INTERNATIONAL TELECOMMUNICATION UNION
TELECOMMUNICATION STANDARDIZATION SECTOR
STUDY PERIOD 2022-2024

FGAI4H-P-019-A01
ITU-T Focus Group on AI for Health
Original: English

WG(s): Plen Helsinki, 20-22 September 2022

DOCUMENT

Source: TG-Psy Topic Driver

Title: Att.1 – TDD update (TG-Psy) [same as Meeting K]

Purpose: Discussion

Contact: Nicolas Langer
ETH Zurich, Switzerland
Email: n.langer@psychologie.uzh.ch

Contact: Stefan Haufe
Charité – Universitätsmedizin Berlin
Germany
Email: stefan.haufe@charite.de

Contact: Tim Hahn
University of Münster
Germany
Email: hahn@wwu.de

Abstract: This topic description document (TDD) specifies a standardized benchmarking for AI-based Psychiatry. It covers all scientific, technical, and administrative aspects relevant for setting up this benchmarking (and follows the template structure defined in document FGAI4H-J-105). The creation of this TDD is an ongoing iterative process until it is approved by the Focus Group on AI for Health (FG-AI4H) as deliverable No. 10.11. This draft will be a continuous input- and output document.

This version of the TDD is the same as seen in Meeting K (FGAI4H-K-019-A01), reproduced for easier reference as a Meeting N document.

Change notes: Version 2 (submitted as FGAI4H-1-019-A01-R01 to meeting 7-8 May 2020)

– 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

Version 1

- First version of TDD

..

Contributors

Nicolas Langer University of Zurich Switzerland	Tel: +416353414 Email: n.langer@psychologie.uzh.ch

CONTENTS

Page

List of Tables

Page

List of Figures

FG-AI4H Topic Description Document

Topic group-Psychiatry

1 Introduction

Psychiatric disorders are among the most common and debilitating illnesses across the lifespan and begin usually prior to age 24, which emphasizes the need for increased focus on studies of the developing brain. The majority of existing studies have focused on differentiating between children with an isolated psychiatric disorder and typically developing children. However, this line of research 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 (multimorbidity). Furthermore, most of the previous studies employ traditional univariate statistics on relatively small samples. Multivariate machine learning/AI approaches have a great potential to overcome the limitations of this approach. The present proposal offers a unique large-sample dataset that provides a wide array of different psychiatric developmental disorders. The goal is to classify the multimorbidity of children and adolescents based on resting electroencephalography (EEG), magnetic resonance imaging (MRI), demographics, and cognitive behavioral data. 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.

This topic description document specifies the standardized benchmarking for Psychiatry systems. It serves as deliverable No. 10.11 version 2.0 of the ITU/WHO Focus Group on AI for Health (FG-AI4H).

1 About the FG-AI4H topic group on Psychiatry

The introduction highlights the potential of a standardized benchmarking of AI systems for Psychiatry to help solving important health issues and provide decision-makers with the necessary insight to successfully address these challenges.

To develop this benchmarking framework, FG-AI4H decided to create the TG-Psychiatry at the meeting 3rd (22-24 January, 2019) in Lausanne.

FG-AI4H assigns a *topic driver* to each topic group (similar to a moderator) who coordinates the collaboration of all topic group members on the TDD. During FG-AI4H meeting 4th Shanghai, Nicolas Langer from University of Zurich was nominated as topic driver for the TG-Psychiatry.

1.1 Documentation

This document is the TDD for the TG-Psychiatry. It introduces the health topic including the AI task, outlines its relevance and the potential impact that the benchmarking will have on the health system and patient outcome, and provides an overview of the existing AI solutions for Psychiatry. It describes the existing approaches for assessing the quality of Psychiatry systems and provides the details that are likely relevant for setting up a new standardized benchmarking. It specifies the actual benchmarking methods for all subtopics at a level of detail that includes technological and operational implementation. There are individual subsections for all versions of the benchmarking. Finally, it summarizes the results of the topic group's benchmarking initiative and benchmarking runs. In addition, the TDD addresses ethical and regulatory aspects.

The TDD will be developed cooperatively by all members of the topic group over time and updated TDD iterations are expected to be presented at each FG-AI4H meeting.

The final version of this TDD will be released as deliverable "DEL 10.11 Psychiatry (TG-Psychiatry)." The topic group is expected to submit input documents reflecting updates to the work on this deliverable **(Table 1)** to each FG-AI4H meeting.

Table 1: Topic group output documents

Number	Title
FGAI4H-I-019-A01-R01	Latest update of the Topic Description Document of the TG-Psychiatry
FGAI4H-H-019-A02	Latest update of the Call for Topic Group Participation (CfTGP)
FGAI4H-I-019-A03	The presentation summarizing the latest update of the Topic Description Document of the TG-Psychiatry

The working version of this document can be found in the official topic group SharePoint directory.

- https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Psy.aspx

Select the following link:

- https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-I-019-A01-R01.docx

1.2 Status of this topic group

The following subsections describe the update of the collaboration within the TG-Psychiatry for the official focus group meetings.

1.2.1 Status update for meeting [27-29 January 2021 Geneva (E-Meeting)]

- Wrote on the J-105 document
- web page for Kaggle competition (https://methlabuzh.wixsite.com/mysite; tested with Chrome)
- derived new neural networks to predict psychiatric disorders
- the challenge now includes MRI data
- Progress data cleaning (behavioral data and EEG)
- Development of R-shiny app (to get easy access to demographics and completeness of the data)
- Prof. Dr. Tim Hahn joined the topic group Psychiatry
- Outlook: work on the benchmarking software

1.3 Topic group participation

The participation in both, the Focus Group on AI for Health and in a TG is generally open to anyone (with a free ITU account). For this TG, the corresponding ‘Call for TG participation’ (CfTGP) can be found here:

- https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/tg/CfP-TG-Psy.pdf

Each topic group also has a corresponding subpage on the ITU collaboration site. The subpage for this topic group can be found here:

- https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Psy.aspx

For participation in this topic group, interested parties can also join the regular online meetings (Monday 6pm MEZ). For all TGs, the link will be the standard ITU-TG ‘zoom’ link:

- https://itu.zoom.us/my/fgai4h

All relevant administrative information about FG-AI4H—like upcoming meetings or document deadlines—will be announced via the general FG-AI4H mailing list fgai4h@lists.itu.int.

All TG members should subscribe to this mailing list as part of the registration process for their ITU user account by following the instructions in the 'Call for Topic Group participation' and this link:

- https://itu.int/go/fgai4h/join

In addition to the general FG-AI4H mailing list, each topic group can create an *individual mailing list:*

- fgai4htgpsy@lists.itu.int

Regular FG-AI4H workshops and meetings proceed about every two months at changing locations around the globe or remotely. More information can be found on the official FG-AI4H website:

- https://itu.int/go/fgai4h

2 Topic description

This section contains a detailed description and background information of the specific health topic for the benchmarking of AI in Psychiatry and how this can help to solve a relevant 'real-world' problem.

Topic groups summarize related benchmarking AI subjects to reduce redundancy, leverage synergies, and streamline FG-AI4H meetings. However, in some cases different subtopic groups can be established within one topic group to pursue different topic-specific fields of expertise. The TG-Psychiatry currently has one subtopic "Prediction of psychiatric multimorbidity in a large pediatric sample". Future subtopics might be introduced.

2.1 Subtopic Prediction of psychiatric multimorbidity in a large pediatric sample

2.1.1 Definition of the AI task

This section provides a detailed description of the specific task the AI systems of this TG are expected to solve. It is *not* about the benchmarking process (this will be discussed more detailed in chapter 4). This section corresponds to DEL03 *"AI requirements specifications*," which describes the functional, behavioural, and operational aspects of an AI system.

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 an 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).

2.1.2 Current gold standard

This section provides a description of the established gold standard of the addressed health topic. 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. 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. 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.

Commented [NL1]: Although psychiatric disorders are a research area in brain science, most psychiatrists still diagnose them based on subjective experience rather than by gaining insights into the pathophysiol- ogy of the diseases [7,8]. As a result, psychiatrists may misdiagnose diseases and incorrectly delineate distinct paths of treatment. Therefore, it is urgent to develop a clear understanding of the etiology and pathogenesis of major psychiatric diseases in order to develop effective treatments and interventions for major brain diseases.

2.1.3 Relevance and impact of an AI solution

This section addresses the relevance and impact of the AI solution (e.g., on the health system or the patient outcome) and describes how solving the task with AI improves a health issue.

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.

Commented [NL2]: •Description of the general status and the maturity of AI systems for the health topic of your TG (e.g., exclusively prototypes, applications, and validated medical devices)
•Which are the currently known AI systems and their inputs, outputs, key features, target user groups, and intended use (if not discussed before)? This can also be provided as a table.
•What are the common features found in most AI solutions that might be benchmarked?
•What are the relevant metadata dimensions characterizing the AI systems in this field and with relevance for reporting (e.g., systems supporting offline functions, availability in certain languages, and the capability to process data in a specific format)?
•Description of existing AI systems and their scope, robustness, and other dimensions.

2.1.4 Existing AI solutions

This section provides an overview of existing AI solutions for the same health topic that are already in operation. It should contain details of the operations, limitations, robustness, and the scope of the

available AI solutions. The details on performance and existing benchmarking procedures will be covered in chapter 6.

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]. 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.

3 Ethical considerations

The rapidly evolving field of AI and digital technology in the fields of medicine and public health raises a number of ethical, legal, and social concerns that have to be considered in this context. They are discussed in deliverable DEL01 "*AI4H ethics considerations,"* which was developed by the working group on "Ethical considerations on AI4H" (WG-Ethics). This section refers to DEL01 and should reflect the ethical considerations of the TG-Psychiatry.

While AI technologies can provide numerous practical benefits, there are also important ethical issues that require careful thought and consideration by the designers, manufacturers, and end users

[1] https://paris-saclay-cds.github.io/autism_challenge/

(i.e., clin- icians and healthcare organizations) of these technologies. The use of new technologies in health care, including many of the AI technologies and approaches described in this book, can cause new ways that existing ethics requirements (and laws) can be violated or challenged. Some of the issues concern additional risks that impact patient privacy, safety, autonomy, and trust. The use of autonomous intelligent care providers is an emerging opportunity that requires particular attention. Consider a virtual human or robot that is designed to provide treatment interventions and monitoring of the symptoms of psychiatric patients. The system will need to be able to make decisions and select appropriate courses of action that are consistent with applicable professional ethics codes and standards during interactions with care seekers. The system must also be capable of resolving complex ethical dilemmas that it encounters during the course of care. The designers, manufacturers, and end users of these technologies must be aware of the ethical issues in order to optimize the design and use of these technologies. The four cornerstones of medical ethics are respect for autonomy, beneficence, nonmaleficence, and justice. Respect for autonomy affirms patients' right to think freely and decide and act on their own free will. This includes patients' rights to self-determination and full disclosure of information so that patients can make informed decisions and choices regarding their care. Beneficence means that healthcare providers will promote patients' general wellbeing. Nonmaleficence requires that the actions of care providers do not harm the patient involved or others in society. This includes providing treatments known to be effective and assuring competence of care. Justice refers to the principle that holds that patients in similar situations should have access to the same care, and that in allocating resources to one group we should assess the impact of this choice on others. These are considered prima facie principles because each principle is considered binding unless it conflicts with another principle, and if it does, it requires us to choose between them (Gillon, 1994). The American Psychiatric Association (APA), American Psychological Association (APA), and the American Counseling Association (ACA) are examples of several of the largest mental healthcare professional organizations in the United States that have published ethical codes and guidelines for their respective professions (see ACA, 2014; APA, 2002, 2013). Several organizations and individuals have proposed guidelines regarding the ethical use and design of intelligent machines. In 2011, the Engineering and Physical Sciences Research Council (EPRSC) and the Arts and Humanities Research Council (AHRC) (Great Britain) published a set of ethical principles for designers, builders, and users of robots along with seven "high-level messages intended to be conveyed" (see Table 11.1). Riek and Howard (2014) also proposed ethical guidelines specific to robots (see Chapter 8), and Luxton (2014) has provided recommendations specifically for ethical use and design of AICPs (see Tables 11.2 and 11.3). Mental healthcare professionals conduct many activities that require general and specific training and skill. For example, they need to know how to appropriately apply therapeutic methods and treatments based on established best practices and standards of care. Competence refers to their ability to appropriately perform these tasks. Competence is of ethical significance because providing services outside of the boundaries of trained skill or expertise could put patients at risk of harm The use of AI systems can present a significant ethical problem if they are not adequately designed to perform tasks in a competent manner or controlled based on the scope of their tested capabilities. Highly autonomous systems would need to demonstrate core competence in the application of interpersonal techniques, treatment protocols, safety protocols (e.g., plan for how to manage situations where care seekers indicate intent to self-harm or harm another person), and cultural competencies. Virtual care providers that become accessible on the Internet and that claim to provide clinical services or benefits when they are not adequate or appropriate to do the stated services may be a particular problem. Professional users of AI systems (i.e., mental healthcare providers, companies that provide AI services) must have an understanding of the appropriate use of, and limitations of, the systems. Future guidelines should include requirements to demonstrate that system users understand the capabilities, scope of use, and limitations of these systems. Ideally, it is best if the burden is not placed on the users of the system, but on the design of the system, with medical ethicists involved, to ensure that it acts in an ethically responsible manner towards the humans it interacts with. Another way that healthcare providers can cause harm to patients is by not

adhering to patient safety protocols, standards, or procedures. Some of the principal things that can go wrong include significant worsening of psychiatric symptoms, adverse emotional reactions (e.g., anger outbursts), self-harm, or unexpected medical emergencies. In order for AI systems to function optimally, they will need to be capable of monitoring risks of harm to the patient or others. This would include monitoring of content of what the patient discloses and ongoing assessment of risk factors, such as significant worsening of symptoms and indication of violent behavior toward the self or others. These systems will also need to be able to apply clinical judgment (e.g., assessing intent, weighing clinical and other factors) to make decisions regarding the most ethically justifiable or acceptable course of action. There should also be a back-up plan if the AI experiences a malfunction during a session or is unavailable (e.g., a network failure occurs). If a human care provider is to step in when an AI is unavailable, the credentials of that care provider need to be established (e.g., professional licensure requirements, appropriate training). Respect of privacy reflects the right to autonomy and dignity and it is seen as essential to individual wellbeing (Moore, 2005). Betrayal of trust due to privacy invasion or unauthorized use of information damages the trust of both individual healthcare providers and their profession. Possible misuse of private data can result from haphazard handling of data (e.g., leaving a lap-top with patient data unsecured in a coffee shop) or unauthorized access to electronic data by hackers. While threats to patient privacy (electronic data security) remain common to many types of technologies used in the practice and management of health care, current and emerging technological capabilities, such as psychological signal detection (e.g., via visual, voice, psychological data collection) as well as recorded conversations (audio, video, or text) between patients and AI create the potential for much more data to be collected about individuals, and without individuals being aware of the collection (Luxton, 2014).

In the present data set, 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.

4 Existing work on benchmarking

This section focuses on the existing benchmarking processes in the context of AI and Psychiatry for quality assessment. It addresses different aspects of the existing work on benchmarking of AI systems (e.g., relevant scientific publications, benchmarking frameworks, scores and metrics, and clinical evaluation attempts). The goal is to collect all relevant learnings from previous benchmarking that could help to implement the benchmarking process in this topic group.

4.1.1 Publications on benchmarking systems

While a representative comparable benchmarking for Psychiatry does not yet exist, some work has been done in the scientific community assessing the performance of such systems. This section summarizes insights from the most relevant publications on this topic. It covers parts of the deliverable DEL07 *"AI for health evaluation considerations,"* DEL07_1 *"AI4H evaluation process description,"* DEL07_2 *"AI technical test specification,"* DEL07_3 *"Data and artificial intelligence assessment methods (DAISAM),"* and DEL07_4 *"Clinical Evaluation of AI for health"*.

In recent years, artificial intelligence (AI)-based applications have rapidly been developed for psychiatric research and diagnosis [9–15]. For example, Jan et al. [16] proposed an AI system to

moni- tor depression that can predict Beck Depression Inventory II (BDI-II) scores from vocal and visual expressions. In addition, Wen et al. [17] extracted multi-type gray–white matter features based on multimodal neuroimaging and used a multicore learning classifier to assign weights to the kernel functions of each feature. However, to the best of our knowledge, there is no systematic review that illustrates the use of these AI-based applications for psychiatric research and diagnosis. Thus, we will briefly review commonly used AI-based applications for psychiatric disorders and illustrate how to apply AI technology to explore biomarkers for psychiatric disorders. AI techniques [18,19] are being progressively introduced for psychiatric disorders. Brain structure and function are the most important biological phenotypes and key diagnostic biomarkers for psychiatric disorders [20]. Therefore, AI-related techniques that can obtain detailed information to characterize different psychia- tric disorders should be used for the diagnosis of these diseases [16]. There are three major techniques for brain observation in the study of psychiatric disorders: magnetic resonance imaging (MRI), electroencephalography (EEG), and kinesics diagnosis [21]. We will subsequently discuss their related AI-based applications.

Magnetic resonance imaging (MRI)

MRI is the predominant technique for behavioral and cognitive neuroscience since it can explore obvious psychiatric abnormali- ties that cannot be detected by computed tomography (CT) [22– 25]. At present, commonly used AI technologies for brain imaging include multitask/multimodal learning, classification, kernel, and deep learning methods [26], which can help in effectively analyz- ing existing disease data for key biomarkers exploration and increasing the capacity for clinical brain disease treatment [24,25].

Although many AI-related applications have been developed to assist MRI [26–28], this section only focuses on convolutional neu- ral networks (CNNs) [29] and deep neural networks (DNNs) [30– 32], which are employed in neuroimaging studies to elucidate the neural correlates of psychiatric disorders [30,33–36]. For exam- ple, Hosseini-Asl et al. [37] proposed a new depth-supervised adaptive three-dimensional (3D) CNN that can automatically extract and recognize Alzheimer's disease features, capture changes caused by Alzheimer's disease, and use these networks to analyze and recognize MRI images. In addition, Koyamada et al. [38] built up a subject-transfer decoder using a DNN. It is trained by a functional MRI (fMRI) dataset in the Human Connec- tome Project (HCP), the decoder of which has been assessed as hav- ing higher decoding accuracy than other methods.

Although MRI is currently an important tool for diagnosis in general, it still has several major shortcomings. First, it requires extensive computer configurations. Second, big data is needed to optimize the key parameters of the model. Third, the imaging pro- cess takes a long time. Thus, the question of how to improve the current AI-based applications to solve these problems for MRI is an important future research direction.

Electroencephalography (EEG)

Diagnosis and treatment of human brain and nervous system diseases can be performed by detecting and recording human EEG signals. EEG signals are critical for both understanding how the human brain processes information and diagnosing psychiatric disorders [39]. In comparison with CT and MRI, EEG has a higher temporal resolution [40]. Therefore, despite the limited spatial resolution of EEG, it is still a valuable tool for research and diagnos- tics, especially when specified studies require time resolution in the millisecond range, such as studies on anxiety, psychosis, and depression [41].

Here, we focus on describing the application of the classical machine learning algorithm for EEGs. Since EEG data is represented by a graph, it is always analyzed by AI-based models [42–45]. For

example, Field and Diego [46] employed linear discriminant anal- ysis to process EEG data and obtained 67% accuracy when classify- ing normal patients and patients with depression. In addition, Iosifescu et al. [47] employed a support vector machine (SVM) to process resting-state EEG data for 88 subjects at the midpoint of the eight-lead connection at the forehead and achieved a 70% clas- sification accuracy. Moreover, Bisch et al. [48] used logistic regres- sion (LR) to classify a nine-lead EEG for depression with an 83.3% classification accuracy.

Although EEGs can simplify the data acquisition process, they encounter information loss. More importantly, too many undetermined factors in EEG data result in a large amount of noise in the classification decision. Therefore, developing a machine learning model that is more suitable for EEG data is a future research direction.

Kinesics

Kinesics data (including behavioral [49], facial [50], and other data [48]) is becoming very important for the study of the pathogenesis, development transition, and diagnosis assistance of psychiatric disorders. AI-based technologies are widely employed to analyze such data to help diagnose and predict psychiatric disorders.

Many AI-related applications have recently been developed for kinesics diagnoses [50–52]. For example, Wang et al. [53] proposed a computational approach to develop probabilistic facial expres- sion profiles for video data, which can automatically quantify the difference in emotional expression between patients with psychi- atric disorders (e.g., schizophrenia) and healthy controls [16]. Zhu et al. [54] implemented automatic diagnosis of depression by means of a deep learning algorithm, which significantly improved the depression prediction performance by reducing the mean absolute error by 30.3%. In addition, Kaletsch et al. [55] examined differences in emotional expression by body movements between patients with major depressive disorder (MDD) and their healthy counterparts, and demonstrated that patients with MDD are more negative than their healthy counterparts.

In addition, Dhamecha et al. [56] proposed an algorithm to investigate human and machine performance for recognizing/ verifying disguised faces [57]. The method can identify disguised face patches and account for this information to obtain improved matching accuracy by automatically localized feature descriptors. The experiments showed that the proposed algorithm can not only outperform popular commercial systems, but also evaluate the dis- guised face images when they are matched.

In general, with the development of AI and precision medicine, collecting and analyzing kinesics data will become easier, cheaper, and more convenient. Moreover, kinesics data could help to improve models' predictive accuracy, reduce the misdiagnosis rate, and assist psychiatrists in diagnosing and treating psychiatric disorders.

Artificial intelligence algorithms

Bayesian model

In AI, the naïve Bayes classifier [58–60] is a general term for a classification algorithm. The naïve Bayesian method is a classification method based on Bayes' theorem and characteristic condition-independent hypothesis.

Recent studies have often employed Bayesian models to diagnose psychiatric disorders. For example, the Strüngmann Forum on Computational Psychiatry [61–63] proposed using Bayesian inference to connect underlying causes (genetics and sociological phenomena [15,64]), latent hypothesized theoretical constructs, and symptoms [65]. Furthermore, Grove et al. [66] used a

Bayesian model comparison approach to explore the relationship between visual integration and general cognition. The results showed that a Bayesian model can draw a comparison of the disease categorization systems and have common psychopathological information from diagnostic groups.

Logistic regression

In statistics, logistic models [67,68] (or logit models) are widely used statistical models, and LR is an important AI algorithm [68,69]. Recent studies often employ LR models to diagnose psychiatric disorders. For example, Hagen et al. [70] evaluated the associations between psychological distress and two cognitive screening tools by means of an LR method. The results demonstrated that performance-based assessment could reduce the impact of psychological distress on cognitive screening. In addition, Barker et al. [71] employed models of multivariable LR to predict 30-day psychiatric readmission. Their findings are considered to be crucial predictors for psychiatric readmission, and have provided a better way of readmission prediction.

Shen et al. [72] generated a risk stratification model to obtain the odds ratio (OR) of psychiatric comorbidities by a classification and regression tree method. Using the LR method, the OR of psychiatric comorbidities was calculated between subjects with and without borderline personality disorder. In general, the accuracy of LR models is so high that they are commonly applied in clinical practice.

Decision tree

A decision tree [73] is a flowchart-like diagram that shows the various outcomes from a series of decisions, including chance event outcomes and utility. Decision trees are one of the most widely and broadly used algorithms for supervised classification learning. In AI, a decision tree is a predictive model that represents a mapping between object properties and object values. Most modern decision tree learning algorithms adopt a purity-based heuris- tic [74]. Next, we detail two commonly used decision tree applications for psychiatric disorders.Carpenter et al. [77] used the decision tree algorithm to test whether individual Preschool Age Psychiatric Assessment (PAPA) items can predict whether a child is likely to have generalized anxiety disorder (GAD) or separation anxiety disorder (SAD). They used a decision tree to identify children who were on the brink of experiencing anxiety disorder, and their results showed that the decision tree can achieve accurate prediction up to 96% for both GAD and SAD.With a decision tree, Sattler et al. [78] analyzed data from the Spence Children's Anxiety Scale (SCAS) and SCAS-P obsessive–com- pulsive disorder subscales, and worked out two screening algo- rithms to diagnose obsessive–compulsive disorder from a combined clinical and community sample of children and families. The results showed that the algorithms that reduced the number of SCAS-P items needed to make a diagnosis of obsessive–compulsive disorder diagnoses up to 67%–83% without sacrificing the nature relative to the full subscales.

Support vector machines

The SVM is a current supervised learning method, the decision boundary of which is the maximum margin hyperplane for solving learning samples [79]. It can be described as follows: Start from a training dataset of n points of the form is used to denote the class labels. The goal is to find the maximum margin hyperplane that divides the group of points.

SVM models have been commonly used for diagnosing psychiatric disorders. For example, in order to describe users' situations, Peng et al. [80] employed a multi-kernel SVM-based model to locate potential users who might suffer from depression by extracting three social methods (user microblog text, user profile, and user behaviors). Based on a multiclass SVM, Al-Shargie et al. [81]

put forward a discriminant analysis method. The results showed that the method could discriminate between different stress levels for EEG with a 94.79% average classification accuracy.

Deep learning

Classic machine learning methods, such as the Bayesian model and SVM, have been widely employed in psychiatry and neuro- science [64–66] studies for a long time. At present, deep learning [82–84], which is a hot machine learning research direction, out- performs the aforementioned AI models by a considerable margin [85–87].

Deep learning refers to a set of algorithms on a multi-layer neu- ral network that uses various machine learning algorithms to solve various problems such as images and text. Combined with low- level features, deep learning can develop more abstract high- level attribute categories or features that can discover distributed feature representations of data. Weight updating can be solved by the stochastic gradient descent method:

The choice of this function is related to the type of learning (such as supervised learning, unsupervised learning, and enhanced learning) and the activation function. Here, we detail two commonly used deep learning applications for psychiatric disorder diagnosis.

By leveraging a DNNs on the TensorFlow framework, Khan et al. [88] proposed a computational tool (integrated mental-disorder GEnome score, or iMEGES) to analyze the whole genome/exome sequencing data on personal genomes. Based on the deep learning framework, this tool creates prioritized gene scores for psychiatric disorders [89]. The findings revealed that the property of this tool are better than that of competing approaches when a large training dataset is available.

In addition, Heinsfeld et al. [39] applied deep learning algorithms on a large brain imaging dataset in order to identify patients with autism spectrum disorder based solely on the patients' brain activation patterns. The findings revealed that 70% accuracy was achieved in the dataset, and that deep learning methods can classify large datasets better than other methods. Furthermore, the results showed the promise of deep learning for clinical datasets and illustrated the future application of AI in the identification of mental disorders.

Although extremely advanced performance has been demonstrated in several fields, deep learning has been under close concern for its lack of transparency during the learning and testing processes [90–92]. For example, deep learning has been referred to as a ‘‘black box.” In comparison, techniques such as LR are simple and easy to understand.

For this reason, recent endeavors in interpretable DNNs are introduced here. For example, in terms of CNN visualization, Springenberg et al. [93] proposed a deconvolution approach that can be used to acquire features from deep learning. Kindermans et al. [94] proposed a method to visualize the regions in the input image that contribute most to the CNN decision-making process. In a study on interpreting neural networks with traditional machine learning models, Zhang et al. [95] proposed a method to interpret the convolution layer characteristics of pretrained CNNs, and used an explanatory graph to reveal the knowledge level hidden in the CNN. In short, a good AI model should be interpretable, generalizable, and more adaptive, and should learn from data, rules, and interactions.

4.1.2 Benchmarking by AI developers

All developers of AI solutions for Psychiatry implemented internal benchmarking systems for assessing the performance. This section will outline the insights and learnings from this work of relevance for benchmarking in this topic group.

The overarching goal of this challenge is to derive accurate predictions of the existence of one or more psychiatric conditions in the presence of multi-morbidities. From an 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 Behaviour 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 favourably:

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 labelling approach employed here represents an effective way of reducing labelling 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 favourably, 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

$$ACC = 1 - \frac{1}{ND}\sum_{n=1}^{N}\sum_{d=1}^{D}\left|Y_{n,d}^{true} - Y_{n,d}^{pred}\right|,$$

where N is the number of test samples, D is the number of disorders in the sample, Y^{true} is a the matrix of labels, where $Y_{n,d}^{true}$ is a binary variable (coded as 0/1) indicating the presence of the d-th disorder in the n-th test subject, and Y^{pred} 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.

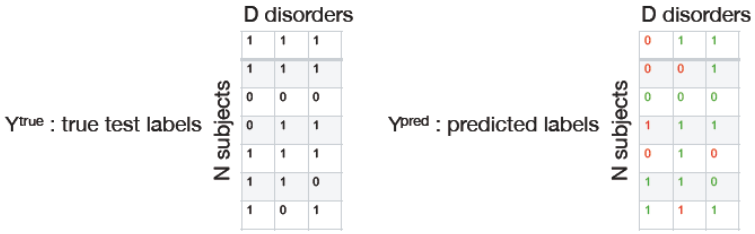


Figure 1: Illustration of Ytrue and Ypred

4.1.3 Relevant existing benchmarking frameworks

Triggered by the hype around AI, recent years have seen the development of a variety of benchmarking platforms where AIs can compete for the best performance on a determined dataset. Given the high complexity of implementing a new benchmarking platform, the preferred solution is to use an established one. This section reflects on the different existing options that are relevant for this topic group and includes considerations of using the assessment platform that is currently developed by FG-AI4H and presented by deliverable DEL07_5 *"FG-AI4H assessment platform"* (the deliverable explores options for implementing an assessment platform that can be used to evaluate AI for health for the different topic groups).

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 EvalAI, AIcrowd, crowdai.org, kaggle.com or ramp.studio. The decision on the final platform is still to be discussed.

5 Benchmarking by the topic group

This section describes all technical and operational details regarding the benchmarking process for the Psychiatry AI task including subsections for each version of the benchmarking that is iteratively improved over time.

It reflects the considerations of various deliverables: DEL05 *"Data specification"* (introduction to deliverables 5.1-5.6), DEL05_1 *"Data requirements"* (which lists acceptance criteria for data submitted to FG-AI4H and states the governing principles and rules), DEL05_2 *"Data acquisition"*, DEL05_3 *"Data annotation specification"*, DEL05_4 *"Training and test data specification"* (which provides a systematic way of preparing technical requirement specifications for datasets used in training and testing of AI models), DEL05_5 *"Data handling"* (which outlines

how data will be handled once they are accepted), DEL05_6 *"Data sharing practices"* (which provides an overview of the existing best practices for sharing health-related data based on distributed and federated environments, including the requirement to enable secure data sharing and addressing issues of data governance), DEL06 *"AI training best practices specification"* (which reviews best practices for proper AI model training and guidelines for model reporting), DEL07*"AI for health evaluation considerations"* (which discusses the validation and evaluation of AI for health models, and considers requirements for a benchmarking platform), DEL07_1 *"AI4H evaluation process description"* (which provides an overview of the state of the art of AI evaluation principles and methods and serves as an initiator for the evaluation process of AI for health), DEL07_2 *"AI technical test specification"* (which specifies how an AI can and should be tested *in silico*), DEL07_3 *"Data and artificial intelligence assessment methods (DAISAM)"* (which provides the reference collection of WG-DAISAM on assessment methods of data and AI quality evaluation), DEL07_4*"Clinical Evaluation of AI for health"* (which outlines the current best practices and outstanding issues related to clinical evaluation of AI models for health), DEL07_5 *"FG-AI4H assessment platform"* (which explores assessment platform options that can be used to evaluate AI for health for the different topic groups), DEL09 *"AI for health applications and platforms"* (which introduces specific considerations of the benchmarking of mobile- and cloud-based AI applications in health), DEL09_1 *"Mobile based AI applications,"* and DEL09_2 *"Cloud-based AI applications"* (which describe specific requirements for the development, testing and benchmarking of mobile- and cloud-based AI applications).

5.1 Subtopic Prediction of psychiatric multimorbidity in a large pediatric sample

The benchmarking of Psychiatry is going to be developed and improved continuously to reflect new features of AI systems or changed requirements for benchmarking. This section outlines all benchmarking versions that have been implemented thus far and the rationale behind them. It serves as an introduction to the subsequent sections, where the actual benchmarking methodology for each version will be described.

Data availability

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.

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.

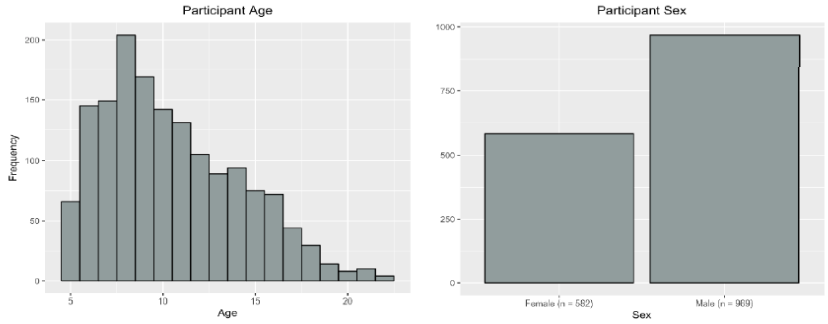


Figure 2: Age and sex distribution of the participants from the HBN data

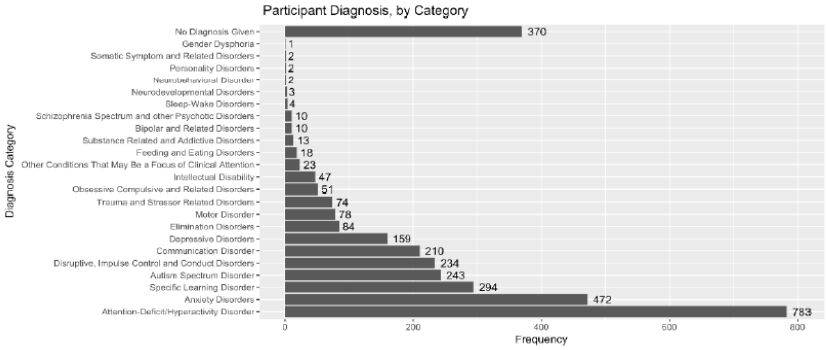


Figure 3: Number of DSM-5 diagnosis by category given to participants from the HBN study

Training and Test Sets

To date, in the ongoing acquisition phase of the HBN effort, data from approximately 1600 subjects have already been recorded and openly released. The raw and preprocessed version of these data as well as the features extracted from these data will also be made publicly available as part of the challenge. Together, these data constitute a huge training set that can be used by all participants to tune their predictive models. It is expected that data of approximately 500 new subjects will become available each year. These new data will serve as test sets and will enable an objective and statistically sound evaluation of all challenge submissions.

Resting State EEG Recording

The EEG data considered in the context of this challenge were acquired during the resting state, in absence of any cognitive task. This state can be viewed as a starting point from which subsequent cognitions are generated and monitored (Langer et al., 2012). Studies have shown that measures of neural activity during the resting state provide valuable biological markers for individual cognitive performance (e.g., Vlahou et al., 2014) as well as for the various mental diseases (for a review see Vecchio et al., 2012). These measures are typically highly test-retest reliable (Näpflin et al.,2007).

Here, a standard procedure of assessing resting state EEG was conducted (c.f. Langer et al., 2012). The EEG data were collected at a sampling rate of 500Hz with a bandpass filter of 0.1Hz to 100Hz. A 128 channel EEG geodesic hydrocel system (EGI) was used for the recordings. Recording reference was set to be at the Electrode Cz. Before starting the recording, each electrode's impedance was examined to ensure that they were below 40 kOhm. During the resting state condition participants had to look at a fixation cross centered on a computer screen. The participants were asked to alternately either have their eyes open or closed. The eyes open (EO) condition lasted 20 seconds while the eyes closed (EC) condition lasted 40 seconds. This sequence of alternating between EO and EC was repeated a total of 5 times. The raw EEG data are available to the participants of the AI challenge (file size: ~500MB/subject).

EEG Preprocessing

The EEG data were preprocessed with the state-of-the-art preprocessing pipeline Automagic (Pedroni, Bahreini, & Langer, 2019). The toolbox facilitates steps like removal of artifacts, interpolating bad channels, and choosing the data with the best quality. One important goal of Automagic is to avoid subjective preprocessing steps like manual bad channel selection and bad independent component selection.

For the preprocessed data in the present challenge a notch filter of 60Hz and a high pass filter of 0.5Hz was used. Bad channels were identified using the PREP pipeline (Bigdely-Shamlo et al. 2015). As a method for the removal of artifacts EOG regression (Parra et al., 2005) and MARA using independent component analysis (Winkler et al., 2011; 2014) was chosen. Bad channels were substituted using a spherical interpolation method.

The preprocessing pipeline also enables objective data quality assessment by categorizing the preprocessed data into either "good", "ok" or "bad". Four different quality criteria can be regulated which have an influence on the quality categorization namely overall high amplitude (OHA), timepoints of high variance (THV), channels of high variance (CVH) and ratio of band channels (RBC). These details are also available to the participants of the challenge. Over 80% of the available data demonstrated "good" or "ok" data quality. The preprocessed EEG data are available to the participants of the AI challenge (file size: ~500MB/subject).

Feature Extraction

We are going to extract various measures from resting state EEG data, which have been related to several psychiatric developmental disorders, including measures derived from the frequency spectrum (including theta-beta ratio and frontal alpha asymmetry, alpha suppression, 1/f slope), the characterization of connectivity between EEG sources (e.g., Nolte et al., 2004; Haufe et al., 2013) as well as the description of the temporal evolution of the EEG signal at the level of the topography (i.e. microstate analysis, Lehmann et al., 1987). These features will be extracted for each sensor or

each pair of sensors (for connectivity measures), respectively. In addition, various solutions to the EEG inverse problem will be computed in order to estimated the underlying activity of the brain from the sensor-level data. These will likely include the popular inverse solutions eLORETA (Pascual-Marqui, 2007) and LCMV (Van Veen et al., 1997). The brain source activity estimated that way will be aggregated across major anatomical structures. All features derived on the sensor level will also be computed at the level of these brain structures and added to the set of features made available to the participants of the AI challenge (file size ~100MB/subject).

Data quality

All the behavioral and cognitive tests were conducted by licensed clinicians. Finally, all test scores from cognitive tests and clinical interviews are double entered into the database by two (different) trained research assistants. Our extensively validated preprocessing pipeline (see section 6) has shown good EEG data quality for over 80% of the subjects. Furthermore, we have demonstrated in a previous publication that the available data are of high quality (Langer et al. 2017). In this publication we have provided a technical validation of a subset of the available resting EEG data. As can be seen from the figure, we obtained the spatial distributions of spectral amplitudes that are expected for resting-state EEG data (Barry et al., 2009, Michel et al. 2009). These spectral measures have been sensitive and successful for describing, for instance, age-related EEG changes or various clinical conditions of developmental disorders (van Dinteren et al., 2008, Magee et al., 2005).

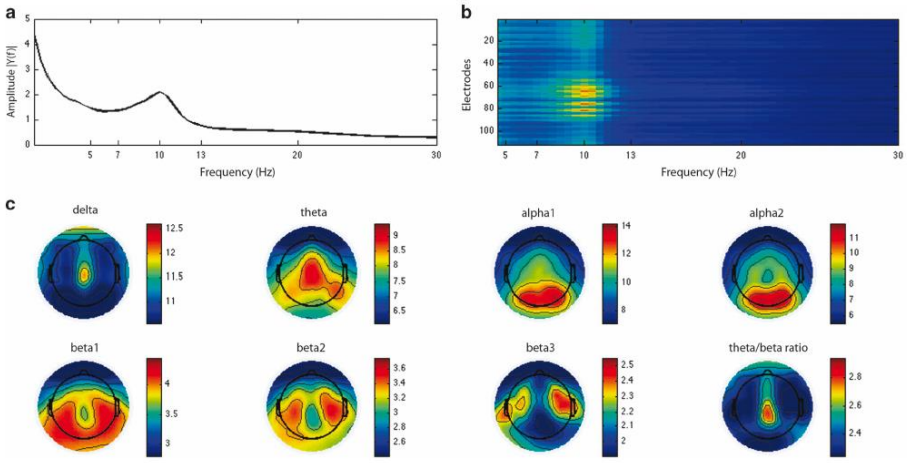


Figure 4: The spectral amplitude

The spectral amplitude was averaged over all subjects and displayed as a mean over all electrodes (a) and for each electrode individually (b). (c) Shows the topographical distribution of the group mean relative power spectra data for the different frequency bands as well as the theta/(beta1+2) ratio

Annotation/label quality

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.

Data provenance

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.

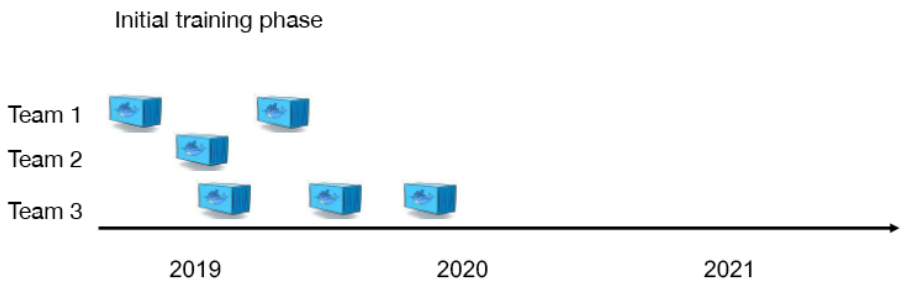
Benchmarking Methodology and Architecture

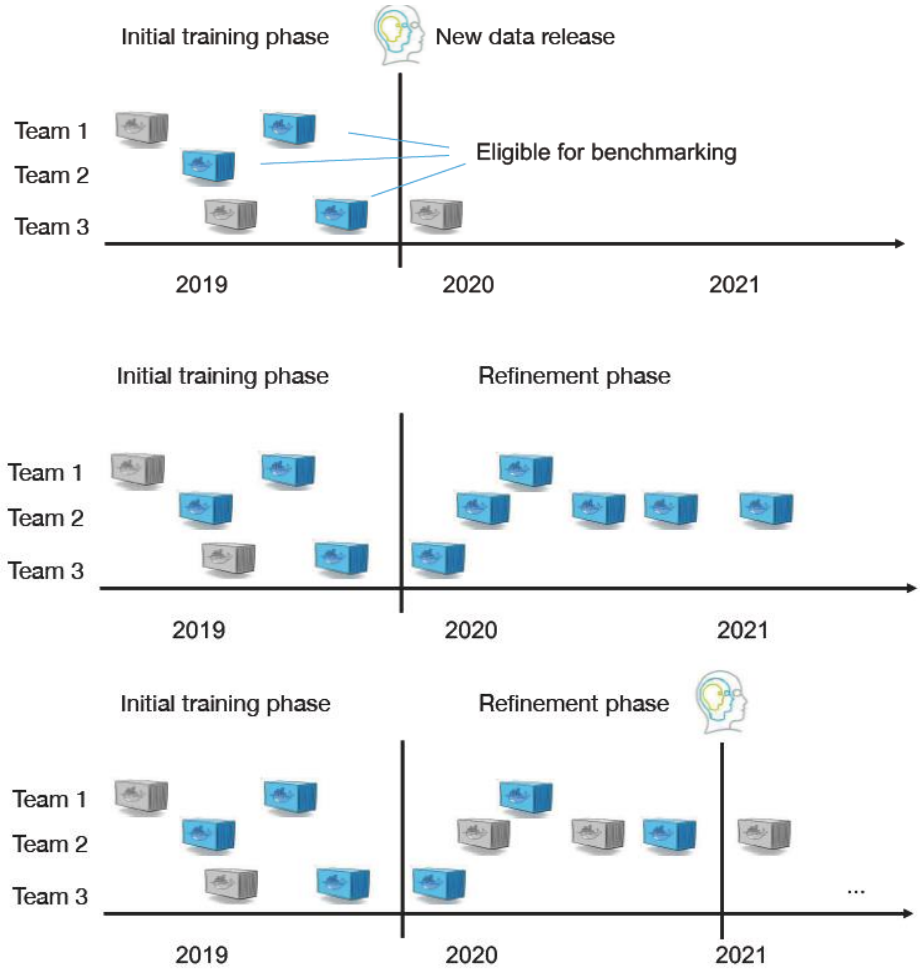
Secret test data (demographic, phenotypic and neurophysiological data and corresponding multi-labels) will be stored at the site of a challenge organizer or a dedicated AI4H server. It is envisioned that challenge participants send executable code for deriving predictions from the demographic, phenotypic 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

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





5.1.1 Benchmarking version

This section includes all technological and operational details of the benchmarking process for the benchmarking version “05_2020_ETHZ”.

5.1.1.1 Overview

This section provides an overview of the key aspects of this benchmarking iteration, version 05_2020_ETHZ.

The overall scope was to perform first benchmarking and investigate feasibility of data access and computational infrastructure.

5.1.1.2 Benchmarking methods

This section provides details about the methods of the benchmarking version05_2020_ETHZ. It contains detailed information about the benchmarking system architecture, the dataflow and the software for the benchmarking process (e.g., test scenarios, data sources, and legalities).

5.1.1.2.1 Benchmarking system architecture

This section covers the architecture of the benchmarking system. For well-known systems, an overview and reference to the manufacturer of the platform is sufficient. If the platform was developed by the topic group, a more detailed description of the system architecture is required.

The details of the benchmarking architecture can be found here:

https://www.dropbox.com/s/vcdng9kazh1mtu7/Final_Report_Greek_Group_Subgroup_Analysis.pdf?dl=0

5.1.1.2.2 Benchmarking system dataflow

This section describes the dataflow throughout the benchmarking architecture.

We have prepared the AI input data for the challenge. In the present challenge, we provide raw and pre-processed 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.

The data and the description of the data are stored here: https://osf.io/ajhgy/.

Data Access

Commented [NL3]: add OSF page

EEG and eye-tracking datasets for participants are currently distributed under the Creative Commons, Attribution Non-Commercial Share Alike License, as they were collected prior to the addition of consent for commercial use to the informed consent. For the high dimensional phenotypic (cognitive, behavioral) data and diagnosis labels the participants of the challenge are required to sign a data usage agreement for access to the data. The intent of the agreement is to ensure that data users agree to protect participant confidentiality when handling data that contains potentially identifying information and that they will agree to take the necessary measures to prevent breaches of privacy. The specific agreement to be employed for the HBN Dataset are those previously defined by the New York State Office of Mental Health and can be found in the Data Usage Agreement (DUA). Institutional review board (IRB) approval is not required for transfer of the data; it will be up to the challenge participants to satisfy any additional requirements specified by their local IRB or ethics committee, prior to using the HBN Dataset. Given that local IRB approval is not required as part of an individual's application for access to the HBN dataset, there is no need for an individual's IRB to have a federal-wise assurance number. Protected behavioral and cognitive assessment data are available on the Collaborative Informatics and Neuroimaging Suite (COINS). The Collaborative Informatics and Neuroimaging Suite (COINS; Scott et al., 2011), developed by the Mind Research Network , was created to facilitate communication and cultivate a data-sharing community by providing researchers with an open source information system that includes web-based tools to manage studies, subjects, imaging, and phenotypic data.

EEG data, organized into folders by participant, may also be accessed through an Amazon Web Services (AWS) S3 bucket . Each file in the S3 bucket can only be accessed using HTTP. If required, the organizers are willing to change the data access procedures.

Data formats

Behavioural data: .csv

Raw and pre-processed EEG data: .mat (MATLAB, http://www.mathworks.com/). We can also provide the data in comma separated values (.csv). (if requested)

AI Output Data Structure

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

Figure 5:AI output data structure (CSV file)

5.1.1.2.3 Safe and secure system operation and hosting

This section addresses security considerations about the storage and hosting of data (benchmarking results and reports) and safety precautions for data manipulation, data leakage, or data loss.

In the case of a manufactured data source (vs. self-generated data), it is possible to refer to the manufacturer's prescriptions.

5.1.1.2.4 Benchmarking process

To be written

This section describes how the benchmarking looks from the registration of participants, through the execution and resolution of conflicts, to the final publication of the results.

- *To be announced*

5.1.1.3 AI input data structure for the benchmarking

This section describes the input data provided to the AI solutions as part of the benchmarking of Psychiatry. It covers the details of the data format and coding at the level of detail needed to submit an AI for benchmarking. This is the only TDD section addressing this topic. Therefore, the description needs to be complete and precise. This section does *not* contain the encoding of the labels for the expected outcomes. It is only about the data the AI system will see as part of the benchmarking.

- How are new benchmarking iterations scheduled (e.g., on demand or quarterly)?
    - continuously
- How do possible participants learn about an upcoming benchmarking?
    - mailing list
- How can one apply for participation?
    - https://methlabuzh.wixsite.com/mysite
- What information and metadata do participants have to provide (e.g., AI autonomy level assignment (IMDRF), certifications, AI/machine learning technology used, company size, company location)?
    - To be discussed
- Are there any contracts or legal documents to be signed?
    - See above: For the high dimensional phenotypic (cognitive, behavioral) data and diagnosis labels the participants of the challenge are required to sign a data usage agreement for access to the data
- Are there inclusion or exclusion criteria to be considered?
    - Exclusion criteria: commercial interest
- How do participants learn about the interface they will implement for the benchmarking (e.g., input and output format specification and application program interface endpoint specification)?
    - https://methlabuzh.wixsite.com/mysite
- How can participants test their interface (e.g., is there a test dataset in case of file-based offline benchmarking or are there tools for dry runs with synthetic data cloud-hosted application program interface endpoints)?
    - *to be announced*
- Who is going to execute the benchmarking and how is it ensured that there are no conflicts of interest?
    - The participants will have the chance to benchmark it themselves
- If there are problems with an AI, how are problems resolved (e.g., are participants informed offline that their AI fails to allow them to update their AI until it works? Or, for online benchmarking, is the benchmarking paused? Are there timeouts?)?
    - IT from AI4H
- How and when will the results be published (e.g., always or anonymized unless there is consent)? With or without seeing the results first? Is there an interactive drill-down tool or a static leader board? Is there a mechanism to only share the results with stakeholders approved by the AI provider as in a credit check scenario?
    - It is planned to have a ranking on our webpage
- In case of online benchmarking, are the benchmarking data published after the benchmarking? Is there a mechanism for collecting feedback or complaints about the data? Is there a mechanism of how the results are updated if an error was found in the benchmarking data?
    - Yes. See answer above

5.1.1.4 AI output data structure

Similar to the input data structure for the benchmarking, this section describes the output data the AI systems are expected to generate in response to the input data. It covers the details of the data format, coding, and error handling at the level of detail needed for an AI to participate in the benchmarking.

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

Figure 6: The AI output data structure

5.1.1.5 Test data label/annotation structure

While the AI systems can only receive the input data described in the previous sections, the benchmarking system needs to know the expected correct answer (sometimes called 'labels') for each element of the input data so that it can compare the expected AI output with the actual one. Since this is only needed for benchmarking, it is encoded separately. The details are described in the following section.

- *To be announced*

5.1.1.6 Scores and metrics

Scores and metrics are at the core of the benchmarking. This section describes the scores and metrics used to measure the performance, robustness, and general characteristics of the submitted AI systems.

5.1.1.7 Test dataset acquisition

Test dataset acquisition includes a detailed description of the test dataset for the AI model and, in particular, its benchmarking procedure including quality control of the dataset, control mechanisms, data sources, and storage.

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.

5.1.1.8 Data sharing policies

This section provides details about legalities in the context of benchmarking. Each dataset that is shared should be protected by special agreements or contracts that cover, for instance, the data sharing period, patient consent, and update procedure (see also DEL05_5 on *data handling* and DEL05_6 on *data sharing practices*).

see section 6.1.1.2.2.

5.1.1.9 Baseline acquisition

The main purpose of benchmarking is to provide stakeholders with the numbers they need to decide whether AI models provide a viable solution for a given health problem in a designated context. To achieve this, the performance of the AI models needs to be compared with available options achieving the same clinically meaningful endpoint. This, in turn, requires data on the performance of the alternatives, ideally using the same benchmarking data. As the current alternatives typically involve doctors, it might make sense to combine the test data acquisition and labelling with additional tasks that allow the performance of the different types of health workers to be assessed.

5.1.1.10 Reporting methodology

This section discusses how the results of the benchmarking runs will be shared with the participants, stakeholders, and general public.

- What is the general approach for reporting results (e.g., leader board vs. drill down)?

  - Leader board
- How can participants analyse their results (e.g., are there tools or are detailed results shared with them)?
  - Performance metrics (ACC)
- How are the participants and their AI models (e.g., versions of model, code, and configuration) identified?
  - username
- What additional metadata describing the AI models have been selected for reporting?
  - n/a
- How is the relationship between AI results, baselines, previous benchmarking iterations, and/or other benchmarking iterations communicated?
  - Leader board
- What is the policy for sharing participant results (e.g., opt in or opt out)? Can participants share their results privately with their clients (e.g., as in a credit check scenario)?
  - n/a
- What is the publication strategy for the results (e.g., website, paper, and conferences)?
  - website
- Is there an online version of the results?
  - yes, website
- Are there feedback channels through which participants can flag technical or medical issues (especially if the benchmarking data was published afterwards)?
  - To be discussed
- Are there any known limitations to the value, expressiveness, or interpretability of the reports?
  - tba

5.1.1.11 Result

This section gives an overview of the results from runs of this benchmarking version of your topic. Even if your topic group prefers an interactive drill-down rather than a leader board, pick some context of common interest to give some examples.

Version 05_2020_ETHZ

A research group from the ETH Zurich has decomposed the multi-class, multi-label problem into several binary classification tasks. They have used the following machine learning approaches: Support Vector Machines (SVM), Logistic Regression (LGR) and Random Forest (RF).

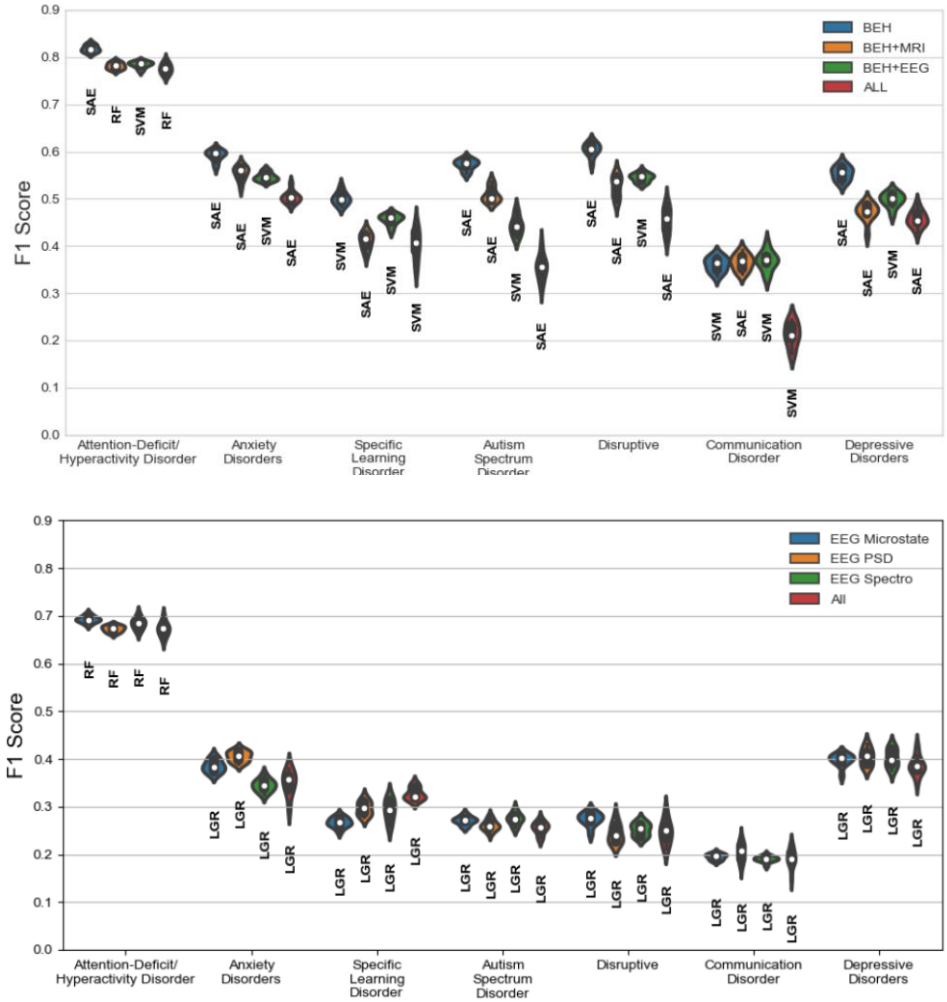


Figure 7: Version 05_2020_ETHZ

5.1.1.12 Discussion of the benchmarking

This section discusses insights of this benchmarking iterations and provides details about the 'outcome' of the benchmarking process (e.g., giving an overview of the benchmark results and process).

One of the greatest limitations in this study is the limited number of patients in the dataset, especially when compared with the number of available features. At the time of this writing, the dataset is in its 6th version and in every new release additional subjects are introduced but also

some inconsistencies among data sources are resolved. For example, there may be subjects who have behavioural records and were adviced by the doctor to undertake an MRI or EEG examination but the results have not yet been released. Hence, it would be interesting to deploy the methods outlined in this paper on a wider population and observe how the predictive performance of our models varies.

It is also important to bear in mind the characteristics of the population involved in this study. Subjects in the dataset are children whose parents or guardians sought for psychiatric consultation. Hence, these subjects do not represent the average healthy population in a community. Generalisation of this study to a wider population would necessitate the inclusion of average healthy subjects in the dataset to avoid the introduction of possible biases in the dataset. Having a completely automated pipeline process would require that the data distribution is properly balanced to reflect the statistics of the general population. This is a matter of data collection which is mainly restricted by time and cost.

An interesting as well as challenging task is predicting the order of diagnoses for each patient. As previously mentioned, each subject is diagnosed with up to ten different disorders and the order of each diagnosis signifies the importance of the corresponding disorder. This is something that was slightly overlooked during this study. A possible way to tackle this task is to treat multiple diagnoses per subject as different time steps and attempt to predict both the order as well as the diagnoses using a recurrent model (e.g. a recurrent neural network). The task resembles that of sequence generation with recurrent neural networks described by Graves in [39]. Input at each time step could be a one-hot encoding representation of the corresponding disorder and the state of the system could be initialised with the latent representation of the subject (inferred by the SAE). The model would be trained using teacher forcing whereas inference would be done by sampling from the output unit at each time step. We narrowly experimented with this model to predict the same set of labels along with their order but the results were (as expected) slightly worse. This task is particularly challenging and prone to subjectivity in doctor diagnosis as different clinicians may give rise to different priority of disorders in the diagnosis. We believe that this architecture could yield more promising results when trained on a larger sample of patients.

Finally, we would like to mention that this is still a work in progress with the potential to publish some of our findings in a relevant venue.

5.1.1.13 Retirement

This section addresses what happens to the AI system and data after the benchmarking activity is completed. It might be desirable to keep the database for traceability and future use. Alternatively, there may be security or privacy reasons for deleting the data. Further details can be found in the reference document of this section DEL04 "*AI software lifecycle specification"* (identification of standards and best practices that are relevant for the AI for health software life cycle).

The models are saved on a github webpage.

5.1.2 Benchmarking version NEW

This section includes all technological and operational details of the benchmarking process for the benchmarking version [X].

6 Overall discussion of the benchmarking

This section discusses the overall insights gained from benchmarking work in this topic group. This should not be confused with the discussion of the results of a concrete benchmarking run (e.g., in 5.1.1.12).

This section will be written, at the end of the project.

7 Regulatory considerations

Commented [NL4]: This is not our area of expertise, nor interest. So, we will not work on this.

For AI-based technologies in healthcare, regulation is not only crucial to ensure the safety of patients and users, but also to accomplish market acceptance of these devices. This is challenging because there is a lack of universally accepted regulatory policies and guidelines for AI-based medical devices. To ensure that the benchmarking procedures and validation principles of FG-AI4H are secure and relevant for regulators and other stakeholders, the working group on *"Regulatory considerations on AI for health" (WG-RC)* compiled the requirements that consider these challenges.

The deliverables with relevance for regulatory considerations are DEL02 *"AI4H regulatory considerations"* (which provides an educational overview of some key regulatory considerations), DEL02_1 *"Mapping of IMDRF essential principles to AI for health software",* and DEL02_2 *"Guidelines for AI based medical device (AI-MD): Regulatory requirements"* (which provides a checklist to understand expectations of regulators, promotes step-by-step implementation of safety and effectiveness of AI-based medical devices, and compensates for the lack of a harmonized standard). DEL04 identifies standards and best practices that are relevant for the "*AI software lifecycle specification."* The following sections discuss how the different regulatory aspects relate to the TG-Psychiatry.

7.1 Existing applicable regulatory frameworks

Most of the AI systems that are part of the FG-AI4H benchmarking process can be classified as *software as medical device* (SaMD) and eligible for a multitude of regulatory frameworks that are already in place. In addition, these AI systems often process sensitive personal health information that is controlled by another set of regulatory frameworks. The following section summarizes the most important aspects that AI manufacturers need to address if they are developing AI systems for Psychiatry.

7.2 Regulatory features to be reported by benchmarking participants

In most countries, benchmarked AI solutions can only be used legally if they comply with the respective regulatory frameworks for the application context. This section outlines the compliance features and certifications that the benchmarking participants need to provide as part of the metadata. It facilitates a screening of the AI benchmarking results for special requirements (e.g., the prediction of prediabetes in a certain subpopulation in a country compliant to the particular regional regulatory requirements).

7.3 Regulatory requirements for the benchmarking systems

The benchmarking system itself needs to comply with regulatory frameworks (e.g., some regulatory frameworks explicitly require that all tools in the quality management are also implemented with a quality management system in place). This section outlines the regulatory requirements for software used for benchmarking in this topic group.

7.4 Regulatory approach for the topic group

Building on the outlined regulatory requirements, this section describes how the topic group plans to address the relevant points in order to be compliant. The discussion here focuses on the guidance and best practice provided by the DEL02 *"AI4H regulatory considerations."*

8 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 Pre-processing 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 Pre-processing 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. *Behavioural and Brain Functions*, *7*(1), 30.

Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., … Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, *17*(1), 67-75. doi:10.1002/wps.20491

Deshpande, G., Wang, P., Rangaprakash, D., & Wilamowski, B. (2015). Fully Connected Cascade Artificial Neural Network Architecture for Attention Deficit Hyperactivity Disorder Classification From Functional Magnetic Resonance Imaging Data. *IEEE Transactions on Cybernetics*, *45*(12), 2668-2679. doi:10.1109/tcyb.2014.2379621

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering Cognitive Behaviour Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health*, *4*(2), e19. doi:10.2196/mental.7785

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254-280. doi:10.1016/j.techfore.2016.08.019

Hoffmann, D. E., & Tarzian, A. J. (2003). The Girl Who Cried Pain: A Bias Against Women in the Treatment of Pain. *SSRN Electronic Journal*. doi:10.2139/ssrn.383803

Just, M. A., Pan, L., Cherkassky, V. L., McMakin, D. L., Cha, C., Nock, M. K., & Brent, D. (2017). Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nature Human Behaviour*, *1*(12), 911-919. doi:10.1038/s41562-017-0234-y

Lovejoy, C. A., Buch, V., & Maruthappu, M. (2019). Technology and mental health: The role of artificial intelligence. *European Psychiatry*, *55*, 1-3. doi:10.1016/j.eurpsy.2018.08.004

Meredith, S. (2018, March 23). Here's everything you need to know about the Cambridge Analytica scandal [Web log post]. Retrieved from https://www.cnbc.com/2018/03/21/facebook-cambridge-analytica-scandal-everything-you-need-to-know.html

Meyer-Lindenberg, A. (2018). Künstliche Intelligenz in der Psychiatrie – ein Überblick. *Der Nervenarzt*, *89*(8), 861-868. doi:10.1007/s00115-018-0557-6

Neighborhood Psychiatry. (2018, February 13). Can artificial intelligence improve psychiatric diagnosis? Retrieved from https://www.psychologytoday.com/intl/blog/psychiatry-the-people/201802/can-artificial-intelligence-improve-psychiatric-diagnosis

Place, S., Blanch-Hartigan, D., Rubin, C., Gorrostieta, C., Mead, C., Kane, J., … Azarbayejani, A. (2017). Behavioural Indicators on a Mobile Sensing Platform Predict Clinically Validated Psychiatric Symptoms of Mood and Anxiety Disorders. *Journal of Medical Internet Research, 19*(3), e75. doi:10.2196/jmir.6678

Reece, A. G., & Danforth, C. M. (2017). Erratum to: Instagram photos reveal predictive markers of depression. *EPJ Data Science*, *6*(1). doi:10.1140/epjds/s13688-017-0118-4

Rutledge, R. B., Chekroud, A. M., & Huys, Q. J. (2019). Machine learning and big data in psychiatry: toward clinical applications. *Current Opinion in Neurobiology*, *55*, 152-159. doi:10.1016/j.conb.2019.02.006

Staff, W. (2019, March 17). How Cambridge Analytica sparked the great privacy awakening [Web log post]. Retrieved from https://www.wired.com/story/cambridge-analytica-facebook-privacy-awakening/

Stark, H. (2017, September/October 30). Artificial intelligence is here and it wants to revolutionize psychiatry. *Forbes*

Torous, J. (2014). Mobile technology and global mental health. *Asian Journal of Psychiatry*, *10*, 69-70. doi:10.1016/j.ajp.2013.07.004

Torous, J., Onnela, J., & Keshavan, M. (2017). New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices. *Translational Psychiatry*, *7*(3), e1053-e1053. doi:10.1038/tp.2017.25

Vieira, S., Pinaya, W. H., & Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioural Reviews*, *74*, 58-75. doi:10.1016/j.neubiorev.2017.01.002

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting Risk of Suicide Attempts Over Time Through Machine Learning. *Clinical Psychological Science*, *5*(3), 457-469. doi:10.1177/2167702617691560

Wang, Y., & Kosinski, M. (2017). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. doi:10.31234/osf.io/hv28a

Wardenaar, K. J., & De Jonge, P. (2013). Diagnostic heterogeneity in psychiatry: towards an empirical solution. *BMC Medicine, 11*(1). doi:10.1186/1741-7015-11-201

World Health Organization. (n.d.). Gender and women's mental health. Retrieved from https://www.who.int/mental_health/prevention/genderwomen/en/

Zarley, D. (2019, January 28). Meet the scientists who are training AI to diagnose mental illness [Web log post]. Retrieved from https://www.theverge.com/2019/1/28/18197253/ai-mental-illness-artificial-intelligence-science-neuroimaging-mri

Annex A:
Glossary

This section lists all the relevant abbreviations, acronyms and uncommon terms used in the document.

Acronym/Term	Expansion	Comment
TDD	Topic Description Document	Document specifying the standardized benchmarking for a topic on which the FG AI4H Topic Group works. This document is the TDD for the Topic Group [YOUR TOPIC GROUP]
TG	Topic Group	
WG	Working Group	
FGAI4H	Focus Group on AI for Health	
AI	Artificial intelligence	
ITU	International Telecommunication Union	
WHO	World Health Organization	
DEL	Deliverable	
CfTGP	Call for topic group participation	
AI4H	Artificial intelligence for health	
IMDRF	International Medical Device Regulators Forum	
MDR	Medical Device Regulation	
ISO	International Standardization Organization	
GDPR	General Data Protection Regulation	
FDA	Food and Drug administration	
SaMD	Software as a medical device	
AI-MD	AI based medical device	
LMIC	Low-and middle-income countries	
GDP	Gross domestic product	
API	Application programming interface	
IP	Intellectual property	
PII	Personal identifiable information	
[…]		

Annex B:
Declaration of conflict of interests

Company/Institution/Individual XYZ

_____