Machine Learning for Large-Scale Quality Control of 3D Shape Models in Neuroimaging

MICCAI MLMI 2017

Presenter: Dmitry Petrov

September 10, 2017

Imaging Genetics Center, University of Southern California, Los Angeles
Institute for Information Transmission Problems, Moscow

Full list of authors

Dmitry Petrov[1], Boris A. Gutman[1], Shih-Hua (Julie) Yu, Theo G.M. van Erp, Jessica A. Turner, Lianne Schmaal, Dick Veltman, Lei Wang, Kathryn Alpert, Dmitry Isaev, Artemis Zavaliangos-Petropulu, Christopher R.K. Ching, Vince Calhoun, David Glahn, Theodore D. Satterthwaite, Ole Andreas Andreasen, Stefan Borgwardt, Fleur Howells, Nynke Groenewold, Aristotle Voineskos, Joaquim Radua, Steven G. Potkin, Benedicto Crespo-Facorro, Diana Tordesillas-Gutiérrez, Li Shen, Irina Lebedeva, Gianfranco Spalletta, Gary Donohoe, Peter Kochunov, Pedro G.P. Rosa, Anthony James, Udo Dannlowski, Bernhard T. Baune, André Aleman, Ian H. Gotlib , Henrik Walter, Martin Walter, Jair C. Soares, Stefan Ehrlich, Ruben C. Gur, N. Trung Doan, Ingrid Agartz, Lars T. Westlye, Fabienne Harrisberger, Anita Riecher-Rössler, Anne Uhlmann, Dan J. Stein, Erin W. Dickie, Edith Pomarol-Clotet, Paola Fuentes-Claramonte, Erick Jorge Canales-Rodríguez, Raymond Salvador, Alexander J. Huang, Roberto Roiz-Santiañez, Shan Cong, Alexander Tomyshev, Fabrizio Piras, Daniela Vecchio, Nerisa Banaj, Valentina Ciullo, Elliot Hong, Geraldo Busatto, Marcus V. Zanetti Mauricio H. Serpa, Simon Cervenka, Sinead Kelly, Dominik Grotegerd, Matthew D. Sacchet, Ilya M. Veer, Meng Li, Mon-Ju Wu, Benson Irungu, Esther Walton and Paul M. Thompson, for the ENIGMA consortium

[1] these authors contributed equally

Table of contents

1. Quality check: the problem

2. Automated QC: method

3. Data and experiments

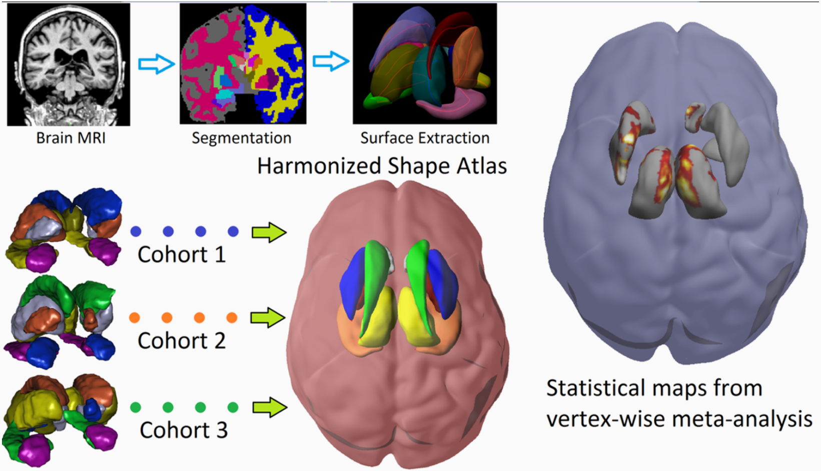
4. Results

5. Discussion and future work

6. Acknowledgements

Quality check: the problem

Subcortical structures analysis



Brain MRI

Segmentation

Surface Extraction

Harmonized Shape Atlas

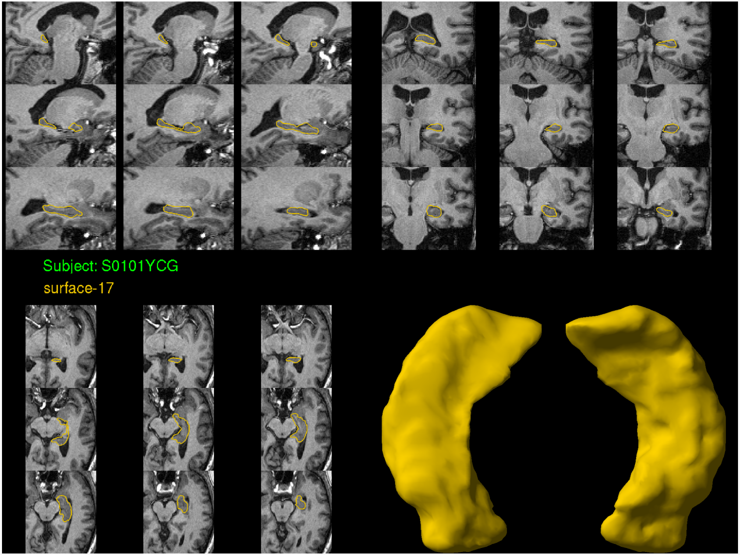
Cohort 1

Cohort 2

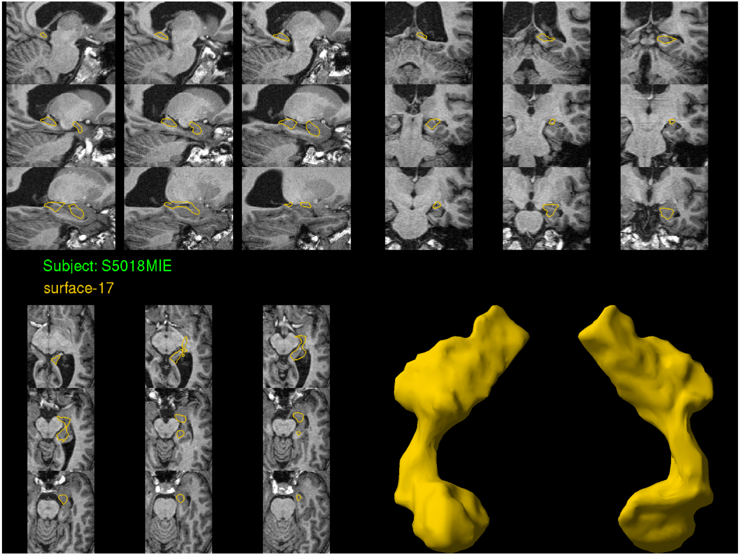
Cohort 3

Statistical maps from vertex-wise meta-analysis

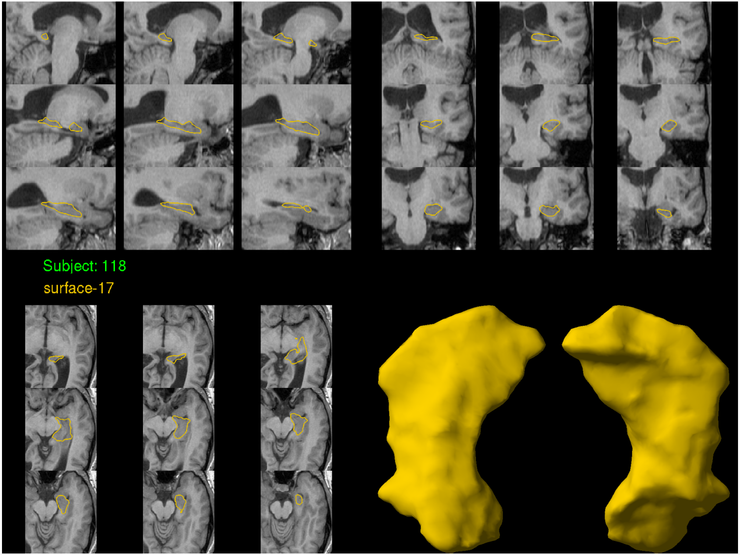
Quality check: the good



Quality check: the bad



Quality check: the borderline



Quality check: the challenge

- QC is the practical bottleneck in big-data neuroimaging, especially for the coming big datasets like UK Biobank
- QC for for 100 subjects takes ~7-15 hours
- Each time you rerun segmentation, you need to rerun QC
- We don't know bias introduced by raters

biobank uk Imaging study

ENIGMA

HUMAN Connectome PROJECT

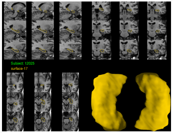
Automated QC: method

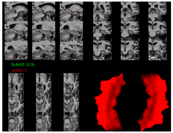
Shape classification: idea

- Use human ratings and shape descriptors to train binary classifier to distinguish shapes which passed QC (PASS) and those which didn't (FAIL)

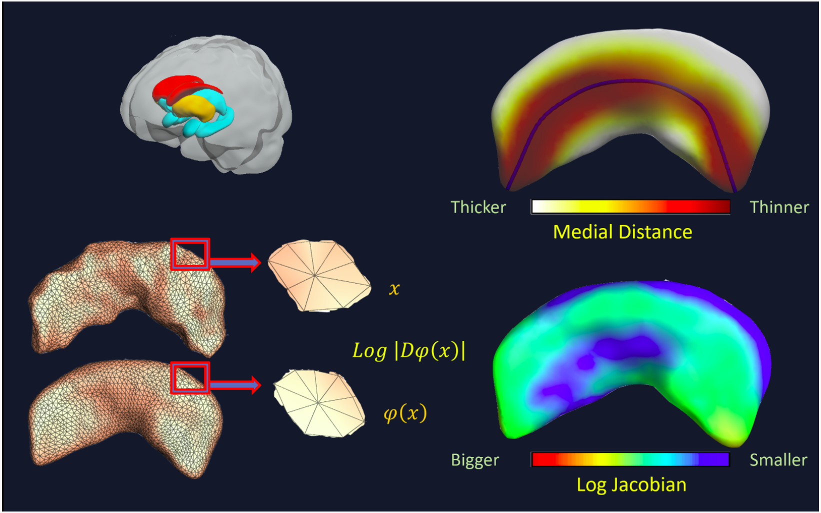
- Tweak classifier to catch as many FAILs as possible (i.e. set low probability threshold)

- Test results for robustness on a distribution which differs from train distribution





Shape features: intuition



Shape features: formulae

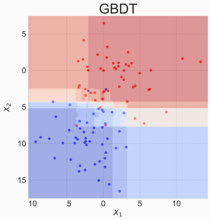
Each ROI is approximated with mesh with ~2,500 or ~1,250 vertices. Each vertex p of mesh model M is endowed with two shape descriptors:

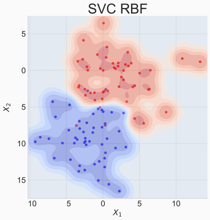
- **Medial Thickness**, $D(p) = \|c_p - p\|$, where c_p is the point on the medial curve c closest to p.
- **LogJac(p)**, Log of the Jacobian determinant J arising from the template mapping, $J : T_{\phi(p)}M_t \to T_pM$.
- **Two global features**: the shape-wide feature median, and the shape-wise 95th percentile feature threshold.

Shape classifiers

- **Gradient Boosted Decision Trees (GBDT)**. In our experiments we used the Xgboost implementation due to speed and regularization heuristics, with the logistic loss function

- **Support Vector Classifier (SVC)** with the radial basis function (RBF) kernel. We used scikit-learn's implementation of SVC





Performance metrics

TF = TRUE FAIL, FF = FALSE FAIL, TP = TRUE PASS, and FP = FALSE PASS.

$$\textbf{F-recall} = \frac{TF}{TF + FP},$$

proportion of FAILS caught — ↑ is better.

$$\textbf{F-share} = \frac{TF + FF}{\text{Number of observations}},$$

share of the test sample labeled as FAIL — ↓ is better.

$$\textbf{Modified F-score} = 2 \times \frac{\text{F-recall} \times (1 - \text{F-share})}{\text{F-recall} + (1 - \text{F-share})},$$

allows to compare models — ↑ is better.

Data and experiments

Datasets: ENIGMA SCZ and ENIGMA MDD

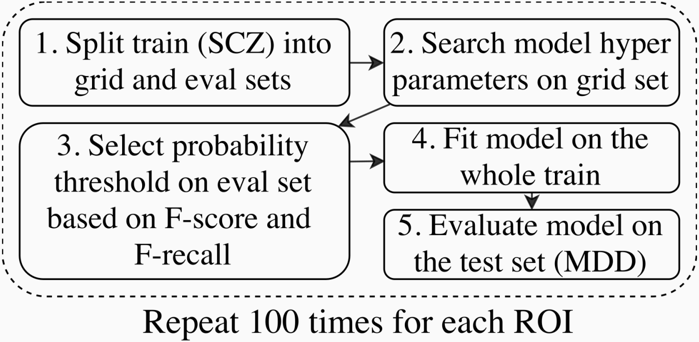
We used the ENIGMA Schizophrenia (train, 21 sites) and Major Depressive Disorder (test, 4 sites) working groups' data.

	FAIL %	accumbens	caudate	hippocampus	thalamus	putamen	pallidum	amygdala
Train	mean±std	3.4±4.6	1.4±1.9	3.2±3.0	1.5±2.3	0.7±0.9	3.6±4.7	0.8±0.8
	max	16.4	8.7	11.4	9.2	2.9	15.5	2.6
	min	0.0	0.0	0.5	0.0	0.0	0.0	0.0
	size	3017	3018	3018	3018	3017	3018	3018
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
	min	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	size	12931	12933	12936	12936	12936	12935	12936

Sample sizes for each ROI vary slightly due to FreeSurfer segmentation failure.

Experiment: overall scheme

For each of seven ROI we combined left and right hemisphere data and trained FAIL/PASS classifier



1. Split train (SCZ) into grid and eval sets

2. Search model hyper parameters on grid set

3. Select probability threshold on eval set based on F-score and F-recall

4. Fit model on the whole train

5. Evaluate model on the test set (MDD)

Repeat 100 times for each ROI

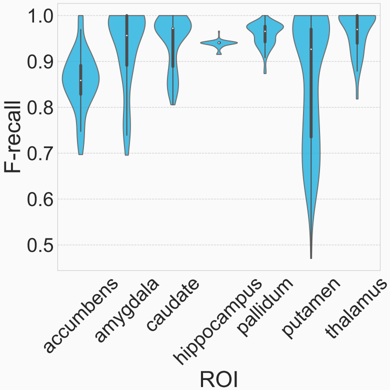
Important note. Results in our paper are reported for one grid/eval split. Since submission we've decided to investigate the robustness of our models.

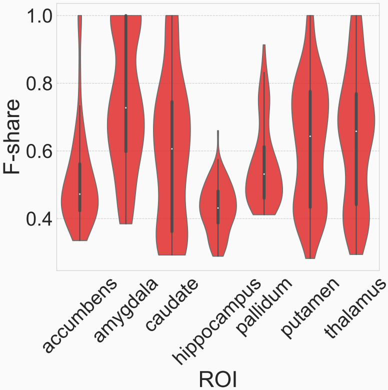
Experiment: details

- Grid/eval set splits were 50/50 and stratified by sites and target
- For grid search we maximized ROC AUC on stratified 5-fold or Leave-One-Site-Out (LOSO) cross-validations
- We tried normed/non-normed by volume features
- On the evaluation set we tested 0.1, 0.2, ..., 0.9 quantile thresholds of classifier probabilities
- For final testing we chose the threshold with best F-score and F-recall ≥ 0.8

Results

F-recall and F-share distributions on test data

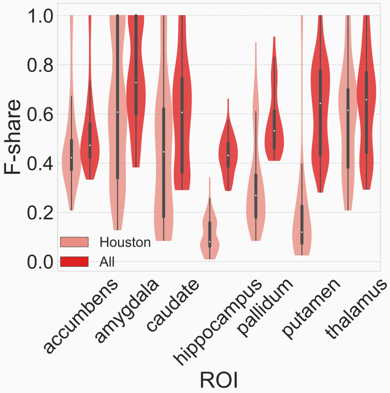




F-recall: ↑ is better

F-share: ↓ is better

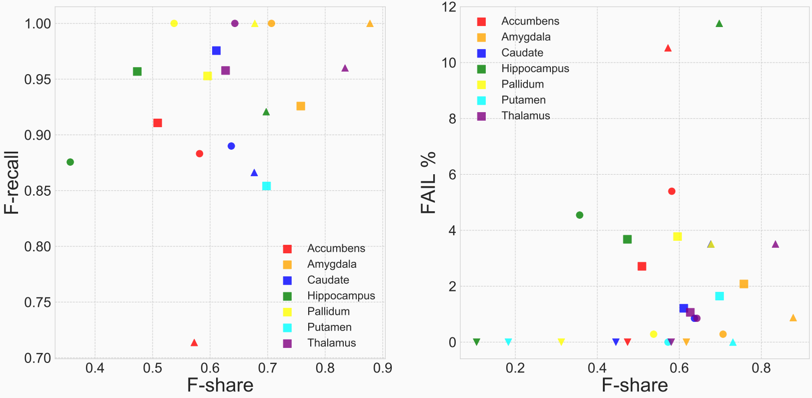
Houston F-share distributions



F-share: ↓ is better

- Houston site is 12.3% of test data
- It has no TRUE FAILs, so F-recall is not available
- Models have overall lower F-share on it, especially 'better' ROIs

F-share vs F-recall and FAIL %



Mark shapes: ◯ - CODE-Berlin (N=176); □ - Münster (N=1033); △ - Stanford (N=105); ▽ - Houston (N=195).

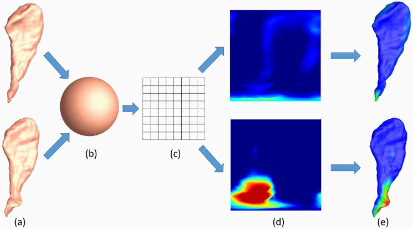
Discussion and future work

Conclusion: what we did

- Presented a preliminary study of potential solutions for semi-automated QC of subcortical structures
- Showed that ML can reduce human visual QC time by 30-50% for for six out of the seven regions in question
- Tested our results on diverse MRI datasets and populations and provided a baseline for future researchers in this area

Future work: directions

- Increase the robustness of our models (lower F-share std)
- Convolutional and geometrical neural nets
- Visualization of models decisions



Concept: mockup attention map

Acknowledgements

Thanks

We thank people and organizations from ENIGMA Schizophrenia and MDD projects for gathering, preprocessing and quality checking data.

```
http://enigma.ini.usc.edu/ongoing/
enigma-schizophrenia-working-group/
```

```
http://enigma.ini.usc.edu/ongoing/
enigma-mdd-working-group/
```

Funding

This work was funded in part by NIH BD2K grant U54 EB020403, Russian Science Foundation grant 17-11-01390 and other agencies worldwide.

Questions?

to.dmitry.petrov@gmail.com

Thank you!

to.dmitry.petrov@gmail.com