European Child & Adolescent Psychiatry (2021) 30:1143–1146
https://doi.org/10.1007/s00787-021-01836-0

EDITORIAL

Taming the chaos?! Using eXplainable Artificial Intelligence (XAI) to tackle the complexity in mental health research

Veit Roessner[1] · Josefine Rothe[1] · Gregor Kohls[1] · Georg Schomerus[2] · Stefan Ehrlich[1,3] · Christian Beste[1]

Published online: 8 July 2021

Mental disorders cause a significant degree of burden to affected individuals and to society at large. Reasons for this are their high prevalence (one in every two people suffers from a mental disorder at some point in their lifetime), their usually early onset (three in four patients fall ill before the age of 23), and-particularly if left untreated-their mostly chronic course, precipitating numerous disease-related disabilities and poor health outcomes [1–3]. In addition, a substantial percentage of non-responders and non-compliant patients exists. Notably, particularly for youth under the age of 18, access to diagnosis, prevention/intervention, and care for mental health problems is still relatively limited compared to common somatic issues [4]. In the following, we will explain the reason for this discrepancy and provide a possible solution.

In general, the combination of clinical and translational science within medicine has steadily increased over the years, and this has led to tremendous progress also in mental healthcare [5, 6]. Nevertheless, mental disorders are very heterogeneous, dynamic, and multi-causal phenomena. Despite the widespread recognition of their inherent complex nature (including gene-environment and psyche-soma interactions as well as developmental and other experience-based changes over the life span), progress in understanding mental disorders as multifaceted bio-psycho-social conditions remains rather slow. In addition, even if acknowledged as such, our knowledge about etiopathophysiology, diagnosis, and management of mental disorders is still incomplete. The reasons that we still lack a more comprehensive picture of mental disorders are manifold, including the dichotomy of hypothesis-driven versus exploratory data-driven research methods and resulting findings, which all have their own pros and cons [7].

Furthermore, there is still an ongoing discussion to what extent the classification systems, such as DSM or ICD, are valid for diagnosing mental disorders [8]. For example, despite great research efforts clinically usable biomarkers that could potentially improve the early identification of mental disorders or that could be utilized for early intervention strategies (e.g., as predictors for treatment response) are still lacking. Among other reasons, the nosological classification systems primarily define mental disorders categorically according to a set of core symptoms, thereby neglecting the substantial dimensional, multifactorial, and heterogeneous clinical presentation and emergence of mental disorders as well as their symptomatic overlap. Consequently, dimensional transdiagnostic approaches have been introduced into the research arena, including the Research Domain Criteria (RDoC) project [9], but these approaches are still in their infancy, and they are no less hotly debated than the "traditional" ones [10].

This increase in complexity is accompanied by substantial technological and methodological advances in the areas of (epi)genetics, neuroimaging, psychophysiology, and others. However, these highly promising new leads that each attempts to identify the (neuro)biological correlates of mental disorders-or specific valid disease subtypes-have yet failed to convincingly resolve the issue of within-and across-disorder heterogeneity as they often lack specificity [8]. For instance, despite thorough research in the area of neuroimaging, with the exception of a few and relatively rare conditions, we do not have a single indicator available grounded in brain biology that can reliably distinguish patients with a specific mental disorder from typical controls, let alone

✉ Veit Roessner
veit.roessner@uniklinikum-dresden.de

[1] Department of Child and Adolescent Psychiatry, Faculty of Medicine, TU Dresden, Fetscherstrasse 74, 01307 Dresden, Germany

[2] Department of Psychiatry, Leipzig University Medical Center, Semmelweisstr. 10, 04103 Leipzig, Germany

[3] Division of Psychological and Social Medicine and Developmental Neurosciences, Faculty of Medicine, TU Dresden, Fetscherstrasse 74, 01307 Dresden, Germany

differentiate between different mental disorders or their subtypes.

Technical innovations have also led to an increased quantity and diversity of data that can be measured to help disentangling the complex nature of mental disorders. In addition to "classic" data sets from questionnaires, performance tests, and clinical interviews, three sources of data are particularly crucial in modern mental health research [11]: (1) social media data (e.g., content and color analytics of social network usage); (2) facility data (e.g., electronic health records from different digital health information systems, but also data from animal models, or genetics); and (3) sensory data (e.g., real-time monitoring of human physiological measures, such as glucose, or heart rate; see Fig. 1).

Nevertheless, collecting these diverse sets of data alone—even if they are high in quality, quantity, and ideally in validity, will not substantially improve our understanding of mental disorders as multifaceted bio-psycho-social conditions. To achieve such aim sophisticated and meaningful analytic approaches are required, including, amongst others, Machine Learning (ML)/Artificial Intelligence (AI).

The unique potential of ML/AI for characterizing complex data structures has already been successfully demonstrated in the context of neuroscientific research [12]. For example, the identification of relevant information from EEG patterns for neural classification of brain functionality has been significantly improved by ML techniques [13]. Within child and adolescent psychiatry, ML has been used, for instance, to predict the risk of psychosis [14, 15]. However, even though complex classification patterns can be identified using such exploratory data-driven ML/AI approaches, these patterns are only partially useful, because often researchers do not obtain knowledge of their internal workings (concept of a "black box" in ML/AI), and therefore, the meaning and relevance of the results can rarely be explained (but there are actually several attempts to solve this problem, e.g., [16]). This might be the main reason why the use of these approaches in the field of mental disorders (i.e., computational psychiatry) has not yet been more fruitful although their potential is obvious [7].

By contrast, eXplainable AI (XAI) methods follow the three principles transparency, interpretability and explainability. Thereby, it is possible to examine which feature(s) in the data set contribute(s) most to a specific classification pattern. One such approach is to use saliency maps [17], which are designed to visualize the relative weight or importance of feature(s) in the data that are intuitively fed into deep learning algorithms (e.g., allotting values between zero and one denoting the importance of a feature). Although XAI approaches per se cannot provide causal mechanistic insights into how the brain accomplishes a particular function or complex behaviors [7], both experimental studies on mechanisms (i.e., cognitive, affective, and neural mechanistic models of mental health) combined with results of XAI can be related to multi-modal data (e.g., EEG [17], neuroimaging, clinical, and environmental data [18]). By harvesting predictive inter-relationships among different data types, these approaches are able to outperform unimodal data models in terms of classification accuracy. This offers new opportunities, for instance, for diagnostic purposes in the realm of mental disorders in children and beyond. In particular, using XAI with everyday social media, (health) facility, and sensory data (see above) would clearly advance our understanding of the mechanisms from mental health to disorders that will help predicting risk and disease trajectories which then would allow developing personalized and scalable detection and prevention/intervention tools (e.g., eHealth and mHealth). Consequently, this would move the field forward to a transdisciplinary, integrative, context-sensitive and person-centered healthcare model [5].

However, to develop such mental healthcare model(s) even further the entire translational continuum spanning from basic science discovery, early human studies, clinical trials, implementation, evaluation, and optimization in practice and communities is required [6, 19]. For example, in attention deficit hyperactivity disorder (ADHD), basic cognitive neuroscience research, including results of EEG studies, is on its way to provide more scalable interventions, such as optimized neurofeedback training, addressing the gap between basic biomedical research and mental healthcare. Moreover, the application of neurofeedback@home brings healthcare innovations deeper into the community, including both urban and rural areas.

In contrast to the "traditional" professionally driven, more one-directional translational continuum, where scientists develop diagnostic and intervention tools for clinicians, who in turn use them in their patients, several continuous feedback loops as part of a living lab approach are needed [20]. Such feedback loops help to constantly re-adjust ongoing research priorities. The concept of living labs describes a user-centered environment for open innovation that combines multiple methods with participatory research in a real-world setting. Research particularly on child and adolescent mental health has to be closely connected to the wider community it serves to implement easy-to-access help for target groups both in urban and rural contexts, support self-management and help young people build and utilize local support networks. In this context utilizing locally established partnerships and networks as well as performing an adaptive research process with constant stakeholder input and feedback is pivotal to both the protection and the improvement of the mental health of children and adolescents in a changing social environment, particularly during critical developmental phases, including the transition from childhood to adolescence into adulthood. This altogether will help to establish models of close collaboration between academia,

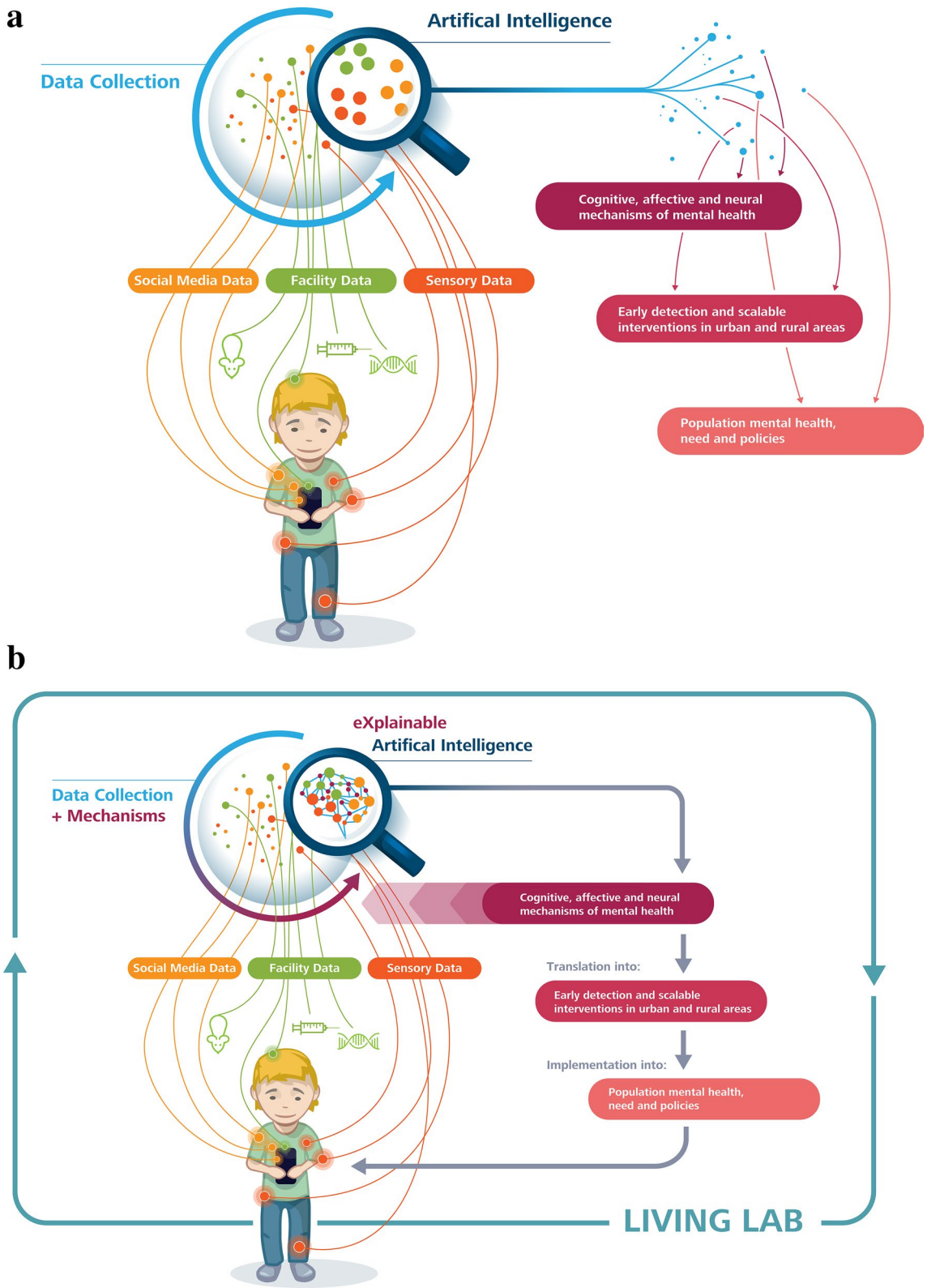

Fig. 1 **a** Existing mental healthcare and prevention approaches using ML/AI often suffer from their lack of explainability, leading to single pieces of a puzzle, but not a meaningful picture. **b** Future holistic mental healthcare and prevention approaches (see text for details): should be based on data collection using social media data, (health) facility data, and real-time monitoring of human sensory data-analyzed by using eXplainable Artificial Intelligence (XAI)-that is guided by evidence about underlying mechanisms; and this will help to develop, implement, evaluate and optimize scalable healthcare and prevention/intervention approaches, including "classic" Health as well as eHealth and mHealth, as part of a living lab approach

regional and (inter)national partners to deliver cutting-edge research, innovative clinical services, evidence-based training, and policy development that will ensure continuous improvement in the access to diagnosis, prevention/intervention and care provided to young people suffering from impairing mental disorders.

Funding Open Access funding enabled and organized by ProjektDEAL.

References

1. Whiteford HA, Ferrari AJ, Degenhardt L, Feigin V, Vos T (2015) The global burden of mental, neurological and substance use disorders: an analysis from the global burden of disease study 2010. PLoS ONE 10:e0116820. https://doi.org/10.1371/journal.pone.0116820
2. Committee on Psychosocial Aspects of Child and Family Health (2001) American Academy of Pediatrics. The new morbidity revisited: a renewed commitment to the psychosocial aspects of pediatric care. Committee on Psychosocial Aspects of Child and Family Health. Pediatrics 108:1227–30. https://doi.org/10.1542/peds.108.5.1227.
3. Murphy M, Fonagy P. Chief Medical Officer annual report 2012: children and young people's health. GOVUK 2013. https://www.gov.uk/government/publications/chief-medical-officers-annual-report-2012-our-children-deserve-better-prevention-pays. Accessed 7 June 2021
4. Saunders NR, Gandhi S, Chen S, Vigod S, Fung K, De Souza C et al (2020) Health care use and costs of children, adolescents, and young adults with somatic symptom and related disorders. JAMA Netw Open 3:e2011295–e2011295. https://doi.org/10.1001/jamanetworkopen.2020.11295
5. Waldman SA, Terzic A (2010) Clinical and translational science: from bench-bedside to global village. Clin Transl Sci 3:254–257. https://doi.org/10.1111/j.1752-8062.2010.00227.x
6. Hegyi P, Petersen OH, Holgate S, Erőss B, Garami A, Szakács Z, et al. Academia Europaea position paper on translational medicine: the cycle model for translating scientific results into community benefits. J Clin Med 2020;9. https://doi.org/10.3390/jcm9051532.
7. Fellous J-M, Sapiro G, Rossi A, Mayberg H, Ferrante M. Explainable artificial intelligence for neuroscience: behavioral neurostimulation. Front Neurosci 2019;13. https://doi.org/10.3389/fnins.2019.01346.
8. Owen MJ (2014) New approaches to psychiatric diagnostic classification. Neuron 84:564–571. https://doi.org/10.1016/j.neuron.2014.10.028
9. Weine SM, Langenecker S, Arenliu A (2018) Global mental health and the National Institute of Mental Health Research Domain Criteria. Int J Soc Psychiatry 64:436–442. https://doi.org/10.1177/0020764018778704
10. Katahira K, Yamashita Y (2017) A theoretical framework for evaluating psychiatric research strategies. Comput Psychiatry Camb Mass 1:184–207. https://doi.org/10.1162/CPSY_a_00008
11. Liang Y, Zheng X, Zeng DD (2019) A survey on big data-driven digital phenotyping of mental health. Inf Fusion 52:290–307. https://doi.org/10.1016/j.inffus.2019.04.001
12. Vu M-AT, Adalı T, Ba D, Buzsáki G, Carlson D, Heller K et al (2018) A shared vision for machine learning in neuroscience. J Neurosci Off J Soc Neurosci 38:1601–1607. https://doi.org/10.1523/JNEUROSCI.0508-17.2018
13. Craik A, He Y, Contreras-Vidal JL (2019) Deep learning for electroencephalogram (EEG) classification tasks: a review. J Neural Eng 16:031001. https://doi.org/10.1088/1741-2552/ab0ab5
14. Pina-Camacho L, Garcia-Prieto J, Parellada M, Castro-Fornieles J, Gonzalez-Pinto AM, Bombin I et al (2015) Predictors of schizophrenia spectrum disorders in early-onset first episodes of psychosis: a support vector machine model. Eur Child Adolesc Psychiatry 24:427–440. https://doi.org/10.1007/s00787-014-0593-0
15. Bourgin J, Duchesnay E, Magaud E, Gaillard R, Kazes M, Krebs M-O (2020) Predicting the individual risk of psychosis conversion in at-risk mental state (ARMS): a multivariate model reveals the influence of nonpsychotic prodromal symptoms. Eur Child Adolesc Psychiatry 29:1525–1535. https://doi.org/10.1007/s00787-019-01461-y
16. Haufe S, Meinecke F, Görgen K, Dähne S, Haynes J-D, Blankertz B et al (2014) On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87:96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067
17. Vahid A, Bluschke A, Roessner V, Stober S, Beste C. Deep Learning Based on Event-Related EEG Differentiates Children with ADHD from Healthy Controls. J Clin Med 2019;8. https://doi.org/10.3390/jcm8071055.
18. Durstewitz D, Koppe G, Meyer-Lindenberg A (2019) Deep neural networks in psychiatry. Mol Psychiatry 24:1583–1598. https://doi.org/10.1038/s41380-019-0365-9
19. Thornicroft G (2011) Completing the unfinished revolution in mental health. BMJ 343:d7490. https://doi.org/10.1136/bmj.d7490
20. Bergvall-Kareborn B, Stahlbrost A (2009) Living lab: an open and citizen-centric approach for innovation. Int J Innov Reg Dev 1:356–370. https://doi.org/10.1504/IJIRD.2009.022727