Epilepsy & Behavior 123 (2021) 108273

Contents lists available at ScienceDirect

Epilepsy & Behavior

journal homepage: www.elsevier.com/locate/yebeh

Review

Machine learning models for decision support in epilepsy management: A critical review

Eliot D. Smolyansky [a], Haris Hakeem [b,c], Zongyuan Ge [d], Zhibin Chen [c,e,f], Patrick Kwan [b,c,f,g,*]

[a] Melbourne Medical School, The University of Melbourne, Parkville, Victoria 3010, Australia
[b] Department of Neurology, The Alfred Hospital, Melbourne, Victoria 3004, Australia
[c] Department of Neuroscience, Central Clinical School, Monash University, Melbourne, Victoria 3004, Australia
[d] Monash eResearch Centre, Monash University, Clayton 3800, Australia
[e] Clinical Epidemiology, School of Public Health and Preventive Medicine, Monash University, Melbourne 3004, Victoria, Australia
[f] Departments of Medicine and Neurology, The Royal Melbourne Hospital, Melbourne, Victoria 3050, Australia
[g] Chongqing Key Laboratory of Neurology, The First Affiliated Hospital, Chongqing Medical University, Chongqing, China

ARTICLE INFO

Article history:
Received 30 June 2021
Revised 13 August 2021
Accepted 14 August 2021
Available online 8 September 2021

Keywords:
Machine learning
Clinical decision support
Anti-seizure medication
Drug-resistant epilepsy
Epilepsy surgery
Outcomes

ABSTRACT

Purpose: There remain major challenges for the clinician in managing patients with epilepsy effectively. Choosing anti-seizure medications (ASMs) is subject to trial and error. About one-third of patients have drug-resistant epilepsy (DRE). Surgery may be considered for selected patients, but time from diagnosis to surgery averages 20 years. We reviewed the potential use of machine learning (ML) predictive models as clinical decision support tools to help address some of these issues.
Methods: We conducted a comprehensive search of Medline and Embase of studies that investigated the application of ML in epilepsy management in terms of predicting ASM responsiveness, predicting DRE, identifying surgical candidates, and predicting epilepsy surgery outcomes. Original articles addressing these 4 areas published in English between 2000 and 2020 were included.
Results: We identified 24 relevant articles: 6 on ASM responsiveness, 3 on DRE prediction, 2 on identifying surgical candidates, and 13 on predicting surgical outcomes. A variety of potential predictors were used including clinical, neuropsychological, imaging, electroencephalography, and health system claims data. A number of different ML algorithms and approaches were used for prediction, but only one study utilized deep learning methods. Some models show promising performance with areas under the curve above 0.9. However, most were single setting studies (18 of 24) with small sample sizes (median number of patients 55), with the exception of 3 studies that utilized large databases and 3 studies that performed external validation. There was a lack of standardization in reporting model performance. None of the models reviewed have been prospectively evaluated for their clinical benefits.
Conclusion: The utility of ML models for clinical decision support in epilepsy management remains to be determined. Future research should be directed toward conducting larger studies with external validation, standardization of reporting, and prospective evaluation of the ML model on patient outcomes.

1. Introduction

1.1. Challenges in epilepsy management

There remain major challenges for the clinician in managing patients with epilepsy effectively. Ideally, for a patient with newly diagnosed epilepsy, drug selection would be personalized to provide them with the best chance of responding and becoming seizure free. There are now more than 20 anti-seizure medications (ASMs) available for clinicians to choose from [1]. However, a high level evidence-base to guide personalized drug selection is lacking

Abbreviations: AI, artificial intelligence; AUC, area under the curve; ASM, anti-seizure medication; CDS, clinical decision support; DRE, drug-resistant epilepsy; DTI, diffusion tensor imaging; EEG, electroencephalography; iEEG, intracranial electroencephalography; ML, machine learning; MRI, magnetic resonance imaging; PPV, positive predictive value; NPV, negative predictive value; NSF, not seizure free; SF, seizure free; SVM, support vector machine; VBM, voxel-based morphometry.

* Corresponding author at: Department of Neuroscience, Central Clinical School, Monash University, Level 6, The Alfred Centre, 99 Commercial Road, Melbourne, Victoria 3004, Australia.
E-mail address: Patrick.Kwan@monash.edu (P. Kwan).

https://doi.org/10.1016/j.yebeh.2021.108273
1525-5050/

[2,3]. Therefore, ASM selection often relies on trial and error, a process which may expose patients to persistent seizures and medication adverse effects [4].

Furthermore, predicting drug-resistance is a major challenge. One-third of patients will be unable to achieve seizure freedom with current pharmacotherapy [5]. Of these patients with drug-resistant epilepsy (DRE), a proportion will be surgical candidates [6]. A recent meta-analysis concluded that patients with a shorter duration of epilepsy prior to surgery were more likely to be seizure free at follow-up, highlighting the need for prompt referral [7]. Yet, the time from epilepsy diagnosis to surgery averages 20 years in adults [8–11]. Some patients seem to be drug-resistant at epilepsy onset, while in others pharmacoresistance appears to emerge over time [10,12]. Predicting the likelihood of drug-resistance at epilepsy diagnosis may prompt earlier recourse to non-drug treatments such as surgery, sparing patients potentially many years of unsuccessful drug trials, and attendant adverse effects, as well as excess morbidity and mortality.

However, facilitating recourse to surgery for patients with DRE poses a major challenge. It is estimated that only 1% of patients with DRE are referred to epilepsy centers in the United States [13]. Prompt identification of surgical candidates and the prediction of surgical outcomes may help to bridge the divide between the vast numbers of potentially eligible candidates, and the small numbers of patients actually undergoing surgery. As unrealistic patient perceptions of harm may contribute to the underutilization of surgery [9,14,15], surgical outcome prediction would likely help both the clinician and patient in weighing-up risks and benefits.

In this article, we review recent research in individualized prediction of ASM responsiveness, the prediction of DRE, the identification of surgical candidates, and the prediction of surgical outcome, using machine learning (ML) models. We discuss these models' potential as clinical decision support (CDS) tools and their limitations, and propose areas for future research.

1.2. Outcome prediction models for clinical decision support

Prediction models have long been used clinical decision support (CDS) tools in clinical practice [16]. CDS is a broad term that encompasses a variety of tools intended to assist clinicians in workflows and improve patient outcomes. Established examples include alert systems, computerized ECG interpretation, automated assistance with dose adjustment for patients with renal impairment, diagnostic tools, and models that can assist with decision-making (e.g. choosing medications) and predict outcomes [16–18]. However, the data revolution within medicine and science is spurring the expanding interest in CDS and personalized prediction [19,20]. This is being advanced by the digitization of health care systems and the resultant proliferation of administrative databases in insurance and pharmaceutical claims, electronic health records, clinical registries, and biological databases containing genetic and tissue information, to name but a few sources of personal data [16,19,20].

In this context, attempts have been made to individualize the prediction of ASM responsiveness, the prediction of DRE, the identification of surgical candidates, and the prediction of surgical outcome, using traditional statistical approaches. For example, a previous study devised a nomogram scale based on 5 clinical and EEG predictors to predict the 6- and 12-month probability of seizure freedom with ASMs in magnetic resonance imaging (MRI)-negative epilepsy [21]. Huang et al. [22] used multivariate logistic regression to develop a model for predicting DRE in a cohort of children based on only 4 features. Hughes et al. [23] used multivariate generalized linear mixed models developed on the Standard versus New Antiepileptic Drugs (SANAD) [24,25] study dataset to identify patients who would not achieve a period of 12 months of seizure freedom within 5 years of starting treatment. Jehi et al. [26] developed nomograms based on clinical features and type of surgery to provide personalized prediction of probability of seizure freedom or Engel score I [27] at 2 and 5 years post epilepsy surgery.

1.3. Machine learning in clinical medicine

There has also been increasing awareness that current statistical methods may have limited capacity to handle such vast quantities of data, and a growing interest in other techniques such as ML [28]. In contrast to traditional approaches which are based on assuming a data model, algorithms developed by ML are learned directly from the data [29,30]. Examples of ML models include decision trees, support vector machine (SVM), k-nearest neighbor, random forests, artificial neural networks, and K-means [17,29,31–33]. ML is divided into supervised learning and unsupervised learning [32,34]. Supervised learning entails "training" ML algorithms on datasets consisting of inputs (or features) and "labeled" outputs. Unsupervised learning does not employ labeled data, but instead attempts to extract underlying patterns in a dataset [17,31,32,34].

In broad terms, ML models may be classification algorithms or regression algorithms. Classification algorithms predict categorical variables (e.g. seizure free (SF) or not seizure free (NSF)), while regression algorithms predict quantitative variables (e.g. probability of seizure freedom) [32]. Performance metrics such as accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), area under the curve (AUC) of the receiver operating characteristic, and F-measure (harmonic mean of sensitivity and PPV) may be calculated [17,33,34]. Recently, much attention has been gained by classification algorithms for medical artificial intelligence (AI) applications. Abramoff et al. [35] designed a classification algorithm based on multilayered convolutional neural networks (a deep learning method; see below) to detect diabetic retinopathy from fundal photos, and without requiring clinician review. In 2018 it became the first autonomous AI diagnostic tool approved by the US Food and Drug Administration (FDA) [35]. By 2020, 29 AI/ML devices or algorithms had been approved by the FDA, including software for the identification of wrist fractures, the identification of large vessel occlusion on CT angiography, and the identification of suspicious lesions on mammography [36].

A typical ML model training and validation pipeline is as follows. First, in order to build a predictive model relevant inputs must first be selected in a process known as "feature selection" [32]. Some ML algorithms may perform this automatically while others require manual selection [17]. ML algorithms can then be trained, whereby they are exposed to examples of the input features and corresponding labeled outputs, and through this process devise the functions by which to map inputs to outputs [37]. After training, ML models may be tweaked on a tuning dataset and their performance assessed on a test dataset to ensure they are internally valid; which is to say that they are reproducible, or that they can accurately predict outcomes in patients not used in model development, but who are similar and from the same setting. Ultimate assessment of performance, however, requires external validation on a testing set of never-before-seen data of an independent cohort of patients, ideally from a completely different setting, to ascertain the model's generalizability [17,33,38]. Validation is essential to ensure models are not overfit to the data on which they have been trained, a phenomenon in which the model is fit to noise in the data rather than the signal, resulting in a model that may not be reproducible or generalizable [17,33]. Models may be particularly prone to being overfit if they contain too many input features for the sample size on which they were trained [33,34]. Common

ways data sets are generated for training and internal validation are split-sample validation and k-fold cross validation. In split-sample validation a data set is divided (e.g. 50:50 or 70:30) into training and test data subsets [33]. Typically larger datasets are required for this to be a feasible method. Furthermore, random splitting may lead to "lucky" or "unlucky" test sets and thus an unreliable estimation of model performance [17,33]. K-fold cross validation is often useful when datasets are smaller, as it enables recycling of portions of the dataset for training and testing. The data set is divided into k (e.g. 5 or 10) folds or portions of equal size. The model is trained on k-1 folds and tested on the left-out fold. This is repeated k times so that each fold is used once as the test set [33]. The final performance of the model is taken as the average of the performance on each fold [17,33]. Leave-one-out cross validation is a form of k-fold cross validation in which k equals the total sample size. Here the model is trained on k-1 patients and tested on the left-out patient; the procedure is repeated k times such that the model is tested on each patient once [33]. Methods may also be combined such that a dataset may be split into training and testing sets and k-fold cross validation performed on the training set. This enables performance to be assessed with cross validation and also with final validation on the testing set. Discrepancies in performance between results may alert to the possibility of the model being overfit to the training data [34].

A more advanced type of ML, called deep learning, does not depend on features being manually selected, but can self-select features from raw data [39,40]. These models use algorithms called multilayered artificial neural networks. Multilayered artificial neural networks have many "layers" through which data are transformed in complex non-linear ways to ultimately map inputs into outputs [39,40]. This allows the model to uncover relevant associations between features of interest and patterns that are masked from traditional statistical models [41]. Such models have particularly achieved success in medical image analysis [42].

1.4. Machine learning in epilepsy

In epilepsy there has been growing research interest in machine learning applications. ML has been used for seizure detection from EEG data [43] and video footage [44], imaging analysis for the detection of epileptogenic lesions [45], lateralizing temporal lobe epilepsy [46,47], differentiating between epileptic and nonepileptic seizures using autonomic data from wearable devices [48], as well as in a number of studies on outcome prediction for medical and surgical management of epilepsy [49]. There have also been a number of reviews on the use of ML in epilepsy [43,49–51]. These reviews have mostly focused on the use of machine learning for seizure analysis and detection; however, some reviews have discussed applications of machine learning for epilepsy medical and surgical management [49,51].

2. Materials and methods

2.1. Search methods

We searched Medline and Embase databases using the "advanced search" toolbar on September 6, 2020 for relevant articles published in English from the year 2000 onward. A combination of key words and Medical Subject Heading terms were employed. Examples of key words used were "epilep*", "decision support", "machine learning", "predict* model*", "drug resistant epilepsy", "epilepsy surgery", and various combinations of these search terms. Articles with "EEG" or "electroe*" or "electrop*" in their titles were excluded using the "NOT" Boolean operator in the search strategy, to remove the overwhelming numbers of articles related to EEG analysis and seizure prediction. The complete search strategy used for Medline and Embase is provided in the Supplementary Material. References from both databases were uploaded into Endnote X9.3.3 where duplicates were removed; references were then imported into Covidence [52] for title and abstract screening, full text review, and data extraction.

2.2. Inclusion and exclusion criteria

Articles were included if they related to the use of machine learning predictive models to assist with epilepsy management decision-making relevant to the following 4 areas:

1. the prediction of ASM responsiveness,
2. the prediction of DRE,
3. the identification of candidates for epilepsy surgery, and
4. the prediction of surgical outcomes.

Studies on pediatric and adult populations were included. Articles related to seizure or epilepsy diagnosis, EEG analysis, localization of epileptogenic foci, or neurostimulation (e.g., vagal nerve stimulation), were excluded. Articles were also excluded if they were not full text articles, such as abstracts or conference reports, if data was not presented, if they were review articles, if they related to animal models, or if they constructed predictive models without using machine learning techniques.

3. Results

Medline and Embase generated 739 and 711 articles, respectively (Fig. 1). After duplicates were removed 1203 papers remained. Title and abstract screening resulted in 1031 articles being excluded, leaving 172 for full text review. On full text review 159 articles were excluded. Searching through reference lists of included papers yielded another 11 relevant articles; hence, the total number of included articles was 24 (Fig. 1). Of the 24 articles, 6 related to predicting ASM outcomes, 3 to predicting DRE, 2 to identifying candidates for epilepsy surgery, and 13 to predicting surgical outcomes. The methods and results of these articles are summarized in Table 1.

4. Discussion

4.1. ASM selection and predicting treatment response

We identified 6 studies that modeled ASM treatment response using ML approaches [53,54,56–59]. A variety of ML algorithms, input features, and validation strategies were employed. With the exception of one study [56], all employed classificatory ML models with binary outcome measures. The best performing model [59] employed an eXtreme Gradient Boosting (XGBoost) algorithm, achieving near perfect discriminatory ability in differentiating between SF and NSF individuals (AUC 0.98). Although the result is promising, two related studies [53,54] exemplify the importance of assessing external validity in model evaluation. In these studies, a k-NN model taking 5 single nucleotide polymorphisms as inputs and trained on an Australian cohort, was not generalizable to 2 independent UK cohorts.

Another study [56] constructed a prognostic regression ML model, using the random forests algorithm, to predict the probability of requiring treatment change within 12 months of commencing an ASM regimen, using the need for treatment change as a surrogate indicator of the persistence of seizures. This study asked the question: for a given patient prescribed a given ASM regimen,

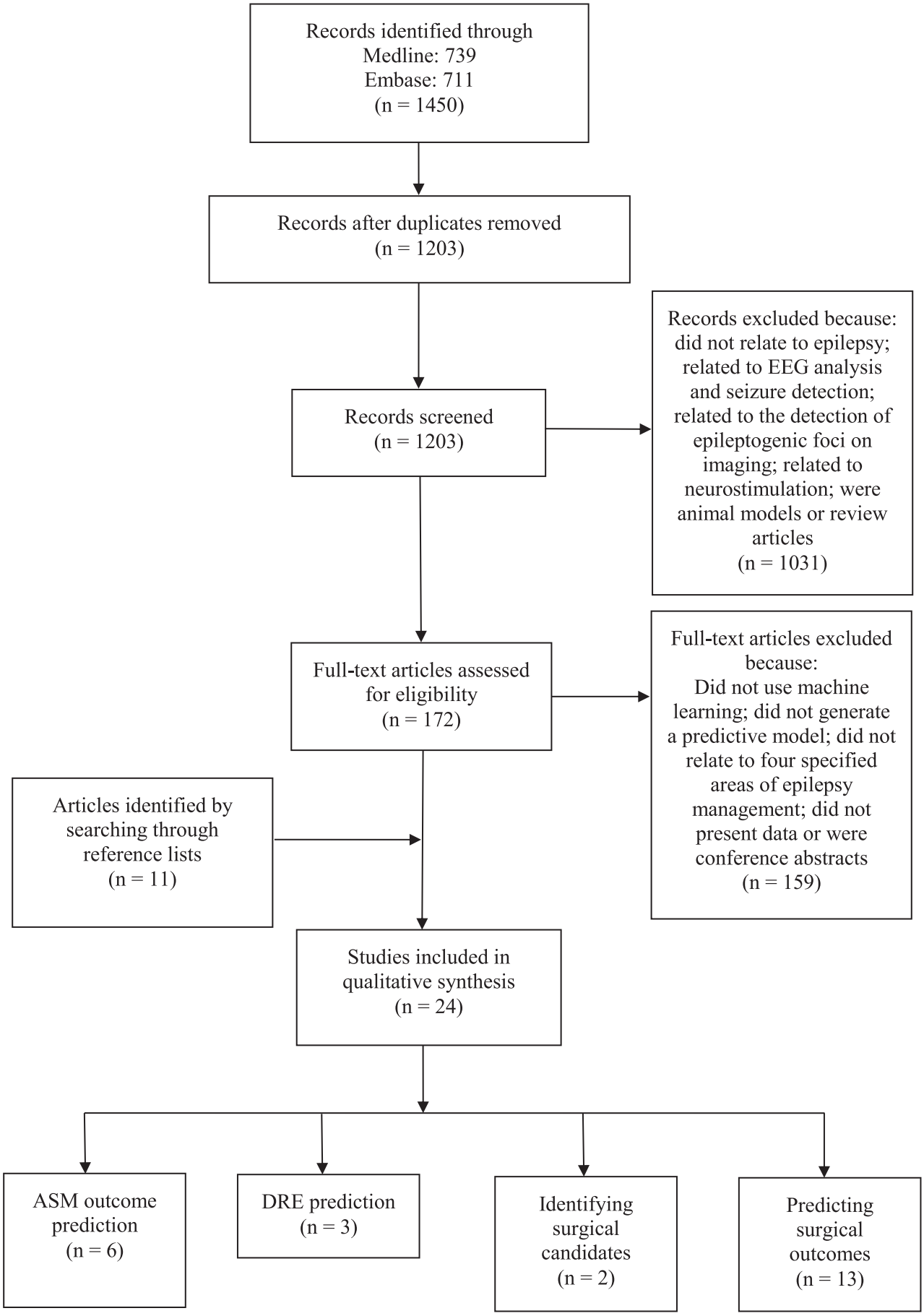

Fig. 1. Flow diagram of the literature search.

Table 1
Studies included for review.

Authors (year)	Aims	Population/Methods	ML models used/compared	ML Input	ML Output	Internal validation	External validation	Results
ASM outcome prediction								
Petrovski et al. [53] (2009)	To predict treatment response to first ever ASM	- Training set of 115 newly treated patients with epilepsy from Australian hospital clinics - Testing on 63 newly treated patients from same hospital population as training cohort - Testing on a separate cohort of 108 community-treated patients with chronic epilepsy from Australia	k-NN	5 SNPs	SF or NSF at 1-year post treatment initiation	- 5-fold cross validation on training set - On testing set of 63 newly treated patients	On 108 patients with chronic epilepsy	- Newly treated patients: sensitivity 91%, specificity 53%, PPV 84%, NPV 69% - Patients with chronic epilepsy: sensitivity 81%, specificity 50%, PPV 81%, NPV 50%
Shazadi et al. [54] (2014)	To determine external validity of Petrovski et al. [53] model in predicting treatment response to first ever ASM	- Training set same as in Petrovski et al. [53] - Testing on 2 independent UK cohorts with newly treated epilepsy: Glasgow cohort of 281 patients [55]; SANAD cohort of 491 patients (subset from SANAD trial [24,25])	k-NN	5 SNPs	SF or NSF at 1-year post treatment initiation	Not applicable	On 2 UK cohorts	Model unable to predict treatment response in either UK cohort
Devinsky et al. [56] (2016)	To predict probability of ASM regimen success for an individual patient	- US claims data collected for patients > 16 years - 34,990 patients in training set and 8292 patients in testing set	RF	Patient features (e.g. demographics and comorbidities) and treatment features (e.g. ASM regimen, mechanism of action, number of ASMs in regimen)	Probability of requiring treatment change from a given ASM regimen	On testing set	Not performed	AUC 0.72
Ouyang et al. [57] (2018)	To predict treatment response after ASM start/change	- 20 pediatric patients - EEGs taken before and 1–3 months after ASM start/change used for training and feature selection - EEGs taken before and 6 months after ASM start/change used for model testing	SVM	6 quantitative EEG features from before and after ASM start/change	Effective (>50% reduction in seizure frequency) or ineffective (<50% reduction in seizure frequency) response	- 10-fold cross validation on training set - On testing EEGs	Not performed	- Using testing EEGs: sensitivity 85.7%, specificity 76.9%, PPV 90.9%, NPV 66.7%, accuracy 83.3%, balanced accuracy 81.3%
Zhang et al. [58] (2018)	To predict seizure freedom with levetiracetam monotherapy	- 46 newly diagnosed patients with epilepsy given levetiracetam as initial therapy - 80/20 training/testing split	SVM	Clinical and quantitative EEG features	SF or NSF	5-fold cross validation on training set and final validation on 20% testing set	Not performed	AUC 0.96, accuracy 90%
Yao et al. [59] (2019)	To predict treatment response with ASMs in newly diagnosed epilepsy	- 287 patients with newly diagnosed epilepsy - 207 in remission (SF for at least 1 year) and 80 not (never SF for any whole year) - 80/20 training/testing set split	Decision trees, RF, SVM, logistic regression, XGBoost	Demographics, clinical features, presence of MRI or EEG abnormalities	SF or NSF	5-fold cross validation on the training set and final validation on 20% testing set	Not performed	XGBoost model best performing: AUC 0.98, sensitivity 97%, PPV 92%
DRE prediction								
Silva-Alves et al. [60] (2017)	To predict DRE or drug responsiveness	- 122 adults with mesial temporal lobe epilepsy; 38 ASM responsive and 84 with DRE - DRE defined as per ILAE; drug-responsive defined as seizure free as per ILAE [61]	RF	SNPs and clinical variables (e.g. presence of HS)	DRE or drug responsive	LOOCV	Not performed	AUC 0.82, sensitivity 95%, specificity 21%

(*continued on next page*)

Table 1 (*continued*)

Authors (year)	Aims	Population/Methods	ML models used/compared	ML Input	ML Output	Internal validation	External validation	Results
An et al. [62] (2018)	To predict DRE or non-DRE	- US claims data from pharmacies, hospitals, clinics for patients > 16 years - 292,892 patients included: 38,382 with DRE (13.1%) - Dataset split 60/20/20 for training/tuning/testing - DRE defined as failure of at least 3 ASMs; non-DRE as maintenance for ≥ 1 year on first ASM	SVM, RF, multivariate logistic regression	Demographics, comorbidities, insurance policy, treatments, encounters (e.g. hospitalizations)	DRE or non-DRE	On 20% tuning dataset and 20% testing dataset	Not performed	- RF model best performing with AUC 0.76 and best calibrated - DRE predicted mean 2 years before patients failed 2 ASMs
Delen et al. [63] (2020)	To predict DRE or non-DRE	- Pooled EMR database from US hospitals and clinics - 37,024 patients included: 806 with DRE (2.2%) - Dataset split 70/30 for training/testing - DRE defined as latest treatment being a non-ASM; non-DRE as latest treatment being ASMs	Decision trees, RF, GBT	Demographics, comorbidities, initial epilepsy diagnosis	DRE or non-DRE	On 30% testing set	Not performed	GBT model best performing: AUC 0.83, accuracy 75%
Surgical candidate identification								
Cohen et al. [64] (2016)	To identify surgical candidates using a natural language processing ML tool	- 200 pediatric patients (100 had epilepsy surgery; 100 SF on medication and so non-candidates) used for training and testing of ML models. - Multiple models generated to assess impact of varying algorithm, features, training set size, data balance on performance - Comparison with gold-standard evaluations of 2 epileptologists	NB, SVM	Combinations (depending on model generated) of unigrams, bigrams, and codes for ASMs taken from free text of EMR notes	Surgical candidate or not surgical candidate	10-fold cross validation	Not performed	- SVM model best performing with F-measure 0.82 - Performance of baseline NB model (F-measure 0.74) comparable to epileptologists' (F-measure 0.71) and able to identify surgical candidates up to 12 months sooner than epileptologists
Wissel et al. [65] (2020)	To prospectively evaluate ability of a natural language processing ML tool to identify potential surgical candidates	- 6395 adult and pediatric patients - Training set updated weekly: mean of 519 patients - Of 6395 patients, 58 randomly selected patients had prospectively assigned surgical candidacy scores compared to gold-standard assessment of an epileptologist	SVM	Unigrams, bigrams, trigrams, and codes for ASMs taken from free text of EMR notes	Surgical candidacy score	- 10-fold cross validation on training set - Prospective testing on 58 patients	Not performed	- AUC 0.79 - At optimal surgical candidacy score cutoff: sensitivity 80%, specificity 77%, PPV 25% (for scores ≥ cutoff), and NPV 98% (for scores < cutoff)
Surgery outcome prediction								
Antony et al. [66] (2013)	To predict surgical outcome in patients with TLE	- 23 adult patients with drug-resistant TLE who underwent anterior temporal lobectomy - 1-year follow-up	SVM	Interictal stereo-EEG functional connectivity data	SF or NSF	LOOCV	Not performed	Accuracy 87%
Armañanzas et al. [67] (2013)	To predict surgical outcomes in patients with TLE	- 19 patients > 16 years old with unilateral TLE and HS - Minimum 3-year follow-up	NB, logistic regression with ridge estimators, k-NN	Clinical and neuropsychological features	SF (Engel I) or NSF (Engel II or III) [27]	LOOCV	Not performed	All 3 models with equal 89.5% accuracy
Feis et al. [68] (2013)	To predict surgical outcomes in patients with TLE	- 49 patients (19 males/30 females) with unilateral left mesial TLE and HS who underwent amygdalohippocampectomy - Separate analyses for males and females - Minimum 1-year follow-up	SVM	VBM data of white matter segments from presurgical MRI	Favorable outcome (ILAE 1 or 2) or non-favorable outcome (ILAE 3–6) [69]	Nested LOOCV	Not performed	- Males: AUC 0.93, balanced accuracy 94%, sensitivity 100%, specificity 88%, PPV 92% - Females: AUC 0.95, balanced accuracy 96%, sensitivity 100%, specificity 92%, PPV 95%

Table 1 (*continued*)

Authors (year)	Aims	Population/Methods	ML models used/compared	ML Input	ML Output	Internal validation	External validation	Results
Bernhardt et al. [70] (2015)	To predict surgical outcome in patients with TLE	- 79 patients with drug-resistant TLE who underwent temporal lobe surgery - Minimum follow-up 5 years - Testing on 79 patients and on independent cohort of 27 patients from same institution	K-means to cluster patients into classes followed by LDA within classes	Surface-based mesiotemporal volume features based on pre-surgical MRI	SF (Engel I) or NSF (Engel II)	- LOOCV on 79 patients - On independent 27 patients	Not performed	- 87% accuracy - 96% accuracy for independent cohort
Memarian et al. [71] (2015)	To predict surgical outcome in patients with TLE	- 20 patients with focal drug-resistant seizures from mesial TLE who underwent anteromesial temporal lobectomy - Average follow-up 54 months	NB, LDA, SVM-rbf, SVM-mlp, least-square SVM	Clinical, demographic, pre-operative MRI, iEEG variables	SF (Engel I) or NSF (Engel $\geq$ II)	LOOCV	Not performed	Least-square SVM was best model: accuracy 95%
Munsell et al. [72] (2015)	To predict surgical outcome in patients with TLE	- 35 patients from the US and 35 patients from Germany with drug-resistant TLE and HS - US patients underwent anterior temporal lobectomy and German patients amygdalohippocampectomy - Minimum 1-year follow-up	SVM	DTI white matter connectome data from presurgical MRI	SF (Engel I) or NSF (Engel $\geq$ II)	10-fold cross validation (when US and German patients combined as training set)	On 35 German patients (when only US patients used as training set)	- 70% accuracy when US and German patients combined - 66% accuracy when US patients used as training set and German patients as testing set
Yankam Njiwa et al. [73] (2015)	To predict surgical outcome in patients with TLE	- 16 patients with mesial TLE and HS who underwent temporal lobe surgery - Minimum 1-year follow-up	RF	Voxel-wise signal intensities from pre-operative FMZ-PET or FDG-PET whole brain images	SF (Engel IA) or NSF (not Engel IA)	LOOCV	Not performed	RF model unable to distinguish between SF and NSF patients for either FMZ-PET (accuracy 40%) or FDG-PET (accuracy 36%) images
Hong et al. [74] (2016)	To predict surgical outcome in patients with frontal lobe epilepsy	- 41 FCD patients: 13 with type I and 28 with type II - Mean follow-up 3.9 years for type I and 4.9 years for type II	SVM	SBM cortical thickness and curvature variables from presurgical MRI	SF (Engel I) or NSF (Engel II-IV)	LOOCV	Not performed	Accuracy 92% for FCD type I and 82% for FCD type II
He et al. [75] (2017)	To predict surgical outcome in patients with TLE	- 56 patients with drug-resistant unilateral TLE who underwent anterior temporal lobectomy - Outcomes assessed at 1-yr follow-up	SVM	Nodal hubness measures from rs-fMRI	SF (Engel I) or NSF (Engel II-IV)	Split-sample validation, 7-fold cross validation, LOOCV compared	Not performed	Best accuracy 76.8% achieved with LOOCV
Tomlinson et al. [76] (2017)	To predict surgical outcome in pediatric patients	- 17 pediatric patients - 15 with FCD; 1 ganglioglioma; 1 cerebral vascular accident - Minimum 2-year follow-up	SVM	Interictal iEEG global synchrony and local heterogeneity data	SF (Engel I) or NSF (Engel $\geq$ II)	LOOCV	Not performed	Accuracy 94.1% (16/17 correctly classified)
Gleichgerrcht et al. [77] (2018)	To predict surgical outcome in patients with TLE	- 50 patients with unilateral mesial TLE who underwent anterior temporal lobectomy or mesial temporal laser ablation - Minimum 1-year follow-up	Deep learning neural network	Whole-brain structural connectome from pre-operative DTI	SF (Engel I) or NSF (Engel $\geq$ II)	5-fold cross validation	Not performed	PPV 88%, NPV 79%
Taylor et al. [78] (2018)	To predict surgical outcome in patients with TLE	- 53 patients with TLE who underwent anterior temporal lobe surgery - Outcomes assessed at 1-yr follow-up	SVM	Network connectivity change variables derived from pre-operative DTI and post-operative MRI	SF (ILAE 1) or NSF (ILAE 2–6)	LOOCV	Not performed	Accuracy 79.2%, sensitivity 86.1%, specificity 64.7%

(*continued on next page*)

Table 1 (*continued*)

Authors (year)	Aims	Population/Methods	ML models used/compared	ML Input	ML Output	Internal validation	External validation	Results
Larivière et al. [79] (2020)	To predict surgical outcome in TLE	- 30 patients with drug-resistant unilateral TLE and HS who underwent anterior temporal lobectomy - Minimum 1-yr follow-up	Logistic regression	Functional connectivity distance features from pre-operative T1-weighted MRI, diffusion MRI, rs-fMRI	SF (Engel I) or NSF (Engel II-IV)	5-fold cross validation	Not performed	Accuracy 76%

FCD, focal cortical dysplasia; FDG-PET, fluorodeoxyglucose positron emission tomography; FMZ-PET, flumazenil positron emission tomography; DTI, diffusion tensor imaging; EMR, electronic medical records; GBT, gradient boosted trees; HS, hippocampal sclerosis; iEEG, intracranial electroencephalography; ILAE, International League Against Epilepsy; k-NN, k nearest neighbor; LDA, linear discriminant analysis; LOOCV, leave one out cross validation; MRI, magnetic resonance imaging; NB, naïve bayes; NPV, negative predictive value; NSF, not seizure free; PET, positron emission tomography; PPV, positive predictive value; RF, random forest; rs-fMRI, resting-state functional MRI; SANAD, standard versus new antiepileptic drug trial; SBM, surface-based morphometry; SF, seizure free; SNP, single nucleotide polymorphism; SVM, support vector machine; SVM-mlp, support vector machine with multilayer perceptron kernel; SVM-rbf, support vector machine with radial basis function kernel; VBM, voxel-based morphometry; XGBoost, eXtreme Gradient Boosting.

what is the probability the patient will require a treatment change within 12 months? The model enabled researchers to input an ASM regimen (individual ASMs or combinations) and employed a suite of other patient and treatment features as inputs; the ASM regimen which resulted in the smallest probability of requiring treatment change, served as the model-determined optimal ASM regimen for the given patient. The model achieved an AUC of 0.72. The study also examined the health economic impact of using the model-determined optimal ASM regimen compared to the regimen actually used, and ascertained that for 8292 test set patients there would be an average of 281.5 fewer annual epilepsy-related hospitalizations. The study used large training and test sets of 34,990 and 8292 patients, respectively, and an initial set of 5000 input features. A procedure to remove features to avoid overfitting was also used. However, the researchers did not specify how many input features were ultimately used. It is suggested that there should be at least 10 events per predictor to avoid the model overfitting the training data [17,33,34]. Recent research has suggested, however, that random forests algorithms may require large datasets with potentially >200 events per predictor to avoid large discrepancies between AUCs of training and test data and so reduce the chances of model overfitting [80]. The absence of reporting of feature numbers limits the extent of critical appraisal.

Another study [57] used an SVM model to classify 20 children with epilepsy into ASM effective and ineffective groups using differences in quantitative EEG features before and after an ASM start or change as predictors, achieving an accuracy of 83.3%. The study, however, used the same 20 patients for training and testing so the model performance may be overly optimistic. Finally, a study [58] used an SVM classifier to predict seizure freedom with levetiracetam monotherapy. The SVM classifier used both clinical and quantitative EEG variables as input features and outputted whether a patient would be SF or NSF with levetiracetam, achieving 90% accuracy on test data. Additionally, the model did not appear to be overfit to the training data (AUC on training data 0.95; AUC on test data 0.96), in spite of the fact that only 36 patients were used for training, 10 for testing, and that 15 input features were used. Given the promising results, external validation in a large prospective study would be useful. However, the use of quantitative EEG features for model training [57,58] would be a barrier to widespread use.

4.2. DRE prediction

Three studies used ML to predict DRE [60,62,63]. One of the models achieved an AUC of 0.82, a sensitivity of 95%, and a specificity of 21% using clinical variables and single nucleotide polymorphisms from 11 genes involved in drug transport and metabolism [60]. The high sensitivity and low specificity of the model suggests it may perform better in ruling out drug-resistance than ruling it in, potentially supporting its use as a screening tool for DRE, rather than as a diagnostic tool. However, only 122 patients were included for model training and testing, with 57 input features used as outcome predictors. Hence, the AUC found for this model may be overly optimistic.

Another study [62] used methods similar to those of Devinsky et al. [56] but instead compared the performance of 3 algorithms to predict drug-resistance on the date of the first ASM prescription. The best performing model was a random forests model incorporating 635 features which achieved an AUC of 0.76. A potential limitation of the study is that DRE was defined as the failure of 3 ASMs to enable better separation in the data between patients with DRE and without DRE; it may be that the model would perform less well if the conventional definition of DRE were used [61]. A strength of the study is that it reported model calibration, the only one to do so among all studies included in this review. Calibration is an essential measure of accuracy and is often underreported in clinical prediction modeling [81]. Calibration assesses the degree to which model predictions match observed outcomes in the dataset, and is best seen through a calibration plot [17,33,82]. A well-calibrated model ensures a realistic model to support patient-counseling and clinical decision-making [82]. The model was well-calibrated to the dataset, but the utility of this is uncertain because of the definition of DRE employed and that the resultant frequency of DRE (13.1%) arising from this definition diverges from reality, in which the proportion of DRE is closer to one-third [5]. The model [62] did, however, display potential clinical utility in being able to predict the failure of 2 ASMs an average of 2 years sooner than failure would have occurred in reality, an important result given the higher mortality rate associated with DRE [83].

Delen et al. [63] performed a study to identify DRE at the time of initial epilepsy diagnosis using longitudinal electronic medical records data of 37,024 patients. Their best performing model achieved an AUC of 0.83. However, similar to other database studies [56,62], seizure control status was not directly ascertained from clinical assessment. Instead, the definition of DRE was based on the use of "non-ASM" treatments (e.g. surgery), while patients taking ASMs were assumed to have "non-refractory" epilepsy. This does not conform with the recognized definition of DRE, nor reflect the reality that many patients with DRE may not have had non-ASM treatment and remain on ASMs [61,84]. This might explain why only 806 (2.2%) patients were classified as having refractory epilepsy, which is likely a gross underestimation.

4.3. Surgical candidacy identification

Two studies [64,65] used ML and natural language processing to aid in identifying surgical candidates. A Naïve Bayes baseline model in Cohen et al. [64], using unigrams from the electronic medical record only, was able to perform comparably to 2 epileptologists in identifying surgical candidates (F-measures 0.74 and 0.71, respectively). Model performance was similar to that of the epileptologists up to 12 months prior to actual surgical referral, leading to a referral for surgery earlier than the epileptologists would have provided. The highest F-measures occurred when unigrams, bigrams, and ASM names were used as features with an SVM model, resulting in an F-measure of 0.82. Confidence in the study, however, is tempered by the wide 95% confidence intervals around reported F-measure values.

Rather than a definite surgical candidacy classification, Wissel et al. [65] used an SVM model to prospectively generate a surgical candidacy score (higher scores indicated a higher likelihood of being a surgical candidate) based on electronic medical record notes. Model performance was evaluated by comparing the model's classification to that of an epileptologist, for a defined cutoff score. The model achieved an AUC of 0.79 (95% confidence interval 0.62–0.96), with a sensitivity of 0.8 and specificity of 0.77 at the optimal cutoff score. As pointed out by the authors, because the model produces a score it would be up to the clinician to define a threshold at which to separate potential candidates from non-candidates. Although this system may alert clinicians about potential surgical candidates, it is unclear whether this would translate to increased numbers of candidates undergoing surgery. It has been suggested previously that using tools to identify surgical candidates may lead primary care physicians and general neurologists to falsely identify some patients with DRE as not being surgical candidates [85]. This is a criticism that would apply to the use of clinical decision support tools in general, however, which should be used to assist clinical decision-making, not replace it. Furthermore, surgical candidacy is sometimes not agreed upon even by epileptologists [15]. In Cohen et al. [64] the interrater agreement between epileptologists is only moderate (Cohen's kappa 0.63). In Wissel et al. [65] interrater agreement between 2 epileptologists was assessed for scores above a high cutoff score, chosen specifically to maximize the PPV, and here too, Cohen's kappa was only 0.61. This indicates that experts themselves may sometimes not agree on surgical candidacy and potentially miss out on identifying eligible surgical candidates. Ultimately, studies evaluating their effects on patient outcomes are needed to resolve the issues surrounding the benefit of these tools.

4.4. Prediction of surgical outcomes

Thirteen studies used a variety of methodologies to predict epilepsy surgery outcomes (SF or NSF); 1 relying on clinical and neuropsychological inputs [67], 9 on imaging data [68,70,72–75,77–79], 2 on intracranial electroencephalography (iEEG) [66,76], and 1 on a variety of these [71]. One study [67] included the neuropsychological features personality style and performance intelligence quotient, alongside the side of seizure onset, and found that these measures contributed most to model performance. This finding highlights the importance of a neuropsychological evaluation in presurgical workup.

Feis et al. [68] used voxel-based morphometry (VBM) white matter segment data with an SVM model to achieve a balanced accuracy of 94% for males and 96% for females, but the model was not externally validated.

Bernhardt et al. [70] used K-means unsupervised machine learning to cluster patients with TLE into 4 distinct classes based on surface-based differences in volume changes of mesiotemporal lobe structures compared to controls. A linear discriminant analysis supervised ML model was then used to predict surgical outcome within classes, achieving 87% accuracy. The model notably was further tested on an independent but similar cohort of patients from the same institution scanned on a different MRI scanner, achieving an accuracy of 96%. The classes into which patients were clustered displayed different rates of seizure freedom, suggesting the ability of unsupervised learning to identify anatomical subtypes of TLE that may assist in outcome prediction and patient counseling.

Hong et al. [74] used surface-based morphometry cortical thickness and curvature features with an SVM model to predict outcome in focal cortical dysplasia type I and II, achieving 92% accuracy for type I and 82% accuracy for type II.

A number of studies looked at the ability of network connectivity measures based on imaging data to predict surgical outcomes. Munsell et al. [72] used diffusion tensor imaging (DTI) white matter connectome data with an SVM model and achieved an accuracy of 70% that dropped only to 66% accuracy at external validation, indicating it was fairly generalizable.

Gleichgerrcht et al. [77] used whole brain structural connectome data from DTI in a deep learning neural network model and was the only study in this review to use deep learning methods. In typical deep learning approaches the algorithm automatically extracts features from the input data during the learning process. Such an approach was not followed in this study. Instead, to reduce computational load, the authors used methods to extract connectivity features relevant to outcomes prior to neural network training. This methodology appeared to be superior to using a neural network design alone (PPV 88% vs 65% and NPV 79% vs 51%). To address the risk of overfitting, given the small sample size and high dimension input layer, the study authors used dropout layers in their neural network design. Here network nodes are randomly dropped during training, a procedure which has been shown to reduce model overfitting and improve generalizability [86].

Taylor et al. [78] used changes in network connectivity as a result of surgery to predict outcomes using an SVM model. They achieved an accuracy of 79.2%. While the model requires change in network connectivity inputs, which depend on knowledge of how surgery has impacted network connectivity, the model could still be used for presurgical prediction by simulating the resection to be performed and from this calculating the relevant inputs [78].

Larivière et al. [79] used functional connectivity distance features from preoperative resting-state functional MRI with a logistic regression model, predicting surgical outcomes with 76% accuracy.

He et al. [75] used an SVM model to predict surgical outcomes in TLE using nodal hubness features derived from resting-state functional MRI. Nodal hubness provides indication of the importance of a node to network function [87]. The authors found increases in nodal hubness in bilateral thalami in NSF patients, with increased connectivity to brain regions contralateral to the side of seizure onset, a finding which may help explain postsurgical seizure recurrence in these patients. Their model achieved an accuracy of 76%.

In contrast to these models, the study by Yankam Njima et al. [73] was unable to distinguish between SF and NSF patients using positron emission tomography imaging.

Memarian et al. [71] used demographic, clinical, MRI, and iEEG features with a least-square SVM algorithm, and achieved an accuracy of 95% in predicting surgical outcome.

Antony et al. [66] used iEEG metrics to predict surgical outcome for adult patients with drug-resistant TLE, while Tomlinson et al. [76] used other iEEG metrics to predict surgical outcome in a pediatric population predominantly with focal cortical dysplasia. Both

studies were able to achieve high accuracies and show promise in the potential use of EEG data to predict surgical outcomes. However, one major limitation is restricted access to iEEG to only a few specialized epilepsy centers across the globe.

4.5. Limitations and future directions

This review has highlighted the emerging use of machine learning predictive models in epilepsy management decision-making. At the moment there exist a number of barriers to the implementation of these models in practice. The first is that the field is in its nascency. We identified only 24 studies and most were in single settings (18 of 24) with small sample sizes (median number of patients considering total of training and testing sets for each study was 55). Increased sharing of research databases and larger multi-institution collaborations could help to resolve some of these limitations.

Another limitation is the lack of external validation, which was performed by only 3 studies [53,54,72]. External validation is essential if models are to be broadly implemented. Use of quantitative EEG variables, advanced imaging modalities, and individual genetic profiles for model training may prevent widespread use of the models because of limited accessibility of these investigations. Although this review did not specifically assess the quality of reporting of ML models, there is also a lack of standardization in reporting. The number of model input features was not always specified and only one study reported model calibration [62]. Furthermore, confidence or credible intervals around performance metrics were mostly not reported, with only 5 papers including them [54,62,64,65,68]. Reporting standards will need to be improved if models are to be reproduced and validated. Efforts are already underway to standardize reporting for ML prediction models [88–90]. The development of a checklist to enable consistency in assessing adherence to future guidelines might also be useful [91].

The types of ML models studied are also quite limited. Only one study used deep learning methods [77], which are becoming increasingly prominent in neurological research [40].

Crucially, clinical utility of these models in improving patient outcomes has not been demonstrated. Prospective studies, ideally using a randomized controlled trial design, are needed to assess the clinical and health economic benefits in the adoption of these models.

5. Conclusion

There have been emerging studies in using ML that show potential in being able to support clinicians in making management decisions in epilepsy. However, studies are generally limited by small sample sizes and a lack of external validation. Larger scale collaborative research, standardization and transparency in reporting, and prospective evidence of improvement in outcomes will need to take place before ML models can be incorporated into daily workflows, and lead to improvements in the lives of people with epilepsy.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

Zongyuan Ge is supported by NVIDIA and Airdoc Research grant. His institution has received research grants from Australian Research Data Commons (ARDC). Zhibin Chen is supported by the NHMRC Early Career Fellowship and has received research grant from University of Melbourne Early Career Researcher Grant Scheme. His institution has received research funding from UCB Pharma. This funding is unrelated to this study. Patrick Kwan is supported by a Medical Research Future Fund Practitioner Fellowship (MRF1136427). His institution has received research grants from Biscayne Pharmaceuticals, Eisai, GW Pharmaceuticals, LivaNova, Novartis, UCB Pharma, and Zynerba outside the submitted work; he has received speaker fees from Eisai, LivaNova, and UCB Pharma outside the submitted work. Eliot David Smolyansky and Haris Hakeem have no interests to declare.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.yebeh.2021.108273.

References

[1] Rho JM, White HS. Brief history of anti-seizure drug development. Epilepsia open 2018;3(S2):114–9.
[2] Perucca E, Tomson T. The pharmacological treatment of epilepsy in adults. Lancet Neurol 2011;10(5):446–56.
[3] Chen Z, Brodie MJ, Kwan P. What has been the impact of new drug treatments on epilepsy? Current Opinion in Neurology 2020;33.
[4] Schmidt D, Schachter SC. Drug treatment of epilepsy in adults. BMJ: Br Med J 2014;348.
[5] Chen Z, Brodie MJ, Liew D, Kwan P. Treatment outcomes in patients with newly diagnosed epilepsy treated with established and new antiepileptic drugs: a 30-year longitudinal cohort study. JAMA Neurol 2018;75(3):279. https://doi.org/10.1001/jamaneurol.2017.3949.
[6] Vaughan KA, Ramos CL, Buch VP, Mekary RA, Amundson JR, Shah M, et al., An estimation of global volume of surgically treatable epilepsy based on a systematic review and meta-analysis of epilepsy. 2018;130: 1127.
[7] Bjellvi J, Olsson I, Malmgren K, Wilbe Ramsay K. Epilepsy duration and seizure outcome in epilepsy surgery. A systematic review and meta-analysis 2019;93: e159-e166.
[8] Martínez-Juárez IE, Funes B, Moreno-Castellanos JC, Bribiesca-Contreras E, Martínez-Bustos V, Zertuche-Ortuño L, et al. A comparison of waiting times for assessment and epilepsy surgery between a Canadian and a Mexican referral center. Epilepsia Open 2017;2(4):453–8.
[9] Solli E, Colwell NA, Say I, Houston R, Johal AS, Pak J, et al. Deciphering the surgical treatment gap for drug-resistant epilepsy (DRE): a literature review. Epilepsia 2020;61(7):1352–64.
[10] Berg AT, Langfitt J, Shinnar S, Vickrey BG, Sperling MR, Walczak T, et al. How long does it take for partial epilepsy to become intractable? Neurology 2003;60(2):186–90.
[11] Choi H, Carlino R, Heiman G, Hauser WA, Gilliam FG. Evaluation of duration of epilepsy prior to temporal lobe epilepsy surgery during the past two decades. Epilepsy Res 2009;86(2-3):224–7.
[12] Brodie MJ, Barry SJE, Bamagous GA, Norrie JD, Kwan P. Patterns of treatment response in newly diagnosed epilepsy. Neurology 2012;78(20):1548–54.
[13] Engel J. What can we do for people with drug-resistant epilepsy? The 2016 Wartenberg Lecture. Neurology 2016;87(23):2483–9.
[14] Hrazdil C, Roberts JI, Wiebe S, Sauro K, Vautour M, Hanson A, et al. Patient perceptions and barriers to epilepsy surgery: evaluation in a large health region. Epilepsy Behav 2013;28(1):52–65.
[15] Steinbrenner M, Kowski AB, Holtkamp M. Referral to evaluation for epilepsy surgery: reluctance by epileptologists and patients. Epilepsia 2019;60 (2):211–9.
[16] Middleton B, Sittig DF, Wright A. Clinical Decision Support: a 25 Year Retrospective and a 25 Year Vision. Yearbook of Medical Informatics 2016;25(Suppl 1):S103–16.
[17] Kubben P, Dumontier M, Dekker A, editors. Fundamentals of clinical data science. Cham: Springer International Publishing; 2019.
[18] Osheroff JA, Teich JM, Middleton B, Steen EB, Wright A, Detmer DE. A roadmap for national action on clinical decision support. J Am Med Informatics Assoc: JAMIA 2007;14(2):141–5.
[19] Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. J Big Data 2019;6:54.
[20] Shilo S, Rossman H, Segal E. Axes of a revolution: challenges and promises of big data in healthcare. Nat Med 2020;26(1):29–38.
[21] Yang S, Han X, Wang Na, Gu R, Chen W, Wang E, et al. Predicting seizure freedom with AED treatment in newly diagnosed patients with MRI-negative epilepsy: A large cohort and multicenter study. Epilepsy Behav 2020;106:107022. https://doi.org/10.1016/j.yebeh.2020.107022.
[22] Huang L, Li S, He D, Bao W, Li L. A predictive risk model for medical intractability in epilepsy. Epilepsy Behav 2014;37:282–6.

[23] Hughes DM, Bonnett LJ, Czanner G, Komárek A, Marson AG, García-Fiñana M. Identification of patients who will not achieve seizure remission within 5 years on AEDs. Neurology 2018;91(22):e2035–44.
[24] Marson AG, Al-Kharusi AM, Alwaidh M, Appleton R, Baker GA, Chadwick DW, et al., The SANAD study of effectiveness of valproate, lamotrigine, or topiramate for generalised and unclassifiable epilepsy: an unblinded randomised controlled trial. Lancet (London, England) 2007;369: 1016-1026.
[25] Marson AG, Al-Kharusi AM, Alwaidh M, Appleton R, Baker GA, Chadwick DW, et al.. The SANAD study of effectiveness of carbamazepine, gabapentin, lamotrigine, oxcarbazepine, or topiramate for treatment of partial epilepsy: an unblinded randomised controlled trial. Lancet (London, England) 2007;369: 1000-1015.
[26] Jehi L, Yardi R, Chagin K, Tassi L, Russo GL, Worrell G, et al. Development and validation of nomograms to provide individualised predictions of seizure outcomes after epilepsy surgery: a retrospective analysis. Lancet Neurol 2015;14(3):283–90.
[27] Engel Jr J. Outcome with respect to epileptic seizures. Surgical treatment of the epilepsies 1993: 609-621.
[28] Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319(13):1317. https://doi.org/10.1001/jama.2017.18391.
[29] Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12–22.
[30] Breiman L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statist Sci 2001;16:199–231.
[31] Deo RC. Machine learning in medicine. Circulation 2015;132(20):1920–30.
[32] Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Method 2019;19:64.
[33] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. 2nd ed. Springer International Publishing; 2019.
[34] Liu Y, Chen P-H, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. JAMA 2019;322(18):1806. https://doi.org/10.1001/jama.2019.16489.
[35] Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. npj Digital Med 2018;1:39.
[36] Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. npj Digital Med 2020;3:118.
[37] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380(14):1347–58.
[38] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130:515–24.
[39] Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, et al. Deep learning: a primer for radiologists. RadioGraphics 2017;37(7):2113–31.
[40] Valliani A-A, Ranti D, Oermann EK. Deep learning and neurology: a systematic review. Neurol Therapy 2019;8(2):351–65.
[41] Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science 2015;349(6245):255–60.
[42] Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19(1):221–48.
[43] Siddiqui MK, Morales-Menendez R, Huang X, Hussain N. A review of epileptic seizure detection using machine learning classifiers. Brain informatics 2020;7: 5-5.
[44] Karayiannis NB, Xiong Y, Tao G, Frost JD, Wise MS, Hrachovy RA, et al. Automated detection of videotaped neonatal seizures of epileptic origin. Epilepsia 2006;47(6):966–80.
[45] Jin B, Krishnan B, Adler S, Wagstyl K, Hu W, Jones S, et al. Automated detection of focal cortical dysplasia type II with surface-based magnetic resonance imaging postprocessing and machine learning. Epilepsia 2018;59:982–92.
[46] Chen S, Zhang J, Ruan X, Deng K, Zhang J, Zou D, et al. Voxel-based morphometry analysis and machine learning based classification in pediatric mesial temporal lobe epilepsy with hippocampal sclerosis. Brain Imaging and Behavior 2020;14(5):1945–54.
[47] Zhou B, An D, Xiao F, Niu R, Li W, Li W, et al. Machine learning for detecting mesial temporal lobe epilepsy by structural and functional neuroimaging. Front Med 2020;14(5):630–41.
[48] Zsom A, LaFrance WC, Blum AS, Li P, Wahed LA, Shaikh MA, et al. Ictal autonomic activity recorded via wearable-sensors plus machine learning can discriminate epileptic and psychogenic nonepileptic seizures. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). p. 3502–6.
[49] Abbasi B, Goldenholz DM. Machine learning applications in epilepsy. Epilepsia 2019;60(10):2037–47.
[50] Si Y. Machine learning applications for electroencephalograph signals in epilepsy: a quick review. Acta Epileptol 2020;2:5.
[51] Yang S, Wang B, Han X. Models for predicting treatment efficacy of antiepileptic drugs and prognosis of treatment withdrawal in epilepsy patients. Acta Epileptol 2021;3:1.
[52] Covidence systematic review software. In. Melbourne, Australia: Veritas Health Innovation.
[53] Petrovski S, Szoeke C, Sheffield L, D'Souza W, Huggins R, O'Brien T. Multi-SNP pharmacogenomic classifier is superior to single-SNP models for predicting drug outcome in complex diseases. Pharmacogenetics and genomics 2009;19: 147-152.
[54] Shazadi K, Petrovski S, Roten A, Miller H, Huggins RM, Brodie MJ, et al. Validation of a multigenic model to predict seizure control in newly treated epilepsy. Epilepsy Res 2014;108(10):1797–805.
[55] Szoeke C, Sills GJ, Kwan P, Petrovski S, Newton M, Hitiris N, et al., Multidrug-resistant genotype (ABCB1) and seizure recurrence in newly treated epilepsy: Data from international pharmacogenetic cohorts. Epilepsia 2009;50: 1689-1696.
[56] Devinsky O, Dilley C, Ozery-Flato M, Aharonov R, Goldschmidt Y, Rosen-Zvi M, et al. Changing the approach to treatment choice in epilepsy using big data. Epilepsy Behav 2016;56:32–7.
[57] Ouyang C-S, Chiang C-T, Yang R-C, Wu R-C, Wu H-C, Lin L-C. Quantitative EEG findings and response to treatment with antiepileptic medications in children with epilepsy. Brain and Development 2018;40(1):26–35.
[58] Zhang J-H, Han X, Zhao H-W, Zhao Di, Wang Na, Zhao T, et al. Personalized prediction model for seizure-free epilepsy with levetiracetam therapy: a retrospective data analysis using support vector machine. Br J Clin Pharmacol 2018;84(11):2615–24.
[59] Yao L, Cai M, Chen Y, Shen C, Shi L, Guo Yi. Prediction of antiepileptic drug treatment outcomes of patients with newly diagnosed epilepsy by machine learning. Epilepsy Behav 2019;96:92–7.
[60] Silva-Alves MS, Secolin R, Carvalho BS, Yasuda CL, Bilevicius E, Alvim MKM, et al. A prediction algorithm for drug response in patients with mesial temporal lobe epilepsy based on clinical and genetic information. PLoS ONE [Electronic Resource] 2017;12(1):e0169214.
[61] Kwan P, Arzimanoglou A, Berg AT, Brodie MJ, Allen Hauser W, Mathern G, et al., Definition of drug resistant epilepsy: Consensus proposal by the ad hoc Task Force of the ILAE Commission on Therapeutic Strategies. Epilepsia 2010;51: 1069-1077.
[62] An S, Malhotra K, Dilley C, Han-Burgess E, Valdez JN, Robertson J, et al. Predicting drug-resistant epilepsy – a machine learning approach based on administrative claims data. Epilepsy Behav 2018;89:118–25.
[63] Delen D, Davazdahemami B, Eryarsoy E, Tomak L, Valluru A. Using predictive analytics to identify drug-resistant epilepsy patients. Health Informatics J 2020;26(1):449–60.
[64] Cohen KB, Glass B, Greiner HM, Holland-Bouley K, Standridge S, Arya R, et al. Methodological issues in predicting pediatric epilepsy surgery candidates through natural language processing and machine learning. Biomed Informatics Insights 2016;8:BII.S38308. https://doi.org/10.4137/BII.S38308.
[65] Wissel BD, Greiner HM, Glauser TA, Holland-Bouley KD, Mangano FT, Santel D, et al. Prospective validation of a machine learning model that uses provider notes to identify candidates for resective epilepsy surgery. Epilepsia 2020;61 (1):39–48.
[66] Antony AR, Alexopoulos AV, González-Martínez JA, Mosher JC, Jehi L, Burgess RC, et al. Functional connectivity estimated from intracranial EEG predicts surgical outcome in intractable temporal lobe epilepsy. PLoS ONE 2013;8(10): e77916.
[67] Armañanzas R, Alonso-Nanclares L, DeFelipe-Oroquieta J, Kastanauskaite A, de Sola RG, DeFelipe J, et al. Machine learning approach for the outcome prediction of temporal lobe epilepsy surgery. PLoS ONE [Electronic Resource] 2013;8(4).
[68] Feis D-L, Schoene-Bake J-C, Elger C, Wagner J, Tittgemeyer M, Weber B. Prediction of post-surgical seizure outcome in left mesial temporal lobe epilepsy. NeuroImage. Clin 2013;2:903–11.
[69] Wieser HG, Blume WT, Fish D, Goldensohn E, Hufnagel A, King D, et al., ILAE Commission Report. Proposal for a new classification of outcome with respect to epileptic seizures following epilepsy surgery. Epilepsia 2001;42: 282-6.
[70] Bernhardt BC, Hong S-J, Bernasconi A, Bernasconi N. Magnetic resonance imaging pattern learning in temporal lobe epilepsy: classification and prognostics. Ann Neurol 2015;77(3):436–46.
[71] Memarian N, Kim S, Dewar S, Engel J, Staba RJ. Multimodal data and machine learning for surgery outcome prediction in complicated cases of mesial temporal lobe epilepsy. Comput Biol Med 2015;64:67–78.
[72] Munsell BC, Wee C-Y, Keller SS, Weber B, Elger C, da Silva LAT, et al. Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data. Neuroimage 2015;118:219–30.
[73] Yankam Njiwa J, Gray KR, Costes N, Mauguiere F, Ryvlin P, Hammers A. Advanced [18F]FDG and [11C]flumazenil PET analysis for individual outcome prediction after temporal lobe epilepsy surgery for hippocampal sclerosis. NeuroImage: Clin 2015;7:122–31.
[74] Hong S-J, Bernhardt BC, Schrader DS, Bernasconi N, Bernasconi A. Whole-brain MRI phenotyping in dysplasia-related frontal lobe epilepsy. Neurology 2016;86(7):643–50.
[75] He X, Doucet GE, Pustina D, Sperling MR, Sharan AD, Tracy JI. Presurgical thalamic "hubness" predicts surgical outcome in temporal lobe epilepsy. Neurology 2017;88(24):2285–93.
[76] Tomlinson SB, Porter BE, Marsh ED. Interictal network synchrony and local heterogeneity predict epilepsy surgery outcome among pediatric patients. Epilepsia 2017;58(3):402–11.
[77] Gleichgerrcht E, Munsell B, Bhatia S, Vandergrift WA, Rorden C, McDonald C, et al. Deep learning applied to whole-brain connectome to determine seizure control after epilepsy surgery. Epilepsia 2018;59(9):1643–54.

[78] Taylor PN, Sinha N, Wang Y, Vos SB, de Tisi J, Miserocchi A, et al. The impact of epilepsy surgery on the structural connectome and its relation to outcome. NeuroImage. Clin 2018;18:202–14.
[79] Larivière S, Weng Y, Vos de Wael R, Royer J, Frauscher B, Wang Z, et al. Functional connectome contractions in temporal lobe epilepsy: Microstructural underpinnings and predictors of surgical outcome. Epilepsia 2020;61(6):1221–33.
[80] van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Method 2014;14:137.
[81] Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Method 2014;14(1). https://doi.org/10.1186/1471-2288-14-40.
[82] Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 2016;74:167–76.
[83] Burneo JG, Shariff SZ, Liu K, Leonard S, Saposnik G, Garg AX. Disparities in surgery among patients with intractable epilepsy in a universal health system. Neurology 2016;86(1):72–8.
[84] Dalic L, Cook MJ. Managing drug-resistant epilepsy: challenges and solutions. Neuropsychiatr Dis Treat 2016;12:2605–16.
[85] Engel Jr J. The current place of epilepsy surgery. Curr Opin Neurol 2018;31:192–7.
[86] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 2014;15:1929–58.
[87] van den Heuvel MP, Sporns O. Network hubs in the human brain. Trends Cogn Sci 2013;17(12):683–96.
[88] Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. The Lancet 2019;393(10181):1577–9.
[89] Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016;18(12):e323. https://doi.org/10.2196/jmir.5870.
[90] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162(1):W1. https://doi.org/10.7326/M14-0698.
[91] Heus P, Damen JAAG, Pajouheshnia R, Scholten RJPM, Reitsma JB, Collins GS, et al. Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies. BMJ Open 2019;9(4):e025611. https://doi.org/10.1136/bmjopen-2018-025611.