Decoding activity in Broca's area predicts the occurrence of auditory hallucinations across subjects

Thomas Fovet, Pierre Yger, Renaud Lopes, Amicie de Pierrefeu, Edouard Duchesnay, Josselin Houenou, Pierre Thomas, Sébastien Szaffarczyk, Philippe Domenech, Renaud Jardri

Biological Psychiatry

Archival Report

Decoding Activity in Broca's Area Predicts the Occurrence of Auditory Hallucinations Across Subjects

Thomas Fovet, Pierre Yger, Renaud Lopes, Amicie de Pierrefeu, Edouard Duchesnay, Josselin Houenou, Pierre Thomas, Sébastien Szaffarczyk, Philippe Domenech, and Renaud Jardri

ABSTRACT

BACKGROUND: Functional magnetic resonance imaging (fMRI) capture aims at detecting auditory-verbal hallucinations (AVHs) from continuously recorded brain activity. Establishing efficient capture methods with low computational cost that easily generalize between patients remains a key objective in precision psychiatry. To address this issue, we developed a novel automatized fMRI-capture procedure for AVHs in patients with schizophrenia (SCZ).

METHODS: We used a previously validated but labor-intensive personalized fMRI-capture method to train a linear classifier using machine learning techniques. We benchmarked the performances of this classifier on 2320 AVH periods versus resting-state periods obtained from SCZ patients with frequent symptoms (n = 23). We characterized patterns of blood oxygen level–dependent activity that were predictive of AVH both within and between subjects. Generalizability was assessed with a second independent sample gathering 2000 AVH labels (n = 34 patients with SCZ), while specificity was tested with a nonclinical control sample performing an auditory imagery task (840 labels, n = 20).

RESULTS: Our between-subject classifier achieved high decoding accuracy (area under the curve = 0.85) and discriminated AVH from rest and verbal imagery. Optimizing the parameters on the first schizophrenia dataset and testing its performance on the second dataset led to an out-of-sample area under the curve of 0.85 (0.88 for the converse test). We showed that AVH detection critically depends on local blood oxygen level–dependent activity patterns within Broca's area.

CONCLUSIONS: Our results demonstrate that it is possible to reliably detect AVH states from fMRI blood oxygen level–dependent signals in patients with SCZ using a multivariate decoder without performing complex preprocessing steps. These findings constitute a crucial step toward brain-based treatments for severe drug-resistant hallucinations.

https://doi.org/10.1016/j.biopsych.2021.08.024

Hearing distressing voices that other people do not [called auditory-verbal hallucinations (AVHs) (1)] becomes a therapeutic impasse for more than 30% of patients with schizophrenia (SCZ) (2). These complex sensory experiences are highly variable (3,4), making the characterization of their neurobiological basis especially challenging. This situation creates a technical challenge for people who need to detect/decode AVH states from brain activity for therapeutic purposes.

AVHs were first explored using functional magnetic resonance imaging (fMRI) trait studies, which compared SCZ patients with and without AVHs (5). These studies reported inconsistent alterations in brain connectivity and activity, either increased or decreased, within functional networks associated with language, memory, or error monitoring in patients with AVHs (6,7). Even if heterogeneity may result from a subtle combination of nonorthogonal causes, one possible explanation could be that, unlike other symptoms of SCZ, AVHs are intermittent experiences (ON/OFF), an aspect not efficiently addressed with trait designs.

Hence, alternative approaches (called symptom-capture fMRI designs) were developed in an attempt to reduce the inherent complexity of AVHs by focusing on the transient neural changes associated with AVH onsets and offsets (8–11). In these capture studies, hallucinators typically signal AVH occurrence online by pressing a button, revealing a wide range of overactive sensory cortical regions that reflects the high phenomenological interindividual variability in hallucinatory experiences (12–14). Among the brain regions most frequently reported as being associated with AVHs, we can mention Broca's area (BA), the superior temporal gyrus, the temporoparietal junction, and the hippocampal complex (11), which are all part of an associative speech-related network more loosely linked with the sensory content of the AVH experience.

SEE COMMENTARY ON PAGE 164

ISSN: 0006-3223

However, it remains unclear whether the brain networks identified using online self-report capture designs are involved 1) in the AVH experience itself, or 2) in the metacognitive/motor processes required to detect and signal the onset of AVHs. Furthermore, these studies referred to a massively univariate activation-based statistical framework known to potentially lose sensitivity by not considering the covariance between voxels (15), which is not particularly well-suited for analyses at the subject level.

In previous studies, we attempted to address these concerns by building upon these paradigms and proposing a button-press free fMRI-capture design based on an independent component analysis, a data-driven multivariate technique (13,16). This approach was proven to reliably detect AVHs from a post-fMRI interview without creating a dual-task situation (i.e., experiencing a vivid AVH and at the same time, pressing a button). However, despite its demonstrated effectiveness, the application of this symptom-capture approach to new patients still required intensive manual labor to individually tailor the analysis pipeline, limiting its translational potential and its applicability/replicability to nonexpert centers. Tackling this issue depends critically on the development of an accurate and easily generalizable method to automatically detect AVHs from brain activity of new SCZ patients with a low additional processing cost.

Here, we propose a novel approach, abstracting out the highly variable content of AVHs by considering these experiences as stereotyped discontinuous mental events, characterized by core modality-independent properties (17,18). To do so, we combined our previously validated fMRI-capture method with supervised machine learning to characterize the informational mapping of AVH states in highly symptomatic patients with schizophrenia, both within and between subjects (19). We found that AVH occurrences can be accurately and reliably decoded from individual fMRI blood oxygen level–dependent (BOLD) signals and that this signature was robust to concurrent mental processes while being generalizable to new data. This predictive signature is time-selective and appears to mainly rely on the BOLD pattern in the BA. In contrast to previous brain imaging studies that emphasized the distributed nature of AVH brain representations, our results show that a multivariate pattern of neural response in a single hub, namely, the BA, is sufficient to robustly predict whether a patient is experiencing AVH, paving the way for the therapeutic use of fMRI-based neurofeedback or brain-computer interfaces for closed-loop neuromodulation.

METHODS AND MATERIALS

Participants

Two independent groups of right-handed patients with SCZ (DSM-IV-TR) were recruited and scanned on two different MRI scanners. Twenty-three patients were enrolled in sample 1 (SCZ#1), and 34 in sample 2 (SCZ#2). They all experienced very frequent and drug-resistant AVHs (i.e., more than 10 episodes/hour). AVH severity was assessed with the P3-item of the Positive and Negative Syndrome Scale (20). See Supplemental Methods for a full list of exclusion criteria. The clinical characteristics of these samples, including the average dosage of antipsychotic medication (in chlorpromazine equivalent), are summarized in Table 1. The study received approval from a national ethical committee (CPP Nord-Ouest, France IV, #2009-A00842-55), and written informed consent was obtained for each participant enrolled in the study upon inclusion.

Procedure

Each patient underwent a single MRI session after clinical evaluation. This acquisition included two runs: 1) an anatomical MRI and 2) an AVH-capture fMRI. For full details of the experimental procedure, see the Supplement.

Data Labeling

For each patient, a cortex-based independent component analysis (ICA) was performed. ICA allowed us to blindly extract nth components (with n equal to 10% of the total number of volumes) from the fMRI BOLD signal time series recorded from cortical voxels. To identify AVH periods, we applied the two-step method summarized in Figure 1 (13,16,21). Because ICA does not naturally order the resulting components according to their relevance, our first step was to manually sort the ICs capturing only noise or recording artifacts and those capturing neurophysiological sources (IC-fingerprint method). Then, among the components capturing neurophysiological sources, we searched for those with a temporal dynamic compatible with the post-fMRI interview data in terms of number of episodes and times of occurrence. Previous reports confirmed the high accuracy of the postscan information in such a context (16).

We finally checked whether these ICs contained brain regions previously identified during AVHs (e.g., speech-related network) (Figure 1A, B). As a second step, three labels were defined on the normalized signal time course of these AVH-related ICs (Figure 1C): ON for the AVH experience (per-AVH periods feature an increased BOLD signal [z score >0] maintained for at least 12 seconds; SCZ#1: 2320 and SCZ#2: 2000 volumes labeled ON), OFF for periods without AVHs (periods with decreased BOLD signals [z score <0] that occurred prior to the ON periods and persisted for at least 6 seconds; SCZ#1: 997 and SCZ#2: 1302 volumes labeled OFF), and REST for wider (noisier) resting-state periods, distant from any hallucinatory event (SCZ#1: 2974 and SCZ#2: 13,688 volumes labeled REST) (see Figure S1).

Data Analysis

We used the Python scikit-learn library to implement linear support vector machine (lSVM) classifiers (22) using the previously described labels (Figure 1D). The workflow described below was conducted independently for each sample (SCZ#1 and SCZ#2). For a full description of multivoxel pattern analysis preprocessing, see Supplemental Methods.

Supervised Analysis With lSVM. To avoid overfitting and to limit the complexity of our classifier, we used linear SVMs to perform multivoxel pattern analysis analyses. Using the labels extracted from the cortex-based ICA analysis (see Data Labeling), we trained classifiers to distinguish between MRI volumes labeled ON versus OFF (or between MRI volumes labeled ON vs. ALL [OFF + REST]). These classifiers

Table 1. Demographic and Descriptive Characteristics of the Recruited Participants

Characteristics	SCZ Dataset #1 (*n* = 23)	SCZ Dataset #2 (*n* = 34)	CTL Dataset (*n* = 20)
Age, Years, Mean ± SD	34.3 ± 8.3	35.2 ± 9.8	29.9 ± 10.1
Sex, Female, *n* (%)	11 (43.5%)	10 (29.4%)	6 (40%)
Duration of SCZ, Years, Mean ± SD	16.8 ± 10.5	17.1 ± 10.8	–
PANSS Total Score, Mean ± SD	79.1 ± 23.3	82.4 ± 20.3	–
PANSS Positive Score, Mean ± SD	21.4 ± 5.8	21.6 ± 5.5	–
P3 Item, Mean ± SD	5.1 ± 0.9	5.6 ± 1.2	–
AHRS Score, Mean ± SD	24.1 ± 6.8	26.5 ± 6.1	–
CPZ-Eq, mg, Mean ± SD	353.6 ± 273.6	324.5 ± 246.4	–

AHRS, Auditory Hallucinations Rating Scale; CPZ-Eq, medication dosage in chlorpromazine equivalent; CTL, control; P3 Item, third item of the positive subscale of the PANSS; PANSS, Positive and Negative Syndrome Scale; SCZ, schizophrenia.

were trained either within subject (using only a subset of the labels from a given subject at a time for training, and testing on the remaining labels) (Figure 2A–C) or across subjects (using a subset of the labels from all subjects pooled together and testing on the remaining pooled labels) (Figure 2B–D). Corresponding maps are available on NeuroVault. Areas under the curve (AUCs) were calculated for each classifier by randomly taking half of the sample data as the training set and half of the sample data as the test set. Receiver operating characteristic (ROC) curves were computed on the test set. To assess the statistical significance of the results, we applied two complementary strategies: 1) in the case of within-subject classifiers, we performed a Monte Carlo cross-validation with 1000 random 2-fold splits; 2) in the case of between-subject classifiers, we performed a leave-one-subject-out cross-validation on a per-subject basis. Hyperparameter optimization is presented in Figure S2. Discriminative weight-maps illustrating the spatial patterns that best discriminate between AVH states (ON or OFF) were extracted and projected onto glass brains (voxel clusters including more than 10 connected voxels, in the 8-neighbor sense).

Contribution of Local Multivariate BOLD Patterns to AVH-State Prediction. To assess the contribution of nonuniform response signs to AVH-state predictions (i.e., the mixture of activation/deactivation in neighboring voxels within a macroscopic region), we ran an additional multiscale sensitivity analysis, in which voxel coordinates were shuffled within the target brain regions prior to training and testing lSVMs to decode ON and OFF AVH states. This procedure intended to destroy all local spatial information within these areas while preserving the target overall voxel BOLD activity distribution. The loss of information induced by shuffling voxel location was quantified by building the corresponding ROC curves and computing AUCs, as was done in the main multivoxel pattern analysis.

Generalization of the Classifiers (I). The false positive rates of the classifiers were computed using an additional control dataset described below. After a brief training session, 20 healthy volunteers (later called the control group; mean age = 29.9 years, age range = 23–49 years) (see Table 1) underwent an fMRI experiment on the same 1.5T scanner as the SCZ#1 sample, during which they performed a verbal mental

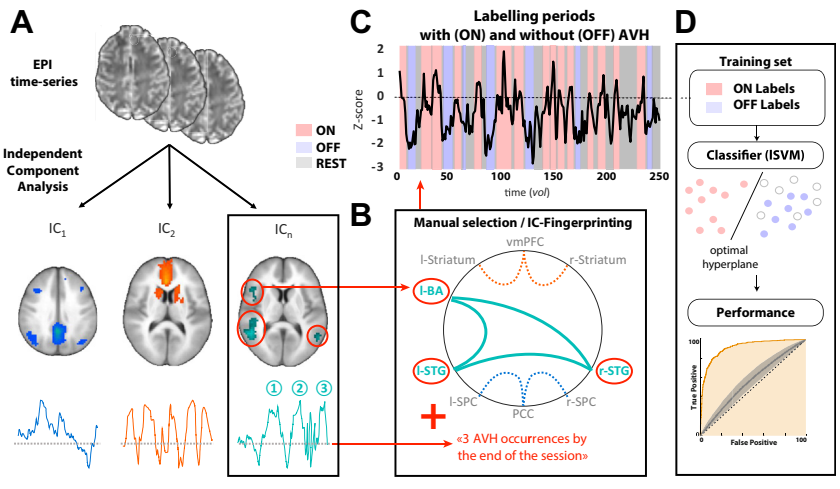

Figure 1. Labeling of the functional magnetic resonance imaging (fMRI) volumes in ON/OFF/REST periods to train linear classifiers. As previously published (13,16,21), we performed **(A)** a spatial independent component (IC) analysis of the fMRI time series collected from patients with frequent auditory-verbal hallucinations (AVHs), resting in the scanner (from schizophrenia samples #1 and #2). **(B)** IC fingerprinting. For each patient, we selected the component (IC_{1-n}) whose spatial network topography matched known AVH-related functional networks (11) and whose temporal activity matched the reported frequency of AVHs based on post-fMRI interviews. **(C)** Finally, we labeled each fMRI volume as being ON (AVH+), OFF (AVH−), or REST, based on the selected component *z*-scored temporal dynamics. We labeled fMRI volumes in which the *z*-scored component was positive for at least 12 consecutive seconds as being ON. Conversely, we labeled volumes as OFF if at least 3 consecutive time points were negative and if they were at least 6 seconds away from an ON volume. Finally, we labeled the remaining volumes as REST. **(D)** Illustration of the training protocol. ON and OFF labels were used to train a linear support vector machine (lSVM) classifier and find the optimal hyperplane separating these two classes. The independent test set consisted of all voxels from the ON, OFF, and REST periods. BA, Broca's area; EPI, echo-planar imaging; IC_{1-n}, independent component 1-to-n; l, left; PCC, posterior cingulate cortex; r, right; SPC, superior parietal cortex; STG, superior temporal gyrus; vmPFC, ventromedial prefrontal cortex.

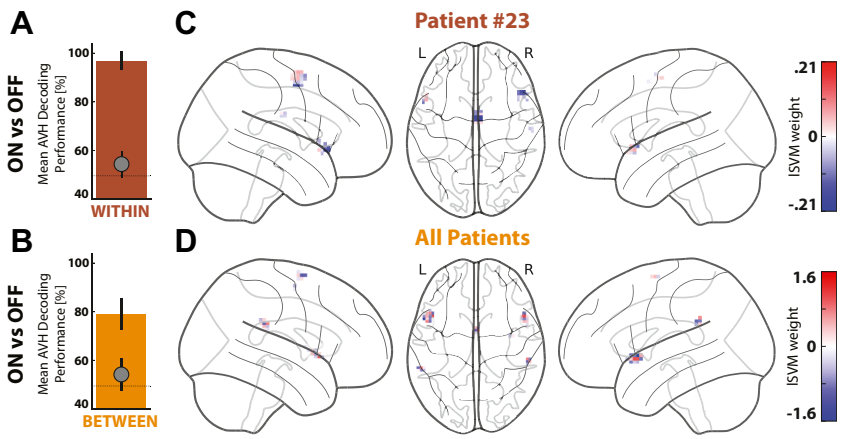

Figure 2. Decoding auditory-verbal hallucinations (AVH) using a linear support vector machine (lSVM) trained within subjects **(A, C)** and between subjects **(B, D)**. **(A)** Group average of within-subject AVH decoding performances (ON vs. OFF). **(B)** Group average of between-subject AVH decoding performance (ON vs. OFF, first schizophrenia dataset). The black circle indicates chance level, as estimated using Monte Carlo simulation (with 1000 permutations). Error bars indicate between-subject SEM **(A, B)**. **(C)** An example of a within-subject contribution map (patient #23). **(D)** Between-subjects contribution map. The 100 most informative voxels are color coded to illustrate their contribution to the classifier (lSVM weight). L, left; R, right.

imagery task. We chose this condition, as verbal imagery was previously found to share neural correlates with AVHs (23,24) (see Supplemental Methods). Data were preprocessed following the steps as described in Supplemental Methods, and functional volumes from this experimental condition were used to challenge the lSVM specificity for AVHs. To take into account the fact that AVH periods typically occur during several time volumes in a row, we convolved the lSVM output probabilities applied to these controls with a flat window of varying size to minimize false positives. The optimal window size was taken as the one minimizing the false positive rates both when the classifier was trained on ON/OFF labels of SCZ#1 and tested on all volumes of SCZ#2, and conversely (Figure S3).

Generalization of the Classifiers (II). Finally, an out-of-sample cross-validation step was added. Performance generalization of the decoding model built and optimized on the SCZ#1 dataset was tested using fMRI data from the SCZ#2 sample kept in a lockbox (25). Conversely, we checked the effect of training/optimizing a decoding model built from SCZ#2, later generalized on the SCZ#1 dataset (kept in the lockbox) for the final performance generalization step.

RESULTS

We first trained lSVM classifiers to detect BOLD activation patterns concomitant with AVHs, either within subject (one classifier per subject) (Figure 2A–C) or between subjects (one classifier for the whole SCZ#1 population) (Figure 2B–D). On average, the within-subject classifier used 100 ON versus 43 OFF volumes/subject (Figure S1). Its mean decoding accuracy for AVHs was 0.96 ± 0.04 across patients, which was significantly above chance (Monte Carlo simulations of the null distribution, 1000 permutations) (see Figure 2A). The between-subjects classifier used a total of 2320 ON versus 997 OFF volumes. Its decoding accuracy for AVHs was also significantly above chance at 0.79 ± 0.06 (1000 permutations) (Figure 2B). Although MRI data were normalized using only a standard linear transformation, this value was in the same range as the within-subject classifier accuracy. The ROC curve also indicated a high probability for correct classification for the between-subjects classifier (AUC_{H0} = 0.55 vs. AUC = 0.85, $p < 1 \times 10^{-3}$) (Figure 3A).

Then, we built contribution maps from the predictive weight given to each voxel by the ON versus OFF lSVM classifier, which revealed AVH-related BOLD activity patterns in the bilateral inferior frontal gyri (i.e., the BA and its right homolog), the supplementary motor area (SMA) and the pre-SMA (SMA will be used to designate these two structures), and the bilateral supramarginal gyri (Figure 2C, D). These brain regions have previously been shown to be involved in AVH pathophysiology (6,11,23).

As a sanity check, we ensured that these performances did not depend on the ON/OFF volume ratio across subjects or on frame displacements (Figure S4). We confirmed that these findings and the performance of the classifier did not depend on the threshold applied for feature selection (i.e., the number of voxels included in the training set up to 10,000 voxels (see Table S1) and that the between-subjects decoding performances were stable and reliable for all training sample sizes larger than 10 patients, indicating that the multivariate pattern of BOLD activity used by the classifier seems especially robust and well conserved between subjects (Figure 3B).

We also ran a series of additional analyses aimed at assessing the specificity and robustness of this between-subject ON versus OFF classifier. First, we estimated its false positive rate when applied to new data. We applied the classifier trained with SCZ#1 ON/OFF labels to noisier volumes (ON vs. [OFF + REST]) taken either from the same dataset (SCZ#1, n = 23) or from a different dataset (SCZ#2, n = 34). As shown in Figure 3C (blue bars), false positive rates were initially high in both cases (i.e., 35% on average). However, by applying a smoothing kernel to the output of the classifier to take into account the fact that ON and OFF periods should cover consecutive volumes, the false positive rate dropped at 21% (with an AUC of 85% for SCZ#1 → SCZ#1 and 81% for SCZ#1 → SCZ#2). It is noteworthy that the application of the smoothing kernel mainly acted by limiting the risk of false positives over the 3 volumes preceding or succeeding actual AVH states (Figure S3C, D).

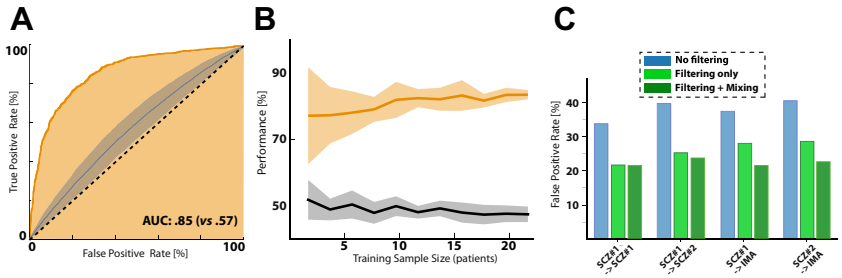

Figure 3. Reliability and performances of between-subject decoding of auditory-verbal hallucinations using a linear support vector machine classifier (ON vs. OFF). **(A)** The receiver operating characteristic curve indicates the trade-off between false positive and true positive label classification (±SEM). **(B)** Reliability of between-subjects auditory-verbal hallucination decoding performance. Mean between-subjects decoding performance (±SEM) as a function of the number of patients in the training set (orange line). The black line indicates the chance performance level (±SEM) estimated using Monte Carlo simulation (1000 permutations). See also Figure S4 presenting a correlation analysis between performances and the number of ON periods. **(C)** False positive rates of a between-subjects linear support vector machine trained on the first or second schizophrenia dataset (ON vs. OFF labels, SCZ#1 or SCZ #2, respectively) and then applied to the control dataset (could be SCZ#1 or SCZ#2 with ON vs. ALL labels, or IMA where IMA is a third dataset collected while performing a verbal mental imagery task [see Methods and Materials]). False positive rates are shown for raw data (blue bars) when the linear support vector machine scores are filtered, i.e., convolved with a flat time window of 5 seconds to ensure that ON periods are contiguous (orange bars) (see Methods and Materials), or when, in addition to this filtering window, half of the test data are also used to refine the training sets. AUC, area under the curve.

Similarly, when applying the classifier to a third dataset acquired in healthy control subjects performing a verbal imagery task (a mental process potentially overlapping AVH; see control dataset, Methods and Materials), the initial false positive rates were high (37%), but filtering the classifier's output resulted in a decrease in this false positive rate to 25% (Figure 3C) (light-green bars). Finally, retraining the lSVM classifiers on a composite dataset mixing OFF labels with some REST and/or IMA labels from the #SCZ1/2 and/or IMA dataset (i.e., using a subpart of the control dataset to enhance the training sets) reduced false positives to a reasonable level of approximately 20% in all conditions without altering AVH decoding accuracy (Figure 3C) (dark-green bars) (see also Figure S5 presenting the decoding accuracy and contributive maps for a between-subject classifier trained on ON vs. [OFF + REST] with SCZ#1).

To ensure out-of-sample generalization and control for the risk of hyperparameter overfitting, the classifier trained on the SCZ#1 dataset was then challenged to predict the labels of the independent SCZ#2 sample (fully held out from the optimization process and only accessed to generate a new unbiased estimate of the AVH decoder's performance). We switched this procedure by training a classifier on SCZ#2 and predicting labels on SCZ#1. In both cases, the AUCs were found to be significantly different from chance (0.85 and 0.88, respectively), while the resulting contributive voxels overlapped in 75% of the cases within the BA (Figure 4).

Finally, we assessed how much AVH-related information was local, encoded as a spatial mixture of activation/deactivation within each brain region identified in the main analysis (namely, the BA and SMA) while controlling for global patterns of average regional levels of BOLD activity (for example, higher mean BA BOLD activity and lower mean SMA activity during AVHs). We found that shuffling voxel coordinates within the BA and SMA prior to training/testing the classifier had a significant effect on AVH prediction within the BA (Figure 5A) (blue ROC curve, $p < 1 \times 10^{-5}$), but not within the SMA (Figure 5A) (pink ROC curve). Moreover, ROC curves when only BA voxels were shuffled were not statistically significant from the ROC curves when both BA and SMA voxels were shuffled (Figure 5A) (dashed black ROC curve). This suggests that most AVH-related information captured by the linear classifier is represented locally in the spatial BOLD pattern of BA and not elsewhere in the brain.

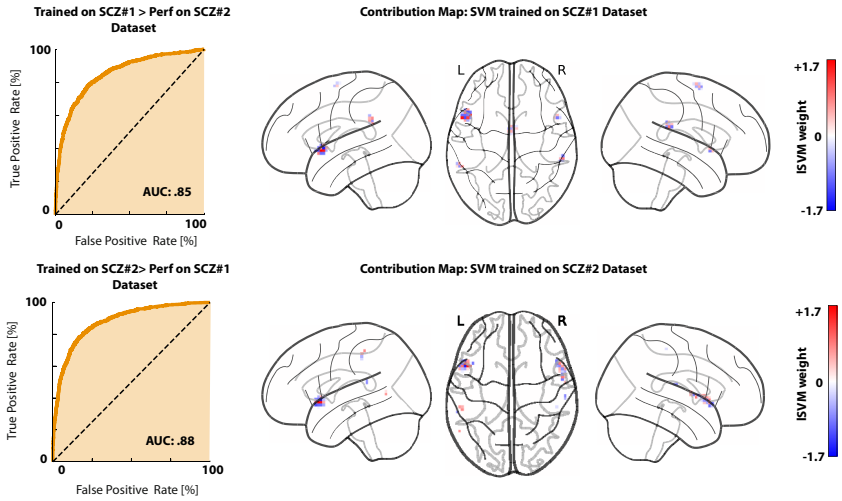

Figure 4. Generalization of the auditory verbal hallucination (AVH) decoder to out-of-sample data. We checked for possible overfitting and overhyping of the AVH linear support vector machine (lSVM) classifier. In conjunction with the cross-validation approach used on the first schizophrenia dataset (SCZ#1) (cf. Figures 2–4), we refer to a lockbox approach, in which the second sample of patients with SCZ scanned during AVH occurrences was set aside from the beginning of the study and used only after hyperparameter optimization of the AVH classifier to check its accuracy on a completely new dataset (upper panel). We repeated the procedure to ensure that we could obtain similar generalization performances when starting training and optimizing the decoder on the SCZ#2 dataset and testing accuracy on the SCZ#1 dataset (lower panel). AUC, area under the curve; L, left; R, right.

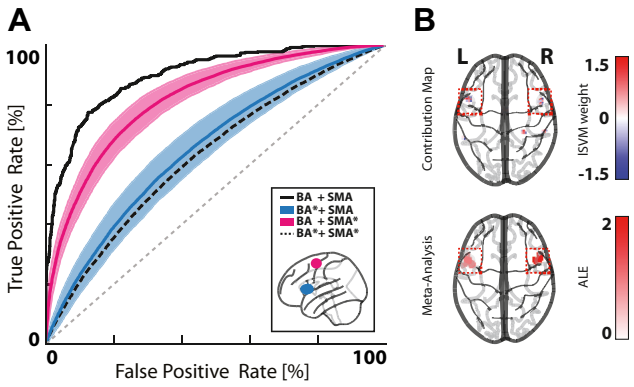

Figure 5. Local vs. global patterns of blood oxygen level–dependent activity associated with auditory verbal hallucinations. **(A)** Receiver operating characteristic curve of a between-subjects ON vs. OFF linear support vector machine (lSVM) indicating the trade-off between false positive and true positive label classification (black line, first schizophrenia dataset). Pink and blue lines indicate the receiver operating characteristic curves after a volume-by-volume shuffling of voxel locations within each of the two regions of interest (ROIs) (pink: supplementary motor area [SMA] shuffled, blue: Broca's area [BA] shuffled). The dashed black line indicates the performance of the classifier when the spatial locations of voxels within both ROIs are simultaneously shuffled. Note that voxels belonging to an ROI are always shuffled to a location within that same ROI. Shaded blue and pink areas indicate between-subjects SEM. **(B)** Contribution map of the between-subjects lSVM (ON vs. OFF) (top panel) against an updated activation likelihood location (ALE) meta-analysis of the literature on per–auditory verbal hallucination blood oxygen level–dependent activity (bottom panel). L, left; R, right.

Taken together, these results shed new light on the pathophysiology of AVHs, which was previously thought to be inherently distributed in a complex network of brain-wide regions, and suggest instead that a nonuniform BA functional pattern is critical to predict AVH occurrences (Figure S6) and set it apart from normal perception.

DISCUSSION

The online detection of spontaneous AVH occurrence has long been very challenging (9,13,16). By demonstrating that a simple linear classifier can robustly decode hallucinatory states from out-of-sample resting-state fMRI data without complex preprocessing, we substantially extended recent work in the field of AVH fMRI capture by departing from a classic activation-based to a multivariate information-based perspective. Notably, we found a 79% between-subjects accuracy in distinguishing hallucinatory and nonhallucinatory periods (ON vs. OFF, 0.85 AUC).

We demonstrated the robustness of our findings by conducting a set of supplemental analyses to precisely characterize our decoder features: 1) repeating the performance measures while increasing the sample size, 2) confirming good performances even when using noisier data, and 3) reducing false positive rates after enforcing temporal regularization and training the algorithm to selectively ignore verbal imagery. We further addressed a crucial issue in the classification literature: replicability of the decoder performances (26). Here, we went beyond conventional leave-one-subject-out cross-validation strategies by replicating our results both ways, using either SCZ#1 or SCZ#2 as a training/test set or as an out-of-sample dataset aside in a lockbox [and thus independent from the optimization process (25)].

We identified a BOLD multivariate signature predictive of AVH in speech-related motor/planning brain regions, such as the Brodmann area 44, part of the BA (27), and the SMA [Brodmann area 6, medially (28)]. This functional signature highlights the special role of BA in hallucinatory experiences among all the regions previously reported in per-AVH activation studies, whether in first-episode psychosis (13,29), SCZ (9,16,30,31), or nonclinical voice hearers (23). Interestingly, this last study reported differences in the timing of SMA activations (relative to BA activations) between AVHs and a verbal imagery condition, which appears fully compatible with our optimization procedure that allowed us to strengthen AVH/imagery discrimination. In addition, consistent with these findings, our BA cluster overlaps with coordinate-based meta-analytic findings of per-AVH hyperactivations in schizophrenia (11) or conditioned hallucinatory mapping obtained from nonclinical participants (32) (Figure 5B; see also Tables S2 and S3 and Figure S7).

The BA and SMA are also known to be involved in error monitoring and inhibition (33), suggesting that AVHs may result from aberrant motor representations/predictions (despite an absence of online self-report in the participants), which may be a core mechanism in the lack of insight typically associated with hallucinations (34). This appears compatible with previous hypotheses of inner speech as a form of action (35). In contrast, hippocampal or temporoparietal structures, also known to be involved in AVH pathophysiology (6,36–38), possibly by reflecting the spatiotemporal, rich, and complex content of these experiences (21,39), were not necessary to reach high decoding performances. Even if anteroposterior dysconnectivity between speech-related areas has been regularly shown to be involved in AVHs (40), this new finding suggests either 1) that the highly variable nature of the information computed by these temporal-hippocampal structures is not stereotyped enough to be decoded using lSVM or 2) that most of the relevant fine-grained information conserved between subjects is encoded in Broca's BOLD activity.

Although the main goal of this study was to demonstrate the feasibility of a reliable and easily deployable multivariate AVH decoder, it also adds several insights to AVH pathophysiology. We know that the performance of a classifier is dependent on the functional features used to train the lSVM. This is why we referred to a valid strategy to determine ON/OFF labels (13,16,21) that proved able to achieve good performances even without special regularization preprocessing steps (41). This may appear surprising at first glance because previous work conducted on more subtle functional profiles (i.e., states preceding AVH onsets) showed that specific classification algorithms with total variation penalty were better at detection than lSVM (39).

In reality, the good performances demonstrated by our classifier reflect the remarkable consistency of the AVH-related BOLD pattern across patients, robust to varying magnetic field strength or sequence parameters (e.g., image resolution, time repetition of the sequence, or differences in number of volumes between SCZ#1 and SCZ#2 datasets). Limited activation studies have previously reported similar spatial stability in per-

AVH activation patterns (16,30), and we propose that such consistency could be due to the involvement of the BA.

The BA is a highly preserved anatomical-functional hub [for an evolutionary perspective, see (42)], which may relate to the coding of a very generic, amodal feature of the AVH experience. In this vein, the BA could be sensitive to intrusiveness into consciousness, irrespective of the highly variable phenomenological content of the voices (4). Even if speculative at this stage, this assumption appears compatible with recent findings showing the involvement of the inferior frontal gyrus in the intrusion of unwanted thoughts more broadly, notably in patients with obsessive-compulsive disorder experiencing severe obsessions (43), while states of mind-blanking were shown to be associated with BA deactivation (44).

In our study, this assumption was also confirmed by voxel permutation tests performed to challenge the local distribution of response signs in contributive maps. Such permutations flattened the lSVM accuracy when applied to the BA, while extending this operation to the SMA only slightly (yet significantly) impacted the classifier. This can be interpreted as a form of redundancy in the information processed by the SMA, while the BA could locally compute crucial elements for AVH intrusion prediction, coded in its microstructure (and not elsewhere in the brain), experimentally accessible only because of the use of multivariate pattern classification methods.

Until now, AVH fMRI capture has been considered to be complex and time-consuming because of its many technical constraints. This situation has limited the use of these methods to offline applications in the lab, which has significantly hindered therapeutic innovations. Thanks to our newly validated and replicable biomarker, reading out hallucinatory states online from resting-state fMRI is now possible, providing a gateway to further validate fMRI-based neurofeedback procedures to relieve severe AVHs and develop brain-computer interfaces for closed-loop neuromodulation (45). Indeed, both approaches require clearly defined cortical targets at key points in the brain networks involved in AVHs (46).

The moderate sample size and the recruitment of only right-handed participants should be acknowledged. However, the high number of ON/OFF labels, replication on a second independent group (SCZ#2), and the good performance stability across sample sizes all suggest that only little improvement can be expected from further increasing the sample size (Figure 3).

The unpredictable nature of brain-state changes over time associated with hallucinations has long remained a major (and supposedly insuperable) challenge in neuroscience, and most therapeutic alternatives to medications have attempted to modulate network activity [see for instance, noninvasive brain stimulation targeting the temporoparietal junction (47)]. We believe that our findings not only uncovered a neurofunctional reconfiguration associated with this fascinating mental experience but also provide a translational way to automatically identify a dynamic neural pattern playing an important, if not critical, role in AVH occurrences (i.e., intrusiveness). This paves the way for the generalization of fMRI capture and the development of new image-guided therapeutic strategies for drug-resistant hallucinations, such as fMRI neurofeedback based on multivoxel patterns of brain activity and closed-loop cortical stimulation.

ACKNOWLEDGMENTS AND DISCLOSURES

This study was supported by a grant from the Agence Nationale de la Recherche (Grant No. ANR-16-CE37-0015 INTRUDE [to RJ]).

The project was supervised by RJ. Data was collected and prepared by TF and RJ. Data were analyzed by PY, PD, and RJ. All the authors contributed to the data interpretation and paper writing.

A previous version of this article was published as a preprint on bioRxiv: https://www.biorxiv.org/content/10.1101/2021.05.21.445102v1.

Resulting maps from this study are available on a public depository using this link: https://neurovault.org/collections/11160/.

RJ and PT have been invited to scientific meetings and expert boards by Lundbeck, Janssen, and Otsuka. None of these links of interest are related to the present work. RJ received research funding from the Programme Hospitalier de Recherche Clinique National (MULTIMODHAL, MOCITRAINING) and the ANR (INTRUDE) national programs, as well as from the H2020 (miniNO) European program. PD received research funding from the Programme de Recherche Médico-Economique (DepVNS) and the Agence Nationale de la Recherche (DUAL-TRACK, MCDM) programs. ED is supported by Big2Small, a teaching and research chair in AI of French National Research Agency. All other authors report no biomedical financial interests or potential conflicts of interest.

ARTICLE INFORMATION

From Plasticity & SubjectivitY team (TF, PY, PT, SS, RJ) and Vascular & Cognitive Deficits team (RL), Lille Neuroscience & Cognition Research Centre, University of Lille, INSERM U1172, Lille; CURE platform (TF, PT, RJ), Psychiatry Department, Fontan Hospital, Centre Hospitalier Universitaire de Lille, Lille; In-vivo Imaging and Functions core facility (RL), Neuroradiology Department, Centre Hospitalier Universitaire de Lille, Lille; Centre National de Ressources et de Résilience (TF), Lille-Paris; Institut de la Vision (PY) and Institut du Cerveau et de la Moelle épinière (PD), Sorbonne Université, INSERM, Centre national de la recherche scientifique, Paris; NeuroSpin (AdP, ED, JH), Univ Paris Saclay, CEA, Gif-sur-Yvette; Neurosurgery, Psychiatry and Addictology Departments (JH, PD), Groupe Hospitalier Universitaire Henri-Mondor, AP-HP, Créteil; and Faculté de Santé UPEC (JH, PD), Université Paris Est Créteil, Créteil, France.

TF and PY contributed equally to this work.

PD and RJ contributed equally to this work.

Address correspondence to Renaud Jardri, M.D., Ph.D., at renaud.jardri@chru-lille.fr.

Received Jun 2, 2021; revised Aug 19, 2021; accepted Aug 31, 2021.

Supplementary material cited in this article is available online at https://doi.org/10.1016/j.biopsych.2021.08.024.

REFERENCES

1. Blom JD (2010): A Dictionary of Hallucinations. New York: Springer-Verlag.
2. Nicolson SE, Mayberg HS, Pennell PB, Nemeroff CB (2006): Persistent auditory hallucinations that are unresponsive to antipsychotic drugs. Am J Psychiatry 163:1153–1159.
3. Schutte MJL, Linszen MMJ, Marschall TM, ffytche DH, Koops S, van Dellen E, *et al.* (2020): Hallucinations and other psychotic experiences across diagnoses: A comparison of phenomenological features. Psychiatry Res 292:113314.
4. Pienkos E, Giersch A, Hansen M, Humpston C, McCarthy-Jones S, Mishara A, *et al.* (2019): Hallucinations Beyond Voices: A conceptual review of the phenomenology of altered perception in psychosis. Schizophr Bull 45(suppl 1):S67–S77.
5. Linden DEJ (2012): The challenges and promise of neuroimaging in psychiatry. Neuron 73:8–22.
6. Allen P, Modinos G, Hubl D, Shields G, Cachia A, Jardri R, *et al.* (2012): Neuroimaging auditory hallucinations in schizophrenia: From neuroanatomy to neurochemistry and beyond. Schizophr Bull 38:695–703.

7. Ćurčić-Blake B, Ford JM, Hubl D, Orlov ND, Sommer IE, Waters F, *et al.* (2017): Interaction of language, auditory and memory brain networks in auditory verbal hallucinations. Prog Neurobiol 148:1–20.
8. Dierks T, Linden DE, Jandl M, Formisano E, Goebel R, Lanfermann H, Singer W (1999): Activation of Heschl's gyrus during auditory hallucinations. Neuron 22:615–621.
9. Sommer IEC, Diederen KMJ, Blom JD, Willems A, Kushan L, Slotema K, *et al.* (2008): Auditory verbal hallucinations predominantly activate the right inferior frontal area. Brain 131:3169–3177.
10. van de Ven VG, Formisano E, Röder CH, Prvulovic D, Bittner RA, Dietz MG, *et al.* (2005): The spatiotemporal pattern of auditory cortical responses during verbal hallucinations. Neuroimage 27:644–655.
11. Jardri R, Pouchet A, Pins D, Thomas P (2011): Cortical activations during auditory verbal hallucinations in schizophrenia: A coordinate-based meta-analysis. Am J Psychiatry 168:73–81.
12. Ffytche DH, Howard RJ, Brammer MJ, David A, Woodruff P, Williams S (1998): The anatomy of conscious vision: An fMRI study of visual hallucinations. Nat Neurosci 1:738–742.
13. Jardri R, Thomas P, Delmaire C, Delion P, Pins D (2013): The neurodynamic organization of modality-dependent hallucinations. Cereb Cortex 23:1108–1117.
14. Dujardin K, Roman D, Baille G, Pins D, Lefebvre S, Delmaire C, *et al.* (2020): What can we learn from fMRI capture of visual hallucinations in Parkinson's disease? Brain Imaging Behav 14:329–335.
15. Hebart MN, Baker CI (2018): Deconstructing multivariate decoding for the study of brain function. Neuroimage 180:4–18.
16. Leroy A, Foucher JR, Pins D, Delmaire C, Thomas P, Roser MM, *et al.* (2017): fMRI capture of auditory hallucinations: Validation of the two-steps method. Hum Brain Mapp 38:4966–4979.
17. Morrison AP, Haddock G, Tarrier N (1995): Intrusive thoughts and auditory hallucinations: A cognitive approach. Behav Cogn Psychother 23:265–280.
18. Waters FAV, Badcock JC, Michie PT, Maybery MT (2006): Auditory hallucinations in schizophrenia: Intrusive thoughts and forgotten memories. Cogn Neuropsychiatry 11:65–83.
19. Kriegeskorte N, Goebel R, Bandettini P (2006): Information-based functional brain mapping. Proc Natl Acad Sci U S A 103:3863–3868.
20. Kay SR, Fiszbein A, Opler LA (1987): The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. Schizophr Bull 13:261–276.
21. Lefebvre S, Demeulemeester M, Leroy A, Delmaire C, Lopes R, Pins D, *et al.* (2016): Network dynamics during the different stages of hallucinations in schizophrenia. Hum Brain Mapp 37:2571–2586.
22. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al.* (2011): Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830.
23. Linden DEJ, Thornton K, Kuswanto CN, Johnston SJ, van de Ven V, Jackson MC (2011): The brain's voices: Comparing nonclinical auditory hallucinations and imagery. Cereb Cortex 21:330–337.
24. Raij TT, Riekki TJJ (2012): Poor supplementary motor area activation differentiates auditory verbal hallucination from imagining the hallucination. Neuroimage Clin 1:75–80.
25. Hosseini M, Powell M, Collins J, Callahan-Flintoft C, Jones W, Bowman H, Wyble B (2020): I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. Neurosci Biobehav Rev 119:456–467.
26. Varoquaux G (2018): Cross-validation failure: Small sample sizes lead to large error bars. Neuroimage 180:68–77.
27. Zilles K, Amunts K (2018): Cytoarchitectonic and receptorarchitectonic organization in Broca's region and surrounding cortex. Curr Opin Behav Sci 21:93–105.
28. Hertrich I, Dietrich S, Ackermann H (2016): The role of the supplementary motor area for speech and language processing. Neurosci Biobehav Rev 68:602–610.
29. Mallikarjun PK, Lalousis PA, Dunne TF, Heinze K, Reniers RL, Broome MR, *et al.* (2018): Aberrant salience network functional connectivity in auditory verbal hallucinations: A first episode psychosis sample. Transl Psychiatry 8:69.
30. Diederen KMJ, Charbonnier L, Neggers SFW, van Lutterveld R, Daalman K, Slotema CW, *et al.* (2013): Reproducibility of brain activation during auditory verbal hallucinations. Schizophr Res 146:320–325.
31. McGuire PK, Shah GM, Murray RM (1993): Increased blood flow in Broca's area during auditory hallucinations in schizophrenia. Lancet 342:703–706.
32. Powers AR, Mathys C, Corlett PR (2017): Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. Science 357:596–600.
33. Behroozmand R, Shebek R, Hansen DR, Oya H, Robin DA, Howard MA, Greenlee JDW (2015): Sensory-motor networks involved in speech production and motor control: An fMRI study. Neuroimage 109:418–428.
34. Powers AR, Kelley M, Corlett PR (2016): Hallucinations as top-down effects on perception. Biol Psychiatry Cogn Neurosci Neuroimaging 1:393–400.
35. Jones SR, Fernyhough C (2007): Thought as action: Inner speech, self-monitoring, and auditory verbal hallucinations. Conscious Cogn 16:391–399.
36. Amad A, Cachia A, Gorwood P, Pins D, Delmaire C, Rolland B, *et al.* (2014): The multimodal connectivity of the hippocampal complex in auditory and visual hallucinations. Mol Psychiatry 19:184–191.
37. Diederen KMJ, Neggers SFW, Daalman K, Blom JD, Goekoop R, Kahn RS, Sommer IEC (2010): Deactivation of the parahippocampal gyrus preceding auditory hallucinations in schizophrenia. Am J Psychiatry 167:427–435.
38. Hare SM, Law AS, Ford JM, Mathalon DH, Ahmadi A, Damaraju E, *et al.* (2018): Disrupted network cross talk, hippocampal dysfunction and hallucinations in schizophrenia. Schizophr Res 199:226–234.
39. de Pierrefeu A, Fovet T, Hadj-Selem F, Löfstedt T, Ciuciu P, Lefebvre S, *et al.* (2018): Prediction of activation patterns preceding hallucinations in patients with schizophrenia using machine learning with structured sparsity. Hum Brain Mapp 39:1777–1788.
40. Geoffroy PA, Houenou J, Duhamel A, Amad A, De Weijer AD, Curčić-Blake B, *et al.* (2014): The Arcuate Fasciculus in auditory-verbal hallucinations: A meta-analysis of diffusion-tensor-imaging studies. Schizophr Res 159:234–237.
41. Xu H, Lorbert A, Ramadge PJ, Guntupalli JS, Haxby JV (2012): Regularized hyperalignment of multi-set fMRI data. In: 2012 IEEE Statistical Signal Processing Workshop (SSP). Ann Arbor: IEEE, 229–232.
42. Ponce de León MS, Bienvenu T, Marom A, Engel S, Tafforeau P, Alatorre Warren JL, *et al.* (2021): The primitive brain of early Homo. Science 372:165–171.
43. Norman LJ, Taylor SF, Liu Y, Radua J, Chye Y, De Wit SJ, *et al.* (2019): Error processing and inhibitory control in obsessive-compulsive disorder: A meta-analysis using statistical parametric maps. Biol Psychiatry 85:713–725.
44. Kawagoe T, Onoda K, Yamaguchi S (2019): The neural correlates of "mind blanking": When the mind goes away. Hum Brain Mapp 40:4934–4940.
45. Humpston C, Garrison J, Orlov N, Aleman A, Jardri R, Fernyhough C, Allen P (2020): Real-time functional magnetic resonance imaging neurofeedback for the relief of distressing auditory-verbal hallucinations: Methodological and empirical advances. Schizophr Bull 46:1409–1417.
46. Orlov ND, Giampietro V, O'Daly O, Lam SL, Barker GJ, Rubia K, *et al.* (2018): Real-time fMRI neurofeedback to down-regulate superior temporal gyrus activity in patients with schizophrenia and auditory hallucinations: A proof-of-concept study. Transl Psychiatry 8:46.
47. Demeulemeester M, Amad A, Bubrovszky M, Pins D, Thomas P, Jardri R (2012): What is the real effect of 1-Hz repetitive transcranial magnetic stimulation on hallucinations? Controlling for publication bias in neuromodulation trials. Biol Psychiatry 71:e15–e16.