Validating the accuracy of administrative healthcare data identifying epilepsy in deceased adults

Title: Validating the accuracy of administrative healthcare data identifying epilepsy in deceased adults: a Scottish data linkage study

Authors: Gashirai K Mbizvo[1], Christian Schnier[2], Colin R Simpson[2,3], Susan E Duncan[1,4], Richard FM Chin[1, 5]

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

[2] The University of Edinburgh, Usher Institute, Edinburgh, UK

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

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

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

Corresponding Author

Dr Gashirai K Mbizvo, University of Edinburgh Centre for Clinical Brain Sciences, Muir Maxwell Epilepsy Centre, Child Life and Health, 20 Sylvan Place, EH9 1UW, Edinburgh, UK

Tel: +44 (0) 131 536 0801, Fax: N/A

Email: gashirai.mbizvo@ed.ac.uk

Other Author Email Contact

Christian Schnier – christian.schnier@ed.ac.uk

Colin Simpson – c.simpson@ed.ac.uk

Susan Duncan – susanxduncan@gmail.com

Richard Chin – rchin@exseed.ed.ac.uk

Abstract

Background: We investigate the case-ascertainment accuracy for potentially active epilepsy of four administrative healthcare datasets used to identify deceased adults in Scotland.

Methods: In this diagnostic accuracy study, unique patient identifiers were used to link administrative healthcare data for adults (aged 16 years and over) who died in Scotland between 01/01/09–01/01/16. Cases were ascertained from linking mortality records, hospital admissions, antiepileptic drug (AED) prescriptions, and primary care attendances. We assessed ICD-10 codes G40 (epilepsy), G41 (status epilepticus), and R56.8 (seizures) listed as causes of death and as hospital admission reasons, various AEDs, and F25 primary care epilepsy Read codes. These epilepsy indicators were searched through 01/01/09–01/01/16, suggesting active epilepsy during a maximal period of seven years before death. They were compared to epilepsy diagnoses made from medical records reviewed by a senior epileptologist, with a second senior epileptologist independently reviewing the medical records in a 10% sample to check for specialist interrater agreement in epilepsy diagnoses. We validated how accurately epilepsy was identified by each dataset alone and when combined, calculating positive predictive value (PPV) and sensitivity (with 95% confidence intervals (CIs)).

Results: 159,032 deceased potential epilepsy cases were captured across the four datasets. Medical records reviewed in a random sample of 936 confirmed that epilepsy was present in 614 and absent in 322. Specialist interrater diagnostic agreement was substantial (100 medical records reviewed in duplicate, kappa = 0.72, CI 0.58–0.86). G40–41 cause of death codes had a PPV of 86% (CI 84–89%) and sensitivity of 73% (CI 69–76%). Adding R56.8 lowered PPV to 69% (CI 65–72%) and raised sensitivity to 87% (CI 84–90%). The optimal algorithm combining two datasets consisted of F25 Read codes paired with AEDs (PPV 86% (CI 80–92%), sensitivity 93% (CI 88–97%)). Also effective was pairing G40–41 and/or R56.8 cause of death codes with AEDs (PPV 91% (CI 89–94%), sensitivity 81% (CI 77–84%)). Whilst algorithms combining three datasets raised PPV to as high as 93–95%, the associated sensitivities were low (71% at most).

Conclusions: Routinely-collected Scottish data can accurately identify epilepsy in deceased adults. It may be necessary to combine the diagnostic coding used with AEDs to ensure optimal case-ascertainment. The results help inform the design of future Scottish epilepsy mortality studies recruiting from administrative data sources.

Keywords: seizures; diagnostic accuracy study; routine data; cause of death, ICD-10; mortality

1. Introduction

Epilepsy is common and contributes 0.7% of the global burden of disease according to various administrative healthcare data sources (Murray et al. 2012; Thijs et al. 2019). Although administrative data are widely available resources for epilepsy research (Garratt E et al. 2010), their accuracy in correctly identifying epilepsy cases requires validation because the potential for epilepsy misdiagnosis remains large (Garratt E et al. 2010 ; ILAE 1993; Kee et al. 2012; Mbizvo et al. 2020). This is because such data were originally collected for routine non-scientific purposes, such as to help organise health insurance claims (England MJ et al. 2013).

People with epilepsy (PWE) are at increased risk of premature death (Escalaya et al. 2015; Levira et al. 2016; Nevalainen et al. 2014; Shackleton et al. 2002; Thurman et al. 2017; Watila et al. 2018). In a recent systematic review of mortality in PWE, we identified a paucity of studies assessing the validity of using administrative data to identify epilepsy cases (Mbizvo et al. 2019a; Mbizvo et al. 2020). Most studies relied on identifying cases using unvalidated epilepsy coding within hospital admissions data or International Classification of Disease (ICD) codes derived from causes of death listed within death certificates (Mbizvo et al. 2019a; Mbizvo et al. 2020; Pickrell and Kerr 2020). The presence of an epilepsy code within the listed causes of death of a patient does not necessarily mean they had epilepsy. Various potential sources of error remain possible, such as an incorrect diagnosis of epilepsy being listed by the death certifying doctor in a patient who did not actually have epilepsy (Death Certifcation Review Service 2016), or human error occurring during the subsequent coding process (O'Malley et al. 2005). For example, the G40–41 ICD-10 codes for epilepsy and status epilepticus are often incorrectly coded as G43.1 (migraine) or G45.x (transient cerebral ischemia) (Jette et al. 2010; Tan et al. 2015). Furthermore, a deceased PWE might not be identified in an epilepsy mortality study if death certificates are the only source for case-ascertainment. This is because the diagnosis of epilepsy would be excluded from their death certificate if epilepsy was not felt to have contributed to causing death (Chief Medical Officer 2018; National Records of Scotland 2020).

In an ongoing, retrospective, population-based, Scottish Epilepsy Deaths Study (SEDS) (Mbizvo et al. 2019b), we identified people who died who may have had active epilepsy before death by searching four linked administrative healthcare databases for indicators of epilepsy occurring within a maximum of seven years before death occurred, suggesting that epilepsy was still active (Fisher et al. 2014). In the current paper, we perform a diagnostic accuracy study to investigate how accurately the coding strategies we used identified epilepsy within these four databases. The outcomes of interest were positive predictive value (PPV) and sensitivity, the most commonly used outcomes in studies validating the accuracy of administrative data in epilepsy (Kee et al. 2012; Mbizvo et al. 2020).

2. Methods

2.1 Environment

Healthcare to the five million population of Scotland is delivered by a single state-provided National Health Services (NHS), administered by 14 regional Health Boards (Pavis and Morris 2015). To access NHS care, patients first register with a general practitioner (GP), who provides them with a Community Health Index (CHI) number. This is a unique patient identifier used whenever they subsequently access any NHS service. As 96.5–99.9% of the Scottish population has a CHI number (Pavis and Morris 2015), it can be used to create nationally linked research datasets derived from almost the entire Scottish population (Katikireddi et al. 2018).

2.2 Study design

We undertook a diagnostic accuracy study to assess how accurately administrative data identified epilepsy cases in adults who had died between 01/01/09–01/01/16 and had been identified by various diagnostic coding and/or antiepileptic drug (AED) strategies used within the following administrative healthcare databases: A) National Records of Scotland (NRS) death records (Information Services Division 2016c); B) Scottish Morbidity Record 01 (SMR01) hospital admissions (Information Services Division 2016b); C) Prescribing Information System (PIS) Scottish prescribing data (Information Services Division 2016a); and GP primary care data. Each database was searched for cases retrospectively between 01/01/09–01/01/16. We used hand-searched medical records as the diagnostic gold standard against which to compare the administrative data for accuracy (Mbizvo et al. 2020). Given the expectation of a large number of deceased participants to be captured by the four datasets, we aimed to review a randomly generated convenience sample focusing on eight of 14 Scottish NHS Health Boards (Ayrshire and Arran, Fife, Forth Valley, Grampian, Glasgow, Lanarkshire, Lothian, and Tayside) (Bossuyt et al. 2015). This study was designed to conform to the Standards for Reporting of Diagnostic Accuracy Studies (STARD 2015) (Bossuyt et al. 2015).

2.3 Study population

Eligible participants were adults (aged ≥16 years) who had died between 01/01/09–01/01/16, were captured within the administrative databases, and had medical records available for review. The focus was on adults as mortality in children with epilepsy has been studied in detail elsewhere (RCPCH 2013) and was therefore not the focus of our epilepsy mortality study (SEDS) – which was the data substrate for the current diagnostic accuracy study.

2.4 Linkage of datasets

The Scottish Information Services Division (ISD) is the legislated producer of NHS healthcare statistics to inform quality assessment and Government resource allocation (Public Health Scotland 2020). Therefore, it is mandatory for NHS care providers to submit their administrative data to ISD. ISD then collate and process these data to produce several key national datasets (Information Services Division 2016b, 2016c, 2016a). These national datasets are also routinely extracted and made available to approved researchers in a secure analytical environment (Safe Haven) as linked and deidentified research datasets (Pavis and Morris 2015). This process is managed by the electronic Data Research and Innovation Service (eDRIS) of ISD, based at Farr Institute Scotland (Pavis and Morris 2015). Primary care data from GPs can also be extracted for research purposes by an approved third party (Albasoft) (Albasoft 2020) and transferred to eDRIS (Douglas et al. 2015). The availability of CHI numbers as unique patient identifiers makes it possible to link primary care data to NRS, SMR01, and PIS as these datasets also hold the CHI variable (Pavis and Morris 2015).

We utilised these existing linkage infrastructures to establish an epilepsy mortality research dataset containing linked national data from ISD and regional primary care data (Figure 1) (eDRIS Team 2020). To facilitate this, eDRIS created a linked dataset in our Safe Haven containing deceased participants recruited according to the coding and/or AED prescription criteria listed in A–D below (see *Datasets*) having occurred between 01/01/09–01/01/16. The codes chosen were based on previous epilepsy validation literature (Kee et al. 2012; Mbizvo et al. 2020). We validated accuracy of the coding and/or prescription conditions listed in A–D both alone and in multiple algorithm combinations (61 in total, see tables 1–3).

2.5 Datasets (each searched through 01/01/09–01/01/16)

a) **NRS death records** (Information Services Division 2016c)**:** those who had died with ≥1 G40–41 (epilepsy and status epilepticus) and/or R56.8 (seizure) code listed within their causes of death in the NRS death records. All Scottish deaths are registered at NRS within eight days of occurrence using death certification (Chief Medical Officer 2018). The causes of death are listed in Parts 1 and 2 of a death certificate as free-text. An automated method of coding, based on a complex set of modification and selection rules, is then used alongside experienced coding staff at NRS to convert the free-text causes of death in Parts 1–2 into their corresponding ICD-10 codes (National Records Scoltand 2016). NRS arrange these ICD-10 codes into a single primary cause of death field and up to nine secondary cause of death fields. Unique to Scotland, the causes of death are reviewed for accuracy by the NRS staff and, where necessary, amended using information from various sources (see appendix A for further details) (Death Certifcation Review Service 2016; Fernie 2019; National Records Scoltand 2016). We

extracted all of the ICD-10-coded causes of death and the corresponding Part 1–2 death certificate free-texts held by NRS.

AND/OR

b) **SMR01 hospital admissions** (Information Services Division 2016b)**:** those who had experienced ≥1 hospital admission diagnosis coded as G40–41 and/or R56.8, regardless of their cause of death. SMR01 is a mandatory national dataset capturing all of the inpatient hospital admissions in Scotland. 1.4 million SMR01 records are generated yearly (Information Services Division 2016b). Data collected include a mandatory diagnostic field containing up to six ICD-10-coded hospital admission diagnoses (Information Services Division 2016b).

AND/OR

c) **AEDs on PIS** (Information Services Division 2016a)**:** those who were prescribed ≥1 AEDs (regardless of their cause of death). PIS is a mandatory national prescribing dataset used to track the community drug history of the entire Scottish population in order to facilitate public drug reimbursement (Information Services Division 2016a). Approximately 100 million prescription data items are uploaded to PIS yearly (Information Services Division 2016a). We used PIS to screen for whether or not any deceased adults in Scotland were prescribed one or more AEDs from a broad list of 36 AEDs provided in appendix B1 during the study period, recruiting any who were. Such a broad list was aimed at increasing sensitivity, creating a file with the potential to have captured nearly all PWE in the country who died (assuming that most were prescribed an AED) (Hamer et al. 2012). However, as this broad list also included several AEDs often prescribed for indications other than epilepsy (e.g. gabapentin and diazepam), we created an additional narrow AED filter in which such agents were excluded (leaving 21 AEDs, see appendix B2) to try and explore whether this would recruit a sample with fewer false positives.

AND/OR

d) **Primary care:** those who had ≥1 epilepsy-related (Read code *F25*) primary care attendance (regardless of cause of death). Read codes are a coded thesaurus of clinical terms used in primary care to create electronic health records (Fonferko-Shadrach et al. 2017). Currently, there is no routine primary care epilepsy dataset held at ISD (Information Services Division 2020). Therefore, we established one consisting of 100 GP practices (a convenience sample of 10% of 900 Scottish GP practices) using an extension of methods detailed elsewhere (see appendix C for Read codes used) (Douglas et al. 2015; Fonferko-Shadrach et al. 2017; Pickrell et al. 2015). To our knowledge, this is a larger sample of Scottish GP practices providing epilepsy data than other established

primary care datasets, including The Health Improvement Network and the Clinical Practice Research Datalink (Herrett et al. 2015; Meeraus et al. 2013).

2.6 Establishing gold standard epilepsy diagnosis from medical records

Gold standard epilepsy diagnosis was established from reviewing all available medical records. Access to medical records was made possible through the support of a national network of consultant neurology colleagues (one in each of the 14 Scottish Health Boards), who were provided with CHI numbers of the validation sample by eDRIS. The medical records were then reviewed by an experienced consultant epileptologist (S.E.D.), who was blinded to the participants' administrative data codes and used the medical records to confirm the presence or absence of epilepsy (Kee et al. 2012; Mbizvo et al. 2020). This was based on corroborative evidence such as the presence of two or more unprovoked seizures or clear documentation of a diagnosis of epilepsy from a neurologist (Fisher et al. 2005; Fisher et al. 2014). The medical records were extensive, and included electronic and paper records from general and/or specialist inpatients, emergency care, outpatients (including neurology clinics), hospital discharge summaries, referral letters from GPs, radiology scans, neurophysiology reports, and medication lists. This meant that although our network of neurology colleagues could provide access to medical records, it was not necessary for the patients themselves to have been followed by a neurologist or under specialist care (i.e. they could have had only non-specialist general inpatient/emergency care attendances in their medical records). A second consultant epileptologist (R.F.M.C), also blinded to the participants' administrative data codes, independently reviewed a random sample of 10% of the validated medical records in NHS Lothian to confirm the presence or absence of epilepsy using the same diagnostic inclusion methods (Fisher et al. 2005; Fisher et al. 2014). This was done to assess for interrater agreement (Landis and Koch 1977; Viera and Garrett 2005). Participants were sampled using electronically-generated random numbers.

2.7 Statistical analysis

The primary focus was to validate cause of death ICD-10 coding, secondarily supported by validating epilepsy hospital admissions coding in SMR01, AEDs prescribed in PIS, and primary care epilepsy coding. Therefore, we ensured randomisation was weighted toward oversampling from the NRS dataset such that 80% of the validation sample were captured by at least one G40–41 and/or R56.8 code within their listed causes of death, using electronically generated random numbers to sample participants. Cohen's kappa statistic was used estimate interrater agreement (above chance) between S.E.D and R.F.M.C for medical record epilepsy diagnoses (as yes, no or unclear due to missing or incomplete medical records). The following cut-offs were used: kappa < 0: no agreement; 0.00–

0.20: slight agreement; 0.21–0.40: fair agreement; 0.41–0.60: moderate agreement; 0.61–0.80: substantial agreement; 0.81–1.00: almost perfect agreement (Landis and Koch 1977; Viera and Garrett 2005).

We validated each of the 61 study algorithms (tables 1–3), reporting a PPV and sensitivity for each (with 95% confidence intervals (CIs) generated using an exact method) (Du Z and Hao y 2019). For clarity, we grouped results into Levels 1, 2 and 3 – presenting the results of a single, two combined, and three combined database coding or AED strategies, respectively. We arranged results in order of highest to lowest PPV. We identified the optimal case-ascertainment algorithm(s) as those with the highest PPV and sensitivity in combination. Categories with five or less events/participants were combined to minimise risk of patient identification (eDRIS Team 2018). Data were analysed using the "reportROC" and "cutpointr" packages in RStudio Version 1.2.1335 (Du Z and Hao y 2019; RStudio Team 2015 ; Thiele C 2020).

2.8 Approvals

The study was approved by South East Scotland Research Ethics Committee 2 (IRAS 181131, 15/SS/0165), and Scottish Public Benefit and Privacy Panel for Health and Social Care.

3. Results

3.1 Population capture and validation sample characteristics

eDRIS removed 11 people (<1%) and 420 people (5%) from the NRS and SMR01 data files, respectively, as they had no CHI number; precluding linkage. Linkage was otherwise complete and a large dataset containing information from 159,032 deceased adults across Scotland was created for this study (see figure 1). From these, a sample of 1,002 participants had their medical records reviewed, confirming the presence of epilepsy in 614, and absence in and 322. The remaining 66 participants were excluded from the validation analysis due to missing or incomplete medical records. Specialist interrater diagnostic agreement was classed as substantial (100 medical records reviewed in duplicate, 87 observed agreements, 54 agreements expected by chance, kappa = 0.72, 95% CI 0.58–0.86) (Landis and Koch 1977). The 936 validated participants had a mean age of 62 years (±19 years SD). 491 were male (52%). Participants were mainly from Lothian (n = 264), Glasgow (n = 215), Tayside (n = 145), and Grampian (n = 142), with the remaining 170 spread between Ayrshire and Arran, Fife, Forth Valley, and Lanarkshire. The denominator for validating the primary care dataset was 155 persons for whom primary care data were available.

3.2 Level 1 – validation of a single database coding or AED strategy

Table 1 summarises Level 1 validation results. Within causes of death in the NRS death records, the optimal coding strategy in terms of PPV was using either G40 (epilepsy) alone or G40–41 (epilepsy and/or status epilepticus): PPV

range 86–90%, with overlapping CIs. This indicates that when present, these causes of death normally reflected a correct diagnosis of epilepsy. However, the associated sensitivities were relatively low (ranging 69–73%, with overlapping CIs), indicating that many positive cases were also missed by each of these coding strategies (high burden of 168–192 false negatives). This is intuitive, as epilepsy and/or status epilepticus might not have appeared on the death certificate of a PWE if they did not die of these (National Records of Scotland 2020), making them appear falsely negative within NRS cause of death coding. The addition of seizures (R56.8) to the G40–41 epilepsy and/or status epilepticus cause of death coding strategy pushed sensitivity up to a much higher 87% (CI 84–90%), lowering the burden of false negatives to 81. However, the PPV for this coding strategy was poor (69%), owing to many false positives (243) also being associated with the addition of seizures (R56.8). Within causes of death in the NRS death records, the poorest PPV and sensitivity results were from lone R56.8 codes for seizures (PPV 39%, sensitivity 17%) and lone G41 codes for status epilepticus (PPV 61%, sensitivity 6%). These results likely reflect appropriate use of the death certificate to indicate instances when people died of acute symptomatic seizures or *de novo* status epilepticus without an underlying diagnosis of epilepsy (Chief Medical Officer 2018; Middleton et al. 2018).

A coding strategy combining G40–41 and/or R56.8 within hospital admissions data from SMR01 had a PPV of 80% and sensitivity of 73%. This indicates that when present, a hospital admission with these codes often reflected a correct diagnosis of epilepsy but many positive cases were also missed (high number of 164 false negatives). This is intuitive, as not all deceased PWE would have had a prior hospital admission due to epilepsy and they would therefore appear falsely negative on SMR01.

Within the PIS dataset, both our broad list and narrow list of AEDs (appendix B) performed similarly well in terms of a very high sensitivity (91–94%, with overlapping CIs), indicating that they both had the ability to capture almost all cases of epilepsy in the sample validated (capturing 560–576 of the 614 cases). However, PPV was much higher for the narrow list AEDs (90%, CI 87–92%) compared to the broad list AEDs (79%, CI 77–82%), indicating that there were many more additional false positive cases captured by the broad list of AEDs (149 false positives) than the narrow list AEDs (64 false positives). The narrow list AEDs were, therefore, more reliable as a case-ascertainment method. The potential impact in this may be illustrated by the large difference in numbers of participants captured by these two AED strategies within the wider PIS dataset (see figure 1). There were 157,509 deceased persons captured by the broad list of AEDs within the wider PIS dataset. Restricting this to only those taking AEDs in the narrow list removed a substantial proportion, 135,049 participants, leaving a remaining 22,460 captured by the narrow list AEDs alone. The PPV and sensitivity figures in the sample validated are highly suggestive most of the 22,460 captured by this narrow list had epilepsy whilst a large number of the additional 135,049 captured by the

broad AED list were probably false positives. This is consistent with our clinical knowledge that the AEDs that differentiated our narrow list from our broad list of AEDs (e.g. diazepam and gabapentin) are frequently used for conditions other than epilepsy and therefore are likely to have contributed many false positives.

F25 epilepsy Read codes within GP primary care data had a low PPV of 77%, suggesting that these were poor at correctly identifying epilepsy.

3.3 Level 2 – validation of algorithms combining two database coding or AED strategies together

Table 2 summarises the Level 2 validation results, assessing 25 algorithms. The highest PPV of 100% was conferred by combining *F25* epilepsy Read codes in primary care with R56.8 seizures codes within the NRS causes of death, but this is unlikely to have been reliable as there were very few people actually captured by this approach and it missed most of the positive cases (7% sensitivity). PPV was otherwise similarly high between 85–94% for the next 16 algorithms (all with overlapping CIs) regardless of the database assessed or coding strategy used within these 16 algorithms. This suggests that the combination of two databases together normally captured a correct diagnosis of epilepsy and was more important than the specific algorithm used within each of those databases. However, there was some variation in the associated sensitivities across these 16 algorithms, with the highest sensitivity (93%, CI 88–97%) conferred by combining *F25* epilepsy Read codes from primary care with AEDs in the narrow list (appendix B2). This strategy had the ability to capture nearly all of the epilepsy cases within the validation sample for whom primary care data were available (111 of 120 epilepsy cases captured). The next most effective coding strategy was combining G40–41 and/or R56.8 causes of death codes with AEDs in the narrow list (sensitivity 81%, CI 77–84%). This combination of both a high PPV and sensitivity strongly suggests the 1,921 cases this algorithm captured within the wider dataset, all of whom died with epilepsy (G40), status epilepticus (G41), or seizure (R56.8) codes listed as their causes of death, were likely to have had epilepsy. Removing R56.8 from this algorithm in NRS did not change PPV as it was 94% (CI 92–96) in G40–41 without R56.8 and 91% (CI 89–94%) with R56.8 included. However, the loss of R56.8 lowered sensitivity greatly to 68% (CI 65–72%) from having been 81% (CI 77–84%) with R56.8 included. This means that in these circumstances, the R56.8 cause of death code was helpful in allowing the algorithm to capture a larger proportion of the deceased cases (494 of 614) without creating many additional false positives. The narrow list of AEDs also helped to increase the PPVs of SMR01 hospital admissions captured by G40–41 and/or R56.8 codes to 93% (CI 90–95%) compared to when hospital admissions were coded alone in Level 1 (PPV 80%, CI 77–84%), although sensitivities remained similarly low: 68% (CI 64%–72%) for the former and 73% (CI 70–77%) for the latter.

3.4 Level 3 – validation of algorithms combining three database coding or AED strategies together

Table 3 summarises the Level 3 validation results, assessing 27 algorithms. The first six achieved PPVs of 100% and included data from primary care. However, these are unlikely to have been reliable as there were very few people actually captured by these algorithms and they missed most of the positive cases (2–6% sensitivity). Although the next 20 algorithms demonstrated a very high PPV between 84–95% (with overlapping CIs), the overall limitation to accuracy came from generally poor sensitivities, ranging between 5–71% across these 20 algorithms (most between 50–60%). This is intuitive in that whilst having epilepsy indicators in three different databases may streamline a patient well into a correct administrative epilepsy diagnosis (high PPV and low false positive rate), not all PWE will have had an opportunity to be captured in three different databases (lowering sensitivity and increasing false negatives).

4. Discussion

We have undertaken the first data-linkage and validation study to focus on establishing how accurately administrative healthcare data identify epilepsy cases in deceased adults (Kerr 2012; Mbizvo et al. 2020). The findings are likely to be relevant mostly to active cases of epilepsy prior to death, as we required the coded seizure and/or AED prescription indicators of epilepsy be present within a maximum of seven years before death. The International League Against Epilepsy defines epilepsy remission as 10 years seizure-free, with the last 5 years off AEDs (Fisher et al. 2014). We have made three key findings that may help future researchers aiming to identify deceased epilepsy cases using administrative data.

First, we show that ICD-10-coded causes of death alone can be used to identify epilepsy in a fairly accurate manner, although the codes need to be selected appropriately to include either G40 (epilepsy) alone or G40–41 (epilepsy and status epilepticus) *without* R56.8 (seizures) to allow PPVs of >90% to be achieved. However, as the associated sensitivity is likely to be ~70%, we caution that such an approach is likely to also miss many positive cases.

Second, if the resources are available to do so, our study indicates that there is additional value in linking a second administrative healthcare dataset to cause of death coding in order to help improve the overall diagnostic accuracy. Where the option is present, the second dataset should consist of a carefully selected group of AEDs more commonly used in epilepsy. When these are used in combination with cause of death codes G40–41 and/or R56.8, PPVs >90% and sensitivities >80% can be achieved. Whilst primary care data alone (*F25* epilepsy Read codes captured by GPs) are potentially unhelpful in identifying epilepsy cases in a deceased population, combining these with a selected group of AEDs more commonly used in epilepsy may be effective, achieving PPVs >85% and sensitivities >90%. There have been few studies validating the accuracy of primary care data in capturing epilepsy cases worldwide

(Fonferko-Shadrach et al. 2017; Mbizvo et al. 2020; Meeraus et al. 2013; Pickrell et al. 2015; Tu et al. 2014), perhaps because of difficulties obtaining such data. Our findings help promote increased use of primary care data to help identify epilepsy cases.

Finally, we caution against using an algorithm linking in a third dataset to two already linked datasets when trying to identify deceased epilepsy cases because although this tips the balance towards a very high PPV (often 93–95%), there are likely to be many missed positive cases and therefore unhelpfully low sensitivity. The maximum sensitivity we were able to achieve across algorithms containing three linked datasets was 71%.

In a previous systematic review, we showed that ICD-10 codes are generally good at identifying epilepsy cases within administrative healthcare datasets of living populations, with PPVs and sensitivities frequently >80% (Mbizvo et al. 2020). The results of our current study are supportive of this, although we caution against making direct comparisons between living and deceased epilepsy populations. More research will be needed to help further ascertain how accurately administrative healthcare data can identify epilepsy cases in deceased populations (Pickrell and Kerr 2020).

Our study's strengths include that the sample validated were taken from a linked dataset containing virtually the entire deceased Scottish population of interest using the three national ISD datasets, supplemented by potentially the largest deceased primary care epilepsy dataset in Scotland (Herrett et al. 2015; Meeraus et al. 2013). This is also the first dedicated diagnostic test accuracy study for cause of death coding in epilepsy (Mbizvo et al. 2020; Pickrell and Kerr 2020). Although the diagnostic accuracy of mortality coding has been extensively validated in other areas, such as in sudden cardiac death (Singh et al. 2019), there is a currently a paucity in epilepsy literature perhaps due to lack of a robust infrastructure to allow access to individual patient medical records as a diagnostic reference standard (Mbizvo et al. 2019a; Pickrell and Kerr 2020). Our study was strengthened by national access to in-depth medical records through a supportive network of neurology colleagues. The sample was not limited to patients under specialist neurology care as all of the available medical records were reviewed, including in cases were only non-specialist attendances to emergency care/inpatients had been made by a patient. This helped to allow the sample to be more representative of the general population, where some patients are not known to a specialist. Our study focuses on validating coding strategies used within potentially active epilepsy cases as future research into epilepsy-related mortality is likely to benefit more from ascertaining active than resolved epilepsy cases.

Our study's main limitation was being unable to provide valid negative predictive value (NPV) and specificity estimates, as all study participants were included based on having at least one epilepsy indicator across one or more

databases. Therefore, any NPV and specificity estimates could not be generalised to a population without epilepsy indicators. This limitation is unlikely to have changed the external validity of our study as in a recent systematic review, we identified that NPVs and specificities in administrative epilepsy data accuracy studies are typically >90% (Mbizvo et al. 2020), meaning our study is perhaps unlikely to have contributed any new information about these outcomes. However, our systematic review also noted that across the 30 available studies validating administrative epilepsy data globally, less than half were able to provide data on NPV and specificity due to similar problems gaining permissions or access to a disease-free cohort (Mbizvo et al. 2020). This is an obstacle that will require further consideration in future studies validating administrative epilepsy data. Another limitation of our study was failure to have all of the medical record cases reviewed in duplicate by two specialists. This was due to study resource limitation. We ameliorated this limitation by checking diagnostic interrater agreement in a 10% sample (n = 100) of duplicate-reviewed medical records, ensuring experienced epileptologists reviewed the medical records and were also blinded to the administrative coding of participants. It remains possible that there may have been bias towards a positive gold-standard epilepsy diagnosis in cases previously managed by a neurologist versus those which were not although this is unlikely to have been significant in this study because the medical records available were extensive and included GP referral letters from primary care, secondary care records, and emergency care records, increasing the chances that the clinical information available was sufficient to allow our experienced team of consultant epileptologists to make an independent diagnosis of epilepsy in the absence of any additional neurology support. A potential limitation was failure to establish a gold-standard diagnosis in 66 participants with incomplete/missing medical records. Overall, this is unlikely to have changed our conclusions given these individuals would have formed only seven percent of the sample validated. Finally, not all countries will have the advantage of administrative datasets with national coverage as in Scotland, perhaps limiting the overall generalisability of our study. However, this is something many countries will be working towards, e.g. in Canada, Holland, and Italy, where there are already vast networks of administrative data (Franchi et al. 2013; Tu et al. 2014; Wassenaar et al. 2018), and so our study helps provide a framework for the methodological avenues researchers could use to validate such datasets.

In conclusion, our study increases confidence in the validity of using administrative data to identify epilepsy cases in mortality studies, particularly when using ICD-10-coded causes of death linked to appropriately chosen AEDs. This helps provide a context within which further work can be done to investigate the rate of epilepsy-related deaths, proportion that were potentially avoidable, and modifiable risk factors (Mbizvo et al. 2019b). Our findings also support the utility of establishing disease-based mortality registries for surveillance, such those available in the UK, US, and Canada (Donner EJ and Devinsky O 2018; Thomas and Osland 2020; Verducci et al. 2019). The study

focuses on validating data from deceased adults and future work will be required to also assess how accurately administrative healthcare data identify epilepsy in deceased children.

Acknowledgements

We are grateful to Saif Razvi (NHS Ayrshire and Arran), Myles Connor (NHS Borders), Ondrej Dolezal (NHS Dumfries and Galloway), Russell Hewett (NHS Greater Glasgow and Clyde), Linda Gerrie and Graham Mackay (NHS Grampian), Martin Zeidler (NHS Fife), Katy Murray (NHS Forth Valley), Kate Taylor (NHS Highland), John Paul Leach (NHS Lanarkshire), Kathleen White and Ian Morrison (NHS Tayside) for arranging access to medical records in the respective areas. We thank Dave Kelly and Albasoft Ltd for facilitating access to primary care data. We would like to acknowledge the support of the eDRIS Team (National Services Scotland) for their involvement in obtaining approvals, provisioning and linking data and the use of the secure analytical platform within the National Safe Haven. We are also grateful to Siddharthan Chandran and Catherine Sudlow for tactical support and advice on the project. We are grateful to Jane Andrews for administerial support in relation to this project.

Declarations of interest

GKM received speaker honoraria from UCB Pharma on two occasions. The work presented was for educational purposes and unrelated to UCB Pharma. RFMC has received speaker honoraria from Zogenix and GWPharma, has provided paid consultancy to Zogenix, Eisai, and GWPharma, and received conference travel grants from Zogneix, Eisai and GWPharma. The remaining authors have no conflicts of interest.

Funding

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.

Ethics approval, informed consent, trial registration

The study was approved by South East Scotland Research Ethics Committee 2 (IRAS 181131, 15/SS/0165), and Scottish Public Benefit and Privacy Panel for Health and Social Care.

5. References

Albasoft 'Albasoft Ltd', <http://www.albasoft.co.uk/>, accessed 01/05/2020.

Bossuyt, P. M., et al. (2015), 'STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies', *Clin Chem,* 61 (12), 1446-52.

Chief Medical Officer (2018), 'Guidance for Doctors Completing Medical Certificates of The Cause of Death (MCCD) and its Quality Assurance '. <https://bit.ly/2yyMzlC>, accessed 03/03/2020.

Death Certifcation Review Service (2016), 'Annual Report 2015–2016 '. <HTTPS://goo.gl/sBDSoU>, accessed 25/06/2020.

Donner EJ and Devinsky O (2018), 'Registries for SUDEP research', in Panelli R, et al. (eds.), *Continuing the global conversation [online]* (SUDEP Action & SUDEP Aware).

Douglas, A., et al. (2015), 'Pilot study linking primary care records to Census, cardiovascular hospitalization and mortality data in Scotland: feasibility, utility and potential', *J Public Health (Oxf)*.

Du Z and Hao y (2019), 'Package 'reportROC': An Easy Way to Report ROC Analysis', *CRAN*. <https://bit.ly/2V3X6OA>, accessed 08/04/2020.

eDRIS Team (2018), 'Researcher guide: requesting outputs from Safe Haven and disclosure control '. <https://bit.ly/2XzD1Bk>, accessed 16/11/2019.

--- (2020), 'How do we link the data?'. <https://bit.ly/39DJ4ZB>, accessed 27/07/2020.

England MJ, et al. (2013), 'IOM Report 2012: Epilepsy Across the Spectrum: Promoting Health and Understanding Abstracts', *Epilepsy Currents,* 13 (1), 2-492.

Escalaya, A. L., et al. (2015), 'Epilepsy and mortality in Latin America', *Seizure,* 25, 99-103.

Fernie, C. G. M. (2019), 'Scotland already has world leading mortality review system', *Bmj-British Medical Journal,* 364.

Fisher, R. S., et al. (2005), 'Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE)', *Epilepsia,* 46 (4), 470-2.

Fisher, R. S., et al. (2014), 'ILAE official report: a practical clinical definition of epilepsy', *Epilepsia,* 55 (4), 475-82.

Fonferko-Shadrach, B., et al. (2017), 'Validating epilepsy diagnoses in routinely collected data', *Seizure,* 52, 195-98.

Franchi, C., et al. (2013), 'Validation of healthcare administrative data for the diagnosis of epilepsy', *J Epidemiol Community Health,* 67 (12), 1019-24.

Garratt E, Barnes H, and Dibben C (2010), 'Health administrative data: Exploring the potential for academic research. St Andrews: Administrative Data Liaison Service'.

Hamer, H. M., et al. (2012), 'Prevalence, utilization, and costs of antiepileptic drugs for epilepsy in Germany--a nationwide population-based study in children and adults', *J Neurol,* 259 (11), 2376-84.

Herrett, E., et al. (2015), 'Data Resource Profile: Clinical Practice Research Datalink (CPRD)', *Int J Epidemiol,* 44 (3), 827-36.

ILAE (1993), 'Guidelines for Epidemiologic Studies on Epilepsy', *Epilepsia,* 34 (4), 592-96.

Information Services Division (2016a), 'Prescribing Information System (PIS)', *National Data Catalogue* (Scotland: ISD Scotland).

--- (2016b), 'General Acute Inpatient and Day Case - Scottish Morbidity Record (SMR01)', *National Data Catalogue* (Scotland: ISD Scotland).

--- (2016c), 'National Records of Scotland (NRS) - Deaths Data', (ISD Scotland).

--- (2020), 'General Practice - SPIRE '. <https://bit.ly/3eldK3B>, accessed 03/03/2020.

Jette, N., et al. (2010), 'How accurate is ICD coding for epilepsy?', *Epilepsia,* 51 (1), 62-9.

Katikireddi, S. V., et al. (2018), 'Assessment of health care, hospital admissions, and mortality by ethnicity: population-based cohort study of health-system performance in Scotland', *Lancet Public Health,* 3 (5), e226-e36.

Kee, V. R., et al. (2012), 'A systematic review of validated methods for identifying seizures, convulsions, or epilepsy using administrative and claims data', *Pharmacoepidemiology and Drug Safety,* 21, 183-93.

Kerr, M. P. (2012), 'The impact of epilepsy on patients' lives', *Acta Neurol Scand Suppl,* (194), 1-9.

Landis, J. R. and Koch, G. G. (1977), 'The measurement of observer agreement for categorical data', *Biometrics,* 33 (1), 159-74.

Levira, F., et al. (2016), 'Premature mortality of epilepsy in low- and middle-income countries: A systematic review from the Mortality Task Force of the International League Against Epilepsy', *Epilepsia*, 1-11.

Mbizvo, G. K., et al. (2019a), 'Epilepsy-related and other causes of mortality in people with epilepsy: A systematic review of systematic reviews', *Epilepsy Res,* 157, 106192.

Mbizvo, G. K., et al. (2019b), 'The Scottish Epilepsy Deaths Study (Seds): Identifying Avoidable Epilepsy-Related Deaths', *Journal of Neurology Neurosurgery and Psychiatry,* 90 (12), E16-E17.

Mbizvo, G. K., et al. (2020), 'The accuracy of using administrative healthcare data to identify epilepsy cases: A systematic review of validation studies', *Epilepsia,* 61 (7), 1319-35.

Meeraus, W. H., et al. (2013), 'Childhood epilepsy recorded in primary care in the UK', *Arch Dis Child,* 98 (3), 195-202.

Middleton, O., et al. (2018), 'National Association of Medical Examiners position paper: Recommendations for the investigation and certification of deaths in people with epilepsy', *Epilepsia,* 59 (3), 530-43.

Murray, C. J., et al. (2012), 'Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010', *Lancet,* 380 (9859), 2197-223.

National Records of Scotland (2020), 'The Medical Certificate of the Cause of Death '. <https://bit.ly/3aPkqnZ>, accessed 03/03/2020.

National Records Scoltand (2016), 'Coding the causes of death', *Death Certificates and Coding the Causes of Death*. <https://goo.gl/oWTxQU>, accessed 03/03/2020.

Nevalainen, O., et al. (2014), 'Epilepsy-related clinical characteristics and mortality: a systematic review and meta-analysis', *Neurology,* 83 (21), 1968-77.

O'Malley, K. J., et al. (2005), 'Measuring diagnoses: ICD code accuracy', *Health Services Research,* 40 (5), 1620-39.

Pavis, S. and Morris, A. D. (2015), 'Unleashing the power of administrative health data: the Scottish model', *Public Health Res Pract,* 25 (4), e2541541.

Pickrell, W. O. and Kerr, M. P. (2020), 'SUDEP and mortality in epilepsy: The role of routinely collected healthcare data, registries, and health inequalities', *Epilepsy Behav,* 103 (Pt B), 106453.

Pickrell, W. O., et al. (2015), 'Epilepsy and deprivation, a data linkage study', *Epilepsia,* 56 (4), 585-91.

Public Health Scotland 'Data and intelligence - Previously ISD Scotland', <https://www.isdscotland.org/>, accessed 29/04/2020.

RCPCH (2013), *Coordinating Epilepsy Care: a UK-wide review of healthcare in cases of mortality and prolonged seizures in children and young people with epilepsies. Child Health Reviews - UK* (London: RCPCH).

RStudio Team (2015), 'RStudio: Integrated Development for R ', (1.2.1335; Boston, MA RStudio, Inc.).

Shackleton, D. P., et al. (2002), 'Survival of patients with epilepsy: an estimate of the mortality risk', *Epilepsia,* 43 (4), 445-50.

Singh, S., et al. (2019), 'Diagnostic Algorithms for Cardiovascular Death in Administrative Claims Databases: A Systematic Review', *Drug Saf,* 42 (4), 515-27.

Tan, M., et al. (2015), 'Development and validation of an epidemiologic case definition of epilepsy for use with routinely collected Australian health data', *Epilepsy & Behavior,* 51, 65-72.

Thiele C (2020), 'cutpointr: Determine and Evaluate Optimal Cutpoints in Binary Classification Tasks '. <https://rdrr.io/cran/cutpointr/man/roc.html>, accessed 01/05/2020.

Thijs, R. D., et al. (2019), 'Epilepsy in adults', *Lancet,* 393 (10172), 689-701.

Thomas, R. H. and Osland, K. (2020), 'Learnings from deaths - the Epilepsy Deaths Register', *Epilepsy Behav,* 103 (Pt B), 106454.

Thurman, D. J., et al. (2017), 'The burden of premature mortality of epilepsy in high-income countries: A systematic review from the Mortality Task Force of the International League Against Epilepsy', *Epilepsia,* 58 (1), 17-26.

Tu, K., et al. (2014), 'Assessing the validity of using administrative data to identify patients with epilepsy', *Epilepsia,* 55 (2), 335-43.

Verducci, C., et al. (2019), 'SUDEP in the North American SUDEP Registry: The full spectrum of epilepsies', *Neurology,* 93 (3), e227-e36.

Viera, A. J. and Garrett, J. M. (2005), 'Understanding interobserver agreement: the kappa statistic', *Fam Med,* 37 (5), 360-3.

Wassenaar, M., et al. (2018), 'Validity of health insurance data to identify people with epilepsy', *Epilepsy Res,* 139, 102-06.

Watila, M. M., et al. (2018), 'Overall and cause-specific premature mortality in epilepsy: A systematic review', *Epilepsy Behav*.

Figure legends

Figure 1: Study design – 1A Flow diagram summarising flow of participants to create a linked research dataset; **1B – Venn diagram** summarising distribution of the deceased population captured across the databases. The primary care database was unable to be included in the Venn diagram due to risk of patient deidentification as some of the overlapping categories with other databases contained five or fewer participants.

Abbreviations: AED – antiepileptic drug; *F25* – primary care diagnostic Read codes for epilepsy; G40–41 – International Classification of Disease 10 (ICD-10) codes for epilepsy and status epilepticus; R56.8 – ICD-10 code for seizures; CHI – Scottish Community Health Index NHS patient identification number; GP – General practitioner