The PsyCourse Study – FAIR Longitudinal, Multi-level Genomic and Phenomic Data for Machine Learning

Urs Heilbronner[1]; Monika Budde[1]; Daniela Reich-Erkelenz[1]; Heike Anderson-Schmidt[2]; Janos Kalman[1,3,4]; Fanny Senner[1,3]; Kristina Adorjan[1,3]; Eva C. Schulte[1,3]; Ashley L. Comes[1,4]; Sergi Papiol[1,3]; Till F. M. Andlauer[5]; Marcella Rietschel[6]; Markus M. Nöthen[7,8]; Peter Falkai[3]; Thomas G. Schulze[1]

[1]Institute of Psychiatric Phenomics and Genomics (IPPG), University Hospital, LMU Munich, Munich, Germany; [2]Department of Psychiatry and Psychotherapy, University Medical Center Goettingen, Goettingen, Germany; [3]Department of Psychiatry and Psychotherapy, University Hospital, LMU Munich, Munich, Germany; [4]International Max Planck Research School for Translational Psychiatry, Max Planck Institute of Psychiatry, Munich, Germany; [5]Department of Neurology, Klinikum rechts der Isar, School of Medicine, Technical University of Munich, Munich, Germany; [6]Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany; [7]Institute of Human Genetics, University of Bonn School of Medicine & University Hospital Bonn, Bonn, Germany; [8]Department of Genomics, Life & Brain Center, University of Bonn, Bonn, Germany

Introduction

The integration of multiple levels of genetic information, longitudinal deep phenotyping, and machine learning holds promise to substantially increase our knowledge on mental disorders. However, well-annotated longitudinal datasets on comprehensively characterized individuals are rare. Also, the workload involved in pre-processing such large datasets can be substantial. Here, we present a phenotype dataset of the longitudinal PsyCourse Study (Budde et al., 2018), that has been extensively annotated and formatted for machine-learning analyses. Metadata of this dataset are publicly available (Heilbronner et al., 2020). Biological OMICS data are available for analyses (genomics, transcriptomics, proteomics, lipidomics) in subsets of participants (see below), these will be extended for additional cross-sectional and longitudinal analyses in the future.

Abstract

- The PsyCourse Study is a longitudinal investigation of the affective-to-psychotic spectrum, including healthy controls, that combines deep phenotyping and collection of DNA, RNA, plasma, and serum (qr_1)
- Multiple OMICS datasets of PsyCourse participants already exist
- The PsyCourse phenotype dataset has been pre-formatted for machine learning analyses, and metadata are publicly available (qr_2)
- Data can be accessed by submitting a short research proposal (qr_3)
- We encourage researchers to collaborate with us!

Methods

The latest version of the phenotype dataset (PsyCourse 4.1) contains n=1,320 clinical (affective-to-psychotic spectrum) and n=466 control participants, who were assessed at the first of four planned study visits (visit 2: n=788/288, visit 3: n=661/280, visit 4: n=589/251), each separated by approximately six months. A comprehensive test battery surveying demographics, psychiatric history, medication, substance abuse, diagnosis, cognition, psychiatric rating scales, and several questionnaires, was assessed. At the time of writing, the following biological data exist: SNP array (n=1,446, all diagnoses and controls, PsychChip), epigenetics (n=96 bipolar disorder, two measurement points, EPIC array), exome sequencing (n=104, bipolar disorder), small-RNAome sequencing (n=1,322, first study visit, all diagnoses and controls), mRNA sequencing (n=538, first study visit, schizophrenia spectrum), plasma proteome profiling (n=220, multiple diagnoses and study visits), serum protein profiling (n= 222, multiple diagnoses and study visits), and lipidomics (n=242 bipolar, n=186 schizophrenia, n=192 controls, multiple study visits). Importantly, missing phenotype data, ubiquitous in longitudinal studies, have been re-coded to identify structurally missing data. Researchers can access the data by filling out a short research proposal (see URL in „Results“).

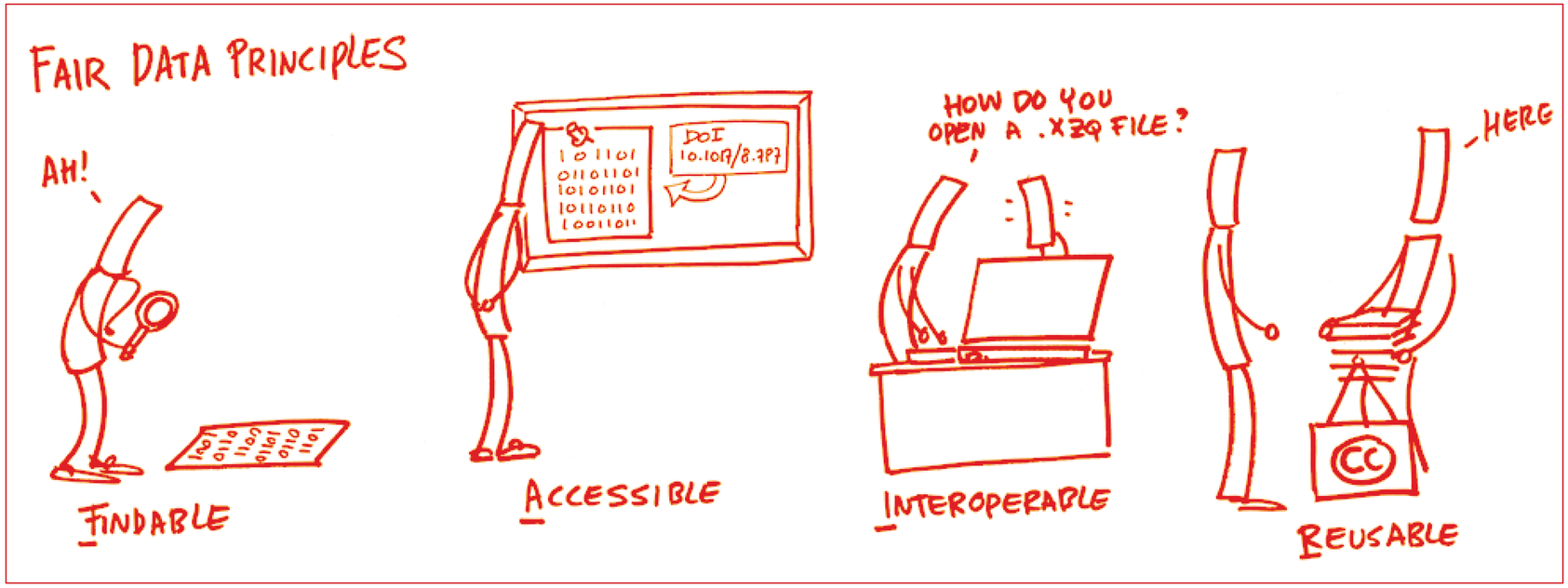

Figure 1. The FAIR Data Principles (https://book.fosteropenscience.eu/).

qr_1 qr_2 qr_3

Results

The framework of FAIR data (Wilkinson et al., 2016) describes a measurable set of principles that have been created to foster the re-use of scientific data (Figure 1). The PsyCourse dataset fulfills these criteria by being:

- ***F**indable* (public metadata),
- ***A**ccessible* (via research proposals),
- ***I**nteroperable* (available in wide and long format R files, also as a text file), and thus
- ***R**eusable*

For details, please refer to the PsyCourse Open Science website: psycourse.de/openscience-en.html.

Discussion

While not in the public domain, PsyCourse data can be re-used by bona-fide researchers world-wide. Accepted research proposals are posted on our website, giving a transparent overview on ongoing research projects. We encourage researchers to collaborate with us!

References

Budde et al. (2018) A longitudinal approach to biological psychiatric research: The PsyCourse study. Am J Med Genet B Neuropsychiatr Genet. DOI: 10.1002/ajmg.b.32639

Heilbronner et al. (2020). The PsyCourse Codebook, Version 4.1. Open Data LMU. DOI: 10.5282/ubm/data.199.

Wilkinson et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. DOI: 10.1038/sdata.2016.18

Grants

TGS: DFG Grants SCHU 1603/5-1 and SCHU 1603/7-1; BMBF Grants IntegraMent and BipoLife; Dr. Lisa-Oehler-Foundation (Kassel, Germany).

HB: DFG Grants BI 576/5-1 and Research Training Group “Scaling Problems in Statistics” RTG 1644.