Questionnaire Response Correlations to Improve Efficiency: Preliminary Evidence From the Healthy Brain Network

Jon Clucas
Jake Son
MATTER Lab
Child Mind Institute
New York, NY 11102, USA
jon.clucas@childmind.org
jake.son@childmind.org

Michael P. Milham
Center for the Developing Brain
Child Mind Institute
New York, NY 11102, USA
Nathan Kline Institute
Orangeburg, NY 10962, USA
michael.milham@childmind.org

Anirudh Krishnakumar
MATTER Lab
Child Mind Institute
New York, NY 11102, USA
Centre de Recherches
Interdisciplinaires, IFFR
Paris, France
anirudh.krishnakumar@childmind.org

Arno Klein
MATTER Lab
Child Mind Institute
New York, NY 11102, USA
arno.klein@childmind.org

Open Access: The author(s) wish to pay for the work to be open access.
Every submission will be assigned their own unique DOI string to be included here.

Abstract
Questionnaires can be detrimentally long for some situations, presumably with dynamically diminishing returns. With an unprecedented set of pediatric questionnaire responses (dozens of questionnaires and eventually 10,000 participants) from the Healthy Brain Network, the Child Mind Institute MATTER Lab is exploring techniques to leverage correlations in responses to reduce the burden of questionnaires in mental health evaluation and monitoring.

Author Keywords
questionnaires; correlation; efficiency; pediatrics; psychiatry

CCS Concepts
•**Applied computing** → **Health informatics;**

Introduction
The Healthy Brain Network, a multimodal pediatric psychiatric biobank [1], includes dozens of questionnaires [3]. In labs and in practice, questionnaires can be burdensome to participants and to administrators. While a response to any individual question is informative, the informative value of each subsequent question will vary. With hundreds of (eventually ten thousand) individuals' responses to many overlapping questionnaires, we are well-positioned to measure the relative information of pairs of questions. Knowing these relative values can afford more efficient question-

naires, allowing administrators to automatically prioritize the most informative questions.

Methods

We analyzed questionnaire responses from the first two Healthy Brain Network releases (n=881 subjects, 79 questionnaires, 2,630 questions, available at http://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network). For each pair of question response vectors, we calculated and inverted Pearson's ρ, dropping any pairs for which $abs(\rho) > 0$. Figure 1 shows each question as a node connected by edges of length $\frac{1}{\rho}$. The code used to generate the figures is available in a Jupyter notebook at https://github.com/ChildMindInstitute/questionnaire-correlations/releases/tag/v0.1.0.

Results

Our initial visual exploration indicated 30 groupings of correlated responses (see Figure 1), often linking questions within a single questionnaire. Two of these clusters contain only two questions each (the Fagerström Test for Nicotine Dependence [5] questions "Are you currently a smoker?" and "Have you been a smoker within the past two years?" clustered only with one another; the Goldman-Fristoe Test of Articulation [4] sounds-in-sentences completion clustered only with accuracy from the same test). One cluster contains 1,876 questions. The second-largest cluster contains 66 questions (excluding the 1,876-question cluster: mean=26, standard deviation=19.5). Most of the clusters contain questions from only one questionnaire each, indicating a sensitivity of this comparison method to artifacts of questionnaire administration. Figure 2 shows a cluster containing only questions from the Extended Strengths and Weaknesses Assessment of Normal Behavior questionnaire [2], but questions about three disorders: Disruptive Mood Dysregulation, Major Depressive and Social Anxiety.

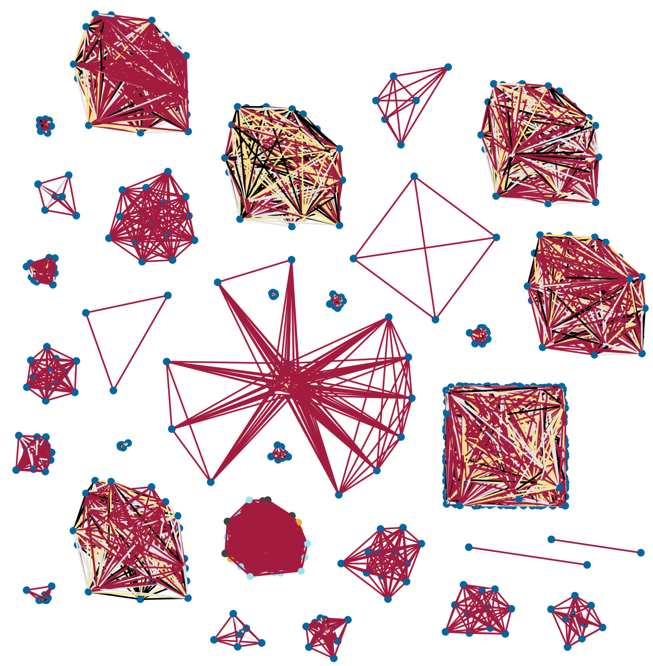

Figure 1: 30 clusters of questions with correlated responses.

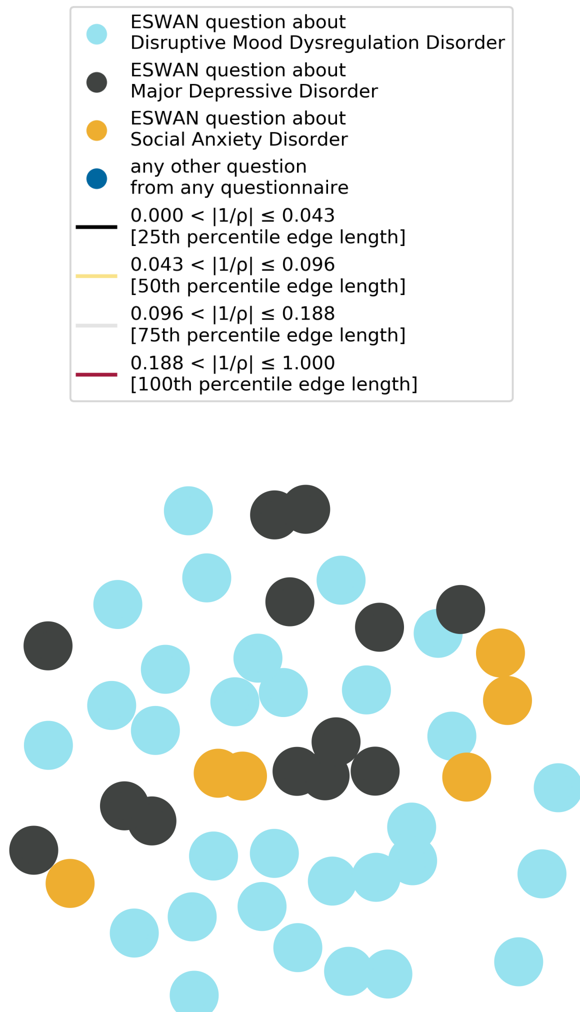

Figure 2: One of the 30 clusters, enlarged, with edges hidden.

Future Work

We have also been employing a variety of methods, including random forests [7][8], randomer forests [9] and probabilistic metamodeling [6], to estimate the most informative of this set of questions for predicting ADHD subtype consensus diagnosis and Autism Spectrum Disorder consensus diagnosis. The code for these analyses is available online at https://github.com/ChildMindInstitute/questionnaire-diagnosis. By employing a variety of methods, we can simultaneously assess the applicability of each method and the strengths of correspondence between categorically distinct data.

REFERENCES

1. Lindsay M. Alexander, Jasmine Escalera, Lei Ai, Charissa Andreotti, Karina Febre, Alexander Mangone, Natan Vega-Potler, Nicolas Langer, Alexis Alexander, Meagan Kovacs, Shannon Litke, Bridget O'Hagan, Jennifer Andersen, Batya Bronstein, Anastasia Bui, Marijayne Bushey, Henry Butler, Victoria Castagna, Nicolas Camacho, Elisha Chan, Danielle Citera, Jon Clucas, Samantha Cohen, Sarah Dufek, Megan Eaves, Brian Fradera, Judith Gardner, Natalie Grant-Villegas, Gabriella Green, Camille Gregory, Emily Hart, Shana Harris, Megan Horton, Danielle Kahn, Katherine Kabotyanski, Bernard Karmel, Simon P. Kelly, Kayla Kleinman, Bonhwang Koo, Eliza Kramer, Elizabeth Lennon, Catherine Lord, Ginny Mantello, Amy Margolis, Kathleen R. Merikangas, Judith Milham, Giuseppe Minniti, Rebecca Neuhaus, Alexandra Levine, Yael Osman, Lucas C. Parra, Ken R. Pugh, Amy Racanello, Anita Restrepo, Tian Saltzman, Batya Septimus, Russell Tobe, Rachel Waltz, Anna Williams, Anna Yeo, Francisco X. Castellanos, Arno Klein, Tomas Paus, Bennett L. Leventhal, R. Cameron Craddock, Harold S. Koplewicz, and Michael P. Milham. 2017a. An open resource for transdiagnostic research in pediatric

mental health and learning disorders. *Scientific Data* 4 (Dec. 2017), 170181. DOI: http://dx.doi.org/10.1038/sdata.2017.181

2. Lindsay M. Alexander, Giovanni Salum, James M. Swanson, and Michael P. Milham. 2017b. Balancing Strengths and Weaknesses in Dimensional Psychiatry. *bioRxiv* (Oct. 2017), 207019. DOI: http://dx.doi.org/10.1101/207019
3. Child Mind Institute. 2016. Complete List of Assessments. (2016). http://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network/assessments/master-list.html
4. Ronald Goldman and Macalyne Fristoe. 2015. *Goldman-Fristoe Test of Articulation 3*. American Guidance Service, Inc., Circle Pines, MN. https://www.pearsonclinical.com/language/products/100001202/goldman-fristoe-test-of-articulation-3-gfta-3.html
5. T. F. Heatherton, L. T. Kozlowski, R. C. Frecker, and K. O. Fagerström. 1991. The Fagerström Test for Nicotine Dependence: a revision of the Fagerström Tolerance Questionnaire. *British Journal of Addiction* 86, 9 (Sept. 1991), 1119–1127.
6. Vikash Mansinghka. 2016. The MIT Probabilistic Computing Project. (Sept. 2016). http://probcomp.csail.mit.edu/
7. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (Oct. 2011), 2825–2830. http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html
8. scikit-learn developers. 2017. Random Forests. In *scikit-learn User Guide*. 1.11.2.1. http://scikit-learn.org/stable/modules/ensemble.html#random-forests
9. Tyler M. Tomita, Mauro Maggioni, and Joshua T. Vogelstein. 2015. Randomer Forests. *arXiv:1506.03410 [cs, stat]* (June 2015). http://arxiv.org/abs/1506.03410 arXiv: 1506.03410.