Veretennikova *et al. Journal of Statistical Distributions and Applications* (2018) 5:8
https://doi.org/10.1186/s40488-018-0086-7

Journal of Statistical Distributions and Applications

RESEARCH Open Access

Parameters of stochastic models for electroencephalogram data as biomarkers for child's neurodevelopment after cerebral malaria

Maria A. Veretennikova[1*], Alla Sikorskii[2] and Michael J. Boivin[3]

*Correspondence:
mveretennikova@hse.ru
[1]Department of Statistics and Data Analysis, Faculty of Economic Science, National Research University, Higher School of Economics, Shabolovka 28/11, 9, 19049 Moscow, Russia
Full list of author information is available at the end of the article

Abstract

The objective of this study was to test statistical features from the electroencephalogram (EEG) recordings as predictors of neurodevelopment and cognition of Ugandan children after coma due to cerebral malaria. The increments of the frequency bands of EEG time series were modeled as Student processes; the parameters of these Student processes were estimated and used along with clinical and demographic data in a machine-learning algorithm for the prediction of children's neurodevelopmental and cognitive scores 6 months after cerebral malaria illness. The key innovation of this work is in the identification of stochastic EEG features that can serve as language-independent markers of the impact of cerebral malaria on the developing brain. The results can enhance prognostic determination of which children are in most need of rehabilitative interventions, which is especially important in resource-constrained settings such as sub-Saharan Africa.

Mathematics Subject Classification: 60E05, 62M10, 62P10

Keywords: Student processes coma EEG wavelets regression regularization

1 Introduction

Cerebral malaria (CM) affects over half a million people annually and has high prevalence in sub-Saharan Africa. Different sources indicate distinct mortality rates, but for children it is above 40 percent (Idro et al. 2010). For those who survive, the sequelae could include neurodevelopmental impairments and metabolic disturbances.

During CM the red blood cells are parasitized, most often by *Plasmodium Falciparum*. *P. Vivax* and *P. Knowlesi* are also known to cause severe malaria, but unlike *P. Falciparum* they do not lead to coma (WHO 2014). Coma is the principal diagnostic difference between severe malaria and CM, with lumbar puncture ruling out bacterial central nervous system (CNS) infection or other cause of coma beyond severe malaria. CM is specific to *P. Falciparum* and is distinctive from other forms of malaria because of the sequestration of infected red blood cells in the microvasculature of the brain and compromise of the blood/brain barrier contributing to an immunopathogenic inflammatory cascade. This cascade together with ischemic and metabolic effects cause coma and often seizures during illness and contribute to the neuropathogenic basis of

subsequent neurological and neurocognitive sequelae (John et al. 2008). Since coma is defined as prolonged unconsciousness and unresponsiveness (usually between 1 h and 4 weeks), clinical language-independent data provide the only opportunity to gauge the extent to which brain injury may impact subsequent neurodevelopment and cognitive function.

Today there is a number of different techniques for the statistical analysis of the brain. Despite some limitations, EEG is still widely used as a noninvasive way to monitor patients, predict seizure onsets and to determine the amount of activity in near-death states. In Awal et al. (2016) the authors use EEG features to predict neurodevelopmental outcomes for term infants with hypoxic ischaemic encephalopathy (HIE). The association between brain activity during coma and trauma outcomes was investigated in Malagurski et al. (2017) and Juan et al. (2015). Statistical analysis of EEG has been used to identify quasi-brain-death from coma (Li et al. 2014) and to confirm brain death (Chen et al. 2008).

Much of the work on the analysis of EEG data has focused on classification (Kirch et al. 2015; Piryatinska et al. 2009; Temko et al. 2011) or prediction of seizures (Duncan et al. 2013) without specification of the model for the underlying stochastic process. The use of spectral methods for the EEG time series, while popular, is problematic due to the evidence of non-stationarity of the process (Ignaccolo et al. 2009). In this paper we propose a stochastic model for the EEG time series, where for each frequency band the increment process is assumed to be a Student process, realized as Lévy driven Ornstein-Uhlenbeck-type process. The parameters of the Student marginal distribution are estimated and entered into a machine learning algorithm to test their association with children's neurodevelopmental and cognitive scores 6 months after cerebral malaria illness. The use of the Student distribution parameters markedly improves explained variation of neurodevelopment and cognition compared to using only demographic and clinical characteristics including plasma and cerebrospinal fluid biomarkers, or prediction based upon frequently used traditional EEG features. Identification of biomarkers such as the parameters of stochastic models for the EEG data has the potential to enhance diagnostic and prognostic determination by complementing the very limited clinical expertise in neurologists able to read and interpret EEG in resource constrained settings. Further, language-independent markers of neurodevelopment and cognition based on stochastic features of EEG data can complement the limited expertise available in sub-Saharan Africa for developmental and cognitive evaluations.

2 Dataset and preprocessing

Data used in this analysis were collected during the observational study of the pathogenesis of severe malaria (cerebral malaria (CM) and severe malarial anemia (SMA)) in surviving children, along community control children from their households who did not have a history of severe malaria (Bangirana et al. 2016). The study was performed at Mulago National Referral and Teaching Hospital in Kampala, Uganda in 2008 − 2015. Children with cerebral malaria, severe malarial anemia, or community control children were enrolled if they were between 18 months and 12 years of age. Cerebral malaria was defined as: 1) coma (Blantyre Coma Score [BCS] $\leq$ 2; 2) Plasmodium falciparum on blood smear; and 3) no other known cause of coma (e.g., hypoglycemia-associated coma reversed by glucose infusion, meningitis, or a prolonged postictal state). Children

were enrolled after obtaining written informed consent in the local language from their parent(s) and signed assent from children 7 years of age and older.

The observational study was approved by the Institutional Review Boards of the Makerere University School of Medicine and the University of Minnesota. Data from CM malaria children were included in this study as EEGs were done only for them and not SMA or community control children. Data from community controls were used to create the z-scores of neurodevelopmental and cognitive outcomes of the CM children as described below.

2.1 EEG data

MATLAB software was used for the processing of the EEG data set that comprised the standard 10–20 EEG recordings for 78 children with the sampling rate of 500 Hz and the average record duration of 30 min. Persyst software (Persyst, Prescott, USA) was used to remove artifacts due to breath, muscle movement and heartbeat from the raw EEG data. For most children there were 19 channels, which means the electrodes were not located very densely, so there was no necessity to use the average reference. Hence we've chosen CZ to be the reference electrode to avoid the laterality bias. CZ is one of the predominant choices for a reference (Teplan 2002). Then we used EEGLAB (Makeig and Delorme 2004) to identify problematic channels based on properties of their voltage measurements, leaving 16 channels for analysis and excluding three (PZ-CZ, C4-CZ, O2-CZ). Data for several of the included 16 channels had substantial numbers of zero observations, which could be due to poor connection between the electrode and the skin. For these channels we extracted features described below with and without zeros, with the rationale that if a feature resulted from an artifact and was not important, it would not be selected by the machine learning algorithm.

We used Daubechies wavelets for splitting the clean signal into frequency bands. Daubechies orthogonal wavelets have a number of vanishing moments, which is used as an index for referencing, e.g. the standard notation Db4 means Daubechies wavelet with 4 vanishing moments. Research indicates particular suitability of Db4 for statistical analysis of EEG. We examined the relative average mean squared error (MSE) between the wavelet signal approximation and the actual signal for different Daubechies wavelets. For these data Db4 yielded the reconstruction error or order 10^{-9}, which is sufficiently low. Also, Db4 frequency band separation resulted in frequency intervals which are very close to the traditional frequency ranges: delta, theta, beta, alpha and gamma bands (Daubechies 1992), see Table 1.

Due to occasional spikes and irregular patterns in the original time series, the idea was not to split them into epochs as it is often done (Fraschini et al. 2016). We hypothesized that some of the statistical features, such as the frequency of flat line measurements

Table 1 Frequency band correspondence

Traditional	Db4 band's central frequency
Delta 0 — 3.5 Hz	2.7 Hz
Theta 3.5 — 7.5 Hz	5.57 Hz
Alpha 7.5 — 13 Hz	11 Hz
Beta 13 — 30 Hz	22.3 Hz
Gamma > 30 Hz	four subsequent bands

relative to the whole EEG record, could be useful as explanatory variables for neurodevelopment and cognition, whilst dividing the record into epochs would complicate extraction of useful information.

Empirical studies show that a low activity level in the gamma frequencies is closely related to the coma state (Chen et al. 2008), and generally, gamma band oscillations are thought to be related to higher cortical functioning, such as consciousness, memory, perception and learning (Uhlhaas et al. 2009). In Deng et al. (2015) it was shown that EEG gamma band activity characteristics are associated with the outcomes of targeted temperature management for brain recovery after cardiac arrest. On the other hand, it is also advised to exclude the highest EEG gamma frequency bands from the analysis, because it is most likely to be noise, rather than the real deterministic signal. In view of these recommendations and the goal of this research to investigate stochastic features of EEG, we excluded only the highest gamma band (D1), keeping the rest. Our rationale was that if the other gamma bands (D2, D3, and D4 in Daubechies' frequency band notation) were indeed useless for the prediction of post-comatose neurodevelopmental and cognitive scores, then this would be empirically determined in statistical algorithms for the extraction of important features.

2.2 Measures of neurodevelopment and cognition

Children had neurodevelopmental assessments (appropriate for those 5 years of age or younger) or cognitive assessments (appropriate for children over 5 years of age) a week after discharge from the hospital (or at enrollment for community control children) and then at 6 and 12 months after enrollment. Data from the assessment at 6 month post-enrollment were used for this analysis.

Neurodevelopmental assessment for children 5 years of age or younger. The Mullen Scales of Early Learning (MSEL) (Mullen 1995) were used to quantify neurodevelopment. MSEL is based on a comprehensive test assessing specific developmental domains: visual reception, gross motor skills, fine motor skills, receptive language, and expressive language. A composite score derived from standardized t-scores of the four domains (excluding gross motor) provides a measure of *g*, the general measure of fluid intelligence.

Cognitive assessment for children over 5 years of age. The Kaufman Assessment Battery for Children, second edition (KABC-II) (Kaufman and Kaufman 2004) evaluates sequential and simultaneous processing, learning, reasoning, and crystallized intellectual ability (knowledge). The knowledge subscale was not administered because it was not suitable in this setting (Bangirana et al. 2009). Summation of scores for the domains of sequential processing, simultaneous processing, learning, and planning yielded the Mental Processing Index (MPI) which was the measure of overall cognitive ability in this age group.

The United States of America (USA) norms were used to arrive at the MSEL composite *g* score and the KABC-II MPI score, because using such norms to adjust for the child's age was necessary to compute these global measures. To obtain a single measure of neurodevelopment or cognition for all children regardless of age, we computed the means and standard deviations of the age appropriate measures, the MSEL composite or the KABC-II MPI, among the community control children. Then for CM children the z-scores in each age group were computed by subtracting the means and dividing by the standard deviation of the community control children.

2.3 Other measures

Home Observation for the Measurement of the Environment (HOME) (Bradley and Caldwell 1979) is a composite measure designed to assess the quality and quantity of stimulation that the child is exposed to in their home environment. A total HOME score was generated by summing the number of *"yes"* responses to a checklist of items; higher HOME scores indicate higher quality of home environment.

Demographic and anthropometric data included age, sex, height-for-age and weight-for-age z-scores computed using the World Health Organization reference norms (Wold Health Organization Growth Standards 2009). Socioeconomic status (SES) was assessed using a checklist of material possessions, housing quality, cooking resources and water accessibility. Clinical variables and biomarker panels from plasma and cerebrospinal fluid were collected during each child's hospitalization for CM.

3 Creating the feature matrices

3.1 Non-EEG features

The non-EEG features included demographic and anthropometric data, SES and HOME scores, the Blantyre Coma Score, and plasma and cerebrospinal fluid biomarker panels, for a total of 54 potential explanatory variables.

3.2 Commonly used EEG features

We have evaluated 362 EEG features that have been commonly used in the past analyses of EEG data. Presence of seizures was reflected by binary variable that was defined using Persyst software indicators (Persyst, Prescott, USA). Frequencies of peaks in the original cleaned time series differing from the nearest measurements from both sides by 1/3, 1, 2 and 3 standard deviations were calculated and denoted by fp1/3, fp1, fp2 and fp3, respectively. Proportion of flat line EEG for each of the 16 channels was evaluated for the original cleaned time series.

For delta (a7), theta (D7), alpha (D6), beta (D5) and gamma (D4, D3, D2) frequency bands for each channel, we calculated amplitude variances and Shannon time entropy using wavelets (De Oliveira 2015; Smolentsev 2014). This version of entropy is defined in MATLAB as

$$S(x_i) = -\sum_{i=1}^{N-1} x_i^2 \log\left(x_i^2\right)$$

where x_i is the i-th measurement in the time series for the signal (MATLAB wavelet packet). Relative frequency band energy was defined as the sum of wavelet coefficients divided by the total sum of the coefficients for all the frequency bands (Smolentsev 2014, Chapter 3.10).

Hjorth complexity and mobility parameters (Hjorth 1970) were calculated for the entire time series based on the second moment as well as the first and second order differences.

3.3 New stochastic features

After splitting the EEG time series into frequency bands using Daubechies wavelets, for each frequency band we constructed the histograms for the increment process at different

time blocks, and evaluated sample means and variances. The histograms were consistently approximately bell-shaped but with peaks higher and tails heavier than normal (Fig. 1).

The means were consistently close to zero, and the variances ratios for different time blocks fell within a rule of thumb range of [0.25, 4] (Montgomery 2012). Based on these empirical features, we selected a modeling approach that uses a stationary stochastic process for the increments of time series. To reflect the leptokurtic distribution seen in the data, we propose a stationary Student process as a model for the increments of the EEG time series for each frequency band. The symmetric scaled Student marginal distribution has the density

$$f_{\nu,\delta}(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\delta\sqrt{\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\left(\frac{x-\mu}{\delta}\right)^2\right)^{-\frac{\nu+1}{2}}, \; x \in \mathbb{R}, \tag{1}$$

$\nu > 0$ is degrees of freedom, $\delta > 0$ is a scale parameter, $\mu \in \mathbb{R}$ is location parameter. We denote this distribution by $T_{\mu,\delta,\nu}$. The expectation exists when $\nu > 1$, the variance exists when $\nu > 2$, and generally the n-th central moment exists when $\nu > n$.

There are several processes with the Student marginal distribution (Heyde and Leonenko 2005), of which we chose the Lèvy driven Ornstein-Uhlenbeck (OU)-type process as the process with the least restrictive potential parameter range. We have considered Student diffusion process (Leonenko and Suvak 2010) as another possible model. However, the mean reversion term needs to be included in the stochastic differential equation (SDE) defining a stationary Student diffusion process. Thus the mean has to exist, restricting the parameter range to $\nu > 1$. The parameter estimates based on these data did not support such restriction across all channels. The Student OU type process provides a flexible model for the increments of the EEG data, and is the solution of SDE

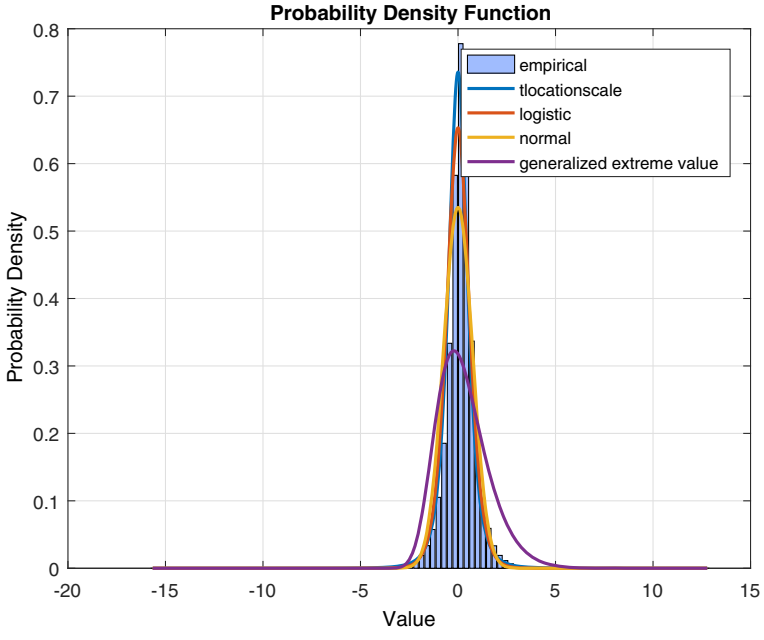

Fig. 1 Alpha frequency band

$$dX(t) = -\lambda X(t)dt + dY(\lambda t), \quad t \geq 0 \tag{2}$$

where $\{Y(t),\ t \geq 0\}$ is the background driving Lévy process (BDLP). This process is defined by specifying the distribution of $Y(1)$ through the cumulant transform

$$\log e^{i\zeta Y(1)} = i\zeta\mu - \delta|\zeta|\frac{K_{\nu/2-1}(\delta|\zeta|)}{K_{\nu/2}(\delta|\zeta|)}, \quad \zeta \in \mathbb{R},\ \zeta \neq 0, \tag{3}$$

where K_s is the modified Bessel function of the third kind:

$$K_s(x) = \frac{1}{2}\int_0^\infty u^{s-1}\exp\left[\frac{1}{2}x\left(u+\frac{1}{u}\right)\right]du,\ x > 0,\ s \in \mathbb{R}.$$

Since the Student distribution is self-decomposable, the distribution of $Y(1)$ is infinitely divisible, and the Student OU-type process exists, as formally stated in the following Theorem (Heyde and Leonenko 2005).

Theorem 1 *There exists a strictly stationary stochastic process* $\{X(t),\ t \geq 0\}$ *that has the marginal* $T_{\mu,\delta,\nu}$ *distribution with the density function (1). The process solves SDE (2) for any* $\lambda > 0$. *The BLDP* $\{Y(t),\ t \geq 0\}$ *has the cumulant transform (3). The solution is given by*

$$X_t = e^{-\lambda t}X_0 + e^{-\lambda t}\int_0^t e^{\lambda s}dY(\lambda s).$$

Note that parametrization in (2) is such that the marginal distribution of the Student OU-type process $X(t)$ does not depend on λ. If $\nu > 1$, then the first moment of the marginal distribution exits, and $\mathbb{E}X(t) = \mu$. If $\nu > 2$, then the correlation function exists and depends only on the parameter λ, namely for $0 < s < t$

$$\mathrm{corr}(X(s), X(t)) = e^{-\lambda(t-s)}.$$

In this work we used the first-order properties of the process $X(t)$ reflected by the parameters of the marginal distribution. We set $\mu = 0$ based on the empirical evidence as all histograms were centered on zero with virtually no variability in this respect among EEG channels. We used quasi-likelihood estimation (Heyde 1997) to evaluate the parameters ν and δ of the stationary Student process. It was possible to use this approach because of the specification of the distribution. A general method for estimation of the tail parameter (ν in this case) without specifying the marginal distribution is discussed in Grahovac et al. (2015) and could be used in other applications. Quasi-likelihood estimation was performed for the lowest four frequency bands for 15 channels, resulting in 120 features. Additional 30 features were derived by repeating the parameter estimation for 15 channels delta frequency band containing excess zeros potentially due to the disconnection of the electrode from the skin. Note that in this set of features there is no CZ-CZ channel at all. The estimated 150 parameters were used as features in the matrix of potential explanatory variables for the neurodevelopment and cognition 6 months post CM illness. We refer to these features as stochastic features since they were derived based on the proposed stochastic process model.

4 Data analysis

Three feature matrices X with 78 rows each, one containing 54 non-EEG features (columns), the second containing 362 traditional EEG features, and the third containing 150 new stochastic EEG features, were prepared as described above for the entry into

a machine learning algorithm to predict the neurodevelopmental and cognitive scores of children surviving cerebral malaria. Out of 54 columns for non-EEG features 20 had at least 19 missing values. The maximum number of empty entries in a column of the non-EEG part of the feature matrix was 23. The pattern of missing data could not be assumed to be missing at random (MAR), because some of the data were missing due to clinical reasons. Therefore we used imputation method that is not dependent upon the MAR assumption. Soft Impute is a matrix completion method based on Singular Value Decomposition (SVD) of a matrix (Mazumder et al. 2010). Application of this algorithm requires the following assumption to hold: the rank of the approximating matrix $rank(Z) << min(n,p)$, which is reasonable in our case. This assumption makes sense due to the nature of features, which may be grouped by correlation into a smaller number of clusters due to inherent synchrony between channels. We present the central lemma behind this method for completeness.

Lemma 1 *Suppose the matrix $W_{m\times n}$ has rank r. The solution to the optimization problem*

$$\min_{Z}\left(\frac{1}{2}\|W-Z\|_F^2+\lambda\|Z\|_*\right) \tag{4}$$

is given by $\hat{Z}=S_\lambda(W)$, where $S_\lambda(W)=UD_\lambda V^T$, with $D_\lambda=diag[\,(d_1-\lambda,\cdots,(d_r-\lambda)_+]$, where UDV^T is the SVD of W, $D=diag[\,d_1,\cdots,d_r]$, and $t_+=\max(t,0)$.

Here $\|A\|_F$ is the Frobenius norm of a matrix A, whilst $\|A\|_*$ is the sum of the singular values of the matrix A. In our case, in computing the Frobenius norm in (4) we only look at the pairs of indices (i,j), for which there are no missing values. The algorithm implemented in SoftImpute iteratively updates the matrix Z through the use of this lemma, until convergence is reached in approximating the matrix of interest. R was used for the imputation of missing values.

Machine learning was performed using the Elastic Net technique that solves the following optimization problem:

$$\min_{\bar{\beta}}\left(\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\alpha\times l_1\sum_{j=1}^{p}|\beta_j|+0.5\times\alpha\times(1-l_1)\sum_{j=1}^{p}\beta_j^2\right), \tag{5}$$

Elastic Net was chosen because it is more suitable than the Least Absolute Shrinkage and Selection Operator (LASSO) in face of multicollinearity (Oyeyemi et al. 2015), which we expected among the selected features. There are two hyper-parameters, and leave-one-out cross validation (LOOCV) was used to select the optimal pair. We have also used LOOCV for the estimation of the mean squared error in the prediction models produced by the algorithm. The Elastic Net algorithm was implemented using Python 3 with Anaconda.

The non-parametric missForest technique for matrix completion (Stekhoven and Bühlmann 2012) is based on averaging over a random forest of regression trees and was used for additional validation of the results. Trees were built based on observed and bootstrapped parts of the training data set.

5 Results

5.1 Imputations

We have run over a grid of different parameter values for α, considering powers of 10, starting with $\alpha = 0.00001$ and finishing with $\alpha = 1000$, whilst for the l_1 ratio coefficient we considered 7 possible values starting with 0.0001 and finishing with 1. In all models, setting the regularization parameter $\lambda = 100$ in SoftImpute gave better mean squared error (MSE) than other values of λ and resulted in the least complex model, after applying the Elastic Net following the matrix completion. This value is used in reporting of the results.

5.2 Results from elastic net for three sets of features

Table 2 summarizes the results of applying the regression methods listed in the previous section.

The lowest LOOCV MSE of 0.15 was obtained with $\alpha = 0.001$, l_1 ratio$= 0.5$ in the objective function subject to minimization:

$$\frac{1}{2*78}\|y - Xw\|_2^2 + \alpha * \text{l1 ratio} * \|w\|_1 + 0.5 * \alpha * (1 - \text{l1 ratio}) * \|w\|_2^2 \tag{6}$$

with leave-one-out cross-validation. This combination of α and l_1 ratio was the best across combinations described in the Section 4.

Similarly to the result for the matrix with 150 features after SoftImpute, using the random forest technique for matrix completion and again, those features in the top 5 in feature importance by Breiman were for the channels FP2-CZ, O1-CZ, F7-CZ, T6-CZ, whilst F8-CZ and T6-CZ appeared most often in the top 12 non-zero coefficients by absolute value. So this result almost copies the outcome with the Elastic Net after SoftImpute

Table 2 Elastic Net - best results, after Soft Impute with $\lambda = 100$

Feature set	LOOCV MSE	Number of nonzero coefficients	Sample features with non-zero coefficients
54 anthropomentic, socio-economic, and medical non-EEG features	0.3982	15	Weight, hemoglobin level, weight, BCS, the HOME score, white blood cell count, cerebrospinal fluid levels of interleukin (IL)-1 receptor antagonist (RA), IL-6, RANTES (an acronym for Regulated on Activation, Normal T Expressed and Secreted), IL-8, and plasma levels of vascular endothelial growth factor and von Willebrand factor.
362 traditional EEG features in frequency bands	0.5285	62	Features include: fp2 in T6-CZ, FZ-CZ and P4-CZ, fp1 F4-CZ, wave energy for theta in channel T5-CZ, variance in theta for FP2-CZ, variance in alpha frequency band for T5-CZ.
150 stochastic features for increment processes in frequency bands	0.1511	85	12 coefficients have the absolute values over 0.5. Top 5 coefficients are for the channels FP2-CZ, O1-CZ, T6-CZ, FZ-CZ. Channels F8-CZ and T6-CZ appear most often among the 12 top coefficients by absolute value, 4 and 3 times respectively. Four of the top 12 are for fitting a stochastic process model for the a7 frequency band.

with $\lambda = 100$ and yields a marginally different Elastic Net LOOCV MSE value. This was anticipated and confirms validity of the matrix completion method for such data.

6 Conclusions

We conclude that stochastic modeling brings a noticeable improvement in explaining the variation in neurodevelpmental and cognitive outcomes of children 6 months after surviving cerebral malaria. Stochastic features alone do an even better job than the tested sets of medical non-EEG or traditional EEG features which are not based on stochastic models for the underlying time series.

Regarding medical non-EEG biomarkers, only tumor necrosis factor alpha (TNF-alpha) in cerebrospinal fluid but not in plasma was predictive of 6-month later cognitive scores of children older than 5 years, but not of neurodevelopment of younger children (Shabani et al. 2017). So our finding of TNF-alpha not being among top predictors is in line with the developing literature on the role of biomarkers collected at the time of acute illness in predicting later neurodevelopment and cognition in children. This paper extends the results of Shabani et al. (2017) to testing more than one biomarker using modern statistical and probabilistic methods. We have identified biomarkers that can be further considered in future research as potentially important prognostic factors for neurodevelopment and cognition.

Regarding the traditional EEG features, their performance in explaining the variation in neurodevelopmental and cognitive outcomes was inferior to that of anthropometric, socio-economic, and non-EEG medical features. This finding may be due to the fact that these features are not based on underlying stochastic models. For example, the traditional computation of Shannon's entropy build into software assumes that the underlying stochastic process is stationary, which could be reasonable in some populations (Piryatinska et al. 2009), but is at odds with other literature (Ignaccolo et al. 2009). Whether or not the underlying stochastic process is stationary in a given population is an empirical question that needs to be addressed in methodology of analyzing EEG data. For the population of Ugandan children in coma from cerebral malaria, we have found that the assumption of stationarity of the time series was unreasonable, while for the increment process it was. Further, stochastic modeling for the increment process had clear advantages, as seen from our results.

When considering the stochastic features for the increment process, the combination of channels for which the stochastic features proved to be particularly useful is FP2-CZ, 01-CZ, T6-CZ, FZ-CZ. Parameters from channels located on the right side dominated the most important features. Taken as a group, these locations are associated with visual and attention processes that are related to visual-spatial simultaneous processing working memory. It would be interesting to see if these channels also arise as important in relation to child neurodevelopment and cognition in other infectious diseases that could affect the brain.

Abbreviations

BCS: Blantyre coma score; CM: Cerebral malaria; Dbi: Daubechies wavelet with *i* vanishing moments; D1,. . ., D7, a7: Frequency bands from the highest gamma (D1) to delta (a7), see the table; EEG: Electroencephalogram; fp1/3, fp1, fp2 and fp3: Frequencies of peaks in the original cleaned time series differing from the nearest measurements from both sides by 1/3, 1, 2 and 3 standard deviations were calculated; HOME: Home observation for the measurement of the environment; Hz: Hertz; KABC-II: Kaufman assessment battery for children, second edition; LOOCV MSE: Leave-one-out cross-validation mean squared error; MAR: Missing at random; MPI: Mental processing index; MSEL: Mullen scales of early learning; SDE: Stochastic differential equation; SES: Socioeconomic status; SMA: Severe malarial anemia; SVD: Singular value decomposition; TNF-alpha: Tumor necrosis factor alpha

Acknowledgments
The authors thank C. C. John and D. Postels for the acquisition and provision of data, as well as G. Shtekh and A. Hall for fruitful discussions and advice with programming. We also thank the company Persyst for providing a trial copy of their software.

Funding
Data collection was supported by the National Institutes of Health (NIH) grant R01 NS055349 (PI: C. John). For M. Veretennikova the study has been funded by the Russian Science Foundation (project number 17–11-01098). M. J. Boivin and A. Sikorskii were supported by the grant NIH R01 HD064416 (PI: M. Boivin).

Availability of data and materials
Matrix of EEG and medical features (without identifiers) and code can be requested by contacting Maria Veretennikova at mveretennikova@hse.ru.

Authors' contributions
Conceptualization (MV, AS, MJB), data curation (MV, AS, MJB), data analysis (MV), funding acquisition (MJB), investigation (MV, AS, MJB), methodology (MV, AS, MJB), manuscript writing and editing (MV, AS, MJB). All authors read and approved the final manuscript.

Competing interests
None of the authors have any competing interests.

Publisher's Note

Author details
[1]Department of Statistics and Data Analysis, Faculty of Economic Science, National Research University, Higher School of Economics, Shabolovka 28/11, 9, 19049 Moscow, Russia. [2]Department of Psychiatry and Department of Statistics and Probability, Michigan State University, 909 Wilson Road, 48824, East Lansing MI, USA. [3]Department of Psychiatry and Department of Neurology and Ophtalmology, Michigan State University, 909 Wilson Road, 48824 East Lansing MI, USA.

Received: 11 June 2018 Accepted: 20 September 2018
Published online: 29 December 2018

References

Awal, M. A., Lai, M. M., Azemi, G., Colditz, P. B.: EEG background features that predict outcome in term neonates with hypoxic ischaemic encephalopathy: A structured review. Clin. Neurophysiol. **125**(1), 285–296 (2016)
Bangirana, P., Opoka, R. O., Boivin, M. J., Idro, R., Hodges, J. S., John, C. C.: Neurocognitive domains affected by cerebral malaria and severe malarial anemia in children. Learn. Individ. Differ. **46**, 38–44 (2016)
Bangirana, P., Seggane, M., Allebeck, P., Giordani, B., John, C. C., Byarugaba, J., Ehnvall, A., Boivin, M. J.: A preliminary investigation of the construct validity of the KABC-II in Ugandan children with prior cerebral insult. Afr. Health Sci. **9**(3), 186–192 (2009)
Bradley, R. H., Caldwell, B. M.: Home observation for measurement of the environment. University of Arkansas Press, Little Rock (1979)
Chen, Z., Cao, J., Cao, Y., Zhang, Y., Gu, F., Zhu, G., Hoong, Z., Wang, B., Cichocki, A.: An empirical EEG analysis in brain death diagnosis for adults. Cogn. Neurodyn. **2**(3), 257–271 (2008)
Daubechies, I.: Ten Lectures on Wavelets. SIAM, Philadelphia (1992)
De Oliveira, H.: Shannon and Renyi Entropy of Wavelets. International Journal of Mathematics and Computer Science. **10**, 13–26 (2015)
Deng, R., Koenig, M. A., Young, L. M., Jia, X.: Early Quantitative Gamma-Band EEG Marker is Associated with Outcomes After Cardiac Arrest and Targeted Temperature Management. Neurocrit. Care. **23 2**, 262–73 (2015)
Duncan, D., Talmon, R., Zaveri, H., Coifman, R.: Identifying preseizure state in intracranial EEG data using diffusion kernels. Math. Biosci. Eng. **10**, 579–90 (2013)
Fraschini, M., Demuru, M., Crobe, A., Marrosu, F., Stam, CJ., Hillebrand, A.: The effect of epoch length on estimated EEG functional connectivity and brain network organisation. J. Neural Eng. **13**, 036015 (2016)
Grahovac, D., Jia, M., Leonenko, N. N., Taufer, E.: Asymptotic properties of the partition function and applications in tail index inference of heavy-tailed data. Statistics. **49**(6), 1221–1242 (2015)
Heyde, C. C.: Quasi-likelihood and its application: A general approach to optimal parameter estimation. Springer, New York (1997)
Heyde, C. C., Leonenko, N. N.: Student processes. Adv. Appl. Probab. **37**, 342–365 (2005)
Hjorth, B.: EEG analysis based on time domain properties. Electroencephalography and Clinical Neurophysiology. **29**(3), 306–310 (1970)
Idro, R., Marsh, K., John, C. C., Newton, C. R.: Cerebral malaria; mechanisms of brain injury and strategies for improved neuro-cognitive outcome. Pediatr. Res. **68**(4), 267–274 (2010)
Ignaccolo, M., Latka, M., Jernajczyk, W., Grigolini, P., West, B.: The dynamics of EEG entropy. J. Biol. Phys. **36**, 185–96 (2009)
John, C. C., Bangirana, P., Byarugaba, J., Opoka, R. O., Idro, R., Jurek, A. M., Wu, B., Boivin, M. J.: Cerebral malaria in children is associated with long-term cognitive impairment. Pediatrics. **122**(1), e92-99 (2008)
Juan, E., Kaplan, P. W., Oddo, M., Rossetti, A. O.: EEG as an indicator of cerebral functioning in postanoxic coma. J. Clin. Neurophysiol. **32**, 465–47 (2015)
Kaufman, N. L., Kaufman, A. S.: Manual for the Kaufman AssessmentBbattery for Children. 2nd. American Guidance Service Publishing/Pearson Products Inc., Circle Pines (2004)

Kirch, C., Muhsa, B., Ombao, H.: Detection of Changes in Multivariate Time Series With Application to EEG Data. J. Am. Stat. Assoc. **110**(511), 1197–1216 (2015). Taylor and Francis, https://doi.org/10.1080/01621459.2014.957545
Leonenko, N. N., Suvak, N.: Statistical inference for Student diffusion process. Stoch. Anal. Appl. **28**(6), 972–1002 (2010)
Li, L., Wilton, A., Marcora, S., Bowman, H., Mandic, D. P.: EEG-based brain connectivity analysis of states of awareness. **1002–5** (2014)
Makeig, S., Delorme, A.: EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics. J. Neurosci. Methods. **134**(1), 9–21 (2004)
Malagurski, B., Péran, P., Sarton, B., Riu, B., Vardon-Bounes, F., Seguin, T., Geeraerts, T., Fourcade, O., de Pasquale, F., Silva, S.: Neural signature of coma revealed by posteromedial cortex connection density analysis. NeuroImage Clinical. **15**, 315–324 (2017)
MATLAB wavelet packet. https://www.mathworks.com/help/wavelet/ref/wentropy.html. Accessed 21 Sept 2018
Mazumder, R., Hastie, T., Tibshirani, R.: Spectral Regularization Algorithms for Learning Large Incomplete Matrices. J. Mach. Learn. Res. **11**, 2287–2322 (2010)
Montgomery, D. C.: Design of Experiments. 8th. Wiley, Hoboken (2012)
Mullen, E. M.: Mullen Scales of Early Learning. American Guidance Service, Circle Pines (1995)
Oyeyemi, G. M., Ogunjobi, E. O., Folorunsho, A. I.: On performance of shrinkage methods - a Monte Carlo study. Int. J. Stat. Appl. **5**(2), 72–76 (2015)
Piryatinska, A., Terdik, G., Woyczynski, W. A., Lopar, K. A., Scher, M. S., Zlotnik, A.: Automated detection of neonate EEG sleep stages. Comput. Methods Prog. Biomed. **95**(1), 31- 46 (2009)
Shabani, E., Ouma, B. J., Idro, R., Bangirana, P., Opoka, R. O., Park, G. S., Conrov, A. L., John, C. C.: Elevated cerebrospinal fluid tumour necrosis factor is associated with acute and long-term neurocognitive impairment in cerebral malaria. Parasite Immunol. **39**(7) (2017). https://doi.org/10.1111/pim.12438. Epub 2017 May 28
Smolentsev, N. K.: Fundamentals of the theory of wavelets. Wavelets in MATLAB, DMK Publishing Press, Moscow (2014)
Stekhoven, D., Bühlmann, P.: MissForest - Non-parametric missing value imputation for mixed-type data. Bioinformatics (Oxf. Engl.) **28**, 112–118 (2012)
Temko, A., Thomas, E., Marnane, W., Lightbody, G., Boylan, G.: EEG-based neonatal seizure detection with Support Vector Machines. Clin. Neurophysiol. **122**(3), 464–473 (2011)
Teplan, M.: Fundamentals of EEG Measurement. Meas. Sci. Rev. **2**(2), 1–11 (2002)
Uhlhaas, P., Pipa, G., Lima, B., Melloni, L., Neuenschwander, S., Nikolić, D., Singer, W.: Neural synchrony in cortical networks: history, concept and current status. Front. Integr. Neurosci. **3**, 17 (2009)
Wold Health Organization Growth Standards (2009). https://www.who.int/growthref/en/. Accessed 20 Jan 2017
WHO: Tropical Medicine and International Health. Wiley, Hoboken (2014). 19 (Suppl. 1)