Multi-dimensional predictions of psychotic symptoms via machine learning

Jeremy A Taylor[1,2], Kit Melissa Larsen[2-5], Marta I Garrido[1-3,6]

[1] Melbourne School of Psychological Sciences, University of Melbourne, Australia

[2] Queensland Brain Institute, University of Queensland, Australia

[3] Australian Research Council Centre of Excellence for Integrative Brain Function

[4] Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital Hvidovre, Hvidovre, Denmark

[5] Child and Adolescent Mental Health Care, Mental Health Services Capital Region Copenhagen, University of Copenhagen, Denmark

[6] Centre for Advanced Imaging, University of Queensland, Australia

Correspondence

Jeremy Taylor
Melbourne School of Psychological Sciences, University of Melbourne
Redmond Barry Building, Tin Alley, Parkville, Victoria, Australia
jeremy.taylor2@unimelb.edu.au

Keywords

schizophrenia; fMRI; regression; fusion; ensemble; dimensional diagnosis

Acknowledgements

This work was supported by the Australian Research Council Centre of Excellence for Integrative Brain Function (ARC Centre Grant CE140100007), a University of Queensland fellowship (2016000071) and Foundation Research Excellence Award (2016001844) to MIG. We would like to thank Randy Gollub and Margaret King for providing data, as well as Ilvana Dzafic, Saskia Bollmann and Kelly Garner for discussions on fMRI, and the University of Queensland Research Computing Centre (RCC) for access to high performance computing resources.

Conflict of Interest

The authors declare no competing financial interests.

Abstract

Background

The diagnostic criteria for schizophrenia comprise a diverse range of heterogeneous symptoms. As a result, individuals each present a distinct set of symptoms despite having the same overall diagnosis.

Methods

Although machine learning techniques are considered a potential gateway to precision psychiatry, prior work has primarily focused on dichotomous patient-control classification. Instead, we predict the severity of each individual symptom on a continuum. We applied machine learning regression within a multi-modal fusion framework to fMRI and behavioural data acquired during an auditory oddball task in 80 schizophrenia patients.

Results

Brain activity was highly predictive of some, but not all symptoms, namely hallucinations, avolition, anhedonia and attention. Each of these symptoms was associated with alterations in specific functional networks encompassing the ventral and dorsal attention networks, the auditory network, amongst other cortical and subcortical regions. We found that an ensemble of subscale models yielded a two-fold increase in accuracy over single models which predict positive and negative compound scores directly.

Conclusions

Our results suggest that modelling symptoms as an ensemble of subscales is more accurate, specific, and informative than the compound-based approach. We provide functional brain maps of model contributions identifying the networks of regions which pertain to each individual symptom. This approach is transferrable to any other psychiatric condition and may also contribute to the development of precision psychiatry.

Introduction

Schizophrenia diagnoses comprise a diverse range of heterogeneous symptoms which collectively manifest in widespread neuroanatomical and functional differences. Over the past decade, machine learning has been widely used in the field of neuroimaging for mapping symptomatic manifestations onto brain substrates. These methods are generally considered a potential gateway to precision psychiatry as they provide predictions at the individual level, hence going beyond classical univariate methods which can only tell us about overall group effects within a given population. The vast majority of psychiatric machine learning studies are primarily concerned with group membership, in particular the binary classification between patients and controls, patient subgroups, or prognoses (1-3). However, given the wide array of symptoms which characterise schizophrenia, individuals each present with their own distinct set of symptoms despite having the same categorical diagnosis. In an effort to parse these symptomatic differences, there has recently been a shift away from dichotomous labels towards a dimensional approach (4, 5).

As defined by the Diagnostic and Statistical Manual of Mental Disorders (DSM; 6), symptoms are categorised as *positive* or *negative*. Positive symptoms are typically absent in the general population, such as hallucinations and delusions, whereas negative symptoms present more often, including affective flattening and poverty of speech. In clinical practice, standardised psychometric tools are widely used to assess the severity of symptoms, in particular the Scale for Assessment of Positive Symptoms (SAPS; 7), the Scale for Assessment of Negative Symptoms (SANS; 8) and the Positive and Negative Symptom Scale (PANSS; 9). Individual symptoms or *subscales* are assigned numeric scores relative to their severity, ranging from absent to severe. The *composite* score is the sum of all symptom subscales, providing an overall summary of the given category. The limited prior work on predicting schizophrenia symptoms via machine learning has thus far only been performed on the basis of composite symptoms (10, 11), general functioning (12, 13) and polygenic risk scores for schizophrenia (14). Other neuroimaging studies have also reported univariate correlates (15, 16), or lack thereof (17, 18), with symptom severity on the basis of composite summary scores rather than those of the underlying symptoms, an approach which significantly comprises aetiological specificity (19). For example, if we were to compare two patients, one with disorganised thought processes which render them unable to bathe themselves to another with severe alogia who is unable to communicate, these are vastly

different symptoms which in turn are likely to be caused by different sources of dysfunction in different neural networks. By combining the breadth of symptoms under the hypernyms of schizophrenia, positive or negative symptoms, the superposition of features pertaining to each specific symptom may appear more heterogeneous en masse than if these symptoms were addressed separately. Furthermore, a pair of individuals may be assigned the same composite score, and yet have vastly different symptomatology (e.g., a large number of mild symptoms or a smaller subset of high severity symptoms). Given the pervasive issues associated with the heterogeneity of schizophrenia, we suggest the distinction between individual symptoms may be pertinent.

The aim of this study was to predict the severity of schizophrenia symptoms on a continuum using a dimensional diagnosis approach, based on individual neural and behavioural responses to an auditory oddball task performed during a fMRI scan. We applied machine learning regression techniques within a multi-modal fusion framework to predict each individual symptom whilst determining the set of neural and behavioural features which inform each model. In addition, we sought to predict global symptom severity as an ensemble of these subscales and compare this to models which predict the composite scores directly. Finally, we provide maps of the brain regions which contributed toward predictions of specific symptoms.

Methods

Dataset

The data used in this study was provided by the Mental Illness and Neuroscience Discovery Institute Clinical Imaging Consortium (MCIC; 20) via the Collaborative Informatics Neuroimaging Suite (COINS; 21) online data repository. Anonymised medication data was also obtained directly from the curators of the dataset. These data were originally collected across multiple sites — the University of New Mexico, Massachusetts General Hospital and University of Iowa.

Participants and cognitive characterisation

From an initial sample of 118 schizophrenia patients obtained from the COINS database, participants were excluded on the basis of missing data and/or poor task performance (mean – 3SD). The final sample consisted of 80 schizophrenia patients (58 male, 22 female) with ages ranging from 18 to 60 years (mean ± SD, 32.55 ± 11.39 years). Diagnoses were confirmed using the Structured Clinical Interview from DSM-IV or Comprehensive Assessment of Symptoms and History (22) with severity of symptoms assessed using the SAPS and SANS. For a summary of participant symptom scores, refer to Supplemental Figure S1. Both SAPS and SANS use a five point scale for each subscale (0 = absent, 1 = questionable, 2 = mild, 3 = moderate, 4 = marked, 5 = severe) with the summary score the sum of all subscales.

Stimulus paradigm

fMRI data were acquired whilst participants listened to an auditory oddball paradigm devised by Kiehl and Liddle (23), comprising streams of predictable standard tones (1kHz, $p = 0.82$), interspersed with infrequent target (1.2kHz, $p = 0.09$) and novel tones (complex, computer-generated tones, $p = 0.09$). Participants were instructed to respond to target stimuli via button press whilst ignoring standard and novel tones. For details on stimulus presentation, MRI data acquisition and pre-processing, refer to Supplemental Material.

Features

Four distinct feature sets were defined categorically as neural responses to target and novel conditions, behavioural measures, and other potential confounds.

For the target and novel conditions, voxel-wise activity was extracted from a set of 15 regions-of-interest (ROIs) within the fMRI contrast images. Informed by a meta-analysis by Kim (24), these regions were assumed *a priori* to be those where task-relevant and irrelevant oddball effects would be most robust. The complete set of regions is shown in Figure 1 with atlas references available in Supplemental Table S1.

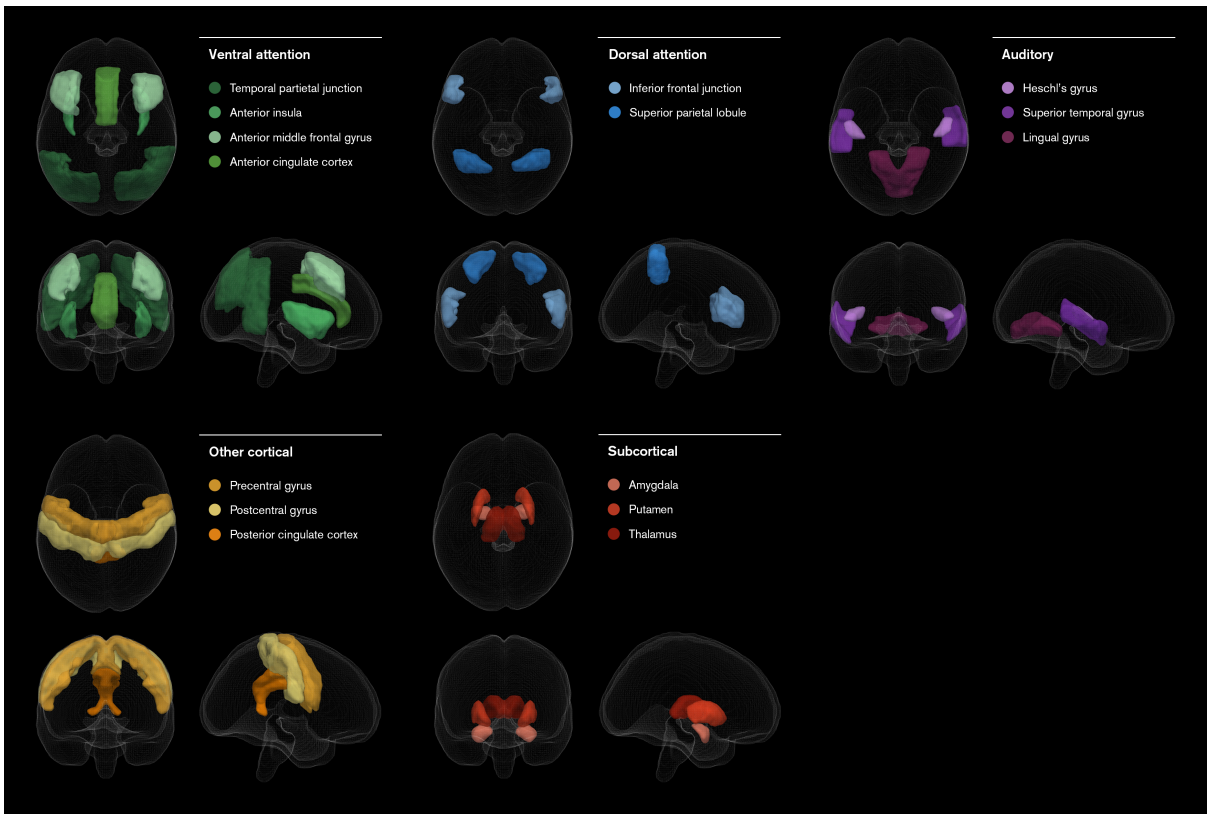

FIGURE 1 — Regions-of-interest and systemic categorisations applied to the fMRI contrasts, defined *a priori* according to a meta-analysis of auditory oddball processing tasks by Kim (24). For Harvard-Oxford cortical and subcortical atlas identifiers, refer to Supplemental Table S1.

The behavioural measures comprised the sensitivity, specificity, precision and mean reaction time of responses to target stimuli, as well as the ex-Gaussian parameters μ, σ and τ (25) summarising the distribution of reaction times over the course of the experiment obtained via the *exgauss* toolbox for MATLAB (version 1.3; 26).

The set of potential confounds comprised age and education, both of which have also been associated with auditory prediction error signals on a univariate basis (17), scanner field strength, which varies between sites, and cumulative antipsychotic drug exposure (20, 27), which is thought to alter neuroanatomy.

Framework

The machine learning framework encompasses an *ensemble* of domain *experts* (28), each trained on a given feature set, which are integrated using a multi-stage fusion tree. The machine learning framework, as illustrated in Figure 2, was implemented using the Scikit-learn (version 0.20.2; 29) and NumPy (version 1.15.4; 30) libraries for the Python programming language (version 3.6.5; Python Software Foundation). All experts and fusion models were trained using the *lasso* algorithm (31), which constrains the size of the model coefficients through regularisation and setting a subset of feature coefficients to zero. The result is a sparse formulation with implicit dimensionality reduction and a high level of model interpretability.

To make predictions for each individual subject, we employed a 10-fold cross-validation scheme with 10 repetitions, stratified by site with the regularisation hyper-parameter α optimised using a nested 9-fold cross-validation scheme.

In Stage 1, a set of experts (denoted f) were each independently trained on data extracted from a single ROI, such that each voxel was a feature and each region was a set of features (X) with one expert per region. This process was repeated for both target and novel contrasts. Each target score (y) was rescaled to a zero to one range, as per the maximum possible score on the given scale. All features were standardised to their respective z-scores, rescaling across participants to zero mean and unit variance. Dimensionality reduction was performed using principal components analysis, projecting the data onto a subset of components which explained 90% of the total variance, such that the number of features was much less than the number of samples ($M \ll N$). Collectively, this set of experts provides a set of region-based predictions ($y_{1\text{-}30a}$ and $y_{1\text{-}30b}$) for both the target and novel conditions (subscripts a and b, respectively).

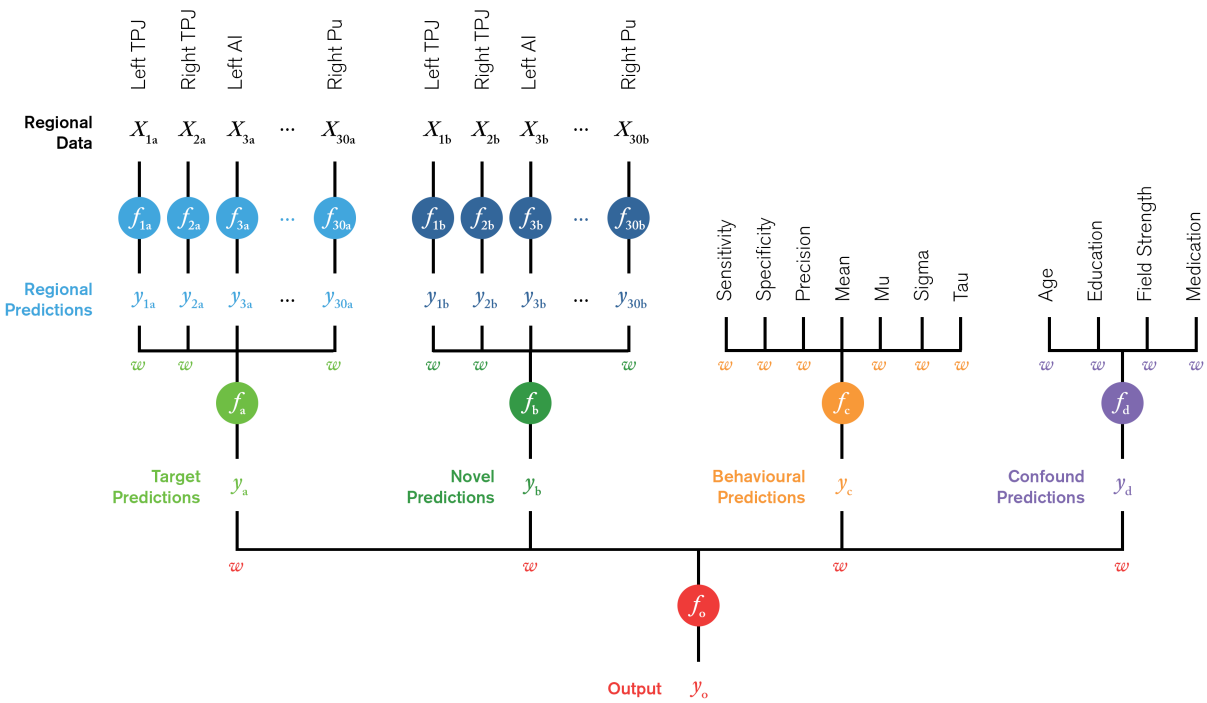

FIGURE 2 — Schematic of machine learning framework. In Stage 1, a set of experts (blue) are independently trained on a subset of fMRI data extracted from one of 30 regions and one of two experimental conditions; target, *a*, and novel, *b*. In Stage 2, the region-based predictions, $y_{1\text{-}30a}$ and $y_{1\text{-}30b}$, are then fused to obtain conditional predictions, y_a and y_b (green). Experts are also trained on the behavioural (orange) and confound (purple) feature sets. In Stage 3, the conditional, behavioural and confound predictions, $y_{a\text{-}d}$, are fused to form a final output prediction, y_o (red). Each fusion model assigns a set of weights, w, to each feature set which is used to intuit the relative feature importance in making predictions.

In Stage 2, these regional predictions were taken as inputs to a pair of secondary fusion models (f_a and f_b) which return a single conditional prediction (y_a and y_b) and a set of weights assigned to each region ($w_{1\text{-}30a}$ and $w_{1\text{-}30b}$). Experts based on the behavioural and confound feature sets (f_c and f_d) are also introduced, again assigning weights to each feature. These non-neuroimaging features were standardised to the z-scores as in Stage 1, however, did not require further dimensionality reduction given $M \ll N$.

In Stage 3, the conditional, behavioural and confound predictions ($y_{a\text{-}d}$) are fused to form a final prediction (y_o) and output weights were assigned to each of the categorical feature sets ($w_{a\text{-}d}$).

This framework was used to train a set of models to predict each of the symptom subscales and summary scores outlined in Table 1. A late fusion approach was then applied to the

subscale predictions by creating summary score ensemble models, taking the expert predictions from each of the subscale models and averaging across them to obtain predictions of the SAPS and SANS.

Model performance was evaluated by comparison of true and predicted scores using the mean-squared error (MSE) and Pearson's correlation coefficient (R). The statistical significance of each model was tested by 1000 permutations of the target variables, with $p < 0.05$ for both metrics indicating that the model has truly learned some pattern within the data, subject to correction for multiple comparisons.

Results

Predicting individual symptom subscales

These data were found to be predictive of some, but not all symptom subscales, as shown in Table 1. The negative symptoms of avolition, anhedonia and attention all had statistically significant correlations between targets and predictions ranging from 0.52 to 0.60 ($p < 0.013$, Bonferroni corrected), whilst the positive symptom of hallucination had the highest correlation of 0.72 ($p < 0.013$, Bonferroni corrected).

All models had a comparable mean-squared error, ranging from 0.047 to 0.054, which translates to approximately 23% error on the original scale. Although the delusions and formal thought disorder models had seemingly significant correlations of 0.35 and 0.40, inspection of the prediction plots shown in Figure 3 indicates that these models were constrained in their predictions, suggesting a possible bias toward the sample means (0.53 and 0.15, respectively).

TABLE 1 — Summary of model performance. Avolition, anhedonia, attention and hallucinations yielded best performance for individual symptoms, whilst the ensemble of subscales outperformed singular summary score models. All *p*-values were computed via 1000 permutations with Bonferroni correction for multiple comparisons.

			R	*p*-value		**MSE**	*p*-value	
				uncorrected	Bonferroni		uncorrected	Bonferroni
Individual dimensions	Negative symptoms	Affective Flattening	**0.03**	0.374	1.000	**0.071**	0.030	0.390
		Alogia	**-0.15**	0.893	1.000	**0.049**	0.504	1.000
		Avolition	**0.59**	0.001	0.013	**0.049**	0.001	0.013
		Anhedonia	**0.52**	0.001	0.013	**0.050**	0.001	0.013
		Attention	**0.60**	0.001	0.013	**0.047**	0.001	0.013
	Positive symptoms	Delusions	**0.35**	0.002	0.026	**0.060**	0.002	0.026
		Hallucinations	**0.72**	0.001	0.013	**0.054**	0.001	0.013
		Bizarre Behaviour	**-0.10**	0.808	1.000	**0.043**	0.724	1.000
		Formal Thought Disorder	**0.40**	0.002	0.026	**0.046**	0.001	0.013
Summary scores	Single model	SAPS	**0.24**	0.021	0.273	**0.062**	0.004	0.052
		SANS	**0.39**	0.001	0.013	**0.019**	0.001	0.013
	Ensemble model	SAPS	**0.51**	0.001	0.013	**0.018**	0.001	0.013
		SANS	**0.49**	0.001	0.013	**0.017**	0.001	0.013

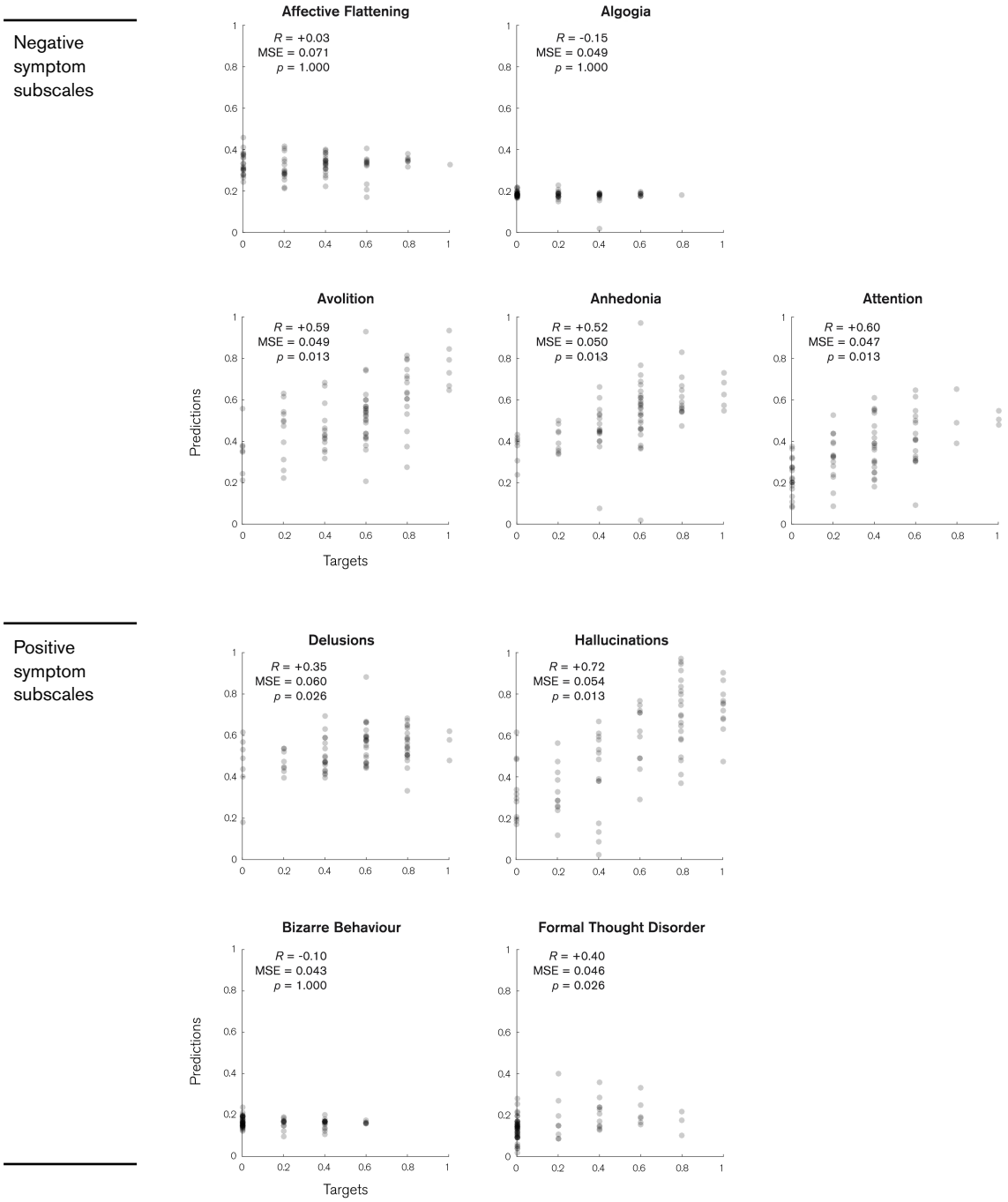

FIGURE 3 — Plots of model predictions and true scores for individual subscales within the SAPS and SANS. Predictions for each subject are shown as grey dots. All scores are rescaled to a zero to one scale. *p*-values are Bonferroni corrected for multiple comparisons.

Predicting symptom summary scores

We applied two different approaches in predicting the SAPS and SANS summary scores. In the first case, a single model was trained directly on the summary scores by using the same framework as depicted in Figure 2. As shown in Table 1 and Figure 4, the composite positive symptom model was not statistically significant and the negative symptoms only yield a modest correlation. However, when considering each of the individual subscale models as an expert on a particular symptom within an ensemble which collectively predicts the summary score, we found a marked improvement in performance. Both of these ensemble models proved statistically significant ($p < 0.013$, Bonferroni corrected), with positive symptoms demonstrating a two-fold increase in correlation (0.24 to 0.51) and a decrease in mean-square error (0.062 to 0.018). On average, this translates to an approximate 13.4% error margin.

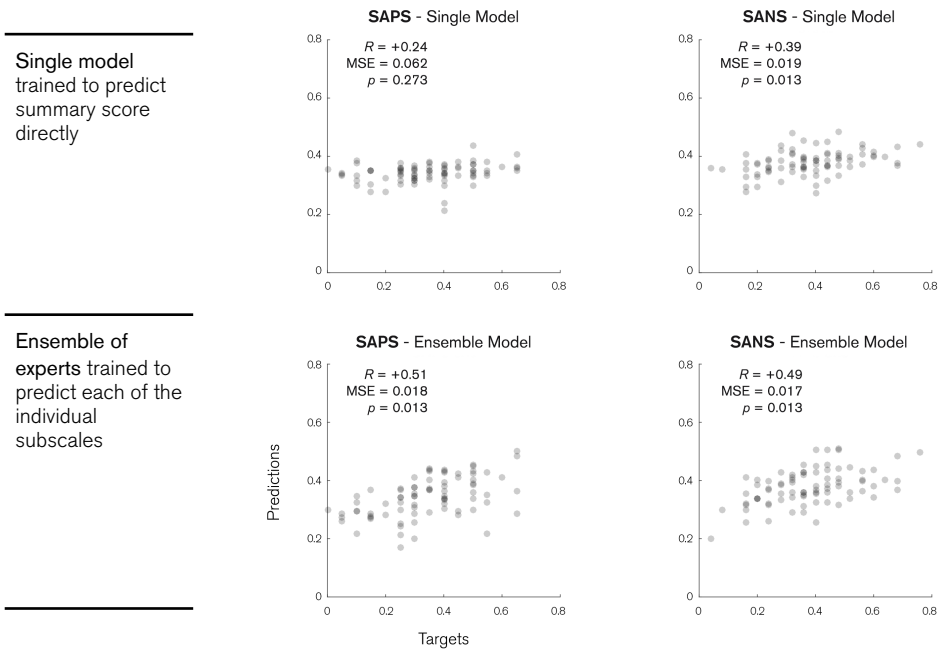

FIGURE 4 — Plots of model predictions and true scores for the SAPS (left) and SANS (right) summary scores. Top row shows the predictions for a single model trained to predict the summary scores directly. Bottom row shows predictions for an ensemble model, where each expert is trained to predict one of the individual subscales. Predictions for each subject are shown as grey dots. All scores are rescaled to a zero to one scale. p-values are Bonferroni corrected for multiple comparisons.

Subscale model explanations

Given that we are able to predict a number of symptoms from the available data, we are also able to intuit the main factors which underpin model performance by examining the contributions of each feature set in a *post hoc* manner. This is achieved by examining the fusion weights assigned to each feature set in Stage 3 and individual features in Stage 2.

Firstly, we wish to establish whether neuroimaging is indeed useful, given the added time and monetary investment necessary to acquire these data. The target and novel feature sets were highly weighted in the Stage 3 output fusion (between 72 and 100%) in comparison with the behavioural and confound feature sets (0 to 17% and 0 to 11%, respectively) in each of our statistically significant models. This indicates that the neuroimaging data was the main driver behind the predictions, above and beyond the behavioural data obtained via the task itself. However, excluding those for the anhedonia model, the behavioural and confound coefficients were not set to zero, therefore these features still had some, albeit negligible contribution to the final predictions (Supplemental Table S2).

To better understand the contribution of individual neuroanatomical features within these neuroimaging feature sets, the output weights from Stage 3 were applied to the categorical fusion weights from Stage 2. For each score, we are then able to obtain a regional weight map, as illustrated in the main 4 × 15 matrices of Figure 5a. Here, each element represents a single feature with colour indicating which features were identified by the algorithm as important, and conversely, which were non-informative. In the top panel, each row represents a region-of-interest, with columns indicating the hemisphere (denoted L and R) and experimental stimulus (target and novel). Furthermore, by adding these elements together, we can collapse across subsets of features to summarise the broader system-wide differences between each brain system (4 × 5 matrix), hemisphere (4 × 1 vector) and experimental condition (bottom 2 × 1 vector). For visualisation purposes, Figure 5b shows the regional weight maps projected onto a three-dimensional representation of the original regions of interest, collapsed across conditions.

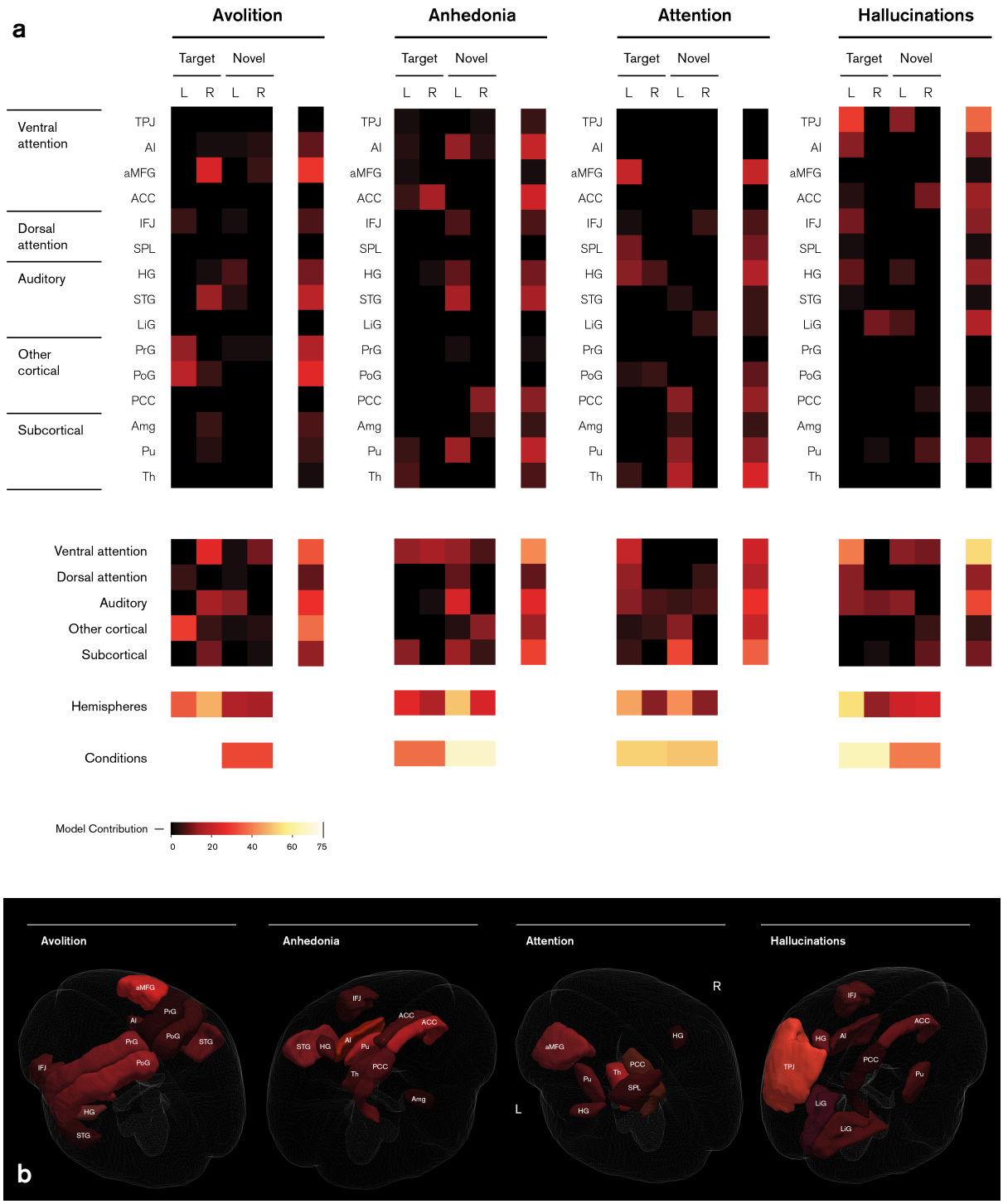

FIGURE 5 — Feature importance for individual subscale models. Colours denote the contribution of each feature toward predictions as a percentage, with black indicating an entire feature set has been marked as irrelevant by the *lasso* algorithm. **(a)** Regional weight maps (top panel) show the relevance of each region (rows) in both left (L) and right (R) hemispheres under target and novel conditions (columns). Categorical weight maps (bottom panel) show the net contribution of each system, hemisphere and condition toward model predictions. Categorical groups of regions are based on Kim (24). **(b)** Regional weight maps projected onto three-dimensional brain structures.

In the avolition model, predictions were mainly driven by the target response, namely the right anterior middle frontal gyrus (aMFG) and other cortical activity in the left hemisphere. For anhedonia, predictions were primarily informed by the anterior cingulate cortex (ACC) target response and left superior temporal gyrus (STG) novel response. Attention predictions were equally driven by both target and novel stimuli, in particular the left aMFG response to target stimuli and subcortical activity in the novel condition. Hallucinations were informed by the left hemispheric target response, principally the ventral attention network and temporal parietal junction (TPJ). Critically, pairwise similarity measures between weight vectors indicated that each symptom had its own distinct pattern of activity across different sets of regions (Supplemental Figure S2).

Summary model explanations

The difference between the two approaches in predicting summary scores is most apparent when comparing the respective weight maps. For the single positive symptoms model (Figure 6a), we can observe that of all the possible combinations of features, the optimal solution computed by the algorithm comprises a single feature — the response to target stimuli in the left precentral gyrus. Given the low performance of this model, this can be attributed to a classic case of underfitting. Conversely, the ensemble model weight map includes the specific contributions toward each of the individual subscales, resulting in a widespread distribution of features across the whole brain. Notably, we observe that the precentral gyrus is not a member of the ensemble weight map, nor any of the constituent subscales. Similarly, the negative symptom models (Figure 6b) demonstrate the same pattern, with the single model reduced to the left inferior frontal junction (IFJ) and right Heschl's gyrus (HG) responses to the target condition, both of which are down-weighted in the equivalent ensemble model.

These comparisons demonstrate that the ensemble of subscales clearly outperforms the single model approach, not only in terms of predictive accuracy, but also in terms of identifying plausible functional neuroanatomical maps — the composite negative and positive symptoms arise from widespread brain networks, whereas individual symptoms pertain to more nuanced sub-networks.

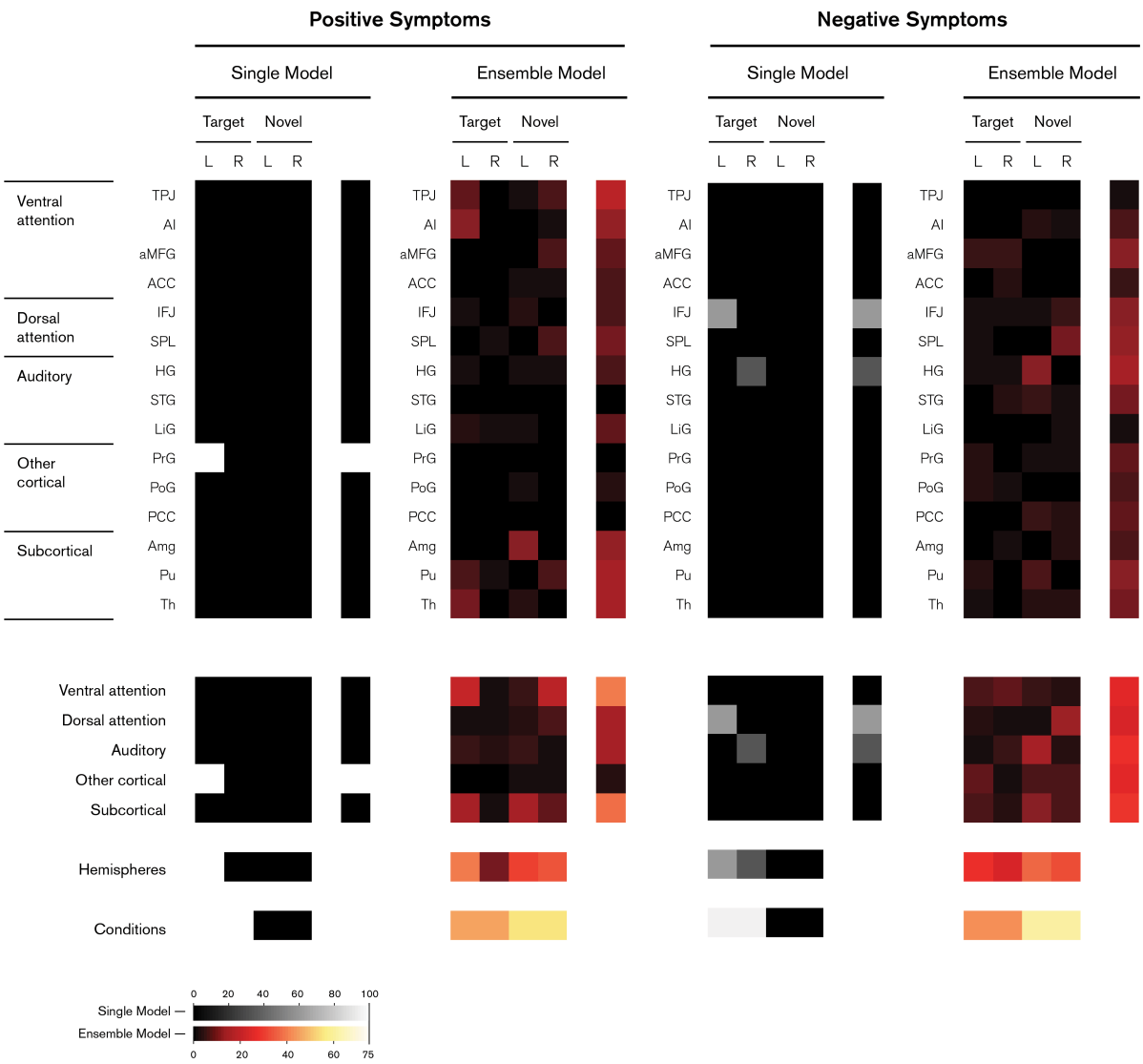

FIGURE 6 — Feature importance for SAPS and SANS summary score models. Ensemble models are shown in warm colour map and single models shown in greyscale. Regional weight maps (top panel) show the relevance of each region (rows) in both left (L) and right (R) hemispheres under target and novel conditions (columns). Categorical weight maps (bottom panel) show the net contribution of each system, hemisphere and condition toward model predictions. Categorical groups of regions are based on Kim (24). Colours denote the contribution of each feature toward predictions as a percentage, with black indicating an entire feature set has been marked as irrelevant by the *lasso* algorithm.

Discussion

In this study, we used multivariate machine learning regression techniques to predict the severity of schizophrenia symptoms on a continuum based on the neural and behavioural responses to an auditory oddball task. By training a set of models to predict each symptom subscale independently, these data were found to be highly predictive of hallucinations, attention, avolition and anhedonia. We also found that by modelling the composite SAPS and SANS summary scores as ensembles of these subscales, the accuracy of predictions significantly increased, whereas single models trained to predict summary scores directly demonstrably underfit to irrelevant features.

Interpreting functional anatomy of psychotic symptoms

Anhedonia is described as a reduced capacity to experience pleasant emotions (32). We found that anhedonia predictions were mainly driven by responses to the novel stimulus in the left hemisphere, in particular the putamen and STG, as well as the target response in the ACC. The ACC is known to play a key role in reward processing (33) and has previously been linked to anticipation of pleasant events (34) and self-referencing (35) in schizophrenia. Deep brain stimulation of the ACC has also been shown to modulate anhedonia-like symptoms (36). To the best of our knowledge, the putamen and STG have not been linked to anhedonia in schizophrenia specifically, however have been previously reported in depression, which has high comorbidity with anhedonia (32). The putamen is connected to the motor cortices and is thought to encode associations between stimuli, actions and rewards (37). Notably, major depressive disorder patients (MDD) with anhedonia have a two-fold age-related putamen volume decrease in comparison with healthy controls (38). Additionally, the STG, primarily involved in auditory and language processing, has a reduced response in first-episode MDD patients with anhedonia when comparing probable vs. improbable rewards (39). STG volume reduction has also been widely reported in schizophrenia patients (40). Collectively, these findings may imply that anhedonia leads to a lower anticipation for rewards upon performing the required action, as reflected in changes within the putamen, STG and ACC.

Predictions of avolition, i.e. a lower pursuit and persistence of goal-directed activities (32), were predominantly informed by the target response in the right aMFG. This region has been previously associated with processing of conflicting information (41) and is reported to have

reduced activity under working memory load in those at ultra-high risk for psychosis (42). Target responses in the left precentral and postcentral gyri also contributed to the prediction, albeit to a lesser degree. Alterations in activity within these sensorimotor regions may suggest that those with avolition are required to make an increased effort in response to stimuli which demand a physical action.

Predictions of attention scores were largely informed by the target response in the left aMFG, a region engaged in tasks requiring divided attention (43). In schizophrenia, activity in the left MFG during sustained attention has been previously been shown to correlate with compound negative symptoms on the PANSS scale (44). Patients with brain tumours in the left MFG also show significant reductions in flexible attention and cognition (45). Volitional or self-initiated shifts in attention in the absence of instructional cues have been associated with both left and right MFG activity (46, 47). Interestingly, the thalamic response to the novel condition was also highly weighted, which is known to filter distracting or conflicting information (48, 49). Given that participants were instructed to ignore novel stimuli, the thalamus involvement here may be inhibiting these distractors and allowing for increased selective attention.

Our model for hallucinations was primarily driven by the target response within the ventral attention network and the left TPJ, a known critical node in the speech perception network implicated in hallucinations (50, 51). The left TPJ has previously been used as a target area for transcranial direct-current stimulation (tDCS), leading to a reduction in hallucinations in schizophrenia patients (52). In turn, hallucinations following tDCS have been shown to correlate with the functional connectivity between the left TPJ and left AI (53). This is consistent with our findings, which identify both of these regions as highly predictive of hallucination severity.

Further studies investigating symptom-specific circuitries may open new possibilities for informing future personalised treatments. In the future, if we were to obtain a robust brain mapping for each symptom based on consistent replications of these findings, brain stimulation or pharmacological interventions could be tailored for an individual based on their symptom profile with dosages relative to the level of severity. This could also provide opportunities for reverse translation into an array of symptom-based animal models of schizophrenia.

Methodological considerations

We employed a multi-modal fusion approach which has previously been applied to neuroimaging data for combining different types of structural and functional images (54-56). When building an integrative model from a dataset comprising multiple distinct feature sets, there are two general approaches for fusing these features to form a single prediction. In an *early fusion* approach, features from each modality can be merged prior to the learning process with the joint representation input to a single model providing a multimodal prediction (57). Alternatively, in a *late fusion* approach, a set of models can be trained on each feature set independently and the unimodal predictions from each are combined to form an overall multimodal prediction, typically through averaging or a secondary linear model. The key methodological advance presented in this study is that we perform fusion not only on the basis of the data structure, but also on the *target variables*. The psychometric tools used in clinical practice, such as the SAPS and SANS, are by definition multidimensional — they comprise a set of symptom subscales which are assessed independently, then combined to obtain a composite summary score. Our results suggest that within a machine learning context, a late fusion of subscale predictions as per the original diagnostic framework provides greater specificity and accuracy than the early fusion equivalent of predicting the summary scores directly. Together with the drawn region-based weight maps, this finding also suggests that each symptom has a distinct functional anatomic pattern, whereas models trained to predict positive and negative composite scores directly were unable to capture this nuanced information. In principle, we envisage that modelling the symptom subscales as an ensemble could be applied to any psychiatric disorder.

In a typical neuroimaging context, the number of available features vastly outweighs the number of samples, a phenomenon known as the curse of dimensionality (58). This often leads to an overfitting of model parameters, which in turn may not generalise to new samples (1). To address this issue, we chose to adopt a multi-tier fusion tree which is conceptually similar to a well-established approach known as stacked regression (59). This enabled us to iteratively reduce the dimensionality of the data from the original voxel space whilst also improving model interpretability by expressing the relevant features in more general terms of brain regions and networks. As such, the functional anatomy which informs model predictions is more interpretable by those familiar with the pathology of the disease and relatable to other univariate studies.

Note that whilst some neuroimaging features may have a greater contribution than others, all features with non-zero weights contribute to the model predictions. Although a highly

weighted feature is unlikely to be a false positive given adequate signal-to-noise ratio (60), the contribution of any one single feature should be interpreted with caution. For example, although left TPJ activity in the target condition may be of greatest importance in our hallucinations model, this feature alone only contributes approximately 24% toward predictions. Given that the optimal solution for this model also includes 17 other regions, the accuracy would likely decrease without these additional features. By collapsing across conditions, hemispheres or networks, we are able to summarise the contributions of subsets of these features, but cannot fully comprehend the interactions between them within the context of the whole feature set.

Conclusion

In conclusion, we suggest that a shift from dichotomous labels toward a multi-dimensional diagnostic approach which delineates the severity of each individual symptom in a piecewise manner may be more useful for parsing the underlying neurobiological causes. In turn, this framework may help improve our understanding of how specific symptoms relate to functional differences within the brain, in schizophrenia or indeed any other psychiatric disorder.

References

1. Arbabshirani MR, Plis S, Sui J, Calhoun VD (2017): Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*. 145, Part B:137-165.
2. Janssen RJ, Mourão-Miranda J, Schnack HG (2018): Making Individual Prognoses in Psychiatry Using Neuroimaging and Machine Learning. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*.
3. Walter M, Alizadeh S, Jamalabadi H, Lueken U, Dannlowski U, Walter H, et al. (2019): Translational machine learning for psychiatric neuroimaging. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*. 91:113-121.
4. Cuthbert BN, Insel TR (2013): Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Medicine*. 11:126.
5. Bzdok D, Meyer-Lindenberg A (2018): Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. 3:223-230.
6. Regier DA, Kuhl EA, Kupfer DJ (2013): The DSM-5: Classification and criteria changes. *World Psychiatry*. 12:92-98.
7. Andreasen NC (1984): The Scale for the Assessment of Positive Symptoms (SAPS). Iowa City: University of Iowa.
8. Andreasen NC (1983): The Scale for the Assessment of Negative Symptoms (SANS). Iowa City: University of Iowa.
9. Kay SR, Fiszbein A, Opler LA (1987): The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. *Schizophrenia Bulletin*. 13:261-276.
10. Tognin S, Pettersson-Yeo W, Valli I, Hutton C, Woolley J, Allen P, et al. (2014): Using Structural Neuroimaging to Make Quantitative Predictions of Symptom Progression in Individuals at Ultra-High Risk for Psychosis. *Frontiers in Psychiatry*. 4.
11. Tolmeijer E, Kumari V, Peters E, Williams SCR, Mason L (2018): Using fMRI and machine learning to predict symptom improvement following cognitive behavioural therapy for psychosis. *NeuroImage: Clinical*. 20:1053-1061.
12. Taylor JA, Matthews N, Michie PT, Rosa MJ, Garrido MI (2017): Auditory prediction errors as individual biomarkers of schizophrenia. *NeuroImage: Clinical*. 15:264-273.
13. Sui J, Qi S, van Erp TGM, Bustillo J, Jiang R, Lin D, et al. (2018): Multimodal neuromarkers in schizophrenia via cognition-guided MRI fusion. *Nature Communications*. 9:3028-3028.
14. Ranlund S, Rosa MJ, de Jong S, Cole JH, Kyriakopoulos M, Fu CHY, et al. (2018): Associations between polygenic risk scores for four psychiatric illnesses and brain structure using multivariate pattern recognition. *NeuroImage: Clinical*. 20:1026-1036.

15. Zhao X, Yao J, Lv Y, Zhang X, Han C, Chen L, et al. (2018): Abnormalities of regional homogeneity and its correlation with clinical symptoms in Naïve patients with first-episode schizophrenia. *Brain Imaging and Behavior*.

16. Vanes LD, Mouchlianitis E, Patel K, Barry E, Wong K, Thomas M, et al. (2019): Neural correlates of positive and negative symptoms through the illness course: an fMRI study in early psychosis and chronic schizophrenia. *Scientific Reports*. 9:14444.

17. Erickson MA, Albrecht M, Ruffle A, Fleming L, Corlett P, Gold J (2017): No association between symptom severity and MMN impairment in schizophrenia: A meta-analytic approach. *Schizophrenia Research: Cognition*. 9:13-17.

18. Carrà G, Crocamo C, Angermeyer M, Brugha T, Toumi M, Bebbington P (2019): Positive and negative symptoms in schizophrenia: A longitudinal analysis using latent variable structural equation modelling. *Schizophrenia Research*. 204:58-64.

19. Tibber MS, Kirkbride JB, Joyce EM, Mutsatsa S, Harrison I, Barnes TRE, et al. (2018): The component structure of the scales for the assessment of positive and negative symptoms in first-episode psychosis and its dependence on variations in analytic methods. *Psychiatry Research*. 270:869-879.

20. Gollub RL, Shoemaker JM, King MD, White T, Ehrlich S, Sponheim SR, et al. (2013): The MCIC Collection: A Shared Repository of Multi-Modal, Multi-Site Brain Image Data from a Clinical Investigation of Schizophrenia. *Neuroinformatics*. 11:367-388.

21. Scott A, Courtney W, Wood D, De la Garza R, Lane S, Wang R, et al. (2011): COINS: An Innovative Informatics and Neuroimaging Tool Suite Built for Large Heterogeneous Datasets. *Frontiers in Neuroinformatics*. 5.

22. Andreasen NC, Flaum M, Arndt S (1992): The Comprehensive Assessment of Symptoms and History (CASH): An Instrument for Assessing Diagnosis and Psychopathology. *Archives of General Psychiatry*. 49:615-623.

23. Kiehl KA, Liddle PF (2001): An event-related functional magnetic resonance imaging study of an auditory oddball task in schizophrenia. *Schizophrenia Research*. 48:159-171.

24. Kim H (2014): Involvement of the dorsal and ventral attention networks in oddball stimulus processing: A meta-analysis. *Human Brain Mapping*. 35:2265-2284.

25. Lacouture Y, Cousineau D (2008): How to use MATLAB to fit the ex-Gaussian and other probability functions to a distribution of response times. *TQMP*. 4:35-45.

26. Bram Z (2014): exgauss: a MATLAB toolbox for fitting the ex-Gaussian distribution to response time data. figshare.

27. Andreasen NC, Pressler M, Nopoulos P, Miller D, Ho B-C (2010): Antipsychotic dose equivalents and dose-years: a standardized method for comparing exposure to different drugs. *Biological Psychiatry*. 67:255-262.

28. Hastie T, Friedman J, Tibshirani R (2001): The Elements of Statistical Learning: Data Mining, Inference and Prediction. 1st ed. New York, NY: Springer.

29. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. (2011): Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 12:2825-2830.

30. Oliphant TE (2015): *Guide to NumPy*. 2nd ed. USA: Continuum Press.

31. Tibshirani R (1996): Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B (Methodological)*. 58:267-288.

32. Marder SR, Galderisi S (2017): The current conceptualization of negative symptoms in schizophrenia. *World Psychiatry*. 16:14-24.

33. Heshmati M, Russo SJ (2015): Anhedonia and the Brain Reward Circuitry in Depression. *Current Behavioral Neuroscience Reports*. 2:146-153.

34. Choi S-H, Lee H, Ku J, Yoon KJ, Kim J-J (2014): Neural basis of anhedonia as a failure to predict pleasantness in schizophrenia. *The World Journal of Biological Psychiatry*. 15:525-533.

35. Lee JS, Kim ES, Kim EJ, Kim J, Kim E, Lee S-K, et al. (2016): The relationship between self-referential processing-related brain activity and anhedonia in patients with schizophrenia. *Psychiatry Research: Neuroimaging*. 254:112-118.

36. Schlaepfer TE, Cohen MX, Frick C, Kosel M, Brodesser D, Axmacher N, et al. (2008): Deep Brain Stimulation to Reward Circuitry Alleviates Anhedonia in Refractory Major Depression. *Neuropsychopharmacology*. 33:368-377.

37. Balleine BW, Delgado MR, Hikosaka O (2007): The role of the dorsal striatum in reward and decision-making. *The Journal of Neuroscience*. 27:8161-8165.

38. Sacchet MD, Camacho MC, Livermore EE, Thomas EAC, Gotlib IH (2017): Accelerated aging of the putamen in patients with major depressive disorder. *Journal of Psychiatry & Neuroscience*. 42:164-171.

39. Yang X-h, Huang J, Lan Y, Zhu C-y, Liu X-q, Wang Y-f, et al. (2016): Diminished caudate and superior temporal gyrus responses to effort-based decision making in patients with first-episode major depressive disorder. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*. 64:52-59.

40. Javitt DC, Sweet RA (2015): Auditory dysfunction in schizophrenia: integrating clinical and basic features. *Nature Reviews Neuroscience*. 16:535.

41. Marini F, Demeter E, Roberts KC, Chelazzi L, Woldorff MG (2016): Orchestrating Proactive and Reactive Mechanisms for Filtering Distracting Information: Brain-Behavior Relationships Revealed by a Mixed-Design fMRI Study. *The Journal of Neuroscience*. 36:988-1000.

42. Fusar-Poli P, Howes OD, Allen P, Broome M, Valli I, Asselin M-C, et al. (2010): Abnormal frontostriatal interactions in people with prodromal signs of psychosis: a multimodal imaging study. *Archives of General Psychiatry*. 67:683-691.

43. Salo E, Salmela V, Salmi J, Numminen J, Alho K (2017): Brain activity associated with selective attention, divided attention and distraction. *Brain Research*. 1664:25-36.

44. Curtin A, Sun J, Zhao Q, Onaral B, Wang J, Tong S, et al. (2019): Visuospatial task-related prefrontal activity is correlated with negative symptoms in schizophrenia. *Scientific Reports*. 9:9575-9575.

45. De Baene W, Rijnen SJM, Gehring K, Meskal I, Rutten G-JM, Sitskoorn MM (2019): Lesion symptom mapping at the regional level in patients with a meningioma. *Neuropsychology*. 33:103-110.

46. Bengson JJ, A. Kelley T, Mangun GR (2015): The neural correlates of volitional attention: A combined fMRI and ERP study. *Human Brain Mapping*. 36:2443-2454.

47. Gmeindl L, Chiu Y-C, Esterman MS, Greenberg AS, Courtney SM, Yantis S (2016): Tracking the will to attend: Cortical activity indexes self-generated, voluntary shifts of attention. *Atten Percept Psychophys*. 78:2176-2184.

48. Pinault D (2011): Dysfunctional Thalamus-Related Networks in Schizophrenia. *Schizophrenia Bulletin*. 37:238-243.

49. Wolff M, Vann SD (2019): The Cognitive Thalamus as a Gateway to Mental Representations. *The Journal of Neuroscience*. 39:3.

50. Vercammen A, Knegtering H, den Boer JA, Liemburg EJ, Aleman A (2010): Auditory Hallucinations in Schizophrenia Are Associated with Reduced Functional Connectivity of the Temporo-Parietal Area. *Biological Psychiatry*. 67:912-918.

51. Chahine G, Richter A, Wolter S, Goya-Maldonado R, Gruber O (2017): Disruptions in the left frontoparietal network underlie resting state endophenotypic markers in schizophrenia. *Human Brain Mapping*. 38:1741-1750.

52. Brunelin J, Mondino M, Gassab L, Haesebaert F, Gaha L, Suaud-Chagny M-F, et al. (2012): Examining Transcranial Direct-Current Stimulation (tDCS) as a Treatment for Hallucinations in Schizophrenia. *American Journal of Psychiatry*. 169:719-724.

53. Mondino M, Jardri R, Suaud-Chagny M-F, Saoud M, Poulet E, Brunelin J (2016): Effects of Fronto-Temporal Transcranial Direct Current Stimulation on Auditory Verbal Hallucinations and Resting-State Functional Connectivity of the Left Temporo-Parietal Junction in Patients With Schizophrenia. *Schizophrenia Bulletin*. 42:318-326.

54. Calhoun VD, Adali T (2009): Feature-based fusion of medical imaging data. *IEEE Transactions on Information Technology in Biomedicine*. 13:711-720.

55. Lahat D, Adali T, Jutten C (2015): Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proceedings of the IEEE*. 103:1449-1477.

56. Calhoun VD, Sui J (2016): Multimodal Fusion of Brain Imaging Data: A Key to Finding the Missing Link(s) in Complex Mental Illness. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. 1:230-244.

57. Snoek CGM, Worring M, Smeulders AWM (2005): Early versus late fusion in semantic video analysis. *13th annual ACM international conference on Multimedia*. Singapore: ACM, pp 399-402.

58. Bellman RE (1961): Adaptive Control Processes: A Guided Tour. Princeton University Press.

59. Breiman L (1996): Stacked Regressions. *Machine Learning*. 24:49-64.

60. Schrouff J, Mourão-Miranda J (2018): Interpreting weight maps in terms of cognitive or clinical neuroscience: nonsense? *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, pp 1-4.