HAL open science

Prediction and detection of human epileptic seizures based on SIFT-MS chemometric data

Amélie Catala, Cecile Levasseur-Garcia, Marielle Pagès, Jean-Luc Schaff, Ugo Till, Leticia Vitola Pasetto, Martine Hausberger, Hugo Cousillas, Frederic Violleau, Marine Grandgeorge

To cite this version:

Amélie Catala, Cecile Levasseur-Garcia, Marielle Pagès, Jean-Luc Schaff, Ugo Till, et al.. Prediction and detection of human epileptic seizures based on SIFT-MS chemometric data. Scientific Reports, 2020, 10 (1), pp.18365. 10.1038/s41598-020-75478-8 . hal-02985868

HAL Id: hal-02985868

https://hal.inrae.fr/hal-02985868

Submitted on 23 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

scientific reports

OPEN

Prediction and detection of human epileptic seizures based on SIFT-MS chemometric data

Amélie Catala[1,2]✉, Cecile Levasseur-Garcia[3], Marielle Pagès[4], Jean-Luc Schaff[5], Ugo Till[6], Leticia Vitola Pasetto[3], Martine Hausberger[2], Hugo Cousillas[2], Frederic Violleau[3,7] & Marine Grandgeorge[2,7]

Although epilepsy is considered a public health issue, the burden imposed by the unpredictability of seizures is mainly borne by the patients. Predicting seizures based on electroencephalography has had mixed success, and the idiosyncratic character of epilepsy makes a single method of detection or prediction for all patients almost impossible. To address this problem, we demonstrate herein that epileptic seizures can not only be detected by global chemometric analysis of data from selected ion flow tube mass spectrometry but also that a simple mathematical model makes it possible to predict these seizures (by up to 4 h 37 min in advance with 92% and 75% of samples correctly classified in training and leave-one-out-cross-validation, respectively). These findings should stimulate the development of non-invasive applications (e.g., electronic nose) for different types of epilepsy and thereby decrease of the unpredictability of epileptic seizures.

With an estimated 50 million people affected worldwide, epilepsy is recognized as a major public health concern[1]. It is characterized by repetitive, unpredictable seizures with the accompanying risk of injury and accident in addition to the associated psychosocial[2] and comorbidity factors[3–5]. This disease affects the everyday life of patients; for example, by restricting their ability to drive any motorized vehicle, by hindering their capacity to obtain and retain meaningful employment, and by producing memory impairment, anxiety, depression, or even suicide. Approximately 30% of people with epilepsy are pharmaco-resistant. The manifest unpredictability of epileptic seizures places an immense psychological burden on patients and substantially compromises their quality of life[6,7]. Accurate prediction of seizures would not only reduce the burden on caregivers but also significantly improve the quality of life of patients, in particular by lessening their psychological stress. In addition, predicting seizures could prevent trauma or life-threatening accidents. Thus, accurate and reliable prediction of seizures has the potential to revolutionize the treatment of epilepsy. Furthermore, given that the clinical phenomenology[8] of epileptic seizures varies widely, no generalized detection system exists to help people monitor their seizures[9]. Although electroencephalography is considered the gold standard for monitoring seizures, it remains difficult to use in everyday situations. Moreover, it generally offers a very short predictive time[10] (on average 19 s before the onset of seizure[11]), allowing patients and caregivers insufficient time to react appropriately[12].

A recent study showed that dogs could be trained to differentiate between the body odor emitted by patients during a seizure and the patients' normal body odor, independently of the type or etiology of the seizure[13]. This discovery reveals that epileptic seizures have an olfactory signature, general to clinical phenomenological variations, which may open new prospects for managing epilepsy. Inspired by this canine study, we evaluate herein the feasibility of seizure detection based on the analysis of volatile organic compounds (VOCs) emitted by epileptic patients and investigate the kinetics thereof.

We use mass spectrometry to analyze the VOC profiles of epileptic patients. Selected ion flow tube mass spectrometry (SIFT-MS) is a gas-phase analytical method based on soft chemical ionization reactions coupled with direct mass spectrometry[14]. This analytical technique provides results in real-time, making it particularly

[1]Association Handi'Chiens, 13 Rue de l'Abbé Groult, Paris, France. [2]Univ Rennes, Normandie Univ, CNRS, EthoS (Éthologie animale et humaine), UMR 6552, 35000 Rennes, France. [3]Laboratoire de Chimie Agro-industrielle (LCA), Université de Toulouse, INRA, University of Toulouse, National Polytechnic Institute of Toulouse, Ecole d'ingénieurs de Purpan, Toulouse, France. [4]Equipe Physiologie, Pathologie et Génétique Végétales (PPGV), University of Toulouse, National Polytechnic Institute of Toulouse, Ecole d'ingénieurs de Purpan, 75 voie du TOEC, BP 57611, 31076 Toulouse Cedex 03, France. [5]Service de Neurologie du CHRU de Nancy, 29, avenue du Maréchal de Lattre de Tassigny, Nancy, France. [6]Laval, France. [7]These authors contributed equally: Frederic Violleau and Marine Grandgeorge. ✉email: amelie.catala@gmail.com

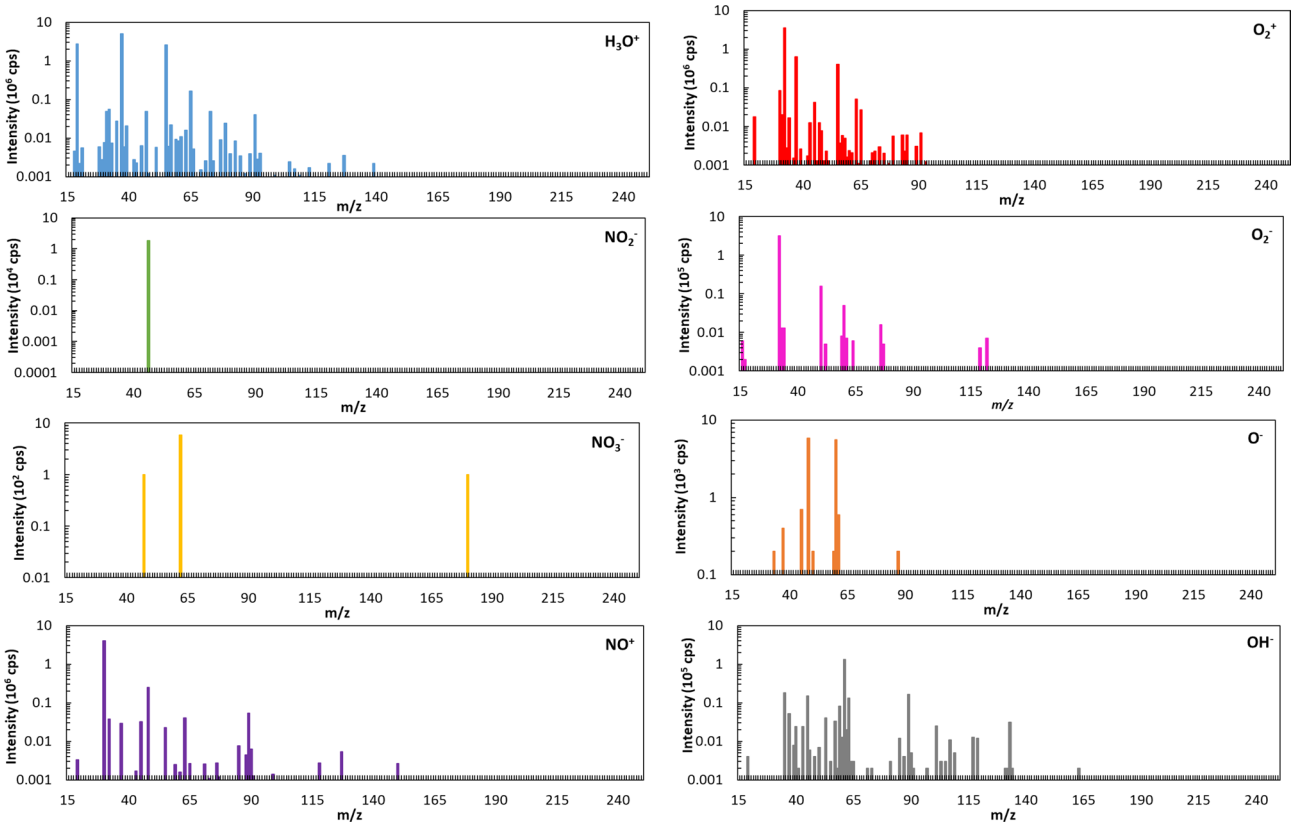

Figure 1. Example of SIFT-MS analysis of odor sample. These spectra reveal the ions generated by the ionization reaction of the sample with each of eight precursor ions (three positive ions: H_3O^+, NO^+, O_2^+, and five negative ions: NO_2^-, NO_3^-, O_2^-, HO^-, O^-).

appealing for use in a clinical setting. For some volatile compounds, the sensitivity of this method is better than parts per trillion[15]; it can also detect several target compounds simultaneously and can record global profiles (Fig. 1). We use a non-specific-recognition, learning-based model that involves relative comparisons between different modalities based on global VOC profiles.

This method has been applied previously in clinical studies to analyze exhaled breath and thereby distinguish between healthy patients and patients suffering from various conditions such as nonalcoholic fatty liver disease[16], oesophagogastric cancer[17], inflammatory bowel disease[18], or pulmonary arterial hypertension[19].

Results

Odor-emitting samples were collected from 14 persons with a confirmed diagnosis of epilepsy, providing a mix of different body odors (see Supplementary Information). To collect the samples, the patients used a sterile cotton pad to wipe their hands, forehead, and the back of their neck. The cotton pad was then placed into a zip-locked bag and the patient was instructed to exhale into the bag before sealing it. These samples were collected in five situations: (a) a "seizure sample," collected during or immediately after (< 5 min) an ictal event; (b) a "physical exercise sample," collected immediately after (< 5 min) moderate exercise; (c) an "$n-1$ sample," which was the last sample collected before a seizure occurred; (d) an "$n-2$ sample," which was the sample obtained just before the $n-1$ sample; and (e) an "inter-ictal sample," collected at least six hours before or after the seizure. All samples were analyzed by SIFT-MS, and the results were processed by chemometric analysis.

Classifying patients with epilepsy. A Welch's modified t-test was applied to the first four principal components and to 61 averaged samples. The only test that approached significance was the one constructed for the third principal component. The SIFT-MS physical exercise and inter-ictal spectra collected from subjects differed sufficiently from the seizure, $n-1$, and $n-2$ spectra to justify classifying them separately ($n=61$, $p=0.056$). Based on these results, the status of each patient was categorized either as "seizure-related" (i.e., seizure, $n-1$, or $n-2$ samples) or "non-seizure-related" (i.e., inter-ictal or physical exercise samples).

Seizure prediction based on classification tree analysis. The 1888 SIFT-MS variables served as input features for the models, and the two patient-status groups served as output. The model was based on a classification tree analysis of the 61 averaged samples. After optimization, the best model selected eight of the 1888 variables of the SIFT-MS spectra. This model correctly classified 92% and 75% of the samples based on the SIFT-MS spectra (see Table 1) in training and leave-one-out-cross-validation, respectively.

To From	Inter-ictal and physical exercise	Seizure, n-1, n-2	Total	Well-classified samples (%)
Training				
Inter-ictal and physical exercise	20	1	21	95.2 (Specificity)
Seizure, n-1, n-2	4	36	40	90 (Sensitivity)
Total	24	37	61	91.8
Cross-validation				
Inter-ictal and physical exercise	16	5	21	76.2 (Specificity)
Seizure, n-1, n-2	10	30	40	75 (Sensitivity)
Total	26	35	61	75.4

Table 1. Confusion matrix for the predicting condition of epileptic patients (classification-tree analysis) based on the SIFT-MS spectra. "Sensitivity" is defined as the percentage of seizure-related spectra that are well predicted by the model, whereas "specificity" is defined as the percentage of non-seizure-related spectra that are correctly rejected.

The SIFT-MS spectra of patients in an inter-ictal phase or during physical exercise were identified correctly by the algorithm in 20 of 21 cases (95.2%) and in 16 of 21 cases (76%) in leave-one-out cross validation. This rate is worth 90% (and 75% in cross-validation) for the "seizure-related" group.

The eight variables retained by the model are H_3O^+ 29 +; H_3O^+ 38 +; H_3O^+ 46 +; H_3O^+ 61 +; H_3O^+ 63 +; H_3O^+ 137 +; NO^+ 79 +, and O_2^+ 77 +. Of the eight precursor ions tested by SIFT-MS, H_3O^+, NO^+, and O_2^+ seemed to be relevant to discriminate between the seizure status of epileptic patients.

Discussion

The results of this classification support the hypothesis that ictal samples differ from inter-ictal and physical exercise samples, which implies that a specific odor, or VOC profile, is associated with epileptic seizures, as suggested by the response of trained dogs to the odor of epileptic patients[13]. These results also indicate that the VOC profile does not depend on the type or etiology of epileptic seizures, reinforcing the hypothesis of an olfactory signature of epileptic seizures.

The rate of correct classification (92.5% of the samples in training and 75% in leave-one-out-cross-validation see Table 1).

Cross-validation has been implemented to ensure that our model does not suffer from overlearning, and that it will be able to make predictions on new data. The number of observations is low, which is why leave-one-out cross-validation is best suited. However, in a leave-one-out cross-validation, very similar training sets are formed, and very different test sets are formed. We will have almost the same model on each fold, and yet the quality of the predictions may vary a lot. Therefore it will be necessary to increase the number of observations to make the model more robust and to be able to validate it with an external data set or a k-folds cross validation.

This result is amongst the highest yet reported[10,11] not only in terms of prediction delay but also in terms of sensitivity and specificity. However, these results are based on VOC profiles instead of on the more classical machine learning techniques applied to electroencephalography spectra. In addition, the results for detection of epileptic seizures are general to all types of seizures or epilepsy etiologies tested and not just to a specific type of seizure or to a precise cerebral location of the seizure onset. This is also one of the reasons why these performances are very encouraging.

The data for $n-2$ samples are more similar to the data for $n-1$ and seizure samples than to the data for inter-ictal and physical exercise samples, which suggests that this chemometric method may be used to predict seizures based on $n-2$ samples. Given that the median time before seizure for $n-2$ samples was 277 min, this result implies that seizures can be predicted over 4 h in advance with this method.

Although other methods[12,20] have given good predictive results (> 90% for some), they are often too complex to process because they can involve significant data processing or *t*RNA expression analysis. In addition, they are usually invasive, requiring an implant, blood sampling, or the collection of other bodily fluids.

The limitations of this study include the small sample size, which was due to practical considerations. A larger sample would improve the learning capacity and validation of the model. In addition, because our sampling method was based on a whole-body approach, we could not determine the precise origin of the VOCs. Further research is thus required to determine the source of VOCs (e.g., skin, breath, sweat) or the possible involvement of sweat glands such as eccrine or apocrine glands. Additionally, we still do not know which biomarker underlies these findings because the pattern of VOCs that produces the VOC signature has yet to be defined.

In conclusion, although further research is needed to fully understand the mechanism of these findings, this first proof-of-concept study shows that epileptic seizures can be detected and predicted simply based on VOC analyses with a whole-spectrum approach. This method suggests that epileptic seizures can be predicted sufficiently in advance to be useful both for people with epilepsy and their caretakers.

Altogether, these results represent an opportunity to improve the life of people with epilepsy, independently of the patients' demographics, type of seizure, or etiology. Furthermore, although this proof of concept is demonstrated herein by using SIFT-MS, these results pave the way for the development of monitoring technologies

for everyday life. In fact, the increasing progress in wearable electronic noses should allow efficient, noninvasive sensors to be developed in a few years to finally guarantee the safety of epileptic patients.

Materials and methods

Ethics. This study was approved by the institutional review board of the OHS (Office of Social Hygiene) of Lorraine (France) for the collection techniques used to obtain odor samples from patients with epilepsy. The present research was noninvasive and did not involve pharmacological interventions. Thus, in accordance with the Ethics Committee's guidelines, the adult patients, or guardians in the case of a minor, were only required to give informed written consent to allow their own or their child's participation in the experiment prior to their inclusion in the study. Thus, Written Informed consent was obtained from participants and from guardians in the case of minors. The National Centre for Scientific Research (CNRS) Data Protection Official was consulted for this study and confirmed the validity of this approach. Information sanitization was fully anonymous.

Odor samples. The patients recruited were from the medical and education institute (MEI) of the OHS Flavigny, Flavigny sur Moselle, France, which is a clinic for developmentally disabled children. Twenty patients were initially invited to participate after medical recommendation. The sampling occurred from March to May 2019. Six participants declined the invitation. Thus, odors were collected from 14 patients (7 women, 7 men) with ages varying from 10 to 43 years (mean: 27.7 ± 11.31 years old). All had a confirmed diagnosis of epilepsy (see Supplementary Information, Table 1) and none experienced psychogenic non-epileptic seizures. Psychogenic non-epileptic seizures, whether exclusive or in addition to epileptic seizures, were a factor of exclusion when recruiting patients on the basis of their medical diagnosis. Since all patients lived in the same MEI at the time of sampling, the diagnosis was further confirmed by the medical personnel. As in a previous study[13], all patients received the same food without dietary restrictions. Since the seizures could occur at any time, the collection of epileptic seizure samples was random (night or day), whereas all other samples were collected during the day. To maintain a normal inter-individual variability corresponding to all types of habits reflective of everyday living conditions, patients were given no specific recommendation regarding hormonal contraception, perfume, smoking, etc.

Odor-sampling procedure. The sampling procedure consisted in collecting several samples per day, with a three-hour interval between each sample. This method allowed us to collect five samples per day, from approximately 7 a.m. to 10 p.m., during inter-ictal or pre-ictal states. In addition, two types of samples were collected: (a) a "seizure sample," collected during or immediately after (< 5 min) an ictal event and (b) a "physical exercise sample," collected immediately after (< 5 min) moderate exercise. For this study, "moderate exercise" was defined as a 1 min interval during which the patient runs approximately 30 m and performs leg and arm movements such as step-touch and punches. Physical exercise samples served as control for movements or arousal, which can occur during a seizure[13,21,22].

The five samples collected per day were divided into three types defined a posteriori: (c) the "$n-1$ sample" is the last sample collected before a seizure occurs, (d) the "$n-2$ sample" is the penultimate sample before a seizure occurs, and (e) the "inter-ictal sample" is collected at least 6 h before or after a seizure (to avoid pre- or post-ictal collection). For each type of sample, only a single sample was analyzed per patient. The average time between the $n-1$ sample and a seizure was 206 min (standard deviation of 11.8 min). If the $n-1$ and the seizure samples were acquired on two different days, they were considered extreme cases and discarded. The median time between the collection of $n-1$ ($n-2$) samples and the collection of seizure samples was 85 (277.5) minutes.

The sampling ended when a seizure occurred if a physical exercise sample had been collected before the seizure; otherwise, the physical exercise sample was collected the following day to avoid contaminating the sample by seizure odor. Thus, sampling lasted between one and five days, depending on the patient.

Odors were collected in accordance with the procedure used in previous studies:[13,23] patients were instructed to use a sterile cotton pad (5×5 cm^2, four-fold) to wipe their hands, forehead, and the back of their neck, allowing a multiplicity of odor origins. The cotton pad was then placed into a zip-locked bag (Ziplock brand, SC Johnson, Racine, WI, USA) and the patient was instructed to exhale into the bag before sealing it. Each bag was labeled with the patient's code, the date, and time of collection. The samples were stored in a refrigerator at 4 °C until use (for an average of 60 days with a standard deviation of 9.7 days).

Selected ion flow tube mass spectrometry. The method used in this study was based on that of Pasetto et al.[24].

In the SIFT-MS device, the eight precursor ions were produced by microwave discharge. A single precursor ion was selected at a time by a first quadrupole mass spectrometer and then injected into a reaction chamber by flowing nitrogen (180 N mL min^{-1}). The sample, heated to 37 °C for 10 h in a controlled-temperature oven, was introduced by a calibrated capillary (20 N mL min^{-1}) into the reaction chamber, which was maintained at 115 °C and 0.07 kPa. The product ions generated from the ionization reaction and the precursor ions were quantified by a second quadrupole mass spectrometer[25]. In the full-mass mode, the second quadrupole mass spectrometer scanned over a large mass range (from 15 to 250 m/z) and calculated a count rate (signal intensity in counts per second) for each unit of m/z. This resulted in a full profile for each of the eight precursor ions, with 236 mass peaks summed for each precursor ion. Altogether, 1888 mass peaks were recorded.

Four full mass scans were recorded for each sample. None of the first scans were used for analysis because they served to purge the system between samples. The other three scans were averaged and used for further analyses.

Statistical analyses. The approach proposed here was based on applying the chemometrics not of individual peaks, but of the entire spectra, as typically done with infrared spectroscopy[26]. This study thus focused on analyzing a global fingerprint.

First, we applied a principal component analysis (PCA), which is an orthogonal transformation, to convert the set of 1888 possibly correlated SIFT-MS variables into a set of linearly uncorrelated variables called principal components (PCs). PCA is defined so that the first PC accounts for the maximum possible variability in the SIFT-MS spectra, with the subsequent PCs accounting for less and less variability[27]. To develop the predictive model, we used a PCA to reduce the multidimensionality of the SIFT-MS spectral data[28]. The PCA was computed by using The Unscrambler (v. X; CAMO A/S, Oslo, Norway).

We selected the first PCs to account for 95% of the initial variability of the SIFT-MS spectra. The PCs were averaged for the triplicated analysis and then compared by using Welch's modified *t*-test[29] regarding the patients' state, as represented by the inter-ictal, physical exercise, seizure, n – 1 and *n* – 2 samples. The analysis was conducted by using Minitab (Minitab Inc, Statistical Software version 19, State College, PA, USA). The null hypothesis is that the average value of the dependent variable is the same for all patient states.

A C&RT classification tree analysis was then used to predict each patient's state. The entropy measure served as a quality measure to split a node. This predictive model was developed from the 1888 variables of the SIFT-MS spectra by using Minitab version 19.2020.1. (Minitab Inc. PA, USA). The model was challenged with a leave-one-out-cross-validation. The quality of the model was evaluated by calculating classification errors and prediction accuracy from a confusion matrix[30,31].

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 7 February 2020; Accepted: 16 October 2020
Published online: 27 October 2020

References

1. World Health Organization. Epilepsy. *World Health Organization website* https://www.who.int/news-room/fact-sheets/detail/epilepsy (2020).
2. O'Donoghue, M. F., Goodridge, D. M., Redhead, K., Sander, J. W. & Duncan, J. S. Assessing the psychosocial consequences of epilepsy: a community-based study. *Br. J. Gen. Pract.* **49**, 211–214 (1999).
3. Devinsky, O. Psychiatric comorbidity in patients with epilepsy: implications for diagnosis and treatment. *Epilepsy Behav.* **4**, 2–10 (2003).
4. Hermann, B. P., Seidenberg, M. & Bell, B. Psychiatric comorbidity in chronic epilepsy: identification, consequences, and treatment of major depression. *Epilepsia* **41**, S31–S41 (2000).
5. Piazzini, A., Canevini, M. P., Maggiori, G. & Canger, R. Depression and anxiety in patients with epilepsy. *Epilepsy Behav.* **2**, 481–489 (2001).
6. French, J. A. *et al.* Efficacy and tolerability of the new antiepileptic drugs II: treatment of refractory epilepsy: report of the Therapeutics and Technology Assessment Subcommittee and Quality Standards Subcommittee of the American Academy of Neurology and the American Epilepsy Society. *Neurology* **62**, 1261–1273 (2004).
7. Schulze-Bonhage, A. & Kühn, A. Unpredictability of seizures and the burden of epilepsy. In *Seizure Prediction in Epilepsy: From Basic Mechanisms to Clinical Applications* 1–10 (2008).
8. Fisher, R. S. *et al.* Operational classification of seizure types by the International League Against Epilepsy: position paper of the ILAE Commission for Classification and Terminology. *Epilepsia* **58**, 522–530 (2017).
9. Ulate-Campos, A. *et al.* Automated seizure detection systems and their effectiveness for each type of seizure. *Seizure* **40**, 88–101 (2016).
10. Kuhlmann, L., Lehnertz, K., Richardson, M. P., Schelter, B. & Zaveri, H. P. Seizure prediction—ready for a new era. *Nat. Rev. Neurol.* **14**, 618–630 (2018).
11. Acharya, U. R., Hagiwara, Y. & Adeli, H. Automated seizure prediction. *Epilepsy Behav.* **88**, 251–261 (2018).
12. Cook, M. J. *et al.* Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study. *Lancet Neurol.* **12**, 563–571 (2013).
13. Catala, A. *et al.* Dogs demonstrate the existence of an epileptic seizure odour in humans. *Sci. Rep.* **9**, 4103 (2019).
14. Smith, D. & Španěl, P. Selected ion flow tube mass spectrometry (SIFT-MS) for on-line trace gas analysis. *Mass Spectrom. Rev.* **24**, 661–700 (2005).
15. Dummer, J. *et al.* Analysis of biogenic volatile organic compounds in human health and disease. *TrAC, Trends Anal. Chem.* **30**, 960–967 (2011).
16. Alkhouri, N. *et al.* Analysis of breath volatile organic compounds as a noninvasive tool to diagnose nonalcoholic fatty liver disease in children. *Eur. J. Gastroenterol. Hepatol.* **26**, 82–87 (2014).
17. Kumar, S. *et al.* Selected ion flow tube mass spectrometry analysis of exhaled breath for volatile organic compound profiling of esophago-gastric cancer. *Anal. Chem.* **85**, 6121–6128 (2013).
18. Dryahina, K. *et al.* Quantification of pentane in exhaled breath, a potential biomarker of bowel disease, using selected ion flow tube mass spectrometry. *Rapid Commun. Mass Spectrom.* **27**, 1983–1992 (2013).
19. Cikach, F. S. *et al.* Breath analysis in pulmonary arterial hypertension. *Chest* **145**, 551–558 (2014).
20. Hogg, M. C. *et al.* Elevation of plasma tRNA fragments precedes seizures in human epilepsy. *J. Clin. Invest.* **129**, 2946–2951 (2019).
21. King, J. *et al.* Dynamic profiles of volatile organic compounds in exhaled breath as determined by a coupled PTR-MS/GC-MS study. *Physiol. Meas.* **31**, 1169–1184 (2010).
22. King, J. *et al.* Isoprene and acetone concentration profiles during exercise on an ergometer. *J. Breath Res.* **3**, 027006 (2009).
23. Hardin, D. S., Anderson, W. & Cattet, J. Dogs can be successfully trained to alert to hypoglycemia samples from patients with type 1 diabetes. *Diabetes Therapy* **6**, 509–517 (2015).
24. Vitola Pasetto, L. *et al.* Aldehydes gas ozonation monitoring: Interest of SIFT/MS versus GC/FID. *Chemosphere* **235**, 1107–1115 (2019).
25. Španěl, P., Dryahina, K. & Smith, D. A general method for the calculation of absolute trace gas concentrations in air and breath from selected ion flow tube mass spectrometry data. *Int. J. Mass Spectrom.* **249–250**, 230–239 (2006).

26. Agelet, L. E. & Hurburgh, C. R. Jr. A tutorial on near infrared spectroscopy and its calibration. *Crit. Rev. Anal. Chem.* **40**, 246–260 (2010).
27. Jobson, J. D. *Applied Multivariate Data Analysis: Volume II: Categorical and Multivariate Methods* (Springer, Berlin, 2012).
28. Martens, H. & Naes, T. *Multivariate Calibration* (Willey, New York, 1989).
29. Welch, B. L. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika* **34**, 28–35 (1947).
30. Visa, S. & Ralescu, A. Issues in mining imbalanced data sets—a review paper. In *Proceedings of the Sixteen Midwest Artificial Intelligence and Cognitive Science Conference*, vol. 2005, 67–73 (sn, 2005).
31. Levasseur-Garcia, C., Couderc, C. & Tormo, H. Discrimination of lactic acid bacteria Enterococcus and Lactococcus by infrared spectroscopy and multivariate techniques. *J. Near Infrared Spectrosc.* **25**, 231–241 (2017).

Acknowledgements
We thank the Healthcare Centre OHS Flavigny, Flavigny sur Moselle, France, as well as its caretaker team and all the children and families involved. We are also thankful to Valérie Lemetter for her help with sample management, and to Katie Collier, Matthieu Bacconnier, and Ann Cloarec for their assistance with the English language.

Author contributions
A.C. and U.T. conceived the study; A.C., U.T., M.G., M.H., H.C., and F.V. designed the experiment; J.-L.S. collected the samples; A.C., M.P., and L.V.P. contributed to the acquisition of the dataset; C. L.-G. performed statistical analyses; A.C., U.T., M.G., F.V., C. L.-G., and M.P. contributed to the interpretation of the results; A.C. and U.T. wrote the first draft; and A.C., U.T., M.G., F.V., C. L.-G., M.P., M.H., and H.C. substantially revised the manuscript.

Funding
We had no specific funding for this study. Adrienne and Pierre Sommer Foundation, Handi'Chiens Association and ANRT provided Amélie Catala's salary and Rennes 1 University and CNRS provided material support. These organizations were not involved in the study.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-75478-8.

Correspondence and requests for materials should be addressed to A.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020