Deep Learning for Quality Control of Subcortical Brain 3D Shape Models

The ENIGMA Consortium

The full author list appears at the end of the paper

Abstract. We present several deep learning models for assessing the morphometric fidelity of deep grey matter region models extracted from brain MRI. We test three different convolutional neural net architectures (VGGNet, ResNet and Inception) over 2D maps of geometric features. Further, we present a novel geometry feature augmentation technique based on parametric spherical mapping. Finally, we present an approach for model decision visualization, allowing human raters to see the areas of subcortical shapes most likely to be deemed of failing quality by the machine. Our training data is comprised of 5200 subjects from the ENIGMA Schizophrenia MRI cohorts, and our test dataset contains 1500 subjects from the ENIGMA Major Depressive Disorder cohorts. Our final models reduce human rater time by 46-70%. ResNet outperforms VGGNet and Inception for all of our predictive tasks.

Keywords: deep learning, subcortical shape analysis, quality checking

1 Introduction

Quality control (QC) has become one of the main practical bottlenecks in big-data neuroimaging. Reducing human rater time via predictive modeling and automated quality control is bound to play an increasingly important role in maintaining and hastening the pace of scientific discovery in this field. Recently, the UK Biobank publicly released over 10,000 brain MRIs (and planning to release 90,000 more); as other biobanking initiatives scale up and follow suit, automated QC becomes crucial.

In this paper, we investigate the viability of deep convolutional neural nets for automatically labeling deep brain regional geometry models of failing quality after their extraction from brain MR images. We compare the performance of VGGNet, ResNet and Inception architectures, investigate the robustness of probability thresholds, and visualize decisions made by the trained neural nets. Our data consists of neuroimaging cohorts from the ENIGMA Schizophrenia and Major Depressive Disorder working groups participating in the ENIGMA-Shape project [1]. Using ENIGMAs shape analysis protocol and rater-labeled shapes, we train a discriminative model to separate FAIL(F) and PASS(P) cases. Features are derived from standard vertex-wise measures.

For all seven deep brain structures considered, we are able to reduce human rater time by 46 to 70 percent in out-of-sample validation, while maintaining

FAIL recall rates similar to human inter-rater reliability. Our models generalize across datasets and disease samples. Our models' decision visualization, particularly ResNet, appears to capture structural abnormalities of the poor quality data that correspond to human raters' intuition.

With this paper, we also release to the community the feature generation code based on FreeSurfer outputs, as well as pre-trained models and code for model decision visualization.

2 Methods

Our goal in using deep learning (DL) for automated QC differs somewhat from most predictive modeling problems. Typical two-class discriminative solutions seek to balance misclassification rates of each class. In the case of QC, we focus primarily on correctly identifying FAIL cases, by far the smaller of the two classes **(Table 1)**. In this first effort to automate shape QC, we do not attempt to eliminate human involvement, but simply to reduce it by focusing human rater time on a smaller subsample of the data containing nearly all the failing cases.

2.1 MRI processing and shape features

Our deep brain structure shape measures are computed using a previously described pipeline [2] [3], available via the ENIGMA Shape package. Briefly, structural MR images are parcellated into cortical and subcortical regions using FreeSurfer. Among the 19 cohorts participating in this study, FreeSurfer versions 5.1 and 5.3 were used. The binary region of interest (ROI) images are then surfaced with triangle meshes and spherically registered to a common region-specific template [4]. This leads to a one-to-one surface correspondence across the dataset at roughly 2,500 vertices per ROI. Our ROIs include the left and right thalamus, caudate, putamen, pallidum, hippocampus, amygdala, and nucleus accumbens. Each vertex p of mesh model $\mathcal{M}$ is endowed with two shape descriptors:

Medial Thickness, $D(p) = \|c_p - p\|$, where c_p is the point on the medial curve c closest to p.

$LogJac(p)$, Log of the Jacobian determinant J arising from the template mapping, $J : T_{\phi(p)}\mathcal{M}_t \to T_p\mathcal{M}$.

Since the ENIGMA surface atlas is in symmetric correspondence, i.e., the left and right shapes are vertex-wise symmetrically registered, we can combine the two hemispheres for each region for the purposes of predictive modeling. Though we assume no hemispheric bias in QC failure, we effectively double our sample.

The vertex-wise features above are augmented with their volume-normalized counterparts: $\{D, J\}_{normed}(p) = \frac{\{D,J\}(p)}{V^{\{\frac{1}{3},\frac{2}{3}\}}}$. Given discrete area elements of the template at vertex p, $A_t(p)$, we estimate volume as $V = \sum_{p \in vrts(\mathcal{M})} 3A_t(p)J(p)D(p)$. We normalize our features subject-wise by this volume estimate to control for subcortical structure size.

2.2 Human quality rating

Human-rated quality control of shape models is performed following the ENIGMA-Shape QC protocol (enigma.usc.edu/ongoing/enigma-shape-analysis). Briefly, raters are provided with several snapshots of each region model as well as its placement in several anatomical MR slices. A guide with examples of FAIL (QC=1) and PASS (QC=3) cases is provided to raters, with an additional category of MODERATE PASS (QC=2) suggested for inexperienced raters. With sufficient experience, the rater typically switches to the binary FAIL/PASS rating. In this work, all QC=2 cases are treated as PASS cases, consistent with ENIGMA shape studies.

2.3 Feature mapping to 2D images

Because our data resides on irregular mesh vertices, we first interpolate the features from an irregular spherical mesh onto an equiangular grid. The interpolated feature maps are then treated as regular 2D images by Mercator projection. Our map is based on the medial curve-based global orientation function (see [5]), which defines the latitude (θ) coordinate, as well as a rotational standardization of the thickness profile $D(p)$ to normalize the longitudinal (ϕ) coordinate. The resulting map normalizes the 2D appearance of $D(p)$, setting the poles to lie at the ends of the medial curve. In practice, the re-sampling is realized as matrix multiplication based on trilinear mesh interpolation, resulting in a 128×128 image for each measure.

2.4 Data augmentation

Although our raw sample of roughly 13,500 examples is exceptionally large by the standards of neuroimaging, this dataset may not be large enough to train generalizable CNNs. Standard image augmentation techniques, e.g. cropping and rotations, are inapplicable to our data. To augment our sample of spherically mapped shape features, we sample from a distribution of spherical deformations, i.e. changes in the spherical coordinates of the thickness and Jacobian features. To do this, we first sample from a uniform distribution of vector spherical harmonic coefficients $\boldsymbol{B}_{lm}, \boldsymbol{C}_{lm}$, and apply a heat kernel operator [4] to the generated field on $T\mathbb{S}^2$. Change in spherical coordinates is then defined based on the tangential projection of the vector field, as in [4]. The width σ of the heat kernel defines the level of smoothness of the resulting deformation, and the maximum point norm M defines the magnitude. In practice, each random sampling is a composition of a large magnitude, smooth deformation ($\sigma = 10^{-1}$, $M = 3 \times 10^{-1}$) and a smaller noisier deformation ($\sigma = 10^{-2}$, $M = 3 \times 10^{-2}$). Once the deformation is generated, it is applied to the spherical coordinates of the irregular mesh, and a new sampling matrix is generated, as above.

2.5 Deep learning models

We train VGGNet [6], ResNet [7] and Inception [8] architectures on our data. We chose these architectures as they perform well in traditional image classification problems and are well-studied.

2.6 Model decision visualization

Deep learning models tend to learn superficial statistical patterns rather than high-level global or abstract concepts [9]. As we plan to provide a tool that both (1) classifies morphometric shapes, and (2) allows a user to visualize what the machine perceives as a 'FAIL', model decision visualization is an important part of our work. Here, we use Prediction Difference Analysis [10], and Grad-CAM [11] to visualize 'bad' and 'good' areas for each particular shape in question.

2.7 Predictive model assessment

We use two sets of measures to evaluate the performance of our models. To assess the validity of the models' estimated 'FAIL' probabilities, we calculate the area under the ROC curve (ROC AUC). We also use two supplementary measures: FAIL-recall and FAIL-share. In describing them below, we use the following definitions. TF stands for TRUE FAIL, FF stands for FALSE FAIL, TP stands for TRUE PASS, and FP stands for FALSE PASS. Our first measure, **F-recall** $= \frac{TF}{TF+FP}$, shows the proportion of FAILS that are correctly labeled by the predictive model with given probability threshold. The second measure, **F-share** $= \frac{TF+FF}{\text{Number of observations}}$, shows the proportion of the test sample labeled as FAIL by the model. Ideal models produce minimal **F-share**, and an **F-recall** of 1 for a given set of parameters.

3 Experiments

For each of the seven ROIs, we performed three experiments defined by three DL models (VGGNet-, ResNet- and Inception-like architecture).

3.1 Datasets

Our experimental data from the ENIGMA working groups is described in **Table 1**. Our predictive models were trained using 15 cohorts totaling 5218 subjects' subcortical shape models from the ENIGMA-Schizophrenia working group. For a complete overview of ENIGMA-SCZ projects and cohort details, see [12].

To test our final models, we used data from 4 cohorts in the Major Depressive disorder working group (ENIGMA-MDD), totaling 1509 subjects, for final out-of-fold testing. A detailed description of the ENIGMA-MDD sites and its research objectives may be found here [13].

	FAIL %	accumbens	caudate	hippocampus	thalamus	putamen	pallidum	amygdala
Train	mean±std	3.4±4.7	0.9±0.7	2.0±1.1	0.8±1.0	0.6±0.6	2.3±3.6	0.9±0.9
	max	16.4	2.1	4.2	3.4	1.5	13.8	2.6
	size	10431	10433	10436	10436	10436	10435	10436
Test	mean±std	4.7±4.5	1.4±1.5	4.9±4.8	1.4±1.5	0.4±0.8	1.9±2.0	0.8±0.9
	max	10.5	3.5	11.4	3.5	1.6	3.8	2.1
	size	3017	3018	3018	3018	3017	3018	3018

Table 1: Overview of FAIL percentage mean, standard deviation and maximum for each site. Minimum is equal to 0 for all regions and sites except for hippocampus on train (FAIL percentage 5%). Sample sizes for each ROI vary slightly due to FreeSurfer segmentation failure.

3.2 Model validation

All experiments were performed separately for each ROI. The training dataset was split into two parts referred to as 'TRAIN GRID' (90% of train data) and 'TRAIN EVAL.' (10% of the data). The two parts contained data from each ENIGMA-SCZ cohort, stratified by the cohort-specific portion of FAIL cases.

Each model was trained on 'TRAIN GRID' using the original sampling matrix and 30 augmentation matrices resulting in 31x augmented train dataset. We also generated 31 instances of each mesh validation set using each sampling matrix and validated models' ROC AUC on this big validation set during the training.

As models produce probability estimates of FAIL (P_{FAIL}), we studied the robustness of the probability thresholds for each model. To do so, we selected P_{FAIL} values corresponding to regularly spaced percentiles values of **F-share**, from 0.1 to 0.9 in 0.1 increments. For each such value, we examined **F-recall** the evaluation set.

Final thresholds were selected based on the lowest **F-share** on the TRAIN EVAL set, requiring that **F-recall** ≥ 0.8, a minimal estimate of inter-rater reliability. It is important to stress that while we used sample distribution information in selecting a threshold, the final out-of-sample prediction is made on an individual basis for each mesh.

4 Results

Trained models were deliberately set to use a loose threshold for FAIL detection, predicting 0.2-0.5 of observations as FAILs in the 'TRAIN EVAL' sample. These predicted FAIL observations contained 0.85-0.9 of all true FAILs, promising to reduce the human rater QC time by 50-80%. These results largely generalized to the test samples: **Table 2** shows our best model and the threshold performance for each ROI. When applied to the test dataset, the models indicated modest over-fitting, with the amount of human effort reduced by 46-70%, while capturing 76-94% of poor quality meshes. The inverse relationship between FAIL

percentage and F-share (**Figure 1**) may indicate model failure to learn generalizable features on smaller number of FAIL examples. ROC AUC and F-recall performance generalize across the test sites. Since 68% of our test dataset is comprised of the Münster cohort, it is important that overall test performance is not skewed by it.

ROI	Model	Eval AUC	Test AUC	Eval F-share	Test F-share	Eval F-recall	Test F-recall
Accumbens	ResNet	0.86	0.8	0.3	0.35	0.83	0.78
Amygdala	ResNet	0.8	0.75	0.5	0.54	0.92	0.8
Caudate	ResNet	0.9	0.84	0.2	0.3	0.82	0.78
Hippocampus	ResNet	0.85	0.93	0.3	0.36	0.81	0.92
Pallidum	ResNet	0.86	0.91	0.3	0.32	0.81	0.91
Putamen	ResNet	0.88	0.7	0.3	0.52	0.93	0.76
Thalamus	ResNet	0.8	0.87	0.4	0.47	0.82	0.94

Table 2: Test performance of the best models for each region. ResNet performs the best in all cases. Overall models' performance generalizes to out-of-sample test data.

Our experiments with decision visualization (see **Fig. 2**) indicate that in most FAIL cases, the attention heat map generated by Grad-CAM corresponds to human raters' intuition while Prediction Difference Analysis tend to concentrate on local 'bumps' on shapes.

5 Discussion and Conclusion

We have presented potential deep learning solutions for semi-automated quality control of deep brain structure shape data. We believe this is the first DL approach for detecting end-of-the-pipeline feature failures in deep brain structure geometry. We showed that DL can robustly reduce human visual QC time by 46-70% for large-scale analyses, for all seven regions in question, across diverse MRI datasets and populations. Qualitative analysis of models decisions shows promise as a potential training and heuristic validation tool for human raters.

There are several limitations of our work. Our planar projection of vertex-wise features introduces space-varying distortions and boundary effects that can affect training, performance and visualization. Recently proposed spherical convolutional neural nets [14] may be useful to fix this issue. Second, our models'

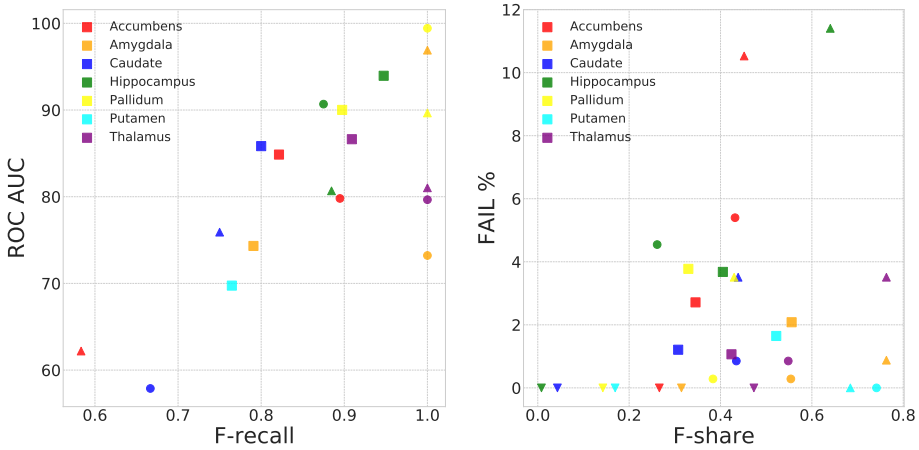

Fig. 1: Scatter plots of F-recall vs. ROC AUC on test datasets and F-share vs. proportion of predicted FAIL cases on test datasets (F-share). **Left:** F-recall vs ROC AUC. **Right:** Fail F-share vs FAIL percentage. F-share was calculated based on thresholds from **Table 2**. Mark size shows the dataset size. Mark shape represents dataset (site): ◯ - CODE-Berlin (N=176); □ - Münster(N=1033); △ - Stanford (N=105); ▽ - Houston(N=195)

.

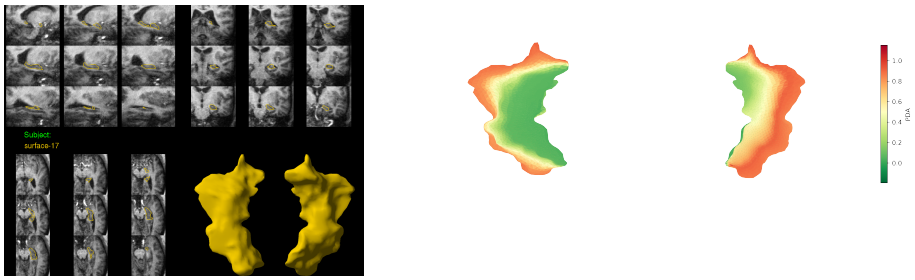

Fig. 2: QC report for human raters (**left**) and decision visualization example based on Grad-CAM for the ResNet model (**right**). Red colors correspond to points maximizing the model's FAIL decision in the last layer. Decision visualization corresponds to the observable deviations from underlying anatomical boundaries indicative of a "FAIL" rating according to an experienced rater.

.

decision visualization only partly matches with human raters' intuition. In some cases, our models do not consider primary "failure" areas, as assessed by a human rater. Finally, our models are trained on purely geometrical features and do not include information on shape boundaries inside the brain. In rare cases, human

raters pass shapes with atypical geometry because their boundaries look reasonable, and conversely mark normal-appearing geometry as failing due to poor a fit with the MR image. Incorporating intensity as well as geometry features will be the focus of our future work.

References

1. Gutman, B., Ching, C., Andreassen, O., Schmaal, L., Veltman, D., Van Erp, T., Turner, J., Thompson, P.M., et al.: Harmonized large-scale anatomical shape analysis: Mapping subcortical differences across the ENIGMA Bipolar, Schizophrenia, and Major Depression working groups. Biological Psychiatry **81**(10) (2017) S308
2. Gutman, B.A., Jahanshad, N., Ching, C.R., Wang, Y., Kochunov, P.V., Nichols, T.E., Thompson, P.M.: Medial demons registration localizes the degree of genetic influence over subcortical shape variability: An n= 1480 meta-analysis. In: Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on, IEEE 1402–1406
3. Roshchupkin*, G.V., Gutman*, B.A., et al.: Heritability of the shape of subcortical brain structures in the general population. Nature Communications **7** (2016) 13738
4. Gutman, B.A., Madsen, S.K., Toga, A.W., Thompson, P.M.: 24. In: A Family of Fast Spherical Registration Algorithms for Cortical Shapes. Volume 8159 of Lecture Notes in Computer Science. Springer International Publishing (2013) 246–257
5. Gutman, B.A., Yalin, W., Rajagopalan, P., Toga, A.W., Thompson, P.M.: Shape matching with medial curves and 1-d group-wise registration. In: Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on. 716–719
6. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
8. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI. Volume 4. (2017) 12
9. Jo, J., Bengio, Y.: Measuring the tendency of cnns to learn surface statistical regularities. arXiv preprint arXiv:1711.11561 (2017)
10. Zintgraf, L.M., Cohen, T.S., Adel, T., Welling, M.: Visualizing deep neural network decisions: Prediction difference analysis. arXiv preprint arXiv:1702.04595 (2017)
11. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. See https://arxiv. org/abs/1610.02391 v3 **7**(8) (2016)
12. van Erp, T.G.M., Hibar, D.P., et al.: Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. Molecular psychiatry (2015)
13. Schmaal, L., Hibar, D.P., et al.: Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the enigma major depressive disorder working group. Mol Psychiatry **22**(6) (2017) 900–909
14. Cohen, T.S., Geiger, M., Koehler, J., Welling, M.: Spherical cnns. arXiv preprint arXiv:1801.10130 (2018)

Full author list

Dmitry Petrov[1,2], Boris A. Gutman[2,3], Egor Kuznetsov[45], Theo G.M. van Erp[4], Jessica A. Turner[6], Lianne Schmaal[24,25], Dick Veltman[25], Lei Wang[5], Kathryn Alpert[5], Dmitry Isaev[1], Artemis Zavaliangos-Petropulu[1], Christopher R.K. Ching[1], Vince Calhoun[40], David Glahn[7], Theodore D. Satterthwaite[8], Ole Andreas Andreassen[9], Stefan Borgwardt[10], Fleur Howells[11], Nynke Groenewold[11], Aristotle Voineskos[11], Joaquim Radua[13,34,35,36], Steven G. Potkin[4], Benedicto Crespo-Facorro[14,38], Diana Tordesillas-Gutirrez[14,38], Li Shen[15], Irina Lebedeva[16], Gianfranco Spalletta[17], Gary Donohoe[18], Peter Kochunov[19], Pedro G.P. Rosa[20,33], Anthony James[21], Udo Dannlowski[26], Bernhard T. Baune[31], Andr Aleman[32], Ian H. Gotlib[27], Henrik Walter[28], Martin Walter[29,41,42], Jair C. Soares[30], Stefan Ehrlich[43], Ruben C. Gur[8], N. Trung Doan[9], Ingrid Agartz[9], Lars T. Westlye[9,37], Fabienne Harrisberger[10], Anita Riecher-Rössler[10], Anne Uhlmann[11], Dan J. Stein[11], Erin W. Dickie[12], Edith Pomarol-Clotet[13,34], Paola Fuentes-Claramonte[13,34], Erick Jorge Canales-Rodrguez[13,34,39,46], Raymond Salvador[13,34], Alexander J. Huang[4], Roberto Roiz-Santiaez[14,38], Shan Cong[15], Alexander Tomyshev[16], Fabrizio Piras[17], Daniela Vecchio[17], Nerisa Banaj[17], Valentina Ciullo[17], Elliot Hong[19], Geraldo Busatto[20,33], Marcus V. Zanetti[20,33], Mauricio H. Serpa[20,33], Simon Cervenka[22], Sinead Kelly[23], Dominik Grotegerd[26], Matthew D. Sacchet[27], Ilya M. Veer[28], Meng Li[29], Mon-Ju Wu[30], Benson Irungu[30], Esther Walton[43,44], and Paul M. Thompson[1], for the ENIGMA consortium

[1] Imaging Genetics Center, Stevens Institute for Neuroimaging and Informatics, University of Southern California, Marina Del Rey, CA, USA
[2] The Institute for Information Transmission Problems, Moscow, Russia
[3] Department of Biomedical Engineering, Illinois Institute of Technology, Chicago, IL, USA
[4] Department of Psychiatry and Human Behavior, University of California Irvine, Irvine, CA, USA
[5] Department of Psychiatry, Northwestern University, Chicago, IL, USA
[6] Psychology Department & Neuroscience Institute, Georgia State University, Atlanta GA, USA
[7] Yale University School of Medicine, New Haven, CT, USA
[8] Department of Psychiatry, University of Pennsylvania School of Medicine, Philadelphia, PA, USA
[9] CoE NORMENT, KG Jebsen Centre for Psychosis Research, Division of Mental Health and Addiction, Oslo University Hospital & Institute of Clinical Medicine, University of Oslo, Oslo, Norway
[10] Department of Psychiatry, University of Basel, Basel, Switzerland
[11] MRC Unit on Risk & Resilience to Mental Disorders, Department of Psychiatry and Mental Health, University of Cape Town, Cape Town, South Africa
[12] Centre for Addiction and Mental Health, Toronto, Canada
[13] FIDMAG Germanes Hospitalaries Research Foundation, Barcelona, Spain
[14] University Hospital Marqus de Valdecilla, IDIVAL, Department of Psychiatry, School of Medicine, University of Cantabria, Santander, Spain

[15] Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN, USA
[16] Mental Health Research Center, Moscow, Russia
[17] Laboratory of Neuropsychiatry, Santa Lucia Foundation IRCCS, Rome, Italy
[18] School of Psychology, NUI Galway, Galway, Ireland
[19] Maryland Psychiatric Research Center, University of Maryland School of Medicine, Baltimore
[20] Department of Psychiatry, Faculty of Medicine, University of São Paulo, São Paulo, Brazil
[21] University of Oxford, Oxford, United Kingdom
[22] Centre for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden
[23] Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA
[24] Orygen, The National Centre of Excellence in Youth Mental Health, Melbourne, Australia
[25] Department of Psychiatry, VU University Medical Center, Amsterdam, The Netherlands
[26] Department of Psychiatry and Psychotherapy, University of Münster, Germany
[27] Department of Psychology, Stanford University, Stanford, CA, USA
[28] Charité Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Department of Psychiatry and Psychotherapy CCM, Berlin, Germany
[29] Clinical Affective Neuroimaging Laboratory, Leibniz Institute for Neurobiology, Magdeburg, Germany
[30] University of Texas Health Science Center at Houston, Houston, TX, USA
[31] Discipline of Psychiatry, Adelaide Medical School, The University of Adelaide
[32] Interdisciplinary Center Psychopathology and Emotion regulation (ICPE), Neuroimaging Center (BCN-NIC), University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[33] Center for Interdisciplinary Research on Applied Neurosciences (NAPNA), University of So Paulo, So Paulo, Brazil
[34] CIBERSAM, Centro Investigacin Biomdica en Red de Salud Mental, Barcelona, Spain
[35] Department of Clinical Neuroscience, Centre for Psychiatric Research, Karolinska Institutet, Stockholm, Sweden
[36] Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, King's College London, United Kingdom
[37] Department of Psychology, University of Oslo, Oslo, Norway
[38] CIBERSAM, Centro Investigacin Biomdica en Red Salud Mental, Santander, Spain
[39] Radiology department, University Hospital Center (CHUV), Lausanne, Switzerland
[40] The Mind Research Network, Albuquerque, NM, USA
[41] Leibniz Institute for Neurobiology, Magdeburg, Germany
[42] Department of Psychiatry and Psychotherapy, University of Tbingen, Tbingen, Germany
[43] Division of Psychological and Social Medicine and Developmental Neurosciences, Faculty of Medicine, TU Dresden, Germany
[44] Psychology Department, Georgia State University, Atlanta, GA, USA
[45] Skolkovo Institute of Science and Technology, Moscow, Russia

[46] Signal Processing Laboratory 5 (LTS5), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland