|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-F-014 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | Plenary | Geneva, 30 May – 1 June 2019 | |
| **DOCUMENT** | | | | |
| **Source:** | | TG-Psy Topic Driver | | |
| **Title:** | | TTD Update: TG-Psy (Psychiatry) | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Nicolas Langer University of Zurich Switzerland | | Tel: +416353414 Fax: N/A Email: [n.langer@psychologie.uzh.ch](mailto:n.langer@psychologie.uzh.ch) |
| **Contact:** | | Stefan Haufe Charité - Universitätsmedizin Berlin Germany | | Tel: +49 30 450 639639 Fax: N/A Email: [stefan.haufe@charite.de](mailto:stefan.haufe@charite.de) |

|  |  |
| --- | --- |
| **Abstract:** | This document specifies a standardized benchmarking for AI-based symptom assessments. It follows the structure defined in [FGAI4H-C-105](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-C-105.docx?d=w50606d7d9bf340198b6423e4d5babbe6) and covers all scientific, technical and administrative aspects relevant for setting up this benchmarking. The creation of this document is an ongoing process until it will be finally approved by the Focus Group. This draft will be a continuous Input- and Output-Document. NOTE – This document is the same one as reviewed in meeting E, FGAI4H-E-015; reproduced for easier reference. |

|  |  |
| --- | --- |
| **Change Notes:** | Version 1.0  This is the initial draft version of the TDD. As a starting point it merges the input documents FGAI4H-C-013-A1, FGAI4H-C-013, FGAI4H-E-005-A08, and fits them to the structure defined in [FGAI4H-C-105](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-C-105.docx?d=w50606d7d9bf340198b6423e4d5babbe6). The focus was especially on the following aspects:   * Introduction to topic and ethical considerations * Workflow proposal for Topic group * Overview of currently available AI-based symptom assessment applications started * Prior works on benchmarking and scientific approaches including first contributions by experts joining the topic. * Brief overview of different ontologies to describe medical terms and diseases. |

Table of Contents

[Table of Contents 2](#_Toc10119538)

[1 Introduction 3](#_Toc10119539)

[1.1 Description of the topic: “Psychiatric Multimorbidity” 3](#_Toc10119540)

[1.1.1 Relevance 3](#_Toc10119541)

[1.1.2 Possible Impact 3](#_Toc10119542)

[1.1.3 Existing Work 3](#_Toc10119543)

[1.1.4 Feasibility 3](#_Toc10119544)

[1.2 Ethical Considerations 3](#_Toc10119545)

[2 Existing AI Solutions 3](#_Toc10119546)

[2.1 Existing Work on benchmarking 3](#_Toc10119547)

[3 AI4H Topic Group 4](#_Toc10119548)

[4 Method 4](#_Toc10119549)

[4.1 AI Input Data Structure 4](#_Toc10119550)

[4.2 AI Output Data Structure 4](#_Toc10119551)

[4.3 Test Data Labels 4](#_Toc10119552)

[4.4 Scores & Metrics 4](#_Toc10119553)

[4.5 Undisclosed Test Data Set Collection 4](#_Toc10119554)

[4.6 Benchmarking Methodology and Architecture 5](#_Toc10119555)

[4.7 Reporting Methodology 5](#_Toc10119556)

[5 Results 5](#_Toc10119557)

[6 Discussion 5](#_Toc10119558)

[6.1 Declaration of Conflict of Interest 5](#_Toc10119559)

[1 Introduction 6](#_Toc10119560)

[1.1 Description of the topic: “Psychiatric Multimorbidity” 6](#_Toc10119561)

[1.1.1 Relevance 6](#_Toc10119562)

[1.1.2 Possible Impact 7](#_Toc10119563)

[1.1.3 Existing Work 8](#_Toc10119564)

[1.1.4 Feasibility 9](#_Toc10119565)

[1.2 Ethical Consideration 10](#_Toc10119566)

[2 Existing AI Solutions 10](#_Toc10119567)

[2.1 Existing Work on benchmarking 10](#_Toc10119568)

[3 AI4H Topic Group 10](#_Toc10119569)

[3.1 Topic Group structure 10](#_Toc10119570)

[3.2 Subtopics 11](#_Toc10119571)

[3.3 Current topic status 11](#_Toc10119572)

[3.4 Topic group contributors 11](#_Toc10119573)

[3.5 3Topic group participation 12](#_Toc10119574)

[3.6 Next Meetings 12](#_Toc10119575)

[4 Method 13](#_Toc10119576)

[4.1 AI Input Data Structure 13](#_Toc10119577)

[4.1.1 Overview 13](#_Toc10119578)

[4.1.2 Subjects 13](#_Toc10119579)

[4.1.3 Data formats 14](#_Toc10119580)

[4.2 AI Output Data Structure 14](#_Toc10119581)

[4.3 Test Data Labels 15](#_Toc10119582)

[4.4 Scores & Metrics 15](#_Toc10119583)

[4.5 Undisclosed Test Data Set Collection 17](#_Toc10119584)

[4.6 Benchmarking Methodology and Architecture 17](#_Toc10119585)

[4.7 Reporting Methodology 17](#_Toc10119586)

[5 Results 19](#_Toc10119587)

[6 Discussion 19](#_Toc10119588)

[6.1 Declaration of Conflict of Interest 19](#_Toc10119589)

[References 20](#_Toc10119590)

**Summary of points to be covered**

# Introduction

## Description of the topic: “Psychiatric Multimorbidity”

### Relevance

### Possible Impact

### Existing Work

### Feasibility

## Ethical Considerations

# Existing AI Solutions

* current systems available with their inputs, output, focus/bias
* existing benchmarking including self-stated performance

## Existing Work on benchmarking

* papers on existing attempts to benchmark solutions on the topic
* clinical evaluation attempts, RCT, etc.
* including existing numbers

# AI4H Topic Group

* Topic group structure
  + Subtopic 1
  + Subtopic 2
* Topic group participation
* Tools/process of TG cooperation
* TG interaction with WG, FG
* Current topic group and topic status
* Contributors so far
* Next meetings

# Method

* Overview of the benchmarking

## AI Input Data Structure

* possible inputs for benchmarking
* ontologies, terminologies
* data format

## AI Output Data Structure

* outputs to benchmark
* ontologies, terminologies
* data format

## Test Data Labels

* label types
* ontologies, terminologies
* data format

## Scores & Metrics

* which metrics & scores to use for benchmarking
* considering relation to parameters stakeholders need for decision making
* considering scores that providers use
* considering the scope providers designed their solutions for
* considering the state of the art in RCT, statistics, AI benchmarking etc.
* considering bias transparency

## Undisclosed Test Data Set Collection

* raw data acquisition / acceptance
* test data source(s): availability, reliability,
* labelling process / acceptance
* bias documentation process
* quality control mechanisms
* discussion of the necessary size of the test data set for relevant benchmarking results
* specific data governance derived by general data governance document (currently C-004)

## Benchmarking Methodology and Architecture

* technical architecture
* hosting (IIC, etc.)
* possibility of an online benchmarking on a public test dataset
* protocol for performing the benchmarking (who does what when etc.)
* AI submission procedure including contracts, rights, IP etc. considerations

## Reporting Methodology

* Report publication in papers or as part of ITU documents
* Online reporting
* public leaderboards vs. private leaderboards
* Credit-Check like on approved sharing with selected stakeholders
* Report structure including an example
* Frequency of benchmarking

# Results

* insert here the reports of the different benchmarking runs

# Discussion

* Discussion of the insights from executing the benchmarking on
  + external feedback on the whole topic and its benchmarking
  + technical architecture
  + data acquisition
  + benchmarking process
  + benchmarking results
  + field implementation success stories

## Declaration of Conflict of Interest

* by each contributor to this document

# Introduction

## Description of the topic: “Psychiatric Multimorbidity”

Psychiatric disorders are among the most common and debilitating illnesses across the lifespan. Epidemiologic studies indicate that 70% of all diagnosable psychiatric disorders begin prior to age 24 (Kessler et al., 2005). This underscores the need for increased focus on studies of the developing brain (Di Martino et al., 2014). Most existing studies employ traditional univariate statistics to identify single indices that show significant differences between healthy subjects and patients with a psychiatric disorder. An example is the theta-beta ratio (TBR) in attention deficit hyperactivity disorder (e.g., Magee et al., 2005). Many such markers, however, tend to be too unstable to enable reliable predictions in new cohorts, and some have recently been scrutinized with respect to their replicability.

Multivariate machine learning (ML) approaches have the potential to overcome limitations of traditional statistics by combining individual data features into composite scores that are optimized for prediction. As the optimal combination is learned automatically from the data, ML approaches are also considered prominent instances of the broader field of artificial intelligence (AI). Optimality with respect to prediction is achieved through several mechanisms, including the subtraction of noise signals with low predictive power and the automatic discovery and use of non-linear interactions between data features, while stability of the derived scores on new data is ensured by automatically constraining the models’ complexity. As such, ML/AI approaches have desirable properties and hold promise to improve the diagnosis of psychiatric and developmental disorders by optimally combining vast numbers of variables derived from behavioral, physiological and other data sources. The prerequisite for employing an AI is, however, the presence of datasets with large sample sizes, from which these complex relationships can be reliably inferred. The acquisition and analysis of such datasets requires large collaborative efforts.

Besides limited sample sizes, a major factor complicating the practical use of AI (e.g., classification algorithms) in psychiatric (developmental) disorders is the presence of multimorbidities. Previous AI approaches show the tendency of researchers to focus on a single disorder (one patient group vs. controls), which limits trans-diagnostic analyses and does not reflect the real-life situation (over 75% of children with a clinical diagnosis have multiple psychiatric disorders), in which a clinician has the task to choose between different diagnoses and/or the combination of multiple diagnoses (i.e., multimorbidity). The present proposal offers a unique large-sample dataset that provides a wide array of different psychiatric developmental disorders. In the present challenge, the goal is to classify the multimorbidity of children and adolescents based on resting electroencephalography (EEG) data, and other data like demographics (age, gender), questionnaires and behavioral data. The sheer sample size of children and adolescents with multimorbidities is worldwide unmatched.

We believe that a community driven effort to derive predictive markers from these data using advanced AI algorithms can help to improve the diagnosis of psychiatric developmental disorders. Specifically, automated assessments of behavioral and neurophysiological measures could support or even replace clinicians’ decisions by providing supplementary evidence from these (arguably more objective) data sources. If reliable signatures of disorders can be found on the provided data alone (as opposed to employing magnetic resonance imaging, for example), such an automated assessment could moreover be obtained in a cost- and relatively time-efficient way.

### Relevance

Children and Adolescents who receive a diagnosis of a psychiatric developmental disorder exhibit enduring impairments throughout adolescence and into adulthood, which can have severe psychological and social consequences (Butterworth and Kovas, 2013; Pastor and Reuben, 2008). Our previous work and other studies indicate that neuroscientific measures provide additional accuracy describing the different clinical conditions (Hoeft et al., 2010; Langer et al., submitted; Supekar et al., 2013). Valid and reliable measurable biomarkers are required to provide biomedical tests for routine clinical practice. Screening of these psychiatric developmental disorders along with neurophysiological measures have the potential to find new possibilities to make diagnoses of psychiatric developmental disorders more objective, reliable and potentially cost-worthy and finally to help refine identification of concurrent occurrences of multiple diagnoses (multimorbidity).

The proposed AI challenge on children with multiple disorders is relevant, because from an internal analysis of 400 children who visited the neuropsychology center at the Child Mind Institute in New York, we observed that over 75% of the patients displayed concurrent multiple psychiatric developmental disorders. However, to the best of our knowledge, no AI project so far has focused on the multimorbidity in children with psychiatric developmental disorders. In the real-life corresponding situation of someone wanting to know if their child has a (or multiple) psychiatric disorders or not, a clinician must be able to determine the diagnosis despite the potential concurrent occurrence of multiple disorders. Therefore, if a biomarker for a psychiatric developmental disorder should be able to support, or even replace, the current diagnosis by a clinician, the marker also has to be valid in the presence of multimorbidities.

In addition to the categorical prediction of different psychiatric developmental disorders, one could also set up a challenge, which has to goal to predict psychiatric developmental disorder on a dimensional level. Dimensional labels could be potentially obtained by auxiliary cognitive and behavioral data from the subjects. The diagnostic classifications of the DSM-V and ICD-10 evolved from conceptualizations that pre-date neuroscience and genetic methods and is based on a categorical classification system. This problem has been extensively discussed (Hyman, 2010; Insel et al., 2010; Kendell and Jablensky, 2003), including excessive comorbidity of disorders, marked heterogeneity of mechanisms, and reification of disorders. In particular the underlying validity of the current diagnostic systems ICD-10 and DSM-V has been scrutinized due to emerging findings from genetics, system neuroscience, and behavioral science that do not accord well onto those independent categorical definitions (Meyer-Lindenberg and Tost, 2012; Morris et al., 2013). Some researchers have argued that a dimensional spectrum may provide a better explanation of the clinical reality (Allardyce et al., 2007; Andrews et al., 2007). Moreover, studies indicate that genes and molecules have to work via mechanistic signal chains, which makes it very improbable that alteration in those signal cascades will have a strong one-to-one mapping with ICD-10 or DSM-V defined developmental disorders (Kapur et al., 2012; Kendler, 2008). In response to this situation, the national institute of mental health of the USA launched the Research Domain Criteria (RDoC) initiative (http://www.nimh.nih.gov/researchpriorities/rdoc/index.shtml) to develop a neuroscience-based nosological framework (“neurophenotyping”) for future research on psychopathology by categorizing individuals using a dimensional approach (Insel et al., 2010). Markon and colleagues synthesized 58 studies addressing the reliability and validity of dimensional versus categorical approaches to the measurement of psychopathology. Reliability and validity for their dimensional approach was increased of 15% and 37%, respectively, compared with the common categorical concept (Markon et al., 2011).

### Possible Impact

The standard practice for diagnosing psychiatric developmental disorders are multiple prolonged interviews conducted by the psychiatrist with the child and its close relatives. The final diagnosis is made according to criteria formulated in classification schemes such as DSM-V and ICD-10. This procedure is relatively costly, as the presence of the clinician is required throughout the entire process. Automated classification based on standardized behavioral or neuroimaging data may require a similar or shorter amount of time, and the acquisition of these data could be carried out by (arguably more abundantly available and less costly) supporting personnel, while the data analysis would be fully computerized. Neuroimaging devices vastly differ in their acquisition and operating costs as well as practical applicability. While structural and functional MR imaging is very expensive and can only be performed in specialized centers, EEG systems are low-cost and can be used anywhere, including private medical practices. Overall, the cost-benefit ratio of automated diagnoses relying on the combination of behavioral and electrophysiological (EEG) data may be competitive to the standard practice provided the predictions made by such a system are accurate.

This consideration is the starting point for the present proposal, which has the central goal of advancing the diagnosis of psychiatric developmental disorders through automated assessment of behavioral and electrophysiological measures. An AI algorithm that can provide/support a diagnosis based on such measures would offer a reliable, objective and cost-worthy diagnostic method compared to the current procedures and finally potentially also shorten the diagnosing time. Such a challenge will help clinicians to decide if a particular AI-based algorithm is suited (or have the potential) to support them in their daily routine and to compare the performance of different AI solution s and to learn about their limits.

Leaving aside economic considerations, the current practice of diagnosis has several additional crucial shortcomings. Even though classification schemes like DSM-V provide precise criteria for each disease, the decision whether each single of these abstract criteria is met to a sufficient degree remains highly subjective. As a consequence, the inter-rater reliability of psychiatric diagnoses has been reported to be low (REF). By utilizing additional behavioural and neurophysiological data sources, which are thought to hold an objective “ground truth” about the underlying pathology, AI algorithms promise to overcome the subjectivity of the manual diagnosis. It is of course true that wrong or contradicting labels (called “label noise”) limits the ability of such algorithms to learn the true data-diagnosis relationship. However, this problem will be addressed here in two ways. First, by requiring a consensus between multiple experts for each diagnosis, our training labels are less noisy than what is typically the case in clinical practice. Second, we advocate the use of state-of-the-art algorithmic approaches that can deal with label noise. Such approaches are either based on identifying mislabeled samples and eliminating them from the training set, or on formulating learning rules that can tolerate a certain amount of label noise (see Frénay et al., 2014 for an overview and Görnitz et al., 2014 for an application of such techniques in a neuroimaging context). Finally, virtually all AI algorithms, even those trained to solve discrete classification problems, provide continuous outputs representing either how prototypical a sample is for a given disease or how certain the algorithm is in its prediction (relative to other possible diseases). These can serve as proxies for the presumed disease severity, and would provide an added value over the binary diagnoses that are common nowadays.

### Existing Work

The identification of imaging biomarkers for psychiatric disorders has been a long-standing goal of clinical neurophysiology. Within this effort, EEG has been identified as a promising technology early on, and several EEG studies have postulated markers such as the theta-beta ratio for ADHD (e.g. Magee et al., 2005, Lenartowicz and Loo., 2014) and frontal alpha asymmetry for depression (e.g. van der Vinne et al., 2017, Olbrich and Arns, 2013) based on “traditional” univariate statistical analyses of (typically) small cohorts. While some of these results have failed to replicate lately (e.g., Gold et al., 2013, similar indices may still constitute starting points for the powerful multivariate big data approaches we will invite with the proposed prediction challenge.

To the best of our knowledge, no AI project so far has focused on the multimorbidity in children with psychiatric developmental disorders. Neuroscientific AI projects on individual disorders (patients vs. control subjects) are more abundant, but these projects almost exclusively focus on structural (and, to a lesser extent, functional) MRI data, which we deem of less practical value in than EEG data in the present context. One example is the ADHD-200 challenge, in which resting state functional magnetic resonance imaging (fMRI) as well as structural MRI data from 776 children was used alongside demographic data to distinguish ADHD patient from healthy controls. Similar efforts in the context of autism spectrum disorder are the ABIDE dataset (Di Martino et al., 2014) and the IMPAC challenge[[1]](#footnote-1).  We are only aware of very few EEG studies using a multivariate predictive approach to address similar classification problems. For example, Abibullaev and An (2012) used a cross-validated support vector machine classification approach on EEG data to classify ADHD in a very small sample of 10 children. In this study they were able to reach an area under the ROC curve of 0.95. However, there were some shortcomings to this study. The sample size was too small to be representative, and to accurately assess the predictive performance, since only 3 children didn’t have ADHD in their sample. The study of Magee et al. (2005) used resting state EEG data to discriminate ADHD from the healthy condition in a larger sample (N = 320). This was possible with a sensitivity of 89% and a specificity of 79.6% with an overall accuracy of 87.0%. These results were, however, obtained in-sample rather than on an independent test set. The study moreover used relatively small sets of pre-selected EEG features (between 4 and 12) for their classification analyses. Thereby, it did not fully exploit the potential of modern AI approaches to autonomously identify the most predictive features from a large pool of candidates. Neither does it fulfill the strict requirements for reporting the statistical performance of AI algorithms that have been put forward by the AI4H focus group.

The proposed data analysis challenge will go beyond the state-of-the-art in multiple ways. What we have in mind is to propose the first public data analysis challenge that has the goal of predicting psychiatric disorders in children using EEG data in combination with behavioral/phenotypical assessments, and that provides the means to objectively evaluate prediction performance in a large sample under standardized conditions. At the same time, we will also initiate the first systematic assessment of multimorbidity in this population. This will be possible by leveraging the, to our knowledge, largest public pediatric clinical EEG database (“Healthy Brain Network”) provided by the Child Mind Institute. As an option, these data will also make it possible to evaluate predictions of disease severity.

### Feasibility

We are convinced that the proposed research project is feasible for the following reasons: First, the planned project benefits from the close collaboration with the ongoing HBN initiative with respect to the granted use of existing and future data. Most importantly, we will profit from the support of Dr. Michael Milham who is a leading expert in the field of pathological developmental neuroscience and has experience with neuroimaging challenges (e.g. ADHD-200, ABIDE). Second, good data quality is guaranteed as it has been shown in an initial release of the HBN initiative (Langer et al., 2017, see also Sections 7 and 8). Third, the expertise regarding signal processing of the EEG and statistical analyses is given by the organizers research groups. Finally, as presented in the previous section, former studies have demonstrated a predictive value of EEG data to discriminate between children with one psychiatric developmental disorder and control children. We expect that the availability of a large standardized training sample will foster the discovery of novel multivariate and multimodal markers for these disorders, and that the presence of large quantities of independent test data will make it possible to objectively quantify the predictive value of these markers under realistic conditions.

## Ethical Consideration

Study participants were taken care of by trained medical professionals throughout the entire process from recruitment to potential follow-ups. The majority of the collected measures (including neuroimaging data) are used routinely in clinical practice. In addition, several cognitive tasks were adopted from the psychological literature. None of the conducted tests has the potential to impose any harm on the study participants. All data were recorded at the Child Mind Institute.

Prior to participation, legal guardians, or participants themselves aged 18 or older, provided written informed consent. This consent explicitly includes the publication of all data in pseudonymized form, and the use of these data by the general public for research purposes.

Ethical approval is given by the Chesapeake Institutional Review Board. The entire project follows the principles of the Declaration of Helsinki.

# Existing AI Solutions

*To be written. For now, refer to some examples in the section 1.1.3. Existing Work.*

*Current Challenges:*

*• Focus on one diagnosis vs. healthy controls*

* *in real life 70% multimorbidity*

*• Univariate vs. multivariate statistics*

*• Small sample size (not enough data)*

*• No objective and standardized preprocessing*

*for EEG data*

*• Unknown reliability of EEG measures*

*• Not enough computing power*

*• Too many variables for human mind/eye*

*- multi dimensional data space*

## Existing Work on benchmarking

*To be written. For now, refer to some examples in the section 1.1.3. Existing Work.*

# AI4H Topic Group

## Topic Group structure

We believe that a community driven effort to derive predictive markers from these data using advanced AI algorithms can help to improve the diagnosis of psychiatric developmental disorders. More details about the activities of the topic group “AI for Psychiatry” can be found in the documents [FGAI4H-C-013](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7b4E497A22-2753-4208-BD31-681F294BB146%7d&file=FGAI4H-C-013.docx&action=default) and [FGAI4H-C-013-A1](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/_layouts/15/WopiFrame.aspx?sourcedoc=%7b604233B9-DF12-4802-8041-908487E1FE0E%7d&file=FGAI4H-C-013-A1.pdf&action=default). These can be accessed with a free ITU account (cf. “Get involved”).

## Subtopics

Currently the Topic Group Psychiatry has only one subtopic: “Prediction of psychiatric multimorbidity in a large pediatric sample”. The topic group “AI for Psychiatry” offers a unique large-sample pediatric dataset that provides a wide array of different psychiatric developmental disorders (*see section XX*). The current goal is to classify the multimorbidity of children and adolescents based on resting state EEG data. In addition, demographic information as well as extensive cognitive and behavioral measures will be permitted to derive predictive models. This restriction is introduced due to the limited real-world practicability and economic viability of MRI and DTI measurements. We would like to invite other researchers and interested stakeholders to complement the current subtopic with additional subtopics. Future subtopics will potentially include other measures, such as task-related EEG and neuroimaging (T1-weighted MRI, DTI, and functional MRI) data to assess whether the previously achieved prediction accuracy can be exceeded using these data.

## Current topic status

We have prepared the AI input data for the challenge. In the present challenge, we provide raw and preprocessed EEG data as well as specifically extracted EEG features, which has been shown relevant to different psychiatric developmental disorders (e.g. theta-beta ratio, frontal alpha asymmetry). In addition, we have prepared an extensive phenotyping protocol of comprehensive psychiatric (clinical classification according to the DSM-V), learning, familial, environmental, and lifestyle assessments. See *section XX* for further details.

## Topic group contributors

Current members of the topic group “AI for Psychiatry” include Prof. Nicolas Langer and Dr. Stefan Haufe. Prof. Langer and Dr. Haufe share interest in using machine learning methods to analyze neurophysiological data in combination with behavioral and cognitive measures to advance the endeavor of biomarkers for psychiatric disorders. Such an integration of different types of brain and behavioral measures requires knowledge about the characteristics of the measurement modalities involved but also about the methodological approaches (e.g. multivariate analysis) to examine information that is encoded in the combination of the measures. During Nicolas Langer’s work with Prof. Nadine Gaab at the Harvard Medical School, Boston, USA, he has built an expertise in studying neurophysiological data and integrating this information behavioral and cognitive data. Prof. Langer has later joined Dr. Michael Milham at the Child Mind Institute in New York, USA, as a co-investigator of the HBN. With the HBN, they initiated a project that closely follows the goal to identify potential biomarkers for psychiatric developmental disorders. Dr. Stefan Haufe is a computer scientist developing machine learning and signal processing techniques for analyzing neuroimaging (in particular, EEG) data. He is currently an ERC junior group leader at Charité - Universitätsmedizin Berlin, Germany, where he is leading a five-year effort to characterize and predict psychiatric and neurological diseases using non-invasive brain electrophysiology. As such, he is very interested in benchmarking EEG-based biomarkers on large public clinical datasets. Prof. Langer and Dr. Haufe are interested to promote and support standardization and benchmarking efforts, which are crucial to the implementation of machine learning for diagnostics of psychiatric disorders.

## 3Topic group participation

The topic group would benefit from further expertise of the medical and AI communities and from additional data. Thus, we like to invite any potential interested stakeholder. For example, clinicians, who could add their clinical expertise, ideally in multimorbidity of psychiatric disorders). In addition, clinical institutions could potentially provide additional (undisclosed) data, such as cognitive, behaviourally, neurophysiological data and clinical diagnosis. Furthermore, we aim to collaborate with researchers from cognitive neuroscience, machine learning experts, engineers and statisticians familiar with neurophysiological data as well as software developers. To join this topic group, please send an e-mail to the focus group secretariat ([tsbfgai4h@itu.int](mailto:tsbfgai4h@itu.int)) and the topic driver ([n.langer@psychologie.uzh.ch](mailto:n.langer@psychologie.uzh.ch)). Please use a descriptive e-mail subject (e.g. "Participation topic group AI for Psychiatry"), briefly introduce yourself and your organization, concisely describe your relevant experience and expertise, and explain your interest in the topic group.

Participation in FG-AI4H is free of charge and open to all. To attend the workshops and meetings, please visit the Focus Group website (<https://www.itu.int/go/fgai4h>), where you can also find the whitepaper, get access to the documentation, and sign up to the mailing list.

## Next Meetings

The Focus Groups meets about every two months at changing locations. The upcoming meetings are:

* E: Geneva, Switzerland; ​29 May-1 June ​2019
* F: Zanzibar, Tanzania; 2-5 September 2019
* G: New Delhi, India; November 2019
* H: Brasilia, Brazi; January 2020

An up to date list can be found at the official [ITU FG AI4H website](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx).

The current plan is that the participator for the topic group psychiatry will attend the meetings in Europe in person. Other meetings will be attended remotely.

# Method

## AI Input Data Structure

### Overview

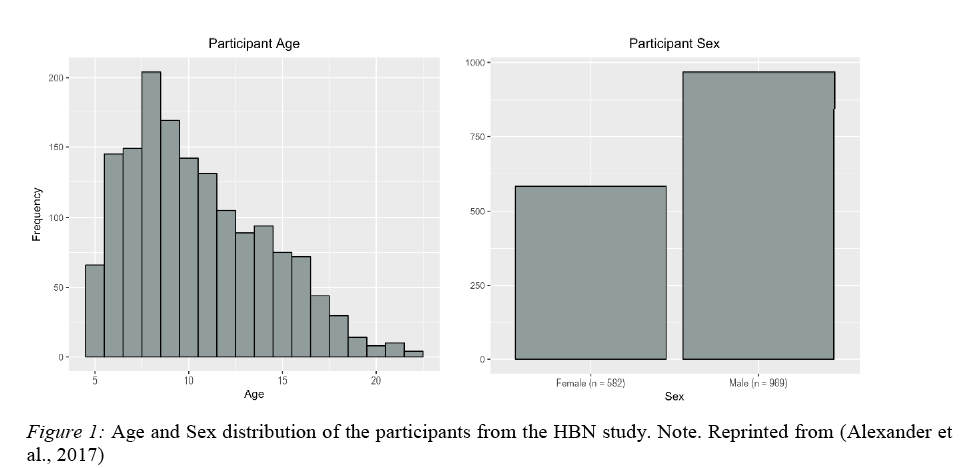
For the present project, we will leverage existing data from the biobank of the Healthy Brain Network (HBN) initiative (Alexander et al., 2017). The data acquisition included multimodal brain imaging (Diffusion Tensor Imaging, structural T1-weighted and functional MRI), electroencephalography (EEG), and an extensive phenotyping protocol of comprehensive psychiatric, learning, familial, environmental, and lifestyle assessments. In addition, clinical classification according to the DSM-V is provided for each subject. The information about the DSM-V diagnosis, demographics, cognitive and behavioral data will be made accessible through a .csv file.

In the context of the present challenge, only resting state EEG data, demographic information as well as extensive cognitive and behavioral measures will be permitted to derive predictive models. This restriction is introduced due to the limited real-world practicability and economic viability of MRI and DTI measurements. Future AI challenges will potentially include other measures, such as task-related EEG and neuroimaging (T1-weighted MRI, DTI, and functional MRI) data to assess whether the previously achieved prediction accuracy can be exceeded using these data. In the present challenge, the organizers will provide raw and preprocessed EEG data as well as specifically extracted EEG features (see below), which has been shown relevant to different psychiatric developmental disorders (e.g. theta-beta ratio, frontal alpha asymmetry). Using this approach, we expect to attract both neuroimaging experts who want to employ their own EEG processing pipeline as well as participants without a neuroscience background, who are more interested in the machine learning aspect of the problem and may be happy to work on pre-extracted features.

All challenge participants have access to the previously published datasets of the Healthy Brain Network Initiative, currently comprising ~1600 samples. In addition, preprocessed data and extracted features will be made available. Together, these resources comprise the training data that can be used by anyone to tune their AI models. Participants are also invited to use their own data for training.

### Subjects

Children and adolescents aged 5-21 years (see figure 1) were recruited through a community referred model. Overall more than 75% of all subjects are diagnosed with multiple psychiatric developmental disorders (ADHD (all subtypes), Anxiety Disorder, Specific Learning Disorder, Autism etc.) (see figure 2). Based on the current data release approximately 48.4% of all subjects are diagnosed with ADHD (all types). 79.2% of subjects diagnosed with ADHD also received one or more additional diagnoses, constituting a sample of great diversity. All subjects were recruited in the greater New York City area, NY, USA. All testings were conducted at the Child Mind Institute.

**

*Figure 1: Age and sex distribution of the participants from the HBN data.*

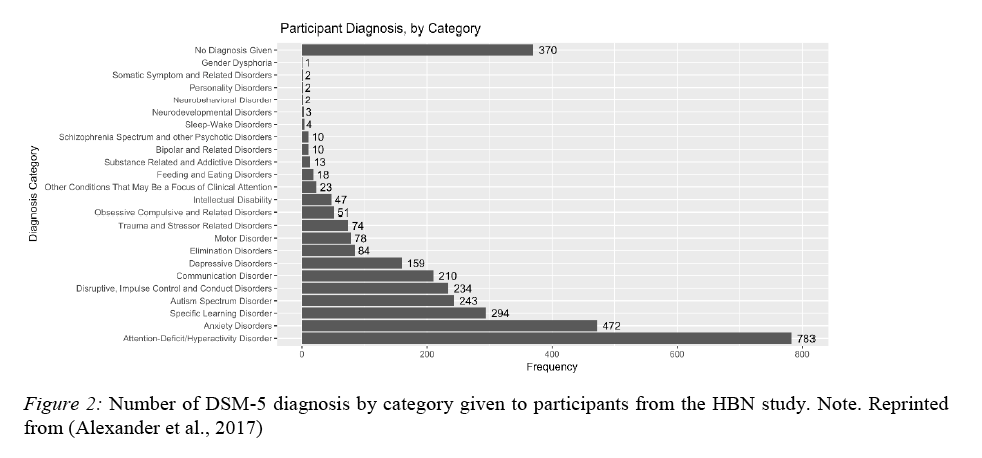
**

Figure 2: Number of DSM-5 diagnosis by category given to participants from the HBN study.

### Data formats

Behavioral data: .csv

Raw and preprocessed EEG data:.mat and .csv

## AI Output Data Structure

The AI output data structure should be a .csv file and structured as follows:

****

## Test Data Labels

We are convinced that the annotations of the data are of high quality for the following reasons: the diagnosis of the children and adolescents is based on a comprehensive diagnostic evaluation reports provided by HBN. A consensus clinical diagnosis was assessed for each child based on the decision of a clinical team which took all interviews and materials conducted as basis for the DSM-5 consensus diagnosis. The clinical staff consists of a combination of psychologists and social workers, with psychopharmacological consultation support provided by psychiatrists. All the tests were conducted by licensed clinicians. Finally, all test scores from clinical interviews are double entered into the database by two (different) trained research assistants.

## Scores & Metrics

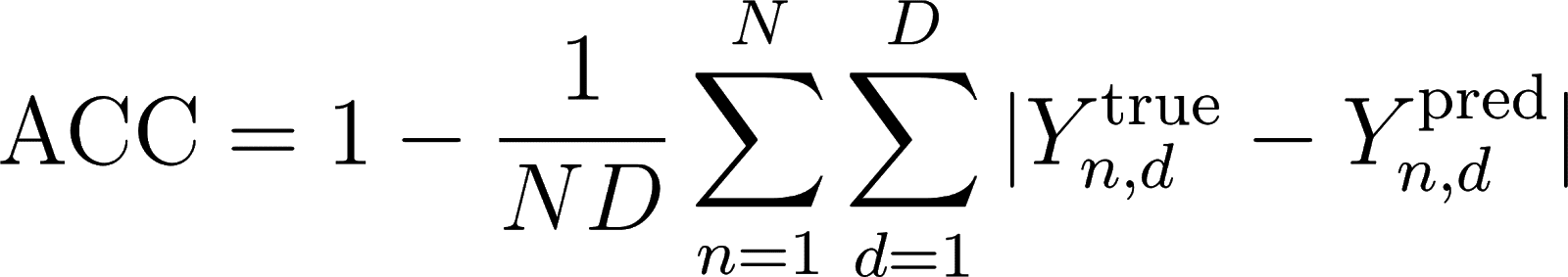
The overarching goal of this challenge is to derive accurate predictions of the existence of one or more psychiatric conditions in the presence of multimorbidities. From a AI standpoint this is a multi-task classification problem, where several related binary classification rules need to be derived from the same data. These predictions should be derived from demographic (age, sex, handedness, etc.), phenotypical (e.g., responses and outcomes of an intelligence scale), and neurophysiological (features derived from resting-state EEG) data. It is also possible to devise a variant in which continuous scores of severity are to be predicted for each disorder, which would turn the problem into a multi-task regression problem. Such continuous labels could be derived from the phenotypical data e.g. using the Strengths and Weaknesses of ADHD Symptoms and Normal Behavior Rating Scales (SWAN, Swanson et al., 2012). In this case, it is understood that the phenotypical data cannot be used any more to derive predictions. Continuous labels for all disorders will be normalized to a common dynamic range (e.g. between 0 and 10).

Overall, we expect that algorithms with the following properties may perform favorably:

1. In psychiatry, the occurrence of multiple morbidities is highly correlated. While it is possible to predict each disease separately (the conventional classification setting), multi-task learning approaches that jointly predict all diseases have an advantage, as they can take correlations of labels and data across distinct diagnoses into account. Technically, such approaches may use parameter sharing to learn common representations of multiple diseases. Two recent paper proposing a multi-task learning approach in joint disease risk prediction and comorbidity discovery as well as general neuroimaging contexts are Wang et al., 2014 and Rahim et al., 2017.
2. Given the to some degree noisy and subjective nature of the categorical diagnoses (although the consensus-based labeling approach employed here represents an effective way of reducing labeling errors), methods coming with a mechanism to deal with label noise (e.g., Görnitz et al., 2014) are expected to perform better than traditional methods.
3. Multi-modal approaches that can properly handle and combine features from different sources (e.g. ordinal questionnaire data, symmetric matrix-valued brain connectivity data, etc.) are expected have an advantage over classical approaches that do not take the structure of the data into account.
4. Approaches employing specific domain-knowledge are expected to perform better than domain-agnostic approaches. As the challenge organizers, we will provide a set of promising candidate features based on the published literature in neurodevelopmental diseases. However, it is conceivable that challenge participants develop additional specific hypotheses based on their own expertise. These can be tested by extracting the corresponding data features directly from the provided raw data.
5. In the absence of specific hypotheses, end-to-end architectures involving multiple layers of nonlinearities (deep learning approaches) may perform favorably, as such models can in theory learn complex non-linear features (such as theta-beta ratios, connectivity metrics) on their own. The relatively large sample size available here may render the application of such approaches feasible.

As can be seen from this list, the posed data analysis problem is challenging and likely requires state-of-the-art AI algorithms to be solved optimally.

For the multi-task classification problem, the primary performance measure (which will be used to rank contributions) will be the “multi-task accuracy” defined as

,

where N is the number of test samples, D is the number of disorders in the sample, Ytrue is a the matrix of labels, where Yn,dtrue is a binary variable (coded as 0/1) indicating the presence of the d-th disorder in the n-th test subject, and Ypred is the corresponding binary matrix of disorder occurrences predicted by the model. As secondary performance measures, we will also evaluate multi-task sensitivity and specificity, as well as accuracy, sensitivity and specificity separately for each disorder. In a similar way, the primary measure for multi-task regression performance will be defined by averaging the squared prediction errors across samples and disorders. Single-disorder mean-squared errors will also be evaluated as secondary performance measures in the continuous label case.

****

## Undisclosed Test Data Set Collection

As the HBN project is still ongoing, test data will become available in regular intervals, the anticipated pace being ~500 new samples per year. A new release is thus expected within the lifespan of the AI4H effort, and will serve as the primary testbed for the proposed challenge. It would be desirable to continue the challenge independently after the completion of AI4H, and to use each new release as test data to assess the performance of the initially submitted algorithms as well as novel, refined, contributions. To this end it would be practical to host critical parts of the challenge infrastructure at a public platform such as crowdai.org, kaggle.com or ramp.studio.

In order to bridge the time before release of the first genuine test set, a pseudo test set will be created by randomly selecting a subset of the published data. The composition of this test set will be kept secret. Submissions to be evaluated on this pseudo test set will need to be submitted no later than six months after public announcement of the challenge. Each participant will be allowed to contribute exactly one AI algorithm. Participants can either be individuals or groups of individuals, but no individual will be allowed to contribute to multiple submissions. Contributions will be ranked according to their test performance in terms of the previously defined primary performance measure. The ranking will be displayed in a public leaderboard at the challenge website.

Before each new batch of (now genuine) test data is publicly released, all participants are invited to submit an updated version of their AI algorithm. These updated algorithms could either incorporate technical refinements or just be trained on a larger corpus of data including the most recent previously released batch. They are in general expected to outperform their predecessors. For each new test set, the current as well as all past submissions of each participant/group will be benchmarked, and all results will be listed together in a separate public leaderboard for this particular test set. This will make it possible to track the progress of each group as well as the field in general. After the evaluation is complete, the test data and their corresponding labels are released to the public and may be used as additional training data in the next round.

Note that with future releases, additional challenge tracks may be established, which may either allow participants to use additional data (e.g. task data, MR imaging data) for prediction, or may pose entirely different prediction tasks.

## Benchmarking Methodology and Architecture

Secret test data (demographic, phenotypic and neurophysiological data and corresponding multi-labels) will be stored at the site of a challenge organizer or a dedicated AI4H server. It is envisioned that challenge participants send executable code for deriving predictions from the demographic, phenotypic and neurophysiological data. The standardized format of training and test data will ensure that code that runs on the public training data will also be applicable to the test data. We favor a flexible approach, in which participants can choose their own environment (e.g. Python, C++, Java, Matlab) for developing the prediction system. Such an approach could be implemented using docker containers. At the organizers’ site(s), derived predictions will be compared to the test labels to derive measures of prediction accuracy.

## Reporting Methodology

*To be written*

**Idea: continuous prediction challenge**

• Participant teams can refine and upload containers any time

• Benchmarking of most recent containers each time new data are released

• Time stamp system allows public release of test set without delay

• Tracking progress over time as new releases become available

****

****

****

****

# Results

*To be written*

# Discussion

*To be written*

## Declaration of Conflict of Interest

References

Abibullaev, B., & An, J. (2012). Decision support algorithm for diagnosis of ADHD using electroencephalograms. *Journal of medical systems*, *36*(4), 2675-2688.

Alexander, L. M., Escalera, J., Ai, L., Andreotti, C., Febre, K., Mangone, A., . . . Milham, M. P. (2017). An open resource for transdiagnostic research in pediatric mental health and learning disorders. Scientific Data, 4, 170181. https://doi.org/10.1038/sdata.2017.181

ADHD-200-Results-Webpage: Adhd-200 global competition results. (2011). Retrieved from http://fcon\_1000.projects.nitrc.org/

Allardyce, J., Suppes, T., and Van Os, J. (2007). Dimensions and the psychosis phenotype. Int J Methods Psychiatr Res 16 Suppl 1, S34-40.

Andrews, G., Brugha, T., Thase, M.E., Duffy, F.F., Rucci, P., and Slade, T. (2007). Dimensionality and the category of major depressive episode. Int J Methods Psychiatr Res 16 Suppl 1, S41-51.

Barry, R. J., Clarke, A. R., Johnstone, S. J. & Brown, C. R. (2009). EEG differences in children between eyes-closed and eyes-open resting conditions. Clinical neurophysiology: official journal of the International Federation of Clinical Neurophysiology 120, 1806–1811.

Bigdely-Shamlo, Nima, Tim Mullen, Christian Kothe, Kyung-Min Su, and Kay A. Robbins. 2015. “The PREP Pipeline: Standardized Preprocessing for Large-Scale EEG Analysis.” Frontiers in Neuroinformatics 9 (June): 16.

Brown, M. R., Sidhu, G. S., Greiner, R., Asgarian, N., Bastani, M., Silverstone, P. H., ... & Dursun, S. M. (2012). ADHD-200 Global Competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. *Frontiers in systems neuroscience*, *6*, 69.

Butterworth, B., and Kovas, Y. (2013). Understanding neurocognitive developmental disorders can improve education for all. Science 340, 300-305.

Chen, A. C., Feng, W., Zhao, H., Yin, Y. & Wang, P. (2008). EEG default mode network in the human brain: spectral regional field powers. NeuroImage 41, 561–574.

Di Martino, A., Yan, C. G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., ... & Deen, B. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, *19*(6), 659.

Di Martino, A. et al. Unraveling the miswired connectome: a developmental perspective. Neuron 83, 1335–1353 (2014).

Frénay, B., & Verleysen, M. (2014). Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, *25*(5), 845-869.

Gold, C., Fachner, J., & Erkkilä, J. (2013). Validity and reliability of electroencephalographic frontal alpha asymmetry and frontal midline theta as biomarkers for depression. *Scandinavian Journal of Psychology*, *54*(2), 118-126.

Görnitz, N., Porbadnigk, A., Binder, A., Sannelli, C., Braun, M., Müller, K. R., & Kloft, M. (2014, April). Learning and evaluation in presence of non-iid label noise. In *Artificial Intelligence and Statistics* (pp. 293-302).

Haufe, S., Nikulin, V. V., Müller, K. R., & Nolte, G. (2013). A critical assessment of connectivity measures for EEG data: a simulation study. *Neuroimage*, *64*, 120-133.

Hoeft, F., McCandliss, B.D., Black, J.M., Gantman, A., Zakerani, N., Hulme, C., Lyytinen, H., Whitfield-Gabrieli, S., Glover, G.H., Reiss, A.L., and Gabrieli, J.D. (2010). Neural systems predicting long-term outcome in dyslexia. Proc Natl Acad Sci U S A 108, 361-366.

Hyman, S.E. (2010). The diagnosis of mental disorders: the problem of reification. Annu Rev Clin Psychol 6, 155-179.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., Sanislow, C., and Wang, P. (2010). Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am J Psychiatry 167, 748-751.

Kendell, R., and Jablensky, A. (2003). Distinguishing between the validity and utility of psychiatric diagnoses. Am J Psychiatry 160, 4-12.

Kessler, R. C. et al. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch. Gen. Psychiatry 62, 593–602 (2005).

Langer, N., Pedroni, A., Gianotti, L. R., Hänggi, J., Knoch, D., & Jäncke, L. (2012). Functional brain network efficiency predicts intelligence. Human brain mapping, 33(6), 1393-1406.

Langer, N., Gorgolewski, C., Becker, B., Benjamin, C., and Gaab, N. (submitted). Examing the comorbid brain using multivariate pattern analysis.

Langer, N., Ho, E. J., Alexander, L. M., Xu, H. Y., Jozanovic, R. K., Henin, S., . . . Kelly, S. P. (2017). A resource for assessing information processing in the developing brain using EEG and eye tracking. Scientific Data, 4, 170040. https://doi.org/10.1038/sdata.2017.40

Lehmann, D., Ozaki, H., & Pal, I. (1987). EEG alpha map series: brain micro-states by space-oriented adaptive segmentation. Electroencephalography and clinical neurophysiology, 67(3), 271-288.

Lenartowicz, A., & Loo, S. K. (2014). Use of EEG to diagnose ADHD. Current Psychiatry Reports, 16(11), 498. https://doi.org/10.1007/s11920-014-0498-0

Magee, C.A., Clarke, A.R., Barry, R.J., McCarthy, R., Selikowitz, M. (2005). Examining the diagnostic utility of EEG power measures in children with attention deficit/hyperactivity disorder. Clinical Neurophysiology. 116, 1033-1040.

Markon, K.E., Chmielewski, M., and Miller, C.J. (2011). The reliability and validity of discrete and continuous measures of psychopathology: a quantitative review. Psychol Bull 137, 856- 879.

Meyer-Lindenberg, A., and Tost, H. (2012). Neural mechanisms of social risk for psychiatric disorders. Nat Neurosci 15, 663-668.

Michel, C., Koenig, T., Brandeis, D., Gianotti, L. R. & Wackermann, J. (2009). Electrical Neuroimaging Cambridge University Press.

Näpflin, M., Wildi, M., & Sarnthein, J. (2007). Test–retest reliability of resting EEG spectra validates a statistical signature of persons. Clinical Neurophysiology, 118(11), 2519-2524.

Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S., & Hallett, M. (2004). Identifying true brain interaction from EEG data using the imaginary part of coherency. Clinical neurophysiology, 115(10), 2292-2307.

Olbrich, S., Anrs, M. (2013). EEG biomarkers in major depressive disorder:Discriminative power and prediction of treatment response. International Review of Psychiatry, 25:5, 604-618, DOI: 10.3109/09540261.2013.816269

Pascual-Marqui, R. D. (2007). Discrete, 3D distributed, linear imaging methods of electric neuronal activity. Part 1: exact, zero error localization. arXiv preprint arXiv:0710.3341.

Pastor, P.N., and Reuben, C.A. (2008). Diagnosed attention deficit hyperactivity disorder and learning disability: United States, 2004-2006. Vital Health Stat 10, 1-14.

Parra, L. C., Spence, C. D., Gerson, A. D., & Sajda, P. (2005). Recipes for the linear analysis of EEG. NeuroImage, 28(2), 326–341. https://doi.org/10.1016/j.neuroimage.2005.05.032

Pedroni, A., Bahreini, A., & Langer, N. (2018). AUTOMAGIC: Standardized Preprocessing of Big EEG Data. bioRxiv; doi: https://doi.org/10.1101/460469

Rahim, M., Thirion, B., Bzdok, D., Buvat, I., & Varoquaux, G. (2017). Joint prediction of multiple scores captures better individual traits from brain images. *NeuroImage*, *158*, 145-154.

Supekar, K., Swigart, A.G., Tenison, C., Jolles, D.D., Rosenberg-Lee, M., Fuchs, L., and Menon, V. (2013). Neural predictors of individual differences in response to math tutoring in primary-grade school children. Proc Natl Acad Sci U S A 110, 8230-8235.

Swanson, J. M., Schuck, S., Porter, M. M., Carlson, C., Hartman, C. A., Sergeant, J. A., ... & Wigal, T. (2012). Categorical and Dimensional Definitions and Evaluations of Symptoms of ADHD: History of the SNAP and the SWAN Rating Scales. *The International journal of educational and psychological assessment*, *10*(1), 51.

Swanson, J., Deutsch, C., Cantwell, D., Posner, M., Kennedy, J., Barr, C., . . . Spence, A. (2001). Genes and attention-deficit hyperactivity disorder. Clinical Neuroscience Re-search. (1), 207–216.

van der Vinne, N., Vollebregt M.A., van Putten, M.J.A.M., Arns, M. (2017). Frontal alpha asymmetry as a diagnostic marker in depression: Fact or

fiction? A meta-analysis. Neuroimage Clinical, (16). 79-87.

van Dinteren, R., Arns, M., Jongsma, M. L. & Kessels, R. P. (2014). P300 development across the lifespan: a systematic review and meta-analysis. PLoS ONE 9, e87347.

Vlahou, E. L., Thurm, F., Kolassa, I. T., & Schlee, W. (2014). Resting-state slow wave power, healthy aging and cognitive performance. Scientific reports, 4, 5101.

Vecchio, F., Babiloni, C., Lizio, R., Fallani, F., V., Blinowska, K., Verrienti, G., Frisoni, G., & Rossini, P. (2012). Resting state cortical EEG rhythms in Alzheimer's disease: toward EEG markers for clinical applications: a review. Supplements to Clinical neurophysiology, 62:223-236.

Van Veen, B. D., Van Drongelen, W., Yuchtman, M., & Suzuki, A. (1997). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. IEEE Transactions on biomedical engineering, 44(9), 867-880.

Wang, X., Wang, F., & Hu, J. (2014, August). A multi-task learning framework for joint disease risk prediction and comorbidity discovery. In Pattern Recognition (ICPR), 2014 22nd International Conference on (pp. 220-225). IEEE.

Winkler, I., Brandl, S., Horn, F., Waldburger, E., Allefeld, C., & Tangermann, M. (2014). Robust artifactual independent component classification for BCI practitioners. *Journal of neural engineering*, *11*(3), 035013.

Winkler, I., Haufe, S., & Tangermann, M. (2011). Automatic classification of artifactual ICA-components for artifact removal in EEG signals. *Behavioral and Brain Functions*, *7*(1), 30.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. https://paris-saclay-cds.github.io/autism\_challenge/ [↑](#footnote-ref-1)