RESEARCH ARTICLE

Evaluating the generalisability of region-naïve machine learning algorithms for the identification of epilepsy in low-resource settings

Ioana Duta[1,2], Symon M. Kariuki[3,4,5], Anthony K. Ngugi[4,6], Angelina Kakooza Mwesige[7], Honorati Masanja[8], Daniel M. Mwanga[6,9], Seth Owusu-Agyei[10,11], Ryan Wagner[12], J Helen Cross[13], Josemir W. Sander[14,15,16], Charles R. Newton[1,3,4,17], Arjune Sen[1], Gabriel Davis Jones[1,2,18] *

1 Oxford Epilepsy Research Group, Nuffield Department of Clinical Neurosciences, John Radcliffe Hospital, Oxford, United Kingdom, 2 Oxford Digital Health Labs, Nuffield Department of Women's and Reproductive Health, The University of Oxford, John Radcliffe Hospital, Oxford, United Kingdom, 3 KEMRI/Wellcome Trust Research Programme, Centre for Geographic Medicine Research–Coast, Kilifi, Kenya, 4 Studies of Epidemiology of Epilepsy in Demographic Surveillance Systems (SEEDS)–INDEPTH Network, Accra, Ghana, 5 Department of Public Health, Pwani University, Kilifi, Kenya, 6 Department of Population Health, Aga Khan University, Nairobi, Kenya, 7 Department of Paediatrics and Child Health, Makerere University College of Health Sciences, Kampala, Uganda, 8 Ifakara Health Institute, Ifakara, Tanzania, 9 Department of Mathematics, University of Nairobi, Nairobi, Kenya, 10 Kintampo Health Research Centre, Kintampo, Ghana, 11 Institute of Health Research, University of Health and Allied Sciences, Ho. Ghana, 12 MRC/Wits Rural Public Health & Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa, 13 Developmental Neurosciences, University College London NIHR BRC Great Ormond Street Institute of Child Health, London, United Kingdom, 14 Department of Clinical & Experimental Epilepsy, UCL Queen Square Institute of Neurology, London, & Chalfont Centre for Epilepsy, Chalfont St Peter, United Kingdom, 15 Stichting Epilepsie Instellingen Nederland, Heemstede, Netherlands, 16 Department of Neurology, West China Hospital, Sichuan University, Chengdu, China, 17 Department of Psychiatry, University of Oxford, Oxford, United Kingdom, 18 The Alan Turing Institute, London, United Kingdom

* gabriel.jones@wrh.ox.ac.uk

Abstract

Objectives

Approximately 80% of people with epilepsy live in low- and middle-income countries (LMICs), where limited resources and stigma hinder accurate diagnosis and treatment. Clinical machine learning models have demonstrated substantial promise in supporting the diagnostic process in LMICs by aiding in preliminary screening and detection of possible epilepsy cases without relying on specialised or trained personnel. How well these models generalise to naïve regions is, however, underexplored. Here, we use a novel approach to assess the suitability and applicability of such clinical tools to aid screening and diagnosis of active convulsive epilepsy in settings beyond their original training contexts.

Methods

We sourced data from the Study of Epidemiology of Epilepsy in Demographic Sites dataset, which includes demographic information and clinical variables related to diagnosing

epilepsy across five sub-Saharan African sites. For each site, we developed a region-specific (single-site) predictive model for epilepsy and assessed its performance at other sites. We then iteratively added sites to a multi-site model and evaluated model performance on the omitted regions. Model performances and parameters were then compared across every permutation of sites. We used a leave-one-site-out cross-validation analysis to assess the impact of incorporating individual site data in the model.

Results

Single-site clinical models performed well within their own regions, but generally worse when evaluated in other regions ($p < 0.05$). Model weights and optimal thresholds varied markedly across sites. When the models were trained using data from an increasing number of sites, mean internal performance decreased while external performance improved.

Conclusions

Clinical models for epilepsy diagnosis in LMICs demonstrate characteristic traits of ML models, such as limited generalisability and a trade-off between internal and external performance. The relationship between predictors and model outcomes also varies across sites, suggesting the need to update specific model aspects with local data before broader implementation. Variations are likely to be particular to the cultural context of diagnosis. We recommend developing models adapted to the cultures and contexts of their intended deployment and caution against deploying region- and culture-naïve models without thorough prior evaluation.

Author summary

Epilepsy disproportionately affects people in low to middle income countries (LMICs). Socioeconmic disadvantages make it hard to access diagnosis and treatment as these rely on resources, personnel and time. Machine learning models may be able to provide cheap, accessible options for diagnosis and screening. Our previous work has demonstrated that such models can perform well. It is, however, crucial that tools are robust, safe, and responsibly deployed, especially in LMICs where poor models may more easily result in adverse impacts.

Models must be trained on data, which are necessarily sourced from certain regions. Models show reduced performance when applied in regions that were not included in the training data. They may also have different optimal parameters, including the thresholds used to determine whether someone is a positive case.

There can also be a trade-off, whereby a model that performs better in regions not included in the training data may perform worse in regions that were, but could make the model more robust.

We recommend applying models in target regions and updating them as necessary. We also caution against generic deployment of models developed and tested in one region into a new area without careful, thorough testing and authentication.

Introduction

Epilepsy is a common neurological condition that disproportionately affects people from disadvantaged socio-economic groups. Studies estimate that up to 75 million people have epilepsy worldwide, with approximately 80% in low- and middle-income countries (LMICs) [1]. Epilepsy accounts for over 13 million disability-adjusted life years annually and over 0.5% of the global burden of disease. While an estimated 70% of people with epilepsy could live seizure-free with anti-seizure medications, over 75% of those living with epilepsy in LMICs cannot obtain a timely and appropriate diagnosis or any treatment [2,3].

The diagnosis of epilepsy requires training, skilled personnel, time and additional resources that are scarce in LMICs, for example access to specialised equipment such as electroencephalograms (EEGs) [4]. A trained clinician's expertise is irreplaceable, but training costs and retention of skilled personnel can impede access to diagnosis in LMICs [4,5]. In such settings, diagnostic tools that require less expertise, experience or specialist training could empower primary healthcare workers to triage and prioritise people who may have epilepsy for referral [6].

Clinical machine learning (ML) models offer a practical solution to aid epilepsy diagnosis in low-resource settings. Such models have demonstrated promising outcomes for epilepsy diagnosis [7] and treatment [8]. Thus, their application in LMICs could reduce the diagnostic and treatment gaps [9]. Given the potential impact of such models, their relevance, robustness and appropriate deployment are crucial. Deployment could be built around the existing infrastructure of personal mobile devices [10,11]

ML models developed on data from one region are not inherently reliable for use elsewhere without prior validation [12]. The measure of model robustness in novel settings is termed 'generalisability'. Failure of a model to generalise sufficiently to a novel setting can be due to, for example, differences in clinical phenomenology or individual self-reporting [13] across the regions. This is particularly applicable in epilepsy, where clinical diagnosis is primarily based on self-reported history and can be nuanced [14]. This phenomenon can also be attributed to the model 'overfitting', where the model learns to make predictions based on biases in the dataset rather than developing a robust method for delineating between the desired diagnoses of interest (or absence thereof) [15]. In the case of diagnostic tools, this can have substantial consequences for the population on whom the model is deployed, resulting in missed cases, over-diagnosis, wasted resources and even mistrust of the technology [15]. The issue of generalisability has been well described in other medical domains, including variability in disease prevalence, differences in treatment protocols, and disparities in healthcare infrastructure, which can all affect how well models developed in one setting perform in another [15]. These factors and the differences in clinical phenomenology and individual self-reporting underscore the need to validate models carefully before they are applied in novel contexts.

We have previously developed a predictive model to support epilepsy diagnosis in LMICs [7], trained on data from five sub-Saharan African regions. In this study we assess the performance of such diagnostic models in novel settings with the aim of evaluating the need for comprehensive cross-regional validation with such models. We investigate the suitability of diagnostic models for deployment in regions which do not contribute data to the models' training. We also consider generalised seizures [14] separately due to their relatively homogeneous presentation, to investigate how incorporating seizure subtype in a model's training influences the model's performance both on sites that were included in training and those that were not.

Methods

Data acquisition, study design and pre-processing

We sourced data from the Study of Epidemiology of Epilepsy in Demographic Surveillance Sites (SEEDS), which assessed the prevalence and risk factors of active convulsive epilepsy (ACE) in five health and demographic surveillance (HDSS) sites across sub-Saharan Africa: Agincourt (South Africa), Ifakara (Tanzania), Iganga (Uganda), Kilifi (Kenya) and Kintampo (Ghana). This dataset has been described in detail elsewhere [16]. The dataset comprised anonymised responses to an ACE-specific questionnaire, with questions administered in local languages specific to each study site: Twi in Kintampo, Xi-Tsonga in Agincourt, Lusoga in Iganga, and Kiswahili in Kilifi and Ifakara. ACE was defined as two or more unprovoked seizures occurring at least 24 hours apart with at least one episode in the preceding year. Convulsive epilepsy was explicitly chosen because convulsions are more easily identified and are associated with higher morbidity, mortality and stigma. Participants were recruited as part of a door-to-door survey conducted by the HDSS at each site. Trained clinicians evaluated the participants, reviewing cases across all sites. To ensure consistency in the diagnostic process, a panel of neurologists reviewed the case report forms and a consensus on the final diagnosis of

each case was reached. The full protocol for the SEEDS study, including details on the percentage of participants recruited at this stage relative to the total screened, has been previously published [16]. The outcome of this study was a clinical diagnosis of ACE (EEG supported where possible) confirmed by an epilepsy-specialised neurologist. Individuals with and without epilepsy (here considered as controls) are included in the dataset. Information on sociodemographic variables, historical risk factors and clinical history was collected from each participant. The resulting dataset comprised sociodemographic information and approximately 170 unique variables for each individual in five domains: clinical history, clinical examination, seizure description and EEG interpretation [7].

The study was approved by the ethics committees of University College London and the London School of Hygiene and Tropical Medicine and by the ethics review boards in each participating country. All participants or guardians gave written informed consent. The data collection protocol is also reported elsewhere [7,16].

Data pre-processing. We selected participants with either a confirmed diagnosis of ACE (cases) or absence of ACE (controls) following evaluation by a trained neurologist, a total of 5108 participants. We used predictors of epilepsy formatted as questions for the person with suspected epilepsy. These predictors have been reported and validated in previous studies from this cohort and were chosen for their maximally discriminative predictive ability to diagnose ACE [7]. The predictors were:

1. During these episodes, have you ever bitten your tongue?

2. Have you ever wet yourself during these episodes?

3. During these episodes, do you lose contact with your surroundings?

4. Has anyone told you that you appear dazed during these episodes?

5. During these episodes, does your body stiffen?

6. Do you experience stomach-ache before these episodes?

7. Do you see odd things (e.g. flashes or bright lights) before these episodes occur?

8. Do you think anything brings on these episodes?

We retained only participants with a confirmed diagnosis (cases) or absence of ACE (controls). This was termed the 'complete dataset'. We also considered a smaller subset in which only participants with either a confirmed 'generalised' seizure type or absence of ACE were retained (referred to as the 'generalised dataset'). The purpose was to minimise inherent heterogeneity in individual symptomatology due to multiple seizure types within a single individual. Selecting for generalised seizures reduces potential confounding across sites due to the inherently heterogeneous symptomatology of focal seizures and non-convulsive generalised epilepsy. The datasets were then separated by study site. Imputation and weighting were performed separately within each site. Predictors with missing values were imputed using the multiple imputation by chained equation method [17]. To adjust for potential confounding between study sites due to participant sex, age and seizure type (general, focal, or 'other') we applied inverse propensity weighting [18].

Study site datasets were then split into model training and testing subsets in a 7:3 ratio. When a model was trained on data from the generalised dataset, it was tested on data from the generalised subset. The data processing procedure and the ten seizure/type site subsets and data processing are summarised in Table 1 and S1 Fig.

Table 1. Dataset breakdown. Table showing the size of each subset of the data as split by diagnostic class, epilepsy type and site. The rightmost column represents a subset of the middle column. Participants with no clinical diagnostic information were excluded from the dataset. ACE = active convulsive epilepsy.

		Seizure types		
		Controls	**Confirmed ACE**	**Confirmed ACE & only generalised seizures**
Site	Agincourt	380	334	184
	Ifakara	678	465	223
	Iganga	498	259	152
	Kilifi	793	791	209
	Kintampo	516	394	233
Total		2865	2243	1001

https://doi.org/10.1371/journal.pdig.0000491.t001

Model development

The models trained were Logistic Regression, Support Vector Machine (SVM) using a linear kernel and Naive Bayes assuming a Bernoulli distribution. These models were selected because they are computationally efficient, well-established in clinical prediction modelling in epilepsy, and because they yield interpretable results [19]. They also have varied approaches to decision boundaries [20]. If there were an observed difference in the performance of these algorithms between study sites, it would likely be more attributable to the data than the model, i.e. an issue with generalisability.

The primary performance metric was the area under the receiver-operator characteristic curve (AUC). AUC is independent of a classification threshold, eliminating the need to identify a threshold parameter and is less sensitive to data imbalance than other metrics, such as accuracy.

To ensure robustness, five-fold cross-validation was used to determine the performance of models on the data they were trained on (internal validation). In this process, training data are randomly split into five balanced subsets; four subsets are used to train the model and the remaining subset is used to test the model and calculate the AUC. This is repeated four more times (five folds total), ensuring each subset is used for testing. The median of the five AUC values and the interquartile range are then used to determine the algorithm's average performance [21].

Models trained on data within only one study site are referred to hereafter as 'single-site'. A model was trained for every permutation of the three algorithms, five sites and two seizure types, resulting in 30 single-site models. Five-fold cross-validation AUC scores were obtained for each single-site model. Each single-site model was tested on the other four study sites, resulting in four external performance AUC scores. These were used to evaluate the performance of the models when tested on data from different regions. The Kolmogorov-Smirnov (K-S) test was used to assess whether there was a significant difference between the single-site AUC when tested on data from within the site (internal performance) and data from the other four study sites (external performance). The magnitudes, signs and ranges of the weights assigned to each predictor in each of the Logistic Regression models were also compared.

For each single-site model, we compared a set of incremental thresholds (20%, 50% and 80%) for classifying the probability output from each algorithm into a designation of ACE or control. For example, cases with less than 20% probability were assigned to controls and otherwise were assigned ACE. These thresholds were chosen to explore the effect on accuracy of changing a threshold when using a diagnostic tool in practice. The accuracies were calculated for the five study sites based on these classifications. We compared the relative change in accuracy between a) each study site from which a single-site model was developed, and b) the accuracy of each separate study site. This analysis adds an additional dimension, extending the

evaluation from AUC to a more clinical scenario where the direct classification of either ACE or control is essential.

We then developed an iteratively inclusive multi-site model to evaluate the effect of adding additional study sites. Beginning with one randomly chosen study site, a model for each of the three algorithms was developed. The performance of each model was assessed on the remaining study sites. Another randomly chosen site was added to the model, and the performance was then re-evaluated for the remaining unincluded study sites. This was repeated until every site except one (four training sites; one testing site) was included in the multi-site model, resulting in a leave-one-site-out (LOSO) model. We repeated this procedure until every permutation of study site order of inclusion was achieved. Lastly, we developed a multi-site model that incorporated all study sites, splitting the data 70:30% into model training and validation subsets.

Statistical analysis

Continuous variables were binned. The two-sided Mann-Whitney test, with a significance level of 0.01, was used to evaluate for differences between internal and external datasets in the single-site models. The two-sample Kolmogorov-Smirnov (K-S) test, with a significance threshold of 0.05, was used to assess the significance of differences in model performances resulting from training data on different sites. Analysis was performed using Python 3 [22]. Libraries used were Pandas [23], NumPy [24], SciPy [25], SKLearn [26], Seaborn [27], Pyplot [28] and Plotnine [29].

Results

The complete dataset comprised 5,108 people with suspected epilepsy: 2,243 with confirmed ACE (44%) and 2,865 (56%) confirmed not to have epilepsy. Missingness in the raw data appeared random (S2 Fig).

Single-site model performance in novel settings

Single-site models performance was reduced on average when tested on new sites. The median AUC when each single-site model was evaluated on novel data acquired from the same site was 0.94 (interquartile range (IQR) 0.90–0.96). However, when these models were tested with data from other study sites, the median AUC decreased significantly to 0.91 (IQR 0.89–0.93, t-test: $p<0.01$). Agincourt and Kilifi demonstrated the greatest difference between internal and external AUCs (Agincourt 0.96 internal, 0.88 external; p value <0.001, difference = -0.08 (7.9%); Kilifi 0.97 internal, 0.84 external, difference = -0.13 (13%), $p<0.001$). Table 2 and Fig 1 compare the AUC values across each study site for internal and external validation.

We performed hypothesis tests comparing the AUCs resulting from the validation of a single-site model on a) the site whose data they were trained on and b) the other. These were statistically significant for eleven of thirty (see Table 1: ten data subsets, three models) cases. There were seven significant differences for the dataset that considered all cases, and four for the generalised seizure dataset (S1 Table). Each site had six tests in total: Agincourt and Kilifi each had five significant tests, while the other sites had at least five insignificant tests.

The weights assigned to the predictors of ACE in the logistic regression model differed between sites (Fig 2). All eight predictors' weights had broad ranges, both when not accounting for seizure type (median range of weights 0.97, [IQR 0.72–1.5]) and when including only cases with generalised seizures (median range of weights 1.01 [IQR 0.86–1.3]). Apart from two variables, all variable weights were negative in some site models and positive in others. These two variables are as follows: only one predictor of ACE had consistently negative weights

Table 2. Median and IQR of the model AUC values across each study site for internal and external validation.

Site	Internal	External	Difference	P value
Agincourt	0.96 (0.95–0.97)	0.88 (0.86–0.88)	-0.08 (7.9%)	<0.001
Ifakara	0.89 (0.89–0.91)	0.92 (0.89–0.96)	0.02 (2.7%)	<0.001
Iganga	0.93 (0.91–0.96)	0.88 (0.87–0.96)	-0.05 (5.1%)	<0.001
Kilifi	0.97 (0.96–0.97)	0.84 (0.78–0.89)	-0.13 (13.0%)	<0.001
Kintampo	0.90 (0.88–0.92)	0.92 (0.91–0.95)	0.02 (2.8%)	<0.001

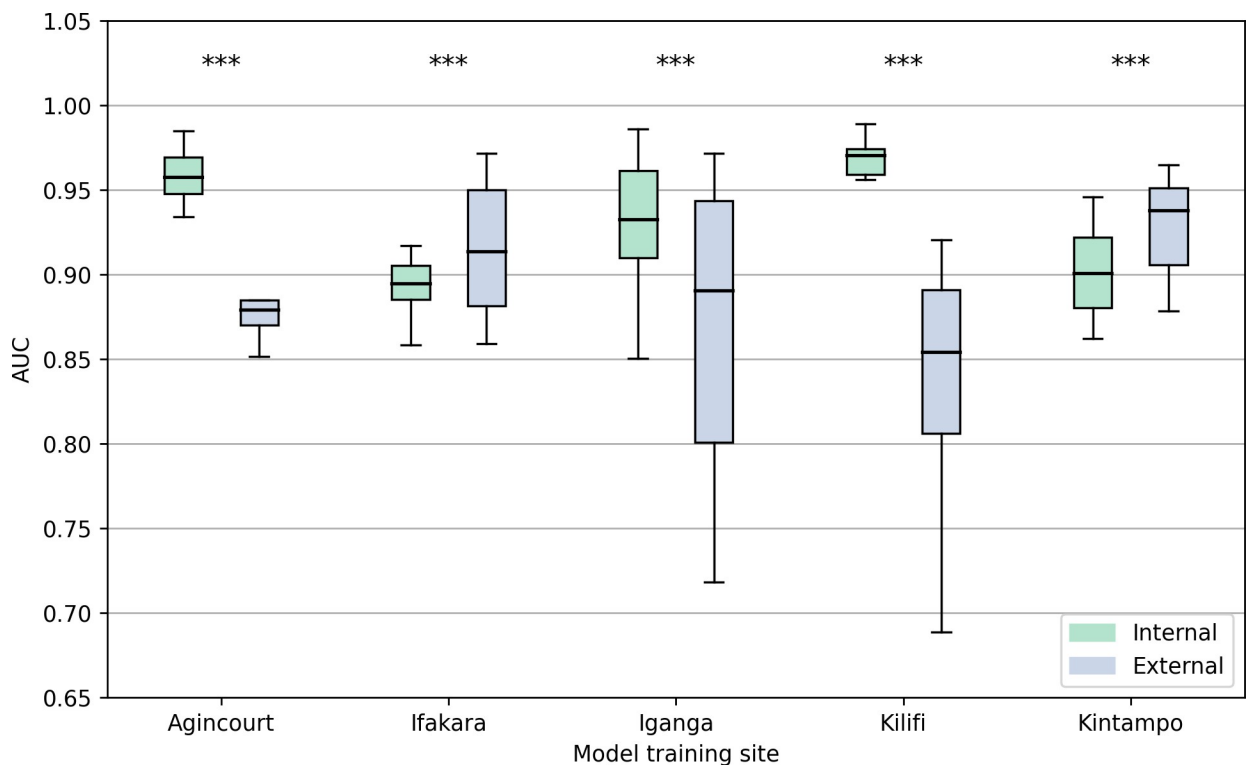https://doi.org/10.1371/journal.pdig.0000491.t002

Fig 1. Comparative Internal and External AUC Performance Across Single-Site Models. The Area Under the Receiver Operating Characteristic Curve (AUC) performance of machine learning models trained and tested on data from single sites. Performance is compared between internal validations (training and testing on the same site) and external validations (training on one site and testing on others). Five distinct study sites in sub-Saharan Africa are evaluated: Agincourt (South Africa), Ifakara (Tanzania), Iganga (Uganda), Kilifi (Kenya), and Kintampo (Ghana). Boxplots describe the distribution of AUC values obtained through bootstrap resampling, indicating the variance within internal and external validations. At Agincourt, the internal AUC is 0.96 (interquartile range (IQR) 0.95–0.97), while the external AUC is 0.88 (IQR 0.86–0.88), demonstrating a statistically significant difference with a decrease of approximately 0.08 in performance when models are externally validated. Similarly, Kilifi shows an internal AUC of 0.97 (IQR 0.96–0.97) against an external AUC of 0.84 (IQR 0.78–0.89), indicating a significant decline in external validation performance. Iganga's internal and external AUCs are 0.93 (0.91–0.96) and 0.88 (IQR 0.87–0.96), displaying a smaller yet significant discrepancy. In contrast, Ifakara and Kintampo exhibit a converse trend, where external AUCs 0.92 (IQR 0.89–0.96) and 0.92 (IQR 0.91–0.95) slightly exceed their internal counterparts 0.89 (IQR 0.89–0.91) and 0.90 (IQR 0.88–0.92), although these differences are also statistically significant. These findings underscore the variability in model generalizability and the importance of external validation when assessing the robustness of predictive models in healthcare settings. *** = p-value <0.001.
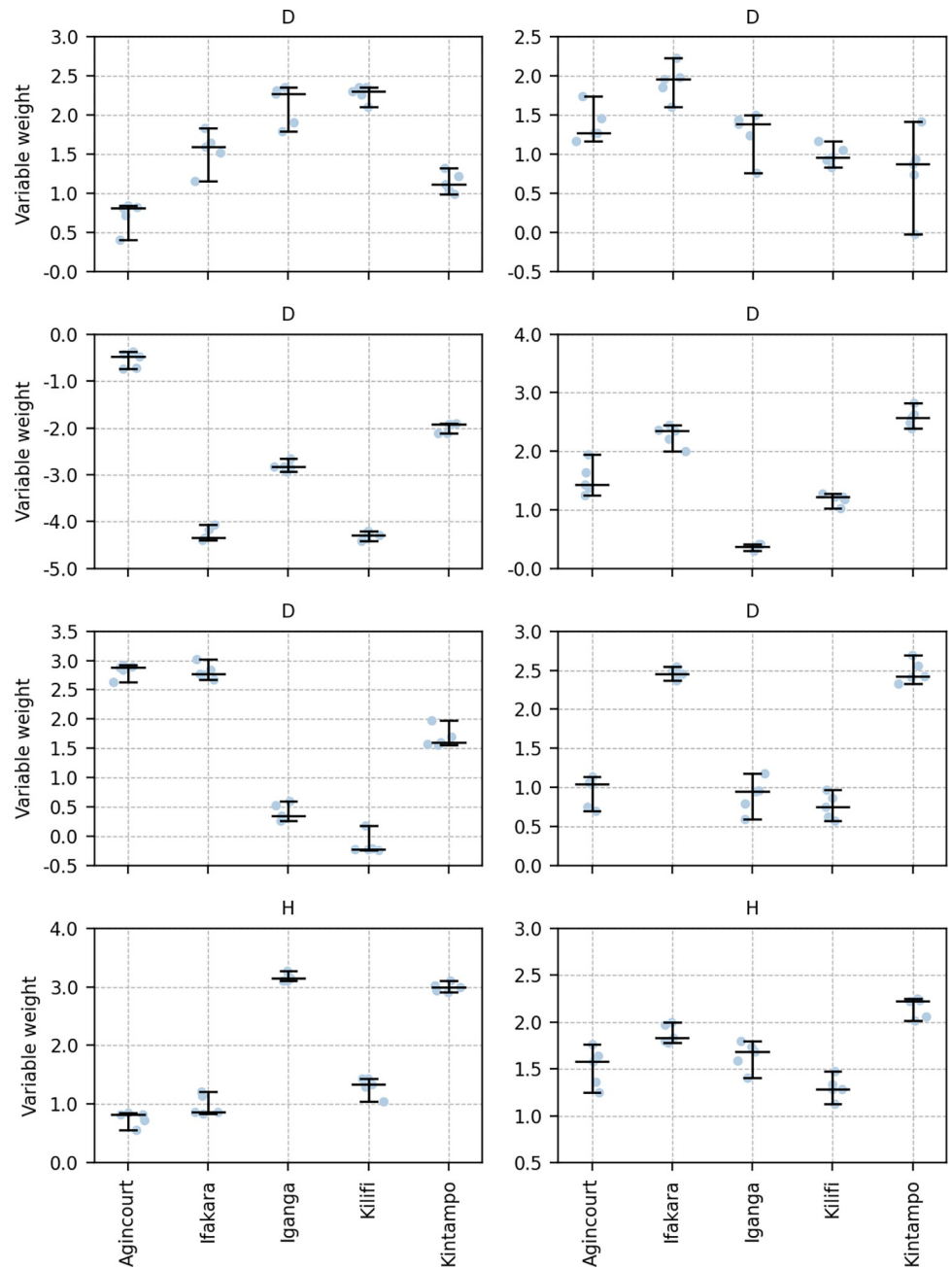
https://doi.org/10.1371/journal.pdig.0000491.g001

Fig 2. Variation of model weights between sites. Boxplot illustrating the values taken by the weights in each logistic regression model trained on a single site. Models were trained on a dataset in which positive cases were limited to participants with generalised epilepsy. All weights show a spread of values, and most change sign between sites. Only one covariate's weights were consistently positive ('Have you ever wet yourself during these episodes?') and only one covariate's weights were consistently negative ('Do you think anything brings on these episodes?'). In this context positive weights correspond to a positive association with epilepsy, and negative weights to negative association. The horizontal line at the origin serves to clarify the threshold between positive and negative weights. Central line is median; lower edge of the box indicates first quartile, upper edge of the box indicates third quartile; points are extreme weight values.
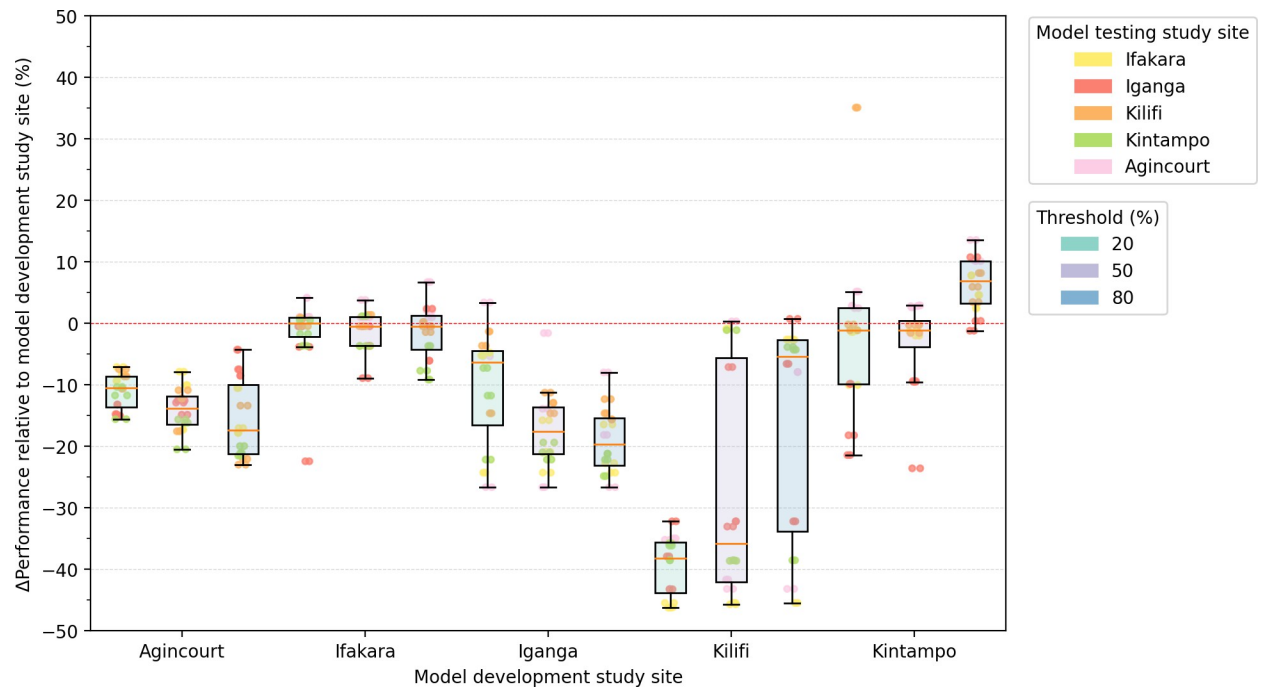
https://doi.org/10.1371/journal.pdig.0000491.g002

Fig 3. Effect of changing thresholds on accuracy. Heatmap showing how the accuracy of the logistic regression models was affected by changing the threshold from 0.2 to 0.8, for each one-site model. Models were trained on only one site's data and tested on each of the other sites in turn. The performance of some site models worsened both internally and externally while the other sites' performance improved. The overall mean change in accuracy was 10% (min 0.4%, max 28%, standard deviation 7.6%).

https://doi.org/10.1371/journal.pdig.0000491.g003

(corresponding to the question 'Do you think anything brings on these episodes?': range 0.31 when considering all controls, 0.75 for generalised seizures), and only one was consistently positive (corresponding to the question 'Have you ever wet yourself during these episodes?': range 0.72 when considering all controls, 0.49 for generalised seizures).

Modulating the probability threshold for a positive diagnosis also resulted in variable accuracy. Increasing the threshold improved the performance in some sites and worsened performance in others. Fig 3 displays the difference in performance observed when models are tested outside of the development site. In Agincourt and Iganga, increasing the threshold from 0.2 to 0.5 to 0.8 resulted in a decrease in relative accuracy (Agincourt median -0.10, -0.13, -0.17, $p<0.01$; Iganga median -0.06, -0.18, -0.19, $p<0.01$), while the inverse was true of Kintampo and Kilifi (Kilifi median -0.38, -0.35, -0.05, $p<0.01$; Kintampo median -0.01, -0.01, 0.06, $p<0.05$).

Incremental site inclusion

As additional sites were incorporated into the model, internal performance declined (initial median AUC 0.94, IQR 0.11; final median AUC 0.93, IQR 0.01; p = 0.06), and external performance improved (initial median 0.90, IQR 0.04; final median 0.92, IQR 0.02; $p<0.01$; Fig 4). At the initial stage of this process, when only one site was included in the training data, internal performance was greater than external (mean AUC difference 0.050, $p<0.01$) and all external performances were lower than all internal performances.

At the final stage, the two measures converged to a small final difference, and all the data points from the external validation were higher than all the internal points (Fig 4). The mean difference between final performance scores was 0.001 (p-value 0.83).
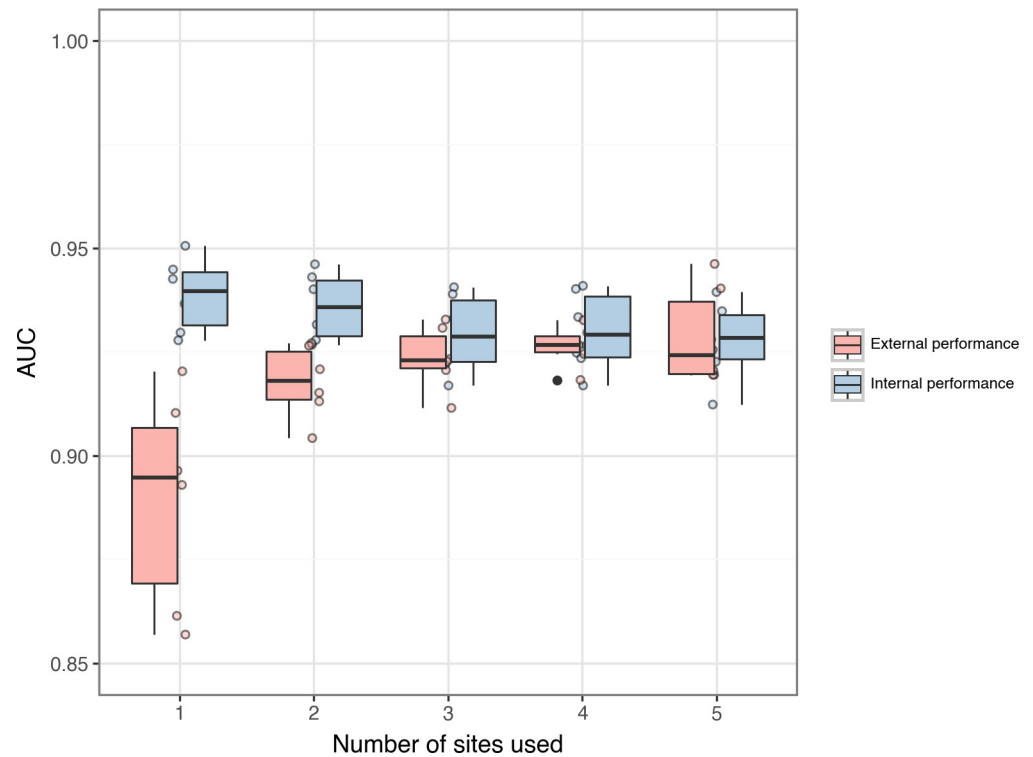
Fig 4. Performance of LOSO models. Boxplot of AUC values resulting from the testing, on each site in turn, of models trained on all but one site. Both internal and external performance is shown. Internal performance was higher in 3 sites (internal median 0.93, external median 0.89). In the other two, AUC values displayed a larger range (internal median 0.92, range 0.04; external median 0.96, range 0.05). AUC = area under receiver operating curve. Internal performance = performance on sites included in training data. External performance = performance on the site that was not included in training data. Central line is median; lower edge of the box indicates first quartile, upper edge of the box indicates third quartile; points are extreme performance scores.

https://doi.org/10.1371/journal.pdig.0000491.g004

Leave one site out

Internal performance was higher in 3 sites: Ifakara (internal median 0.93, IQR 0.00; external median 0.92, IQR 0.06; p = 0.55), Iganga (internal median 0.94, IQR 0.02; external median 0.91, IQR 0.03; p<0.01) and Kintampo (internal median 0.937, IQR 0.019; external median 0.89, IQR 0.006; p<0.01) (Fig 5). Kintampo's external scores had a smaller range (p<0.01) while the other two sites had a larger external range (Ifakara: p<0.01; Iganga: p = 0.80).

Agincourt (internal median 0.924, IQR 0.008; external median 0.95, IQR 0.03; p<0.05) and Kilifi (internal median 0.91, IQR 0.025; external median 0.97, IQR 0.02; p<0.01) had higher external performance, and these values displayed a larger range (Agincourt: p = 0.08; Kilifi: p = 0.65).

Patterns emerge when comparing the one-site models with the LOSO models. For instance, both model types showed that Agincourt and Kilifi had a wider range of external performance scores than internal, with the inverse relationship evident at the other sites. Agincourt and Kilifi demonstrated superior external median performance in the LOSO model relative to their internal performance, and a trend contradicted at the remaining sites. Kintampo consistently presented the lowest external variability in LOSO and one-site models, further underscoring the site-specific patterns inherent in the performance of these diagnostic tools.
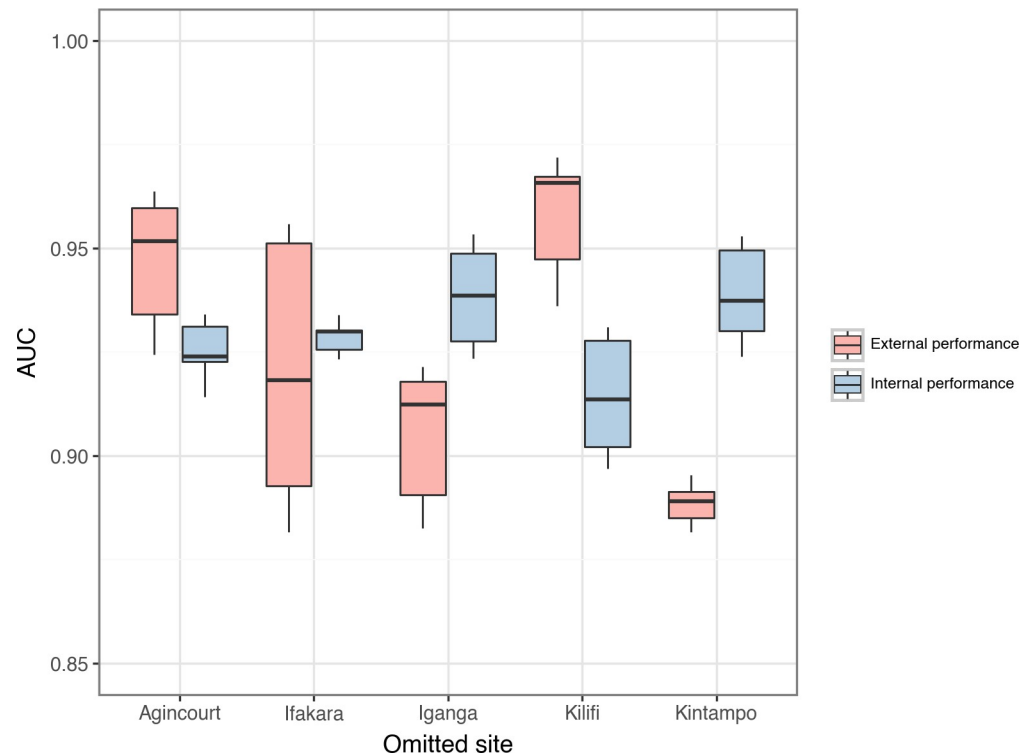
Fig 5. Performance of incremental models. Boxplot (with scatter graph) showing the change in AUC as the number of sites included in training is increased. Both internal and external performance is shown. As sites were added, internal performance worsened and external performance improved. At the stage when only one site was included in the training data, all external performances were lower than all the internal performances. At the final stage, the two measures converged to a small final difference, and all the data points from the external validation were higher than all the internal. AUC = area under receiver operating curve. Internal performance = performance on sites included in training data. External performance = performance on the site that was not included in training data. Central line is median; lower edge of the box indicates first quartile, upper edge of the box indicates third quartile; scatter points are individual performance scores.

https://doi.org/10.1371/journal.pdig.0000491.g005

Discussion

We demonstrate that deploying an epilepsy diagnostic model outside the cultural and geographical region in which it was developed can result in highly unpredictable, frequently suboptimal outcomes.

As with other ML models, the generalisability of clinical epilepsy diagnostic tools is inherently constrained. Extrapolating these models beyond their original regional parameters necessitates a trade-off between internal and external performance. Performance scores within a given region consistently exhibit lower variances and higher medians than those obtained from cross-site applications. This is perhaps especially problematic given the volume of ML-driven diagnostic models for epilepsy that have been developed using single regions [30]. Incorporating data from more sites into the model's training set can help mitigate the risk of ML models making erroneous assumptions and offering incorrect diagnoses. Whilst this approach engenders a model with enhanced robustness and improved external performance, it concomitantly decreases internal performance. This occurs as the model broadens its applicability while becoming less tailored to specific sites. While analysis of generalizability has been performed elsewhere [15], this study presents a novel and significant contribution to the field of epilepsy diagnosis.

The weight of each ACE predictor and the optimal thresholds for a classification vary significantly between sites. At the extreme, a particular symptom might correlate positively with ACE at one site and negatively at another. These disparities may result from the variance in reporting of epilepsy-related symptoms across geographies.

Data underscore the importance of thoughtfully calibrating diagnostic procedures to the unique specificities of each geographic locale. The observed threshold variability also highlights that a simplistic transposition of one site's threshold for ACE classification to another may lead to an unpredictable number of misdiagnoses, most concerningly false negatives, where an individual remains undiagnosed and cannot access treatment. The findings emphasise the need to adjust model parameters to ensure their suitability for application in varying settings.

A variety of explanations may account for differences in performance between the sites. One reason may be that symptoms may be reported differently by those with epilepsy or their carers, depending on the cultural and clinical context [13]. These cannot necessarily be predicted and accounted for and may differ between sites. For example, appearing dazed during episodes was the single most prevalent symptom reported by individuals with seizures in Iganga and one of the least prevalent in Ifakara (S3 Fig).

Based on the results of this study, we suggest that some degree of site-specific validation is essential before a predictive model is deployed in practice. While complete re-training of all model parameters on local data would be ideal and result in optimal performance, this is not always feasible. Studies have shown that merely changing the intercept of a logistic regression model may help recontextualise a model for new settings [31]. It may suffice to adjust only a few parameters.

As performance is continually assessed, we propose that iteratively updating the model throughout deployment can help fine-tune it further, with new improvements building upon previous findings. Such a model would generally require a threshold to determine positive cases, although, as shown, optimal thresholds vary between sites. The results suggest that simply validating a model's threshold may also effectively update its performance in a new setting. In this study, we observed a mean change in accuracy of 10% following a threshold adjustment.

Limitations

This work is limited by the choice of dataset and its contents. The data were acquired from five distinct sites from five different sub-Saharan African countries. We cannot necessarily draw conclusions about, for example, training a model on data from a specific region and deploying it in a nearby location in the same country with similar demographics. The dataset also featured limited accounting of seizure type. The results could be clouded by, for example, the effects of non-convulsive seizure phenomena, which we could not account for.

We attempted to minimise this inherent heterogeneity by separating the generalised seizures into a separate dataset. Generalised convulsive seizures were selected as they demonstrate a higher degree of homogeneity in their clinical presentation than focal seizures [14]. Further work should explore focal seizures to identify better how ML models may also help in their diagnosis.

There also may have been data collection and data entry issues: self-reporting may be influenced by linguistic differences in how questions were asked, cultural differences in how chronic conditions are perceived, or who asked the questions and how [13]. The clinicians making diagnoses or the workers performing data entry may also vary in reliability. While the

tools and questions were adapted to the local context and training was standardised before data collection and monitored during [16], human factors may still play some role.

A potential limitation of our approach is the lack of exploration of deep transfer learning models, which have shown potential in predicting outcomes on datasets distinct from those included in the initial model training [32,33]. These models could mitigate the limitations of small dataset sizes and enhance generalizability across diverse populations.

Conclusions

We demonstrate that, when developing models for epilepsy diagnosis, data collected from one site cannot naïvely be used as representative for all other sites where a model could be deployed. Given the needs of LMICs, ML models can be leveraged to have a significant impact. Nonetheless, careful interrogation and dedication to rooting tools in the setting of their use must be ensured to avoid such models associating with inadvertent harm, including missed diagnoses resulting in delayed initiation of care. Whilst the present analysis has focused on convulsive epilepsy, similar arguments are likely applicable across other seizure types and disorders of mind-brain health more broadly. Future work could explore the application of deep transfer learning models, which have shown the potential to generalise across datasets different from those used in the initial model development. Incorporating such techniques could further enhance the robustness and applicability of epilepsy diagnostic tools in LMICs.

Key points

- Machine learning-driven clinical tools are becoming more prevalent in low-resource settings; however, their general performance across regions is not fully established. Given their potential impact, it is crucial models are robust, safe and appropriately deployed

- Models perform poorly when making predictions for regions that were not included in their training data, as opposed to sites that were

- Models trained on specific regions can have different optimal parameters and thresholds for performance in practice elsewhere

- There is a trade-off between internal and external performance, where a model with better external performance usually has worse internal performance but is generally more robust overall

Supporting information

S1 Fig. Data preparation. Flowchart showing the process of data preparation (see Table 1).
(TIFF)

S2 Fig. Missingness in the raw data. Heatmap showing where there was unexplained missingness in the data before cleaning. Missing values are shown in yellow, others in green. The data are sorted according to assessment date, in ascending order from earliest to latest.
(TIFF)

S3 Fig. Covariate values per site. Stacked bar chart showing the percentage of yes/no answers to the covariate questions. Values taken from the cleaned data, split by site and epilepsy diagnostic class.
(TIFF)

S4 Fig. Histogram showing propensity scores as calculated from the processed data, coloured by whether there was a diagnosis of epilepsy (blue) or not (red). There is a complete overlap–the range of scores for the two diagnostic classes overlaps completely.
(TIFF)

S1 Table. Kolmogorov-Smirnov test p-values. Table summarizing the p-values of the statistical tests of the performances resulting from the testing of each one-site model on each of the other sites in turn. The two samples in each test were the internal performances for that site (from 5-fold cross validation) and the performances of that model on each other site. This was done for each possible combination of the three model types (Logistic Regression, SVM using a linear kernel, and Naive Bayes assuming a Bernoulli distribution) and the two datasets (the whole dataset, and that with the positive cases limited to participants with generalised epilepsy). Thus, there are 6 values for each site. The 2-sample Kolmogorov-Smirnov test has the null hypothesis that the two samples are drawn from the same distribution. Statistical insignificance is taken as insufficient evidence to reject this. There were 7 significant tests for the whole dataset and 4 for the generalised seizure dataset. Agincourt and Kilifi each had 5 significant tests, while the other 3 sites each had at least 5 insignificant tests.
(DOCX)

S2 Table. Summary table of variables present in the data.
(DOCX)

S3 Table. Mean predictor weight values and associated p-values.
(DOCX)

Acknowledgments

Collaborators

- Study of Epidemiology of Epilepsy in Demographic Sites (SEEDS)

- Oxford Martin Programme on Global Epilepsy https://www.oxfordmartin.ox.ac.uk/global-epilepsy/)

- Epilepsy Pathway Innovation in Africa (EPInA, https://epina.web.ox.ac.uk/)

Author Contributions

Conceptualization: Ryan Wagner, Charles R. Newton, Arjune Sen, Gabriel Davis Jones.

Data curation: Symon M. Kariuki, Anthony K. Ngugi, Angelina Kakooza Mwesige, Honorati Masanja, Seth Owusu-Agyei, Charles R. Newton, Gabriel Davis Jones.

Formal analysis: Ioana Duta, Daniel M. Mwanga, Charles R. Newton, Arjune Sen, Gabriel Davis Jones.

Investigation: Charles R. Newton, Gabriel Davis Jones.

Methodology: Daniel M. Mwanga, Gabriel Davis Jones.

Project administration: Gabriel Davis Jones.

Resources: Symon M. Kariuki, Anthony K. Ngugi, Angelina Kakooza Mwesige, Honorati Masanja, Seth Owusu-Agyei, Ryan Wagner, Gabriel Davis Jones.

Supervision: Gabriel Davis Jones.

Validation: Gabriel Davis Jones.

Visualization: Gabriel Davis Jones.

Writing – original draft: Ioana Duta, Charles R. Newton, Arjune Sen, Gabriel Davis Jones.

Writing – review & editing: Symon M. Kariuki, Anthony K. Ngugi, Angelina Kakooza Mwesige, Daniel M. Mwanga, Ryan Wagner, J Helen Cross, Josemir W. Sander, Charles R. Newton, Arjune Sen, Gabriel Davis Jones.

References

1. Ngugi AK, Bottomley C, Kleinschmidt I, Sander JW, Newton CR. Estimation of the burden of active and life-time epilepsy: A meta-analytic approach. Epilepsia. 2010 May; 51(5):883–90. https://doi.org/10.1111/j.1528-1167.2009.02481.x PMID: 20067507

2. World Health Organization. WHO | Epilepsy: a public health imperative. Who. 2019;171.

3. Fiest KM, Sauro KM, Wiebe S, Patten SB, Kwon CS, Dykeman J, et al. Prevalence and incidence of epilepsy. Neurology. 2017 Jan 17; 88(3):296–303.

4. Mbuba CK, Ngugi AK, Newton CR, Carter JA. The epilepsy treatment gap in developing countries: A systematic review of the magnitude, causes, and intervention strategies. Epilepsia. 2008 Sep 1; 49 (9):1491–503. https://doi.org/10.1111/j.1528-1167.2008.01693.x PMID: 18557778

5. Meinardi H, Scott RA, Reis R, Sander JWAS. The treatment gap in epilepsy: the current situation and ways forward. Epilepsia [Internet]. 2001 [cited 2024 Aug 20]; 42(1):136–49. Available from: https://pubmed.ncbi.nlm.nih.gov/11207798/

6. Durkin MS, Elsabbagh M, Barbaro J, Gladstone M, Happe F, Hoekstra RA, et al. Autism screening and diagnosis in low resource settings: Challenges and opportunities to enhance research and services worldwide. Autism Research [Internet]. 2015 Oct 1 [cited 2023 May 15]; 8(5):473–6. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/aur.1575 PMID: 26437907

7. Davis Jones G, Kariuki SM, Ngugi AK, Mwesige AK, Masanja H, Owusu-Agyei S, et al. Development and validation of a diagnostic aid for convulsive epilepsy in sub-Saharan Africa: a retrospective case-control study. Lancet Digit Health [Internet]. 2023 Apr 1 [cited 2023 Mar 27]; 5(4):e185–93. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2589750022002552 https://doi.org/10.1016/S2589-7500(22)00255-2 PMID: 36963908

8. Abbasi B, Goldenholz DM. Machine learning applications in epilepsy. Epilepsia. 2019 Oct 1; 60 (10):2037–47. https://doi.org/10.1111/epi.16333 PMID: 31478577

9. Kwon CS, Wagner RG, Carpio A, Jetté N, Newton CR, Thurman DJ. The worldwide epilepsy treatment gap: A systematic review and recommendations for revised definitions–A report from the ILAE Epidemiology Commission. Epilepsia [Internet]. 2022 Mar 1 [cited 2023 Sep 26]; 63(3):551–64. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/epi.17112 PMID: 35001365

10. Hampshire K, Porter G, Owusu SA, Mariwah S, Abane A, Robson E, et al. Informal m-health: How are young people using mobile phones to bridge healthcare gaps in Sub-Saharan Africa? Soc Sci Med. 2015 Oct 1; 142:90–9. https://doi.org/10.1016/j.socscimed.2015.07.033 PMID: 26298645

11. Zurovac D, Otieno G, Kigen S, Mbithi AM, Muturi A, Snow RW, et al. Ownership and use of mobile phones among health workers, caregivers of sick children and adult patients in Kenya: cross-sectional national survey. Global Health [Internet]. 2013 May 14 [cited 2024 Aug 20]; 9(1):20. Available from: https://link.springer.com/articles/10.1186/1744-8603-9-20 PMID: 23672301

12. Bleeker SE, Moll HA, Steyerberg EW, Donders ART, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research:: A clinical example. J Clin Epidemiol. 2003 Sep 1; 56 (9):826–32. https://doi.org/10.1016/s0895-4356(03)00207-5 PMID: 14505766

13. Park J, Johantgen ME. A Cross-Cultural Comparison of Symptom Reporting and Symptom Clusters in Heart Failure. Journal of Transcultural Nursing. 2017; 28(4). https://doi.org/10.1177/1043659616651673 PMID: 27225884

14. Benbadis S. The differential diagnosis of epilepsy: A critical review. Epilepsy & Behavior. 2009 May 1; 15(1):15–21. https://doi.org/10.1016/j.yebeh.2009.02.024 PMID: 19236946

15. Yang J, Soltan AAS, Clifton DA. Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. npj Digital Medicine 2022 5:1 [Internet]. 2022 Jun 7 [cited 2023 Jul 11]; 5(1):1–8. Available from: https://www.nature.com/articles/s41746-022-00614-9 https://doi.org/10.1038/s41746-022-00614-9 PMID: 35672368

16. Ngugi AK, Bottomley C, Kleinschmidt I, Wagner RG, Kakooza-Mwesige A, Ae-Ngibise K, et al. Prevalence of active convulsive epilepsy in sub-Saharan Africa and associated risk factors: cross-sectional and case-control studies. Lancet Neurol. 2013 Mar 3; 12(3):253. https://doi.org/10.1016/S1474-4422 (13)70003-6 PMID: 23375964

17. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011 Mar 1; 20(1):40–9. https://doi.org/10.1002/mpr.329 PMID: 21499542

18. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med. 2015 Dec 10; 34(28):3661–79. https://doi.org/10.1002/sim.6607 PMID: 26238958

19. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy 2021, Vol 23, Page 18 [Internet]. 2020 Dec 25 [cited 2023 May 15]; 23 (1):18. Available from: https://www.mdpi.com/1099-4300/23/1/18/htm

20. Classifier comparison—scikit-learn 1.2.2 documentation [Internet]. [cited 2023 May 15]. Available from: https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html#sphx-glr-auto-examples-classification-plot-classifier-comparison-py

21. Berrar D. Cross-Validation Call for Papers for Machine Learning journal: Machine Learning for Soccer View project Cross-validation. [cited 2023 Jul 10]; Available from: https://doi.org/10.1016/B978-0-12-809633-8.20349-X

22. Van Rossum G and DFL and others. Python reference manual. Vol. 111. 1995.

23. McKinney W. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference. 2010;56–61.

24. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020 Sep 17; 585(7825):357–62. https://doi.org/10.1038/s41586-020-2649-2 PMID: 32939066

25. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods.

26. Pedregosa FABIANPEDREGOSA F, Michel V, Grisel OLIVIERGRISEL O, Blondel M, Prettenhofer P, Weiss R, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12(85):2825–30.

27. Waskom M. seaborn: statistical data visualization. J Open Source Softw. 2021 Apr 6; 6(60):3021.

28. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007; 9(3):90–5.

29. Kibirige H, Lamp G, Katins J, gdowding, austin, Finkernagel F, et al. has2k1/plotnine: v0.12.1. 2023 May 10 [cited 2023 Jul 10]; Available from: https://zenodo.org/record/7919297

30. Patterson V, Singh M, Rajbhandari H, Vishnubhatla S. Validation of a phone app for epilepsy diagnosis in India and Nepal. Seizure [Internet]. 2015 Aug 1 [cited 2023 Sep 4]; 30:46–9. Available from: http://www.seizure-journal.com/article/S1059131115001314/fulltext https://doi.org/10.1016/j.seizure.2015.05.011 PMID: 26216684

31. Janssen KJM, Vergouwe Y, Kalkman CJ, Grobbee DE, Moons KGM. A simple method to adjust clinical prediction models to local circumstances. Canadian Journal of Anaesthesia. 2009 Mar; 56(3):194–201. https://doi.org/10.1007/s12630-009-9041-x PMID: 19247740

32. Cheplygina V, de Bruijne M, Pluim JPW. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med Image Anal. 2019 May 1; 54:280–96. https://doi.org/10.1016/j.media.2019.03.009 PMID: 30959445

33. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, et al. A Comprehensive Survey on Transfer Learning. Proceedings of the IEEE. 2021 Jan 1; 109(1):43–76.