

## **Developing and validating a clinical prediction model to predict epilepsy-related hospital admission or death within the next year using administrative healthcare data: a population-based cohort study protocol**

Gashirai K. Mbizvo<sup>1,2,3\*</sup>, Glen P. Martin<sup>4</sup>, Laura J. Bonnett<sup>5</sup>, Pieta Schofield<sup>6</sup>, Hilary Garret<sup>7</sup>, Alan Griffiths<sup>7</sup>, W Owen Pickrell<sup>8,9</sup>, Iain Buchan<sup>6</sup>, Gregory Y.H. Lip<sup>1,10,¶</sup>, Anthony G. Marson<sup>2,3,¶</sup>

<sup>1</sup> Liverpool Centre for Cardiovascular Science at University of Liverpool, Liverpool John Moores University and Liverpool Heart & Chest Hospital, Liverpool, United Kingdom

<sup>2</sup> Pharmacology and Therapeutics, Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, United Kingdom

<sup>3</sup> The Walton Centre NHS Foundation Trust, Liverpool, United Kingdom

<sup>4</sup> Division of Informatics, Imaging and Data Science, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, United Kingdom

<sup>5</sup> University of Liverpool Department of Biostatistics, Liverpool, United Kingdom

<sup>6</sup> Department of Public Health, Policy and Systems, Institute of Population Health, University of Liverpool, Liverpool, United Kingdom

<sup>7</sup> Public Involvement Team, Applied Research Collaboration North West Coast (ARC NWC), Liverpool, United Kingdom

<sup>8</sup> Neurology Research Group, Swansea University Medical School, Faculty of Medicine, Health and Life Science, Swansea University, Swansea, Wales, United Kingdom

<sup>9</sup> Morriston Hospital, Swansea Bay University Health Board, Swansea, Wales, United Kingdom

<sup>10</sup> Danish Centre for Health Services Research, Department of Clinical Medicine, Aalborg University, Aalborg, Denmark

¶Joint senior authors

**\* Corresponding author**

Dr Gashirai K Mbizvo –Email: [Gashirai.Mbizvo@liverpool.ac.uk](mailto:Gashirai.Mbizvo@liverpool.ac.uk)

## **Abstract**

### *Introduction*

This retrospective open cohort study develops and externally validates a clinical prediction model (CPM) to predict the joint risk of two important outcomes occurring within the next year in people with epilepsy (PWE): A) seizure-related emergency department or hospital admission; and B) epilepsy-related death. This will provide clinicians with a tool to predict either or both of these common outcomes. This has not previously been done despite both being potentially avoidable, interrelated, and devastating for patients and their families. We hypothesise that the CPM will identify individuals at high or low risk of either or both outcomes. We will guide clinicians on proposed actions to take based on the overall risk score.

### *Methods and analysis*

Routinely collected electronic health data from Clinical Practice Research Datalink (CPRD), Secure Anonymised Information Linkage databank (SAIL), Combined Intelligence for Population Health Action (CIPHA), and TriNetX research platforms will be used to identify PWE aged  $\geq 16$  years having outcomes A and/or B between 2010–2022. Data are held for 60 million patients in England on CPRD, 3.1m in Wales on SAIL, 2.6m in Cheshire and Merseyside on CIPHA, and 250m across 19 countries in TriNetX. Candidate predictors will include demographic, lifestyle, clinical, and management. Logistic regression and multistate modelling will be used to develop a suitable CPM (informed by clinician and public consultation), assessing predictive performance across development (CPRD) and external validation (SAIL, CIPHA, TriNetX) datasets.

### *Conclusions*

This is the largest study to develop and validate a CPM for PWE, creating an internationally generalisable tool for subsequent clinical implementation. It is the first to predict the joint risk of acute admissions and death in PWE. Mortality prediction is highlighted by NICE as a key recommendation for epilepsy research. The study has been co-developed by epilepsy researchers and members of the public affected by epilepsy.

**Keywords:** risk scoring tool, clinical tools, routine data, electronic health data, accident and emergency attendance, inpatient admission, mortality

## **Lay summary**

Some people with epilepsy (PWE) are at high risk of hospital admission or death because of seizures. If we give clinicians a tool to predict who, they'll be in a better position to prevent it. Although statistical methods predicting future events are widely available, they haven't yet been used to predict seizure-related hospital admission or death. Our study is the first to do this.

We'll analyse anonymised electronic research data from thousands of PWE in England. Among them, some will have been admitted to hospital or died because of seizures between 2010–2022. We'll analyse their age, gender, ethnicity, features of their epilepsy, and medical conditions they developed in the year before being admitted to hospital or dying. From this, we'll create a statistical tool to predict the chance of someone else with epilepsy being admitted to hospital or dying within a year. The tool's external accuracy will be checked in Cheshire and Merseyside, Wales, North America, Europe, and other countries.

Giving clinicians the tool should generate substantial impact for PWE. For example, emergency epilepsy clinics tend to be reserved for people experiencing a first seizure. However, given our prediction tool tells clinicians which people with an established diagnosis of epilepsy are at risk of seizure-related hospital admission or death within a year, it would provide strong justification for restructuring services such that these high-risk people are also seen in emergency clinics. A high-risk score could also prompt referral for epilepsy surgery sooner than previously considered. It could also prompt multidisciplinary team meetings between neurology and, for example, cardiology if the newly identified risks were cardiac. Such emergency interdisciplinary discussion would not normally happen for a person with epilepsy without good reason: providing evidence for an increased risk of death or hospital admission within a year due to a newly acquired cardiac problem would be good reason.

## Introduction

### *Clinical prediction models (CPMs)*

CPMs estimate the risk of a future clinical outcome in an individual patient based on information retrieved from their routinely collected health data including age, sex, ethnicity, comorbidities, previous treatments, and other characteristics [1]. Therefore, CPMs help guide and standardise shared clinical decision-making and service delivery by drawing attention to high-risk individuals (needing earlier or more invasive treatments), and low-risk individuals (needing less invasive treatments with fewer side-effects) [2]. This helps personalise clinical management strategies. CPMs also help patients understand their own risks by quantifying them in a succinctly presented tool to aid patient-doctor discussions about risk [3]. For example, our CHA<sub>2</sub>DS<sub>2</sub>-VASc score [4, 5] is a simple 7-item CPM using everyday clinical information to accurately predict risk of stroke after developing atrial fibrillation (AF). This helps guide subsequent anticoagulation treatment decisions and patient discussions. We made CHA<sub>2</sub>DS<sub>2</sub>-VASc freely available online to any clinician in the world at [www.mdcalc.com](http://www.mdcalc.com). This combination of simplicity of use, accuracy, and wide availability has translated into substantial global impact and recommended use of this CPM within National Institute for Health and Care Excellence (NICE) guidelines [6].

### *Seizure-related emergency department or hospital admissions*

As we previously demonstrated [7], people with epilepsy (PWE) are at substantially increased risk of seizure-related emergency department (ED) or hospital admission [8]. Indeed, seizures are the most common neurological cause of ED or hospital admission in England [8], and such admissions can be predictors of subsequent epilepsy-related death [9, 10]. However, undergoing ED or hospital admission for epilepsy is often clinically unnecessary and typically leads to little benefit for management of the epilepsy because treatment decisions in epilepsy are complex and require specialist expertise, training and guidance [8]. The admitted PWE are typically seen by junior doctors and physicians without particular expertise in epilepsy and frequently discharged without specialist consultation or referral [7, 8, 11], representing missed opportunity for risk mitigation at a time when this may be crucial [7]. Such admissions could be avoided by early prediction of high-risk groups (through providing non-specialists such as general practitioners (GPs), paramedics, and physicians with tools for early prediction) because epilepsy is an ambulatory-care-sensitive condition [8, 12, 13]; meaning the high-risk groups are amenable to preventive targeting with scheduled specialist and community resources, diversion to alternative care pathways, optimised self-management, and accelerated medication reviews [14, 15]. Notwithstanding this, there are currently no CPMs for ED or hospital admissions in PWE. Helpful work is planned to develop a risk prediction tool to estimate the benefits a person with epilepsy would receive if conveyed to ED and risks if not [14].

### *Epilepsy-related deaths*

As we previously demonstrated [11, 16], PWE are at significantly increased risk of premature death. Some of those deaths may be entirely unrelated to their epilepsy. However, a substantial proportion are epilepsy-related [16, 17]. These are operationally defined as any death listing epilepsy as an underlying or contributory cause

within death records [11, 18, 19], in line with national guidance [20]. Although sudden unexpected death in epilepsy (SUDEP) is a common epilepsy-related death, we have shown it is equally common for epilepsy-related death to occur through other mechanisms including aspiration pneumonia, cardiac arrest, antiseizure medication (ASM) poisoning, drowning, and alcohol dependence [11, 16]. Most epilepsy-related deaths occur in young adults [11, 16]. When we looked at epilepsy-related deaths as a group, nearly 80% were potentially avoidable [11]. Therefore, it is becoming increasingly pragmatic to investigate epilepsy-related deaths as a group rather than focusing on each cause individually [17, 18, 21].

An effective way to help avoid epilepsy-related deaths is to develop CPMs to identify high-risk groups and thereby aid clinicians in prioritising their care [10]. For example, in non-specialist settings, where such patients often first present [7, 8], high CPM scores could prompt re-discussion about seizure safety, signposting to online resources for support (e.g. [www.epilepsy.org.uk/living/safety](http://www.epilepsy.org.uk/living/safety)), and rapid-access epilepsy clinic referral (rather than reserving such clinics for first-fit patients alone). The potential impact of such approaches is well recognised. For example, the asthma deaths review demonstrated that many deaths were likely to be preventable simply through better proactive clinical management, such as arranging for patients to be seen promptly after a hospital admission, and through following clinical guidelines [22]. In specialist settings, high CPM scores could prompt referral for epilepsy surgery earlier than would have otherwise been considered, new discussion in a multidisciplinary team meeting, further medication reviews, or implementation of seizure alarms and nocturnal supervision.

There are few CPMs of mortality in PWE. Although our recently developed Scottish Epilepsy Deaths Study score [10] predicted epilepsy-related deaths as a group, its reliance on hand-searched medical records limited sample sizes to 224 cases/controls, which widened confidence intervals, reduced precision, and meant only four predictors could be included. The SUDEP-7 and SUDEP-3 inventories focused on predicting SUDEP alone [23], and therefore missed the remaining epilepsy-related deaths as a group. They were also drawn from small sample sizes of 19 and 28 patients, respectively, limiting their reliability. Similarly, the Personalised Prediction Tool [24] focused on SUDEP alone and had only 287 cases and 986 controls. None of the above CPMs have been externally validated, meaning they cannot currently be used in clinical practice [1]. The SUDEP and Seizure Safety Checklist [25] is expert consensus guidance rather than a CPM. Developing a validated CPM for epilepsy-related deaths was highlighted as a key recommendation for epilepsy research in recent NICE guidelines [26].

### ***Aims and benefit to PWE***

We aim to develop and externally validate a CPM to predict the joint risk of two outcomes occurring within the next year in PWE [27]: A) seizure-related ED or hospital admission; and B) epilepsy-related death. We hypothesise the CPM will be able to identify individuals at high or low risk of either or both outcomes. This will give clinicians seeing PWE a validated tool to predict either or both of these important outcomes. This has not been done before despite both outcomes being common, potentially avoidable, devastating for patients and their families, and closely interrelated [7-11, 15]. We will develop guidance for clinicians on proposed actions to take based on the overall risk score. When implemented, we expect this work to contribute to a reduction in the number of seizure-related ED or hospital admissions and epilepsy-related deaths.

## Methods

### *Study design*

We will undertake a retrospective open cohort study [13] of anonymised electronic clinical, administrative, and socio-demographic data from general practices and healthcare organisations (HCOs) contributing to the English CPRD (60 million population) [9, 28] and Welsh SAIL (3.1 million population) [19] research databanks, the CIPHA research platform (2.6 million population, [www.cipha.nhs.uk](http://www.cipha.nhs.uk)): a Trusted Research Environment allowing population analytics across Cheshire and Merseyside (initially developed to combat the COVID-19 crisis [29, 30]), and the TriNetX research platform [31]. TriNetX is an international research platform containing electronic health data from ~250m patients from >120 HCOs across 19 countries predominantly in North America but also South America, Europe, the Middle East, Africa, and Asia Pacific [31, 32]. It holds ~70 billion date- and patient-indexed clinical observations [31]. These datasets (CPRD, SAIL, CIPHA, TriNetX) are demographically representative of their respective general populations [9, 19, 28, 31, 32]. Raw person-level data will be imported from each research platform following the relevant information governance training and data access protocol approval. Imported data will be linked from primary care, secondary care (ED, inpatients, and outpatients), deprivation, and mortality datasets. They will be analysed through 01/01/2010–31/12/2022 to study data from adults aged  $\geq 16$  years, capturing peak mortality risks [11].

Drawing lessons from the limitations highlighted in current CPM literature in the area [10, 23, 24], our study represents a paradigm shift by using large routinely-collected health research datasets, allowing our CPM to be developed using data from hundreds of thousands of PWE. Furthermore, we will undertake both internal and external validation of our CPM within the same project, thereby facilitating clinical implementation directly from our work [1, 4, 5]. We will prioritise variables readily available in primary care, hospital and outpatient settings to help maximise clinical utility both amongst specialists and non-specialists [4, 5, 13]. We will restrict size of the CPM to no more than 7 variables [4, 5] to maximise ease of use for clinicians using the model iteratively [4, 5] (Fig 1A) and avoid statistical overfitting [33]. We will also make the CPM globally accessible by depositing it online [4, 5]. We will focus on including predictor variables that have been newly acquired in the preceding year so that the CPM can also act as a dynamic alert system for the acquisition of important new risks when patients are seen at e.g. annual reviews. Furthermore, we will work with GPs and electronic health record (EHR) vendors such as EMIS® to ensure the CPM is designed in such a way as to facilitate future integration into electronic primary care systems, allowing pre-emptive risk stratification with automated alerts (Fig 1B), with clear action plans developed through GP, specialist, and public consultation.

### *Case ascertainment*

We will apply our previously validated [34, 35] epilepsy symptom and disease codes combined with ASMs to identify PWE within CPRD (Aurum), SAIL, CIPHA, and TriNetX. These have positive predictive values and sensitivities >90% and >80%, respectively, for identifying epilepsy within administrative datasets [34, 35]. Using these codes, our pre-submission feasibility search identified a large cohort of 227,271 PWE aged  $\geq 16$  years within CPRD between 01/01/2010–31/12/2021. This contains data from both incident and prevalent cases, which will be analysed as a single cohort and in respective subgroups.

### *Outcomes of interest*

We will define epilepsy-related death as where seizures or epilepsy are the underlying or contributory causes of death (ICD-10 codes G40–41, R56.8) [11, 19, 34, 35]. The same codes will be used to identify seizure-related ED or hospital admissions [34, 35]. All codes will be combined with a requirement for co-prescribed ASMs to optimise diagnostic accuracy [19, 34, 35].

Two binary outcomes will be assessed between 01/01/2010–31/12/2022: A) first seizure-related ED or hospital admission; and B) epilepsy-related death. Data from PWE experiencing either or both or neither outcomes will be modelled.

### *Predictor variables*

Candidate SNOMED-/Read-/ICD-coded predictors will be derived from existing literature [9, 11, 13, 16, 36] and informed by workshop-based consultation with clinicians and members of the public affected by epilepsy. They will include the following (with continuous variables analysed as such to maximise statistical power, but presented as categories post-hoc to maximise clinical interpretation):

- Age;
- Sex;
- Ethnicity (minority/non-minority);
- Deprivation quintile – Small area level/Welsh Index datasets;
- Cambridge Multimorbidity Score [37];
- Learning difficulties or developmental delay (yes/no);
- Alcohol excess coded in preceding year (yes/no);
- Type of epilepsy (focal-/generalised-/unknown-onset);
- Epilepsy duration (incident  $\leq 1$  year/prevalent  $> 1$  year);
- Management coded in preceding year:
  - Seen by GP about epilepsy (0, 1, 2,  $\geq 3$  times);
  - Seen in a neurology clinic (yes/no) – outpatient dataset;
  - ASM number (1, 2,  $\geq 3$ );
- Emergencies – ED and hospital admissions datasets:
  - Seizure-related ED or hospital admissions in year before study entry (0, 1, 2,  $\geq 3$ );
  - ED or hospital admissions unrelated to seizures in year preceding outcome index (0, 1, 2,  $\geq 3$ );
- Neurological risks coded in preceding year:
  - 0, 1, 2,  $\geq 3$  seizures;

- New stroke/depression/psychosis/dementia/autoimmune encephalitis;
- Other new risks coded in preceding year:
  - *Cardiovascular*: Myocardial infarction/AF/cardiac failure;
  - *Respiratory*: COPD;
  - *Metabolic*: Diabetes/extreme BMI (<18.5, ≥30);
  - *Renal*: Chronic kidney disease stage ≥3;
  - *Infections*: CNS/pneumonia/COVID-19/urinary.

#### *Model development and validation*

*A priori* clinical expertise and data-driven variable selection will refine the variable list for modelling [10]. A blinded variable selection strategy [10] will be implemented *a priori* to reduce the list of CPM predictors down to a clinically pragmatic list of seven variables (or less) [4, 5]. The primary driver of which variables are selected will be consensus amongst the clinical research team and our Public Advisors (people affected by epilepsy) on which predictors are most important clinically and are readily accessible in the datasets (i.e. missing data patterns will be incorporated into the decision-making) [10]. The focus will not be on the strength of the predictor (e.g. using backwards selection) as it is better to select predictors based on a wider body of clinical knowledge than to try to depend on statistical significance of results which may be sensitive to random variation in the data points due to sampling variability, as detailed elsewhere [38]. We will consider penalisation methods (e.g., *lasso*), [39, 40] where appropriate.

CPRD will be used for model derivation. External validation will be undertaken in SAIL, CIPHA and TriNetX, maximising our CPM's international generalisability [1]. Multiple imputation will account for missing data.[10]

Modelling will be done in three ways [2, 27]:

- (i) consider each binary outcome (A = seizure-related ED or hospital admission, and B = epilepsy-related death) separately, and develop individual prediction models for each using logistic regression [2, 10];
- (ii) define a composite outcome of A and B and develop a single logistic regression model for this [2, 10];
- (iii) model the outcomes sequentially through time [27]. For this, we let each outcome (A or B) be a state within a multi-state model [27] where individuals start in an initial (PWE) state, and we then model their risk of moving to A or B states (respecting the temporal order of these). Such an approach also allows us to model the competing risk of death from other causes (co-extracted from death records) and the competing risk of ED or hospital admission unrelated to seizures (also available). Approach (iii) will also allow us to model the interplay between admission and mortality and is thus our preferred approach [27], but we consider (i) and (ii) as computationally easier alternatives.

Internal validation within CPRD will involve 1,000 bootstrap analysis samples. Nagelkerke's  $R^2$  and Brier Scores will estimate overall model performance. Calibration will be estimated with calibration intercept,



slope, and plots. Discrimination will be estimated with area under curve. Calibration and discrimination will be reassessed within SAIL, CIPHA and TriNetX to externally validate model performance [1].

A sample size calculation is shown in Table 1 [33].

**Table 1: Sample size calculation**

<b>Method</b>	<b>We performed a sample size calculation for each outcome based on our previous methodology [33]. We used prevalence figures for each outcome from prior literature [9, 18], supported by our pre-submission CPRD feasibility search.</b>
<b>Logistic regression modelling assumptions</b>	<ul style="list-style-type: none"> <li>(i) The models have a predictive performance of either Nagelkerke’s <math>R^2</math> of 15% or achieve at least a c-statistic of 0.6 (whichever leads to higher sample size to give a conservative estimate).</li> <li>(ii) We target a maximum degree of shrinkage/overfitting of 0.9.</li> <li>(iii) We consider a maximum of 50 candidate predictors per model.</li> </ul>
<b>Outcomes based these assumptions</b>	<ul style="list-style-type: none"> <li><b>A</b> Based on a seizure-related ED or hospital admissions prevalence of 8.8% [9], developing a logistic regression model for this outcome will require a minimum sample size of 44,750 if the model achieves at least a c-statistic of 0.6, or 6,430 if the model achieves a Nagelkerke’s <math>R^2</math> of 15%.</li> <li><b>B</b> Based on an epilepsy-related deaths prevalence of 15.4% [9, 18], developing a logistic regression model for this outcome will require a minimum sample size of 27,528 if the model achieves at least a c-statistic of 0.6, or 4,951 if the model achieves a Nagelkerke’s <math>R^2</math> of 15%.</li> </ul>
<b>Feasibility</b>	We have already undertaken a feasibility search demonstrating a sample size of 227,271 PWE aged $\geq 16$ years within CPRD between 01/01/2010–31/12/2021. Therefore, CPRD will be of more than sufficient size to develop the models for outcome definitions (i) and (ii). At present there are no sample size criteria for developing multi-state prediction models (for outcome definition (iii)) [27]. However, given the size of CPRD, we do not expect this to be a concern. SAIL holds an annual number of approximately 25,000 PWE [19]. Similar numbers are expected in CIPHA. We have identified 271,172 PWE in TriNetX [36]. These numbers will be sufficient for external validation in each dataset [33].

### *Model dissemination*

As per our guide [3], we will explore various presentation formats for the externally validated CPM including points-based systems, graphical score charts, and nomograms. This will be informed by workshop-based consultation with end-users including clinicians, EHR vendors, and the public [3]. A final version of the CPM will be disseminated open-access in peer-reviewed journals and directly to clinicians in workshops, conferences, via charities, and online through [www.mdcalc.com](http://www.mdcalc.com) and GitHub (see e.g. [predictepilepsy.github.io](https://predictepilepsy.github.io) and <https://seds-tool.github.io/seds>), and used to explore automated primary care system integration in future.

### *Public engagement*

Our authorship team includes two members of the public affected by epilepsy, who will co-produce the study with us. They will help select appropriate prediction tool variables and presentation strategies, co-author journal manuscripts, co-lead public engagement workshops, and attend monthly project steering group meetings. Our regular public engagement workshops will also allow others affected by epilepsy to feedback on study design and implementation. We will engage wider public through social media. We have presented this study at a local PPI forum, observing consensus among people affected by epilepsy on need for the study.

## **Discussion**

This is the largest study to develop and validate a CPM for people with epilepsy, creating an internationally generalisable tool for subsequent clinical implementation. The model predicts two important and interrelated outcomes of seizure-related emergency department or hospital admission, and epilepsy-related death. The latter is highlighted as a key recommendation for epilepsy research by NICE [26]. The study has been co-developed by epilepsy researchers and members of the public affected by epilepsy. Whilst the study protocol is limited to developing and validating a CPM only, this will be sufficient to allow external users to consider implementing the CPM into their local clinical care settings, as highlighted in the Prognosis Research Strategy (PROGRESS) 3 guidelines [1].

### *Future directions*

Further work will be required in future to test the model's impact in clinical settings [1]. Whilst externally validating our CPM in geographically distinct Welsh SAIL and international TriNetX datasets will facilitate its clinical implementation sufficiently for external users [1, 4, 5], the use of CIPHA will enhance the potential impact of that clinical implementation further. This is because CIPHA would provide a means for us to feed our externally validated CPM directly back into clinical workflows electronically within Cheshire and Merseyside, allowing us to understand the impact of CPMs like this in future [29, 41, 42]. Through CIPHA, we would be able to implement our CPM for population stratification via electronic dashboards with re-identification of patients directly to attending clinicians, providing them with targeted notifications about risk [41, 42]. Furthermore, CIPHA would allow us to implement long-term monitoring of the predictive performance of our CPM over time, allowing us to adjust it in response to cumulative data, with greater ability to reflect system dynamics such as winter pressures (convolving biological, behavioural and service factors) and thereby combat any future calibration drift in our CPM [43]. To avoid overlap, we would remove Cheshire and Merseyside (~25,000 people with epilepsy (PWE)) from our CPRD development dataset (227,271 PWE) – unaffected of our sample size calculations.

## **Ethics and dissemination**

The University of Liverpool's Research Ethics decision tool was used to determine that ethical approval will not be required as the study consists of a secondary analysis of data that are anonymised by an external party (CPRD, SAIL, CIPHA, TriNetX) and provided to the research team in the fully anonymised format [44]. A final version of the CPM will be disseminated open-access in peer-reviewed journals and made freely available online, as described in more detail earlier.

## **Data management plan**

Data will be curated through University of Liverpool's (UoL) Active Data Storage -a centralised, secure, supported data storage facility with multiple layers of protection. Data are replicated between two secure physical locations and backed up regularly. A regular tape backup is made to a third physical location, and segregated from the public network both physically and logically. Data are encrypted in transit using SSL.

We will use a public repository ([www.github.com](https://www.github.com)) to make all diagnostic and outcome coding algorithms, metadata, and R analysis scripts used publicly available, facilitating external replication and adaptation.

All data storage and use will comply with legal obligations (including GDPR) and UoL's Research Data Management Policy.

### **Study status and timeline**

This study remains at protocol submission stage. We plan to fund the importing of study data and commence data analysis within 12 months of protocol publication.

### **Acknowledgements**

We are grateful to Tonina Takova for support with graphic design.

### **Funding statement**

GKM's salary is funded by an NIHR Clinical Lectureship (CL-2022-07-002). The funders played no role in the design or conduct of this protocol. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising.

### **Competing interests statement**

The authors declare no competing interests.

### **References**

1. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013;10(2):e1001381. Epub 2013/02/09. doi: 10.1371/journal.pmed.1001381. PubMed PMID: 23393430; PubMed Central PMCID: PMC3564751.
2. Bonnett LJ, Kim L, Johnson A, Sander JW, Lawn N, Beghi E, et al. Risk of seizure recurrence in people with single seizures and early epilepsy - Model development and external validation. *Seizure*. 2022;94:26-32. Epub 2021/12/03. doi: 10.1016/j.seizure.2021.11.007. PubMed PMID: 34852983; PubMed Central PMCID: PMC8776562.
3. Bonnett LJ, Snell KIE, Collins GS, Riley RD. Guide to presenting clinical prediction models for use in clinical settings. *BMJ*. 2019;365:l737. Epub 2019/04/19. doi: 10.1136/bmj.l737. PubMed PMID: 30995987.
4. Friberg L, Rosenqvist M, Lip GY. Evaluation of risk stratification schemes for ischaemic stroke and bleeding in 182 678 patients with atrial fibrillation: the Swedish Atrial Fibrillation cohort study. *Eur Heart J*. 2012;33(12):1500-10. Epub 2012/01/17. doi: 10.1093/eurheartj/ehr488. PubMed PMID: 22246443.
5. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*. 2010;137(2):263-72. Epub 2009/09/19. doi: 10.1378/chest.09-1584. PubMed PMID: 19762550.
6. NICE. Atrial fibrillation: diagnosis and management NICE guideline [NG196]. 2021.
7. Dixon PA, Kirkham JJ, Marson AG, Pearson MG. National Audit of Seizure management in Hospitals (NASH): results of the national audit of adult epilepsy in the UK. *BMJ Open*. 2015;5(3):e007325. Epub 2015/04/02. doi: 10.1136/bmjopen-2014-007325. PubMed PMID: 25829372; PubMed Central PMCID: PMC4386236.

8. Dickson JM, Jacques R, Reuber M, Hick J, Campbell MJ, Morley R, et al. Emergency hospital care for adults with suspected seizures in the NHS in England 2007-2013: a cross-sectional study. *BMJ Open*. 2018;8(10):e023352. Epub 2018/10/23. doi: 10.1136/bmjopen-2018-023352. PubMed PMID: 30344177; PubMed Central PMCID: PMC6196871.
9. Wojewodka G, Gulliford MC, Ashworth M, Richardson MP, Ridsdale L. Epilepsy and mortality: a retrospective cohort analysis with a nested case-control study identifying causes and risk factors from primary care and linkage-derived data. *BMJ Open*. 2021;11(10):e052841. Epub 2021/10/27. doi: 10.1136/bmjopen-2021-052841. PubMed PMID: 34697121; PubMed Central PMCID: PMC68547505.
10. Mbizvo GK, Schnier C, Simpson CR, Duncan SE, Chin RFM. Case-control study developing Scottish Epilepsy Deaths Study score to predict epilepsy-related death. *Brain*. 2022:awac463. Epub 2022/12/09. doi: 10.1093/brain/awac463. PubMed PMID: 36477471.
11. Mbizvo GK, Schnier C, Simpson CR, Chin RFM, Duncan SE. A national study of epilepsy-related deaths in Scotland: Trends, mechanisms, and avoidable deaths. *Epilepsia*. 2021;62(11):2667-84. Epub 2021/09/20. doi: 10.1111/epi.17065. PubMed PMID: 34537957.
12. Purdy SG, Tom ; Salisbury, Chris ; Sharp, Debbie. Prioritizing ambulatory care sensitive hospital admissions in England for research and intervention: a Delphi exercise. *Primary health care research & development*. 2010;11(1):41-50.
13. Hippisley-Cox J, Coupland C. Predicting risk of emergency admission to hospital using primary care data: derivation and validation of QAdmissions score. *BMJ Open*. 2013;3(8):e003482. Epub 2013/08/21. doi: 10.1136/bmjopen-2013-003482. PubMed PMID: 23959760; PubMed Central PMCID: PMC3753502.
14. Noble AJ, Mathieson A, Ridsdale L, Holmes EA, Morgan M, McKinlay A, et al. Developing patient-centred, feasible alternative care for adult emergency department users with epilepsy: protocol for the mixed-methods observational 'Collaborate' project. *BMJ Open*. 2019;9(11):e031696. Epub 2019/11/05. doi: 10.1136/bmjopen-2019-031696. PubMed PMID: 31678950; PubMed Central PMCID: PMC6830638.
15. Mathieson A, Marson AG, Jackson M, Ridsdale L, Goodacre S, Dickson JM, et al. Clinically unnecessary and avoidable emergency health service use for epilepsy: A survey of what English services are doing to reduce it. *Seizure*. 2020;76:156-60. Epub 2020/02/25. doi: 10.1016/j.seizure.2020.02.012. PubMed PMID: 32092677.
16. Mbizvo GK, Bennett K, Simpson CR, Duncan SE, Chin RFM. Epilepsy-related and other causes of mortality in people with epilepsy: A systematic review of systematic reviews. *Epilepsy Res*. 2019;157:106192. Epub 2019/09/19. doi: 10.1016/j.epilepsyres.2019.106192. PubMed PMID: 31526975.
17. Devinsky O, Spruill T, Thurman D, Friedman D. Recognizing and preventing epilepsy-related mortality: A call for action. *Neurology*. 2016;86(8):779-86. PubMed PMID: 26674330; PubMed Central PMCID: PMC4763802.
18. Casotto V, Ranzato F, Girardi P, Fedeli U. Increasing epilepsy-related mortality: A multiple causes of death study in Northern Italy. *Seizure*. 2022;94:1-6. Epub 2021/11/22. doi: 10.1016/j.seizure.2021.11.003. PubMed PMID: 34801832.

19. Daniels H, Lacey AS, Mikadze D, Akbari A, Fonferko-Shadrach B, Hollinghurst J, et al. Epilepsy mortality in Wales during COVID-19. *Seizure*. 2022;94:39-42. Epub 2021/12/06. doi: 10.1016/j.seizure.2021.11.017. PubMed PMID: 34864250; PubMed Central PMCID: PMC8626872.
20. Chief Medical Officer. Guidance for Doctors Completing Medical Certificates of The Cause of Death (MCCD) and its Quality Assurance 2018 03/03/2020. Available from: <https://bit.ly/2yyMzIC>.
21. Klovgaard M, Lyng TH, Tsiropoulos I, Uldall PV, Banner J, Winkel BG, et al. Epilepsy-Related Mortality in Children and Young Adults in Denmark: A Nationwide Cohort Study. *Neurology*. 2022;98(3):e213-e24. Epub 2021/11/20. doi: 10.1212/WNL.0000000000013068. PubMed PMID: 34795050.
22. Levy ML, Andrews R, Buckingham R, Evans H, Francis C, Houston R. Confidential Enquiry report. Why asthma still kills: The National Review of Asthma Deaths (NRAD) 2014 19/08/2022. Available from: [www.rcplondon.ac.uk/sites/default/files/why-asthma-still-kills-full-report.pdf](http://www.rcplondon.ac.uk/sites/default/files/why-asthma-still-kills-full-report.pdf)
23. Tarighati Rasekhi R, Devlin KN, Mass JA, Donmez M, Asma B, Sperling MR, et al. Improving prediction of sudden unexpected death in epilepsy: From SUDEP-7 to SUDEP-3. *Epilepsia*. 2021. Epub 2021/06/05. doi: 10.1111/epi.16928. PubMed PMID: 34086290.
24. Jha A, Oh C, Hesdorffer D, Diehl B, Devore S, Brodie MJ, et al. Sudden Unexpected Death in Epilepsy: A Personalized Prediction Tool. *Neurology*. 2021;96(21):e2627-e38. Epub 2021/04/30. doi: 10.1212/WNL.0000000000011849. PubMed PMID: 33910939; PubMed Central PMCID: PMC8205449.
25. SUDEP Action. SUDEP and Seizure Safety Checklist 2017 [07/11/2018]. Available from: <https://bit.ly/347i8iY>.
26. National Institute of Health and Care Excellence (NICE). Epilepsies in children, young people and adults: recommendations for research. NICE guideline [NG217] [Internet]. 2022 16/10/2023. Available from: <https://tinyurl.com/mr44t5jf>.
27. Martin GP, Sperrin M, Snell KIE, Buchan I, Riley RD. Clinical prediction models to predict the risk of multiple binary outcomes: a comparison of approaches. *Stat Med*. 2021;40(2):498-517. Epub 2020/10/28. doi: 10.1002/sim.8787. PubMed PMID: 33107066.
28. CPRD. Latest data from CPRD2021 03/12/2022. Available from: <https://cprd.com/latest-data-cprd-1>.
29. Green MA, Garcia-Finana M, Barr B, Burnside G, Cheyne CP, Hughes D, et al. Evaluating social and spatial inequalities of large scale rapid lateral flow SARS-CoV-2 antigen testing in COVID-19 management: An observational study of Liverpool, UK (November 2020 to January 2021). *Lancet Reg Health Eur*. 2021;6:100107. Epub 2021/05/19. doi: 10.1016/j.lanepe.2021.100107. PubMed PMID: 34002172; PubMed Central PMCID: PMC8114854 conduct of the study. IB declares grants from NIHR, personal fees and other from AstraZeneca, outside the submitted work. SS research is supported by a grant from Wellcome Trust. No other interests to declare.
30. Pattni K, Hungerford D, Adams S, Buchan I, Cheyne CP, Garcia-Finana M, et al. Effectiveness of the BNT162b2 (Pfizer-BioNTech) and the ChAdOx1 nCoV-19 (Oxford-AstraZeneca) vaccines for reducing susceptibility to infection with the Delta variant (B.1.617.2) of SARS-CoV-2. *BMC Infect Dis*. 2022;22(1):270. Epub 2022/03/22. doi: 10.1186/s12879-022-07239-z. PubMed PMID: 35307024; PubMed Central PMCID: PMC8934524.

31. TriNetX. Welcome to the world's largest, living ecosystem of real-world data and evidence for the life sciences and healthcare industries. Global data, for global health 2023 25/08/2022. Available from: [www.trinetx.com](http://www.trinetx.com).
32. Taquet M, Sillett R, Zhu L, Mendel J, Camplisson I, Dercon Q, et al. Neurological and psychiatric risk trajectories after SARS-CoV-2 infection: an analysis of 2-year retrospective cohort studies including 1 284 437 patients. *Lancet Psychiatry*. 2022;9(10):815-27. Epub 2022/08/21. doi: 10.1016/S2215-0366(22)00260-7. PubMed PMID: 35987197; PubMed Central PMCID: PMC9385200.
33. Riley RD, Ensor J, Snell KIE, Harrell FE, Jr., Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441. Epub 2020/03/20. doi: 10.1136/bmj.m441. PubMed PMID: 32188600.
34. Mbizvo GK, Bennett KH, 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;61(7):1319-35. Epub 2020/06/01. doi: 10.1111/epi.16547. PubMed PMID: 32474909.
35. Mbizvo GK, Schnier C, Simpson CR, Duncan SE, Chin RFM. Validating the accuracy of administrative healthcare data identifying epilepsy in deceased adults: A Scottish data linkage study. *Epilepsy Res*. 2020;167:106462. Epub 2020/09/23. doi: 10.1016/j.epilepsyres.2020.106462. PubMed PMID: 32961444.
36. Bucci T, Mbizvo GK, Rivera-Caravaca JM, Mayer J, Marson AG, Abdul-Rahim AH, et al. Epilepsy-Heart Syndrome: Incidence and Clinical Outcomes of Cardiac Complications in patients with Epilepsy. *Curr Probl Cardiol*. 2023:101868. Epub 2023/06/10. doi: 10.1016/j.cpcardiol.2023.101868. PubMed PMID: 37295636.
37. Payne RA, Mendonca SC, Elliott MN, Saunders CL, Edwards DA, Marshall M, et al. Development and validation of the Cambridge Multimorbidity Score. *CMAJ*. 2020;192(5):E107-E14. Epub 2020/02/06. doi: 10.1503/cmaj.190757. PubMed PMID: 32015079; PubMed Central PMCID: PMC7004217.
38. Harrell FE. *Multivariable Modeling Strategies*. 2015 [cited 18/08/2022]. In: *Regression Modeling Strategies* [Internet]. Switzerland: Springer Cham, [cited 18/08/2022]. Available from: <https://link.springer.com/book/10.1007/978-3-319-19425-7>.
39. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1996;58(1):267-88.
40. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med*. 1997;16(4):385-95. Epub 1997/02/28. doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3. PubMed PMID: 9044528.
41. CIPHA. Frequently Asked Questions: Combined Intelligence for Population Health Action (CIPHA): Population Health2020. Available from: <https://bit.ly/3JIIPWr>.
42. Cheshire and Merseyside Health and Care Partnership. Combined Intelligence for Population Health Action (CIPHA)2022 25/08/2022. Available from: [www.cipha.nhs.uk](http://www.cipha.nhs.uk).
43. Jenkins DA, Martin GP, Sperrin M, Riley RD, Debray TPA, Collins GS, et al. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagn Progn Res*. 2021;5(1):1. Epub 2021/01/13. doi: 10.1186/s41512-020-00090-3. PubMed PMID: 33431065; PubMed Central PMCID: PMC7797885.

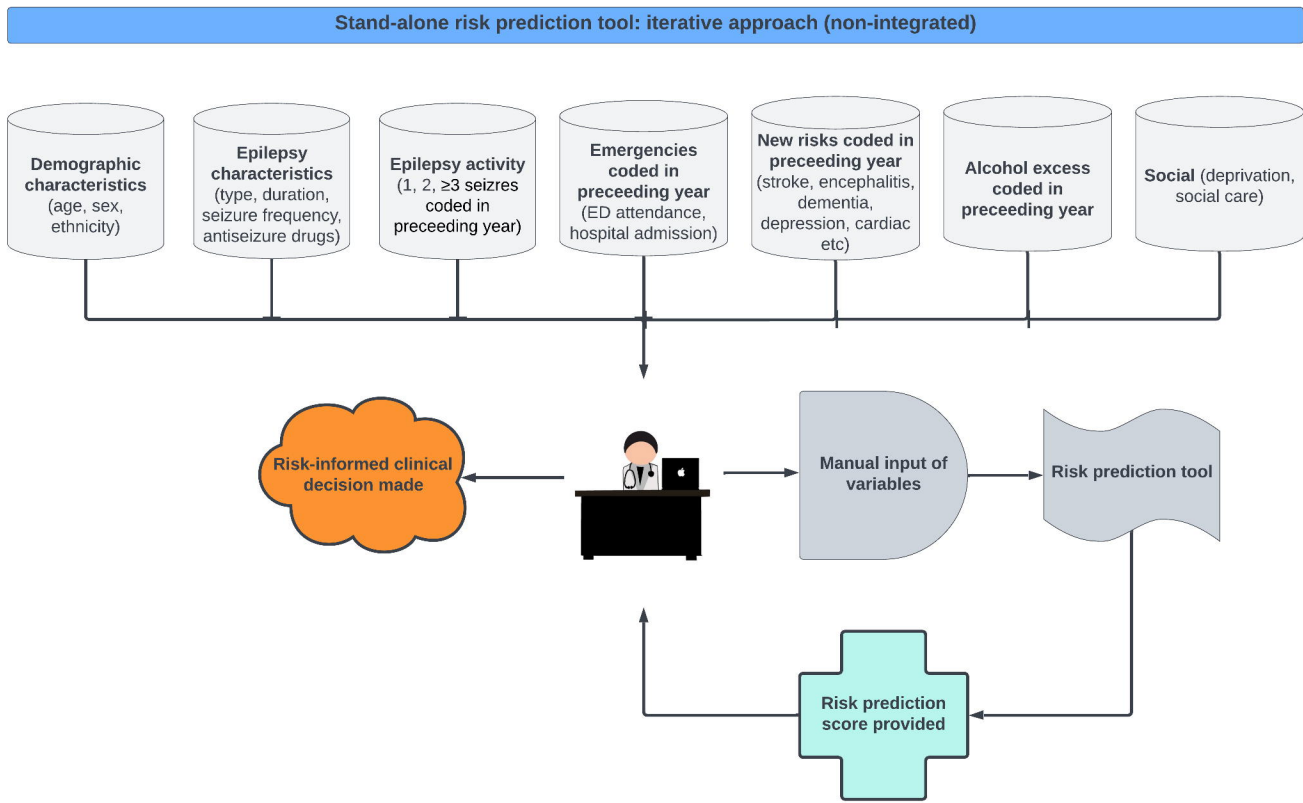
44. University of Liverpool. 'Do I need research ethics approval?' decision tool 2023 14/08/2023.

Available from: <https://liverpool.onlinesurveys.ac.uk/do-i-need-research-ethics-approval-tool>.

### **Figure legends**

**Figure 1:** Future clinical implementation options for the proposed clinical prediction model

*Key:* A = iterative design, B = automated design

**A****B**