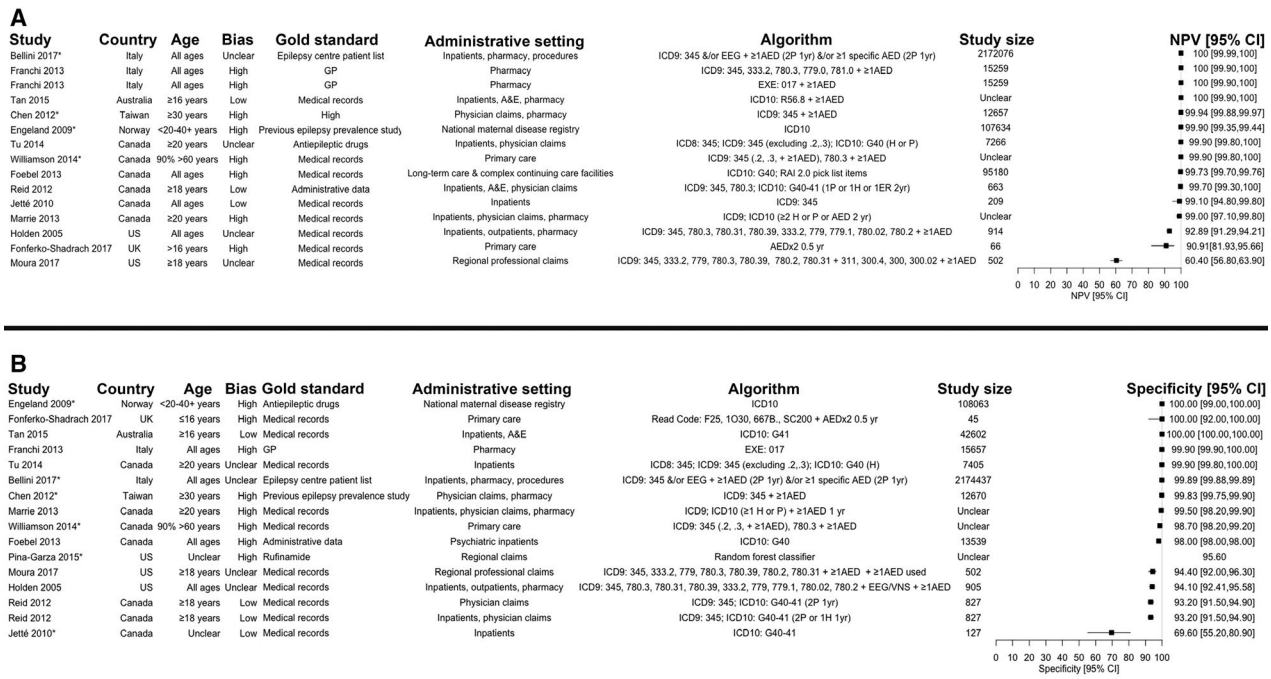


**FIGURE 8** Forest plot showing the optimal diagnostic algorithms (8A = highest negative predictive value [NPV, %], 8B = highest specificity [%]) in each included study that estimated NPV or specificity, alongside various study characteristics. * = Studies that provided one NPV or specificity algorithm only. Bias = Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2)[19] score for overall risk of study bias or applicability concern for that study. A and E: Accident and emergency; AED: Antiepileptic drug; CI: Confidence interval (%); EEG: Electroencephalography; EXE: Exception code; F25, 1O30, 667B., SC200: Epilepsy Read Codes; H: Hospitalization; ICD: International Classification of Diseases; ICD10 F80.3: Acquired aphasia with epilepsy; ICD10 G40: Epilepsy; ICD10 G41; Status epilepticus; ICD8 345: Epilepsy; ICD9 311, 300.4, 300, 300.02: Anxiety and depression codes; ICD9 333.2: Myoclonus; ICD9 345: Epilepsy; ICD9 779: Perinatal illness; ICD9 780.2: Syncope and collapse; ICD9 780.3: Convulsions; ICD9 780.31: Febrile convulsions; ICD9 780.39: Convulsions; P: Physician claim; RAI: Resident Assessment Instrument; VNS: Vagus nerve stimulation; Yr: year

to retain both PPV and sensitivity >80% in several studies (Figure 4A,B). When sensitivity is the priority, our findings suggest that the algorithm should consist of symptom codes with one or more AEDs. We consistently show that an algorithm consisting of one or more AEDs alone, while often able to provide a high sensitivity, is associated with very poor PPVs and should, therefore, not be used. High PPV may be preserved when one or more AEDs alone are used to try and identify children,[14] although this was reported in one study and requires confirmation. For both PPV and sensitivity, there seems to be no added value in adding peripheral epilepsy codes including: ICD-9 333.2 (myoclonus), 779 (convulsions in new-borns), 780 (alteration of consciousness) except 780.3 or 0.39, or ICD-10 F80.3 (acquired aphasia with epilepsy). The algorithms consisting of long combinations of core and peripheral epilepsy codes were no more likely to achieve a higher PPV or sensitivity for epilepsy than those using just core epilepsy disease codes (ICD-10 G40-41, ICD-9 345, Figures 4 and 5, Appendix S7). Although adding in a procedure code (such as EEG) to core epilepsy disease codes optimized sensitivity, it was associated with an unnecessary drop in PPV, and therefore it is unlikely to be helpful. The ICD system is largely used in administrative hospital settings worldwide, and one of its main purposes is to help summarize the incidence and prevalence of diseases on a national or worldwide basis.[57] On the other hand, Read Codes are used primarily to create electronic patient records in the UK primary care system (although they can also be used to summarize disease incidence and prevalence).[57] Taking these differences in function into account, our finding that there was little difference in the diagnostic accuracy of ICD codes and Read Codes for identification of epilepsy cases in administrative healthcare records should be interpreted with caution, particularly as only three studies used Read Codes. Overall, epilepsy coding was particularly good at correctly identifying those without the disease, with estimates of NPV and specificity mostly remaining >90%, despite varying study characteristics.

Our other findings are that no countries appeared to outperform others in terms of overall case-identification accuracy using administrative data. Second, the factors seeming to be most consistently associated with lower optimal PPV or sensitivity estimates globally were modifiable, namely, study design elements that bring about high risk of bias or applicability concern. To lower these risks, future validation studies should aim to clarify/adequately blind the reference standard

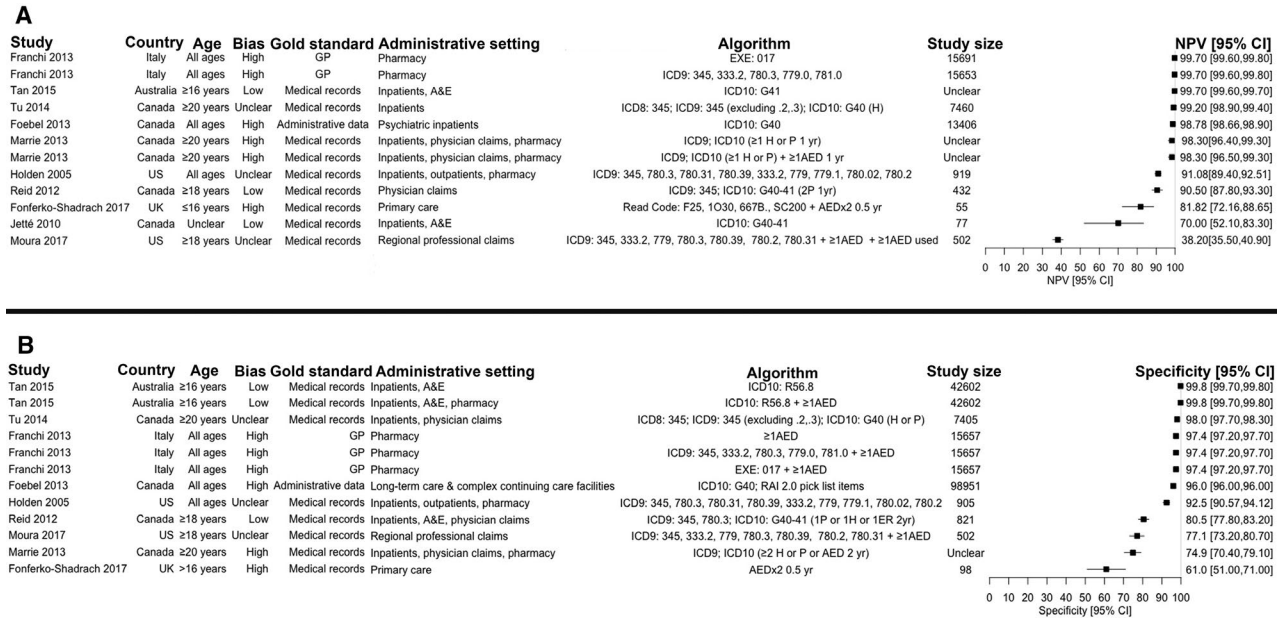


**FIGURE 9** Forest plot showing the worst-performing diagnostic algorithms (9A = lowest negative predictive value [NPV, %], 9B = lowest specificity [%]) in each included study that estimated NPV or specificity, alongside various study characteristics. Bias = Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2)[19] score for overall risk of study bias or applicability concern for that study. A and E: Accident and emergency; AED: Antiepileptic drug; CI: Confidence interval (%); EEG: Electroencephalography; EXE: Exception code; F25, 1O30, 667B., SC200: Epilepsy Read Codes; H: Hospitalization; ICD: International Classification of Diseases; ICD10 F80.3: Acquired aphasia with epilepsy; ICD10 G40: Epilepsy; ICD10 G41; Status epilepticus; ICD8 345: Epilepsy; ICD9 311, 300.4, 300, 300.02: Anxiety and depression codes; ICD9 333.2: Myoclonus; ICD9 345: Epilepsy; ICD9 779: Perinatal illness; ICD9 780.2: Syncope and collapse; ICD9 780.3: Convulsions; ICD9 780.31: Febrile convulsions; ICD9 780.39: Convulsions; P: Physician claim; RAI: Resident Assessment Instrument; VNS: Vagus nerve stimulation; Yr: year

reader, avoid using unvalidated administrative data sets or AEDs as reference standards, administer the same reference standard to all participants, and account for any missing data by performing best- and worst-case sensitivity analyses. A third observation is that administrative settings that specialize in non-neurology and non-general medicine or accident and emergency patients appeared to perform worst in terms of epilepsy PPVs and sensitivities. These were psychiatric inpatients, maternal disease registries, and long-term and complex continuing care facilities. Optimal PPVs were no more than 37% in these settings, and sensitivities were no more than 74% (Figure 4A,B). However, as these studies were also part of the small handful of studies to not use medical records as the diagnostic gold standard (instead using AEDs or other administrative data sets as gold standard), the predominant cause of the lower accuracy estimates becomes less clear and will require further investigation in future studies using gold standard medical records. Finally, the data indicate that administrative setting and age of participants appeared to play little influence on optimal epilepsy case-identification accuracy.

Placing our results in context, the accuracy of administrative healthcare data in identifying epilepsy appears to be similar to that of other neurological diseases with rapidly progressive physical symptoms, such as multiple sclerosis (PPV 86%, sensitivity 84%, NPV 99%, specificity 100%)[58] or motor neuron disease (PPV 55%–92%, sensitivity 75%–93%, NPV/specificity not reported).[10]

## 4.2 | Strengths and limitations

This review expands on findings of the 2012 US/Canada systematic review of this topic[12] by now taking advantage of algorithms tested after that review within the United States or Canada (10 studies),[32,34,40,42,44,45,47,48,52,54] and algorithms tested outside of the United States and Canada (15 studies).[14,27-31,33,35,39,41,49-51,53,55] We also capture 10 more than the single study validating ICD-10,[27,30,33,34,39,40,48,50-52] and we review other non-ICD coding systems for the first time.[14,31,41,49,55] It was concluded previously that a corrective factor would be necessary when relying on ICD-9 to identify epilepsy because the US Parko 2009[43] study had shown a PPV of 90% for a diagnosis of epilepsy or seizures and yet only a PPV of 62% for a diagnosis of epilepsy. However, we now demonstrate several other studies relying on ICD-9 in which no corrective factor was necessary to accurately identify epilepsy (see in Figure 4A, for example, the US study of De Jesus-Alvelo 2013).[32] We also demonstrate this for ICD-10 (eg, Tan 2015[51] in Australia). It was also concluded

previously that multiple epilepsy diagnoses over time are required to identify epilepsy using PPV. Although this is intuitive and was a condition for many of the optimal algorithms with PPVs >80%, our much larger sample of studies demonstrates that it was not a requirement for all (Figure 4A, Appendix S7). This is the first systematic review of epilepsy validation studies to assess included studies for bias and applicability concern. We have been able to demonstrate that studies with a high risk of bias or applicability concern have a tendency toward lower PPV estimates, and we make recommendations on how to reduce these risks. This is also the first systematic review of epilepsy validation studies to incorporate NPV and specificity estimates, demonstrating that these estimates are generally favorable for this condition. Our algorithm look-up table (Appendix S7) is a novel resource that should help future researchers screen epilepsy patients for inclusion into studies recruiting from administrative data sets.

This review demonstrates that there is currently limited scope for detailed meta-analysis of epilepsy validation data, indicating that this an area that will require more investigation in the future as more data become available. The forest plots (Figures 4, 5, 8, 9, Appendix S9) suggest that heterogeneity is low, but this would benefit from further investigation in a meta-regression, which was not possible in the current review because of limitations in the available data. Furthermore, we were unable to assess for publication bias due to the absence of a well-established technique to do so in validation reviews.[10,11]

## 5 | CONCLUSIONS

This review helps researchers in deciding on optimal case-identification methods when using administrative healthcare data to capture persons with epilepsy. It also improves our understanding of the likely accuracy of global estimates for the incidence and prevalence of epilepsy, which have largely been made using administrative data. Our findings are consistent with the hypothesis that administrative epilepsy data accurately identify epilepsy cases, with optimal estimates of PPV, sensitivity, NPV, and specificity >80% in the majority of studies. This is likely to be sufficient for most epidemiological studies, although we show that careful consideration is needed when choosing algorithm structure and code composition for PPV and sensitivity. Although the published algorithms are a useful approximation for accuracy and there are several epilepsy validation studies available around the world, future investigators should still aim to validate putatively selected algorithms within their own data sets first in order to help maintain transparency in the likely accuracy of subsequent study findings.

## ACKNOWLEDGMENTS

This work was charitably supported by Epilepsy Research UK (R44007) and the Juliet Bergqvist Memorial Fund. The funders played no role in the design or conduct of this review. We are also grateful to Kathryn Bush and Tim Wilkinson for support with RStudio, and Sarah Nevitt, Cat Graham, and Jane Hutton for statistical support. We are grateful to Ruth Ann Marrie for providing unpublished study data. We are also grateful to Siddharthan Chandran and Cathie Sudlow for their strategic support.

## CONFLICTS OF INTEREST

None of the authors has any conflict of interest to disclose.

## ETHICAL PUBLICATION STATEMENT

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.

## ORCID

*Gashirai K. Mbizvo* https://orcid.org/0000-0002-9588-2944
*Kyle H. Bennett* https://orcid.org/0000-0002-3922-7056
*Colin R. Simpson* https://orcid.org/0000-0002-5194-8083
*Susan E. Duncan* https://orcid.org/0000-0003-0454-6564
*Richard F.M. Chin* https://orcid.org/0000-0002-7256-3027

## REFERENCES

1. Garratt E, Barnes H, Dibben C. Health administrative data: exploring the potential for academic research. St Andrews, UK: Administrative Data Liaison Service; 2010.
2. Loke YK. Use of databases for clinical research. Arch Dis Child. 2014;99(6):587–9.
3. St Germaine-Smith C, Metcalfe A, Pringsheim T, Roberts JI, Beck CA, Hemmelgarn BR, et al. Recommendations for optimal ICD codes to study neurologic conditions: a systematic review. Neurology. 2012;79(10):1049–55.
4. Abraha I, Montedori A, Eusebi P, Orso M, Grilli P, Cozzolino F, et al. The current state of validation of administrative healthcare databases in Italy: a systematic review. Pharmacoepidemiol Drug Saf. 2012;21:400.
5. Traversa G, Bianchi C, Da Cas R, Abraha I, Menniti-Ippolito F, Venegoni M. Cohort study of hepatotoxicity associated with nimesulide and other non-steroidal anti-inflammatory drugs. BMJ. 2003;327(7405):18–22.
6. Hottes TS, Skowronski DM, Hiebert B, Janjua NZ, Roos LL, Van Caeseele P, et al. Influenza vaccine effectiveness in the elderly based on administrative databases: change in immunization habit as a marker for bias. PLoS ONE. 2011;6(7):e22618.
7. Lalmohamed A, Vestergaard P, Cooper C, de Boer A, Leufkens HGM, van Staa TP, et al. Timing of stroke in patients undergoing total hip replacement and matched controls. Stroke. 2012;43(12):3225–9.

8. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005;58(4):323–37.
9. Krarup LH, Boysen G, Janjua H, Prescott E, Truelsen T. Validity of stroke diagnoses in a national register of patients. Neuroepidemiology. 2007;28(3):150–4.
10. Horrocks S, Wilkinson T, Schnier C, Ly A, Woodfield R, Rannikmäe K, et al. Accuracy of routinely-collected healthcare data for identifying motor neurone disease cases: a systematic review. PLoS ONE. 2017;12(2):e0172639.
11. Wilkinson T, Ly A, Schnier C, Rannikmäe K, Bush K, Brayne C, et al. Identifying dementia cases with routinely collected health data: a systematic review. Alzheimers Dement. 2018;14(8):1038–51.
12. Kee VR, Gilchrist B, Granner MA, Sarrazin NR, Carnahan RM, et al. A systematic review of validated methods for identifying seizures, convulsions, or epilepsy using administrative and claims data. Pharmacoepidemiol Drug Saf. 2012;21:183–93.
13. Benson T. The history of the Read Codes: the inaugural James Read Memorial Lecture 2011. Inform Prim Care. 2011;19(3):173–82.
14. Fonferko-Shadrach B, Lacey AS, White CP, Powell HWR, Sawhney IMS, Lyons RA, et al. Validating epilepsy diagnoses in routinely collected data. Seizure. 2017;52:195–8.
15. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol. 2007;15(7):10.
16. Mbizvo GK, Bennett K, Simpson CR, Duncan SE, Chin RFM. Accuracy and utility of using administrative healthcare databases to identify people with epilepsy: a protocol for a systematic review and meta-analysis. BMJ Open. 2018;8(6):e020824.
17. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ. 2009;21(339):b2535.
18. World Health Organization. Classification of Diseases (ICD). 2016. Available from www.who.int/classifications/icd/en/
19. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–36.
20. Shea BJ, Bouter LM, Peterson J, Boers M, Andersson N, Ortiz Z, et al. External validation of a measurement tool to assess systematic reviews (AMSTAR). PLoS ONE. 2007;2(12):e1350.
21. Shea BJ, Hamel C, Wells GA, Bouter LM, Kristjansson E, Grimshaw J, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. J Clin Epidemiol. 2009;62(10):1013–20.
22. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005;37(5):360–3.
23. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.
24. Deeks JJ, Bossuyt PM, Gatsonis C. Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.02010. [updated 2019 May 19]. Available from https://methods.cochrane.org/sdt/handbook-dta-reviews
25. Reitsma JB, Glas AS, Rutjes AW, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005;58(10):982–90.
26. Doebler P, Holling H. Meta-Analysis of Diagnostic Accuracy with mada 2012. [Accessed 2019 Jan 6]. Available from https://cran.r-project.org/web/packages/mada/vignettes/mada.pdf
27. Aaberg KM, Gunnes N, Bakken IJ, Lund Søraas C, Berntsen A, Magnus P, et al. Incidence and prevalence of childhood epilepsy: a nationwide cohort study. Pediatrics. 2017;139:5.
28. Bellini I, Policardo L, Zaccara G, Palumbo P, Rosati E, Torre E, et al. Identification of prevalent patients with epilepsy using administrative data: the Tuscany experience. Neurol Sci. 2017;38(4):571–7.
29. Chen CC, Chen LS, Yen MF, Chen H-H, Liou H-H. Geographic variation in the age- and gender-specific prevalence and incidence of epilepsy: analysis of Taiwanese National Health Insurance-based data. Epilepsia. 2012;53(2):283–90.
30. Christensen J, Vestergaard M, Olsen J, Sidenius P. Validation of epilepsy diagnoses in the Danish National Hospital Register. Epilepsy Res. 2007;75(2–3):162–70.
31. Coulter A, Brown S, Daniels A. Computer held chronic disease registers in general practice: a validation study. J Epidemiol Community Health. 1989;43(1):25–8.
32. De Jesus-Alvelo I, Labovitz D. How reliable are the ICD9-CM billing codes in the administrative data to estimate the risk of seizures and epilepsy after stroke? Neurology. 2013;12:80.
33. Engeland A, Bjorge T, Daltveit AK, Vollset SE, Furu K. Validation of disease registration in pregnant women in the Medical Birth Registry of Norway Acta Obstet Gynecol Scand. 2009;88(10):1083–9.
34. Foebel AD, Hirdes JP, Heckman GA, Kergoat M-J, Patten S, Marrie RA. Diagnostic data for neurological conditions in interRAI assessments in home care, nursing home and mental health care settings: a validity study. BMC Health Serv Res. 2013;1(13):457.
35. Franchi C, Giussani G, Messina P, Montesano M, Romi S, Nobili A. Validation of healthcare administrative data for the diagnosis of epilepsy. J Epidemiol Community Health. 2013;67(12):1019–24.
36. Frost FJ, Hurley JS, Petersen HV, Gunter MJ, Gause D. A comparison of two methods for estimating the health care costs of epilepsy. Epilepsia. 2000;41(8):1020–6.
37. Holden EW, Grossman E, Nguyen HT, Gunter MJ, Grebosky B, Worley AV, et al. Developing a computer algorithm to identify epilepsy cases in managed care organizations. Dis Manag. 2005;8(1):1–14.
38. Jette N, Reid AY, Quan H, Hill MD, Wiebe S. How accurate is ICD coding for epilepsy? Epilepsia. 2010;51(1):62–9.
39. Lee SY, Chung SE, Kim DW, Eun SH, Kang HChu, Cho YW, et al. Estimating the prevalence of treated epilepsy using administrative health data and its validity: ESSENCE study. J Clin Neurol. 2016;12(4):434–40.
40. Marrie RA, Yu BN, Leung S, Elliott L, Caetano P, Warren S, et al. The utility of administrative data for surveillance of comorbidity in multiple sclerosis: a validation study. Neuroepidemiology. 2013;40(2):85–92.
41. Meeraus WH, Petersen I, Chin RF, Knott F, Gilbert R. Childhood epilepsy recorded in primary care in the UK. Arch Dis Child. 2013;98(3):195–202.
42. Moura LM, Price M, Cole AJ, Hoch DB, Hsu J. Accuracy of claims-based algorithms for epilepsy research: revealing the unseen performance of claims-based studies. Epilepsia. 2017;58(4):683–91.
43. Parko K, Thurman DJ. Prevalence of epilepsy and seizures in the Navajo Nation 1998–2002. Epilepsia. 2009;50(10):2180–5.
44. Pina-Garza JE, Vekeman F, Cheng W, Tuttle E, Giguère-Duval P, Oganisian A, et al. Development of a claims-based classifier to identify lennox-gastaut syndrome. Neurology Conference: 67th American Academy of Neurology Annual Meeting, AAN. 2015;84 (Suppl 14). Conference Abstract.

45. Pugh MJ, Parko K. Research using archival health care data: let the buyer beware. Epilepsia. 2015;56(2):321–2.
46. Pugh MJ, Van Cott AC, Cramer JA, Knoefel JE, Amuan ME, Tabares J, et al. Trends in antiepileptic drug prescribing for older patients with new-onset epilepsy: 2000-2004. Neurology. 2008;70(Issue 22, Part 2):2171–8.
47. Rehman R, Everhart A, Frontera AT, Kelly PR, Lopez M, Riley D, et al. Implementation of an established algorithm and modifications for the identification of epilepsy patients in the veterans health administration. Epilepsy Res. 2016;127:284–90.
48. Reid AY, St Germaine-Smith C, Liu M, Sadiq S, Quan H, Wiebe S, et al. Development and validation of a case definition for epilepsy for use with administrative health data. Epilepsy Res. 2012;102(3):173–9.
49. Shackleton DP, Westendorp RG, Kasteleijn-Nolst Trenite DG, de Boer A, Herings RMC. Dispensing epilepsy medication: a method of determining the frequency of symptomatic individuals with seizures. J Clin Epidemiol. 1997;50(9):1061–8.
50. Syvertsen M, Nakken KO, Edland A, Hansen G, Hellum MK, Koht J. Prevalence and etiology of epilepsy in a Norwegian county-a population based study. Epilepsia. 2015;56(5):699–706.
51. Tan M, Wilson I, Braganza V, Ignatiadis S, Boston R, Sundararajan V, et al. Development and validation of an epidemiologic case definition of epilepsy for use with routinely collected Australian health data. Epilepsy Behav. 2015;51:65–72.
52. Tu K, Wang M, Jaakkimainen RL, Butt D, Ivers NM, Young J, et al. Assessing the validity of using administrative data to identify patients with epilepsy. Epilepsia. 2014;55(2):335–43.
53. Wassenaar M, Carpay JA, Sander JW, Thijs RD. Validity of health insurance data to identify people with epilepsy. Epilepsy Res. 2018;139:102–6.
54. Williamson T, Green ME, Birtwhistle R, Khan S, Garies S, Wong ST, et al. Validating the 8 CPCSSN case definitions for chronic disease surveillance in a primary care database of electronic health records. Ann Fam Med. 2014; 12(4):367–72.
55. Pickrell WO, Lacey AS, Bodger OG, Demmler JC, Thomas RH, Lyons RA, et al. Epilepsy and deprivation, a data linkage study. Epilepsia. 2015;56(4):585–91.
56. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. J Clin Epidemiol. 2012;65(3):343–9.e2.
57. Watson N. Using clinical coding systems to best effect in electronic records. Guidelines in practice [Internet]. 2001. [Accessed 2020 Jan 9]. Available from https://www.guidelinesinpractice.co.uk/using-clinical-coding-systems-to-best-effect-in-electronic-records/305085.article
58. Widdifield J, Ivers NM, Young J, Green D, Jaakkimainen L, Butt DA, et al. Development and validation of an administrative data algorithm to estimate the disease burden and epidemiology of multiple sclerosis in Ontario. Canada Mult Scler. 2015;21(8):1045–54.
59. Aaberg KM, Bakken IJ, Lossius MI, Lund Soraas C, Haberg SE, Stoltenberg C, et al. Comorbidity and childhood epilepsy: a nationwide registry study. Pediatrics. 2016;138:3.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Mbizvo GK, Bennett K, Schnier C, Simpson CR, Duncan SE, Chin RFM. The accuracy of using administrative healthcare data to identify epilepsy cases: A systematic review of validation studies. *Epilepsia*. 2020;00:1–17. https://doi.org/10.1111/epi.16547