

Predicting functional outcome using machine learning and large language models: A transdiagnostic approach in a young, help-seeking population

Pavol Mikolas, Fabian Huth, Kyra Bröckel-Bundt, Christina Berndt, Julia Martini, Birgit Maicher, Michael Marxen, Falk Gerrik Verhees, Paula Marie Henneberg, Christoph Vogelbacher, Andreas Jansen, Tilo Kircher, Irina Falkenberg, Florian Thomas-Odenthal, Martin Lambert, Vivien Kraft, Gregor Leicht, Christoph Mulert, Andreas J. Fallgatter, Thomas Ethofer, Anne Rau, Karolina Leopold, Andreas Bechdorf, Reif Andreas, Silke Matura, Jonathan Repple, Felix Bermpohl, Jana Fiebig, Thomas Stamm, Christoph U. Correll, Georg Juckel, Vera Flasbeck, Isabella C. Wiest, Jakob N. Kather, Udo Dannlowski, Eva Mennigen, Philipp Ritter, Michael Bauer, Andrea Pfennig

Submitted to: JMIR Mental Health
on: September 26, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 37

 Figures 38

 Figure 1..... 39

 Figure 2..... 40

 Figure 3..... 41

 Multimedia Appendixes 42

 Multimedia Appendix 1..... 43

Predicting functional outcome using machine learning and large language models: A transdiagnostic approach in a young, help-seeking population

Pavol Mikolas¹ MD, PhD; Fabian Huth¹ MSc; Kyra Bröckel-Bundt¹ PhD; Christina Berndt¹ MSc; Julia Martini¹ Prof Dr; Birgit Maicher¹ MSc; Michael Marxen¹ PhD; Falk Gerrik Verhees¹ Dr med, MPH; Paula Marie Henneberg¹ MSc; Christoph Vogelbacher² PhD; Andreas Jansen² Prof Dr; Tilo Kircher² Prof Dr; Irina Falkenberg² Prof Dr; Florian Thomas-Odenthal² PhD; Martin Lambert³ Prof Dr; Vivien Kraft³ MSc; Gregor Leicht³ PhD; Christoph Mülert⁴ Prof Dr; Andreas J. Fallgatter⁵ Prof Dr; Thomas Ethofer⁵ Prof Dr; Anne Rau⁵ PhD; Karolina Leopold⁶ PhD; Andreas Bechdolf⁶ Prof Dr; Reif Andreas⁷ Prof Dr; Silke Matura⁷ PhD; Jonathan Repple⁷ Prof Dr; Felix BERPohl⁸ Prof Dr; Jana Fiebig⁸; Thomas Stamm⁸ Prof Dr; Christoph U. Correll⁹ Prof Dr; Georg Juckel¹⁰ Prof Dr; Vera Flasbeck¹⁰ PhD; Isabella C. Wiest¹¹; Jakob N. Kather¹¹ Prof Dr; Udo Dannowski¹² Prof Dr; Eva Mennigen¹ PhD; Philipp Ritter¹ PhD; Michael Bauer¹ Prof Dr; Andrea Pfennig¹ Prof Dr

¹Department of Psychiatry and Psychotherapy Carl Gustav Carus University Hospital TUD Dresden University of Technology Dresden DE

²Department of Psychiatry University of Marburg Marburg DE

³Department of Psychiatry and Psychotherapy University Medical Center Hamburg-Eppendorf Hamburg DE

⁴Centre for Psychiatry Justus-Liebig University Giessen Giessen DE

⁵Tuebingen Center for Mental Health Department of Psychiatry University of Tuebingen Tuebingen DE

⁶Vivantes Hospital Am Urban and Vivantes Hospital Im Friedrichshain Department of Psychiatry, Psychotherapy and Psychosomatic Medicine Charité-Universitätsmedizin Berlin Berlin DE

⁷Department of Psychiatry, Psychosomatic Medicine and Psychotherapy University Hospital Goethe University Frankfurt Frankfurt DE

⁸Charité Campus Mitte Department of Psychiatry and Psychotherapy Charité University Medicine Berlin DE

⁹Department of Child and Adolescent Psychiatry Charité Universitätsmedizin Berlin Berlin DE

¹⁰Department of Psychiatry, Psychotherapy and Preventive Medicine LWL University Hospital, Ruhr-University Bochum DE

¹¹Else Kroener Fresenius Center for Digital Health TUD Dresden University of Technology Dresden DE

¹²Institute for Translational Psychiatry University of Münster Münster DE

Corresponding Author:

Pavol Mikolas MD, PhD

¹Department of Psychiatry and Psychotherapy

Carl Gustav Carus University Hospital

TUD Dresden University of Technology

Fetchenstr. 74

Dresden

DE

Abstract

Background: Functional impairments associated with mental health conditions are on the rise. Predicting functional outcomes may improve the targeting of preventive interventions. While prognostic models have primarily focused on psychosis, early recognition services require a transdiagnostic approach.

Objective: Predict global functioning within 2-year follow-up using baseline clinical and structural MRI data in a population-based sample of young, help-seeking individuals presenting with affective and anxiety symptoms as well as ADHD.

Methods: We classified 357 help-seeking individuals aged 18-35 years recruited from 9 sites as “impaired” (Global Assessment of Functioning (GAF) ≤ 60; N=228) or “non-impaired” (GAF > 60; N=129) at year one and/or year two follow-up. GAF classification group status at follow up was predicted using linear support vector machine (SVM), decision tree and large language model (LLM) Llama-3 using clinical assessments and/or structural MRI. Leave-one-site-out (SVM) or external sample (LLM) were used for validation.

Results: SVM achieved balanced accuracy 69.2% using clinical features only. Items related to baseline occupational functioning, interpersonal relationships, cognitive functioning, psychotic and affective symptoms as well as presence of anxiety

disorder were most predictive. The decision tree further reduced the feature set to 5 predictive items achieving balanced accuracy 76.6%. Although amygdala and hippocampal subregions achieved balanced accuracy 57.1%, structural MRI did not improve the overall prediction. Llama-3 performed comparably well to SVM, (balanced accuracy 72.6%).

Conclusions: Machine learning demonstrated good performance in predicting global functioning. Interestingly, the out-of-the-box LLM performed comparably well without being trained or fine-tuned, highlighting the potential of leveraging free-text data for mental health prognosis.

(JMIR Preprints 26/09/2025:84424)

DOI: <https://doi.org/10.2196/preprints.84424>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <https://preprints.jmir.org/preprint/84424>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://preprints.jmir.org/preprint/84424>

Original Manuscript

Predicting functional outcome using machine learning and large language models: A transdiagnostic approach in a young, help-seeking population

Pavol Mikolas, PhD¹, Fabian Huth, MSc¹, Kyra Bröckel-Bundt, PhD¹, Christina Berndt, MSc¹, Prof. Julia Martini¹, Birgit Maicher, M.Sc.¹, Michael Marxen, PhD¹, Falk Gerrik Verhees¹, Paula M. Henneberg¹, MSc, Christoph Vogelbacher, PhD^{2,3,4}, Prof. Andreas Jansen^{2,3,4}, Prof. Tilo Kircher^{2,3,4}, Prof. Irina Falkenberg^{2,3,4}, Florian Thomas-Odenthal, PhD^{2,3,4}, Prof. Martin Lambert⁵, Vivien Kraft, MSc⁵, Gregor Leicht, PhD⁵, Prof. Christoph Mulert^{4,5,6}, Prof. Andreas J. Fallgatter^{7,8}, Prof. Thomas Ethofer^{7,8,9}, Anne Rau, PhD⁷, Karolina Leopold, PhD¹⁰, Prof. Andreas Bechdolf¹⁰, Prof. Andreas Reif¹¹, Silke Matura, PhD¹¹, Prof. Jonathan Repple^{11,12,13}, Prof. Felix Bermppohl¹⁴, Jana Fiebig¹⁴, Prof. Thomas Stamm^{14,15}, Prof. Christoph U. Correll¹⁶, Prof. Georg Juckel¹⁷, Vera Flasbeck, PhD¹⁷, Isabella C. Wiest, MD¹⁸, Prof. Jakob N. Kather¹⁸, Prof. Udo Dannlowski¹³, Eva Mennigen, PhD¹, Philipp Ritter, PhD¹, Prof. Michael Bauer¹, Prof. Andrea Pfennig¹

¹Department of Psychiatry and Psychotherapy, Carl Gustav Carus University Hospital, TUD Dresden University of Technology, Germany

²Core-Facility Brainimaging, Faculty of Medicine, University of Marburg, Germany

³Department of Psychiatry, University of Marburg, Germany

⁴Center for Mind, Brain and Behavior (CMBB), University of Marburg and Justus Liebig University Giessen, Germany

⁵Department of Psychiatry and Psychotherapy, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

⁶Centre for Psychiatry, Justus-Liebig University Giessen, Germany

⁷Department of Psychiatry, Tuebingen Center for Mental Health, University of Tuebingen, Tuebingen, Germany

⁸German Center for Mental Health (DZPG), partner site Tübingen, Germany

⁹Biomedical Magnetic Resonance, University of Tübingen

¹⁰Department of Psychiatry, Psychotherapy and Psychosomatic Medicine, Vivantes Hospital Am Urban and Vivantes Hospital Im Friedrichshain, Charité-Universitätsmedizin Berlin, Germany

¹¹ Goethe University Frankfurt, University Hospital, Department of Psychiatry, Psychosomatic Medicine and

Psychotherapy, Frankfurt, Germany

¹² Goethe University Frankfurt, Cooperative Brain Imaging Center - CoBIC, Frankfurt, Germany

¹³ Institute for Translational Psychiatry, University of Münster, Münster, Germany

¹⁴ Department of Psychiatry and Psychotherapy, Charité Campus Mitte, Charité University Medicine, Berlin, Germany

¹⁵ Department of Clinical Psychiatry and Psychotherapy, Brandenburg Medical School Theodor Fontane, Neuruppin, Germany

¹⁶ Department of Child and Adolescent Psychiatry, Charité Universitätsmedizin Berlin, Berlin, Germany; Department of Psychiatry, Northwell Health, The Zucker Hillside Hospital, Glen Oaks, NY, USA; Department of Psychiatry and Molecular Medicine, Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY, USA

¹⁷ Department of Psychiatry, Psychotherapy and Preventive Medicine, LWL University Hospital, Ruhr-University, Bochum, Germany

¹⁸ Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany

Corresponding author:

Pavol Mikolas, MD, PhD

pavol.mikolas@uniklinikum-dresden.de

Department of Psychiatry and Psychotherapy

Carl Gustav Carus, University Hospital

Technische Universität Dresden

Fetscherstraße 74

01307 Dresden

Germany

Tel. 0351 458-2662, Fax 0351 458-5397

Key words: Machine Learning, Neuroimaging, Magnetic Resonance Imaging, Longitudinal Studies,

Natural Language Processing, Prognosis, Young Adult, Mental Disorders

Abstract

Background

Functional impairments associated with mental health conditions are on the rise. Predicting functional outcomes may improve the targeting of preventive interventions. While prognostic models have primarily focused on psychosis, early recognition services require a transdiagnostic approach.

Objectives

Predict global functioning within 2-year follow-up using baseline clinical and structural MRI data in a population-based sample of young, help-seeking individuals presenting with affective and anxiety symptoms as well as ADHD.

Methods

We classified 357 help-seeking individuals aged 18-35 years recruited from 9 sites as “impaired”

(Global Assessment of Functioning (GAF) ≤ 60 ; N=228) or “non-impaired” (GAF > 60 ; N=129) at year one and/or year two follow-up. GAF classification group status at follow up was predicted using linear support vector machine (SVM), decision tree and large language model (LLM) Llama-3 using clinical assessments and/or structural MRI. Leave-one-site-out (SVM) or external sample (LLM) were used for validation.

Results

SVM achieved balanced accuracy 69.2% using clinical features only. Items related to baseline occupational functioning, interpersonal relationships, cognitive functioning, psychotic and affective symptoms as well as presence of anxiety disorder were most predictive. The decision tree further reduced the feature set to 5 predictive items achieving balanced accuracy 76.6%. Although amygdala and hippocampal subregions achieved balanced accuracy 57.1%, structural MRI did not improve the overall prediction. Llama-3 performed comparably well to SVM, (balanced accuracy 72.6%).

Conclusions

Machine learning demonstrated good performance in predicting global functioning. Interestingly, the out-of-the-box LLM performed comparably well without being trained or fine-tuned, highlighting the potential of leveraging free-text data for mental health prognosis.

Introduction

Functional impairments due to mental health conditions have been on the rise. For example, in Germany, the proportion of early retirements due to mental disorders increased from 24% to 42% between 2000 and 2022 [1]. According to the WHO, mental disorders accounted for 16% of disability-adjusted life years (totaling 418 million DALYs) worldwide in 2019 [2]. Improving early risk stratification and predicting functional outcomes may enhance the targeting of preventive measures, ultimately mitigating the long-term negative impact. Machine learning methods and large multi-center datasets have facilitated the development and validation of predictive models that incorporate not only clinical variables but also biological data, such as MRI. Recently, large language models (LLMs), probabilistic transformer-based models originally developed to predict text sequences [3], have been increasingly explored for their potential in clinical applications. LLMs outperformed human experts on a range of medical diagnostic and academic tasks [4,5]. In our previous work, we compared the LLM performance to expert ratings while assessing medical text for suicidality, achieving accuracy of 87.5% [6]. The next logical step - the ability to predict clinical outcomes has rarely been tested. LLMs have been effective in predicting hospital readmission due to a general medical condition [7], admission to an intensive care unit [8] or intentional self-harm [9]. Given the fact that mental health care relies heavily on unstructured textual data, the potential of LLMs remains underutilized.

Global assessment tools, such as the Global Assessment of Functioning (GAF) [10] are among the

most commonly utilized instruments for functional outcome [11]. Prognostic studies have mainly focused on psychosis. In individuals at clinical high-risk for psychosis, the inclusion of structural MRI data significantly enhanced the prediction of 1-year functional outcomes, achieving clinically meaningful balanced accuracy exceeding 80% [12]. In first-episode psychosis, demographic and baseline clinical data predicted GAF at 12-month follow-up with a balanced accuracy of 72% [13]. In a sample of individuals with psychosis of varying illness duration, baseline clinical data predicted GAF at 3-year and 6-year follow-ups with accuracies ranging from 63.5% to 67.6% [14]. Additionally, clustering analysis of baseline cognitive profiles identified distinct patient subgroups within first-episode psychosis, who exhibited low clinical functioning at 6- and 12-month follow-ups [15]. While psychosis prediction remains a prominent objective, the non-specificity of prodromal symptoms and relatively low transition rates underscore the necessity for early recognition services to employ a transdiagnostic approach [16]. We aimed to predict functional outcomes within a 2-year follow-up in young, help-seeking individuals with predominantly affective and anxiety symptoms, as well as ADHD, who were recruited within the Early-BipoLife study – a population-based study on bipolar risk. Along with broad baseline clinical characteristics, we explored the potential of baseline structural MRI data to improve prediction accuracy. We employed data-driven feature selection and leave-one-site-out crossvalidation to ensure generalizability over multiple sites. Finally, we used a locally hosted, general purpose LLM to estimate the outcome based on engineered clinical notes derived from the baseline data. To the best of our knowledge, this is the first study to employ large language models (LLMs) to predict global functioning at follow-up in individuals presenting with mental health symptoms.

Methods

Participants

Early-BipoLife [17] is a multicenter, prospective, naturalistic study of participants aged 15–35 years recruited between 2015 – 2019 and assessed over a period of at least 2 years. Help-seeking youth and young adults consulting early detection centers or individuals presenting with at least one of the proposed risk factors for bipolar disorder [17,18] (see Note S1 in Multimedia Appendix 1 for inclusion and exclusion criteria) as well as in- and outpatients with depressive syndrome or ADHD were recruited at nine sites. For detailed data collection procedures see Pfennig et al. (2020) [17], Martini et al. (2024) [19] and Note S1. Comprehensive assessments were performed after 12 and 24 months. All participants received state-of-the-art counselling and treatment. Of the N = 918 participants who completed the 2-year follow-up, N = 31 (3.17%) participants transitioned to bipolar disorder [19]. N = 313 opted to receive a baseline MRI. Given the relatively low transition rates and the absence of a focus on predicting bipolar disorder, this sample is well-suited for predict functional outcome using a transdiagnostic approach.

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2013. All procedures involving human subjects/patients were approved by the Ethics Committee of the Medical Faculty at the Technical University Dresden (No. EK290082014) and all local ethics committees. We obtained a written informed consent after comprehensive information about study aims and procedures.

Baseline and follow-up assessments

Following clinical assessments were included in the analysis of functional outcome: age, sex, medication (none or any current psychotropic medication), migration status (positive when person itself, father or mother were not born in Germany and as negative if all three were born in Germany), first-degree relatives for bipolar disorder, DSM-IV diagnoses and substance use (SKID-I) [20], depressive symptoms (Inventory for Depressive Symptomatology-Clinician - IDSC) [21], early life stress (Childhood Trauma Questionnaire - CTQ) [22], Global Assessment of Functioning (GAF) [10] at baseline and in the past year, Functioning Assessment Short Test (FAST) [23,24], prodromal psychotic symptoms (Prodromal Questionnaire - PQ16) [25], bipolar risk factors EPIbipolar [18], extended Bipolar-at-Risk criteria (BARS) [26] and Bipolar Prodrome Symptom Scale (BPSS-P) [27]. In total 127 items were included as features (for details see Table S1 in Multimedia Appendix 1).

Although more differentiate measures of functional outcome do exist, we used GAF, as it is highly established and largely present in longitudinal datasets (including Early-BipoLife), which enables larger training and validation samples. Rather than predicting GAF as a continuous variable, we dichotomized GAF into impaired and non-impaired for following reasons: (a) to include trajectories, that remained stable on a low functioning level, as well as those that impaired during the follow-ups, (b) include both follow-ups in one outcome variable, (c) reflect a hypothetical clinical scenario, where typically a binary decision should be made between assigning and not assigning to an intervention. We defined the negative outcome as GAF equal or below 60 (i.e. moderate symptoms such as flat affect and circumstantial speech, occasional panic attacks or moderate difficulty in social, occupational, or school functioning, such as few friends, conflicts with peers or co-workers) at least one follow-up. Clinically, functional impairment at this level may be considered as an indication for clinical (out- or inpatient) intervention, as compared to GAF 70 – 61, which denotes good functioning with only mild symptoms and difficulties.

In post-hoc analyses, we aimed to predict following additional outcomes: 1) higher and lower GAF cut-off values (i.e. $GAF \leq 55$, ≤ 65), 2) we used Time Series K-Means clustering implemented in

Scikit-learn v. 1.5.2 to cluster the GAF trajectories in two clusters using 6 time points (GAF values at past year to baseline, baseline, past year to 1-year follow-up, 1-year follow-up, past year to 2-years follow-up, 2-years follow-up) and differentiate between participants belonging to these clusters. 3) maximum GAF values over the past year at follow-up one or two, aiming to capture long-term impairments that may have been overlooked by the GAF value at follow-up assessments.

MRI acquisition, preprocessing and quality assessment

We acquired 1 mm isotropic resolution structural T1-weighted images using Siemens Magnetom MR scanners at 6 sites (Trio, Skyra, Prisma) and a Philips Achieva scanner at 1 site. We standardized the pulse sequence parameters across all sites to the extent permitted by each platform. For a detailed description of the scanning protocol including the detail of MRI scanners, specific hardware configurations and pulse sequence parameters see Vogelbacher et al. (2021) [28].

Prior to preprocessing, we performed the data acquisition and quality assessment using the MRIQC tool [29]. Two authors visually inspected the obtained reports of several metrics including a movement plot and a plot of the background noise. In this way, 23 subjects were excluded from further analysis due to strong movement (N = 18), ghosting (N = 1) or fold-over artifacts (N = 4).

We preprocessed the T1-weighted structural MRI using Freesurfer 6.0 (cortical and subcortical parcellations) and 7.1.1 / 7.2.0 (hippocampal subfields and amygdala nuclei), partially running on the Centre for Information Services and High-Performance Computing (ZIH) by TU Dresden (<https://tu-dresden.de/zih/hochleistungsrechnen>). We obtained regional cortical thicknesses and surface area values for 68 cortical brain areas defined by the Desikan-Killiany atlas [30], 14 volumes of subcortical structures [31], 18 amygdala nuclei, 24 volumes of hippocampal subfields [32,33], i.e. 124 MRI features in total. We performed a standardized quality control of the cortical and subcortical segmentations according to the established protocols of the ENIGMA working group (<http://enigma.ini.usc.edu/protocols/imaging-protocols>). This included a visual inspection of the

segmented regions using the internal and external surface methods, as well as statistical outlier detection. The outliers were subjected for further visual inspection. For details on preprocessing and exclusion of subjects who did not pass the quality control or displayed major segmentation errors (N=8) please see our previous publications [34,35].

Machine learning pipelines

First, we predicted the impaired status at follow up using a linear support vector machine (LIBSVM 3.1.2, C-SVC, L1-loss) with instance weighting, implemented in the Neurominer Toolbox v. 1.2 (retrieved from [neurominer-git/NeuroMiner_1.2.](https://github.com/neurominer/neurominer)) using clinical features in N = 357 subjects after excluding those with missing GAF values at any follow-up and further removing subjects with missing values for 25% or more of the features.

Second, we trained separate SVM classifiers using clinical and MRI features in N = 124 subjects with available baseline MRI. We then applied stacked generalization, where the decision scores from both base models were combined using another linear SVM that utilized the decision scores from the completed analyses as input features to predict the outcome group [12].

All pipelines included feature-wise scaling (range 0 to 1) and imputation of missing values using the median of the seven nearest neighbors (Hamming distance for nominal features with two or fewer unique values, Euclidean distance for ordinal features). Covariates sex, age, and intracranial volume were regressed out using partial correlations exclusively in the MRI feature set, as sex and age were considered informative in the clinical domain. Both the removal of nuisance covariates and the imputation of missing data were carried out separately for test and training sets. The imputed values were derived from the training test and applied to the test set.

We employed a nested cross-validation scheme, with an outer loop (CV2) using six folds (leave-one-site-out, where each site represented a study site) and an inner loop (CV1) using five folds for model selection and C parameter tuning. In each inner fold, we applied a wrapper-based greedy feature

selection approach to retain the ratio of selected features to total participants at 1:5. As a result, $k = 71$ features were selected for the clinical models and $k = 25$ for the MRI and stacking models. The regularization parameter (C) was optimized across nine values ranging from 0.00390625 to 256.

We assessed the contribution of features towards the classification using the sign-based consistency, which was defined as the consistency of the weights assigned by the model to a given feature reduced by the fraction of models that removed that feature during the wrapper-based feature selection. Sign-based consistency of 1 means, that the weights of the feature all share the same sign and the feature has been selected by all classifiers in the inner crossvalidation loop, whereas a value of 0 means, that positive and negative weights occurred equally or the given feature was omitted in all crossvalidation folds [36].

In order to improve clinical interpretability, we analyzed the most predictive set of features retrieved by the SVM analysis (sign-based consistency score $p < 0.05$) using a decision tree classifier implemented in Scikit-learn (v. 1.5.2). Briefly, the decision tree minimizes the entropy by splitting the data so that each subset contains as few mixed class labels as possible (i.e., each leaf ideally contains a single class). For a binary classification, entropy = 1 for a leaf composed of 50 % of class 1 and 50 % of class 2, whereas entropy = 0 for a leaf containing only a single class. Using a 5-fold crossvalidation, we trained and evaluate the model using following parameters: decision criterion = entropy, maximal depth = 3 layers (for interpretability reasons we decided against more layers), minimal sample splits 2-10, class weighting to account for imbalanced classes.

LLM analysis

We used a locally hosted Llama-3 (llama-3.3-70b-instruct-q4km) via the llama.cpp framework on a local hospital computer [6]. For each participant, we converted the clinical assessments into reverse-engineered text notes by substituting item's scores by their descriptions (see Note S2A in Multimedia Appendix 1 for an example). Following items were included: Age, gender, diagnoses, substance use,

IDSC, CTQ, GAF baseline and past year, FAST, information on past and present psychiatric treatment and first-degree relatives for bipolar disorder. Llama-3 was instructed using a prompt to classify the functional outcome group (Note S2B in Multimedia Appendix 1). To further investigate the generalizability of the prompting approach, we repeated the analysis in an external sample of $N = 590$ participants aged 15-35 from the FOR2107 longitudinal sample of persons with affective disorder [37] (see Table S2 in Multimedia Appendix 1 for the details of the validation sample). The reverse-engineered notes included age, gender, main diagnosis, HAMD, CTQ items and GAF baseline.

Results

Demographics

Overall, the global functioning improved within the 2-years follow up ($\text{mean}_{\text{baseline}} = 61.6$, $\text{SD} = 15.7$, $\text{mean}_{\text{year1}} = 69.2$, $\text{SD} = 14.3$, $\text{mean}_{\text{year2}} = 70.6$, $\text{SD} = 14.4$). See Figure 1 for the trajectories.

===Figure 1 here===

The subjects impaired at follow-up fulfilled significantly more frequently criteria for psychiatric diagnoses, particularly affective, anxiety and substance use disorders (Table 1). They also more frequently contacted mental health services. The non-impaired group contained more first-degree relatives of persons with bipolar disorder.

Table 1: Sample demographics. * $p < 0.05$, ** $p < 0.001$. SD – standard deviation.

	Non-impaired	Impaired	Test
N = 357	228	129	n/a
Sex female (%)	128 (56.1)	81 (62.8)	Chi = 1.501, p = 0.22
Age (SD)	25.8 (4.4)	25.32 (4.6)	T = -1.125, p = 0.35
Positive migration status (%)	59 (26.9)	32 (27.4)	Chi = 0.006, p = 0.936
Diagnoses (%)			
Any disorder (SKID)	140 (61.4)	112 (86.8)	Chi = 25.638, p=0.001*
Affective disorder	87 (38.2)	75 (58.1)	Chi = 13.299 p <0.001**
Psychotic disorder	1 (0.004)	1 (0.8)	Chi 0 0.167, p = 0.683
Substance use disorder	14 (6.1)	18 (14.0)	Chi = 5.893, p = 0.015*
Anxiety disorder	56 (24.6)	61 (47.3)	Chi = 18.609, p < 0.001**
ADHD	57 (25.0)	31 (24.0)	Chi = 0.042, p = 0.838
Somatoform disorder	10 (4.4)	10 (7.8)	Chi = 1.765, p = 0.184
Global functioning (SD)			
GAF baseline	66.68 (14.8)	52.6 (13.1)	T = 8.918, p < 0.001**
GAF maximum past 1 year	76.31 (14.1)	64.09 (15.4)	T = 7.548, p = 0.147
GAF 1-year follow-up	76.6 (9.7)	55.67 (11.3)	T = 17.682, p < 0.001**
GAF maximum past 1 year to 1-year follow-up	78.22 (10.5)	63.54 (12.3)	T = 11.441, p < 0.001**
GAF 2-years follow-up	77.93 (10.4)	57.7 (10.8)	T = 17.310, p < 0.001**
GAF maximum past 1 year to 2-years follow-up	79.43 (10.1)	63.8 (11.3)	T = 13.024, p < 0.001**
Contact to mental health services present (%)	140 (61.4)	98 (76.0)	Chi = 7.924, p = 0.005*
Contact to mental health services past (%)	171 (75.0)	115 (89.1)	Chi = 9.988, p = 0.002*
First-degree relatives with bipolar disorder (%)	32 (14.0)	8 (6.2)	5.081, p = 0.024*
Transition to bipolar disorder (%)	4 (1.8)	1 (0.8)	Chi = 0.572, p = 0.449
PQ16 (SD)	3.16 (2.9)	4.6 (3.2)	T = 4.311, p = 0.27

The impairments in global functioning at baseline were more pronounced in the impaired group in all domains and items of the FAST inventory (Table S3 in Multimedia Appendix 1). For distributions of GAF scores at follow-ups see Figure S1 in Multimedia Appendix 1.

Subjects not included in the analysis due to missing data or failed quality assessment with available GAF at baseline (N = 855) displayed better clinical functioning at baseline (mean GAF_{discarded} 64.77, SD 18.87 versus GAF_{included} 61.58, SD 15.73, p = 0.005).

For breakdown of demographic characteristics to single sites see Table S4 in Multimedia Appendix 1.

SVM classification

The linear SVM classifier trained on clinical measures alone (sample size $N = 357$) to predict the impaired status within the 2-years follow-up achieved balanced accuracy 69.2%, sensitivity 64.3% and specificity 71.9%. Training the classifier using exclusively MRI features (sample size $N = 124$) achieved balanced accuracy 54.0%, sensitivity 43.5% and specificity 60.3%. In the same subsample, using only clinical features achieved a balanced accuracy of 73.0% (sensitivity 65.2%, specificity 80.8%). Combining the clinical and the MRI classifiers via stacking did not improve the performance (balanced accuracy 65.6%, sensitivity 54.3%, specificity 76.9%).

The most consistently predictive features were items related to occupational functioning, interpersonal relationships, cognitive functioning, psychotic and affective symptoms as well as presence of anxiety disorder (Table 2, Figure 2, Figure S2 in Multimedia Appendix 1). As GAF baseline was retrieved as the most predictive item, we trained an SVM classifier exclusively using GAF baseline as a single feature, achieving a comparable performance (balanced accuracy 70.2%, sensitivity 74.8%, specificity 65.6%). However, for clinical decision-making, reliable probability estimates are essential, and single-feature classifiers typically yield poor calibration. To evaluate both accuracy and reliability of probability estimates, we constructed calibration curves for single-feature and multi-feature models. The multi-feature model achieved a lower Brier score (0.214 vs. 0.311), indicating more accurate and reliable probability estimates (Note S3 in Multimedia Appendix 1).

===Figure 2 here===

Table 2. The list of significant predictive clinical features as defined by the sign-based consistency (see also Figure S2 in Multimedia Appendix 1).

Ranking	Item	Note
1.	GAF present	Baseline clinical functioning.
2.	PQ16	≥ 6 positive screening for psychosis
3.	FAST 17	Interpersonal relationships: Maintaining a friendship or friendships
4.	FAST 7	Occupational functioning: Working in the field in which you were educated.
5.	GAF maximum	Maximum GAF value over the past year prior to baseline.

6.	Anxiety disorder	DSM-IV diagnosis (SKID-I)
7.	FAST 5	Occupational functioning: Holding down a paid job
8.	FAST 6	Occupational functioning: Accomplishing tasks as quickly as necessary
9.	IDSC 5	Mood (sad)
10.	FAST 12	Cognitive functioning: Ability to solve a problem adequately
11.	FAST 19	Interpersonal relationships: Having good relationships with people close you
12.	FAST 22	Interpersonal relationships: Being able to defend your interests

Decision tree

Training the decision tree using the set of most consistently predictive features (Table 2) retrieved a significant decision tree with following parameters: balanced accuracy 76.6%, sensitivity 78.3% and specificity 74.9%. The stratification using baseline GAF was further refined by combining the sequence with best GAF in past year and items related to social functioning, early life adversity and psychosis screening. For the visualization of the decision tree see Figure 3. Training the decision tree using GAF baseline only achieved balanced accuracy 72.7%, sensitivity 67.4% and specificity 78.0%.

===Figure 3 here===

Classification using LLM

Classification using full text notes yielded balanced accuracy 72.6%, sensitivity 78.3%, specificity 66.9%. After excluding GAF baseline value Llama-3 achieved balanced accuracy 69.7%, sensitivity 88.4%, specificity 51.0%. In the external sample of persons with affective disorders (Table S2 in Multimedia Appendix 1), LLM achieved balanced accuracy 61.2%, sensitivity 55.4%, specificity 70.0%. After excluding GAF Llama-3 achieved balanced accuracy 60.8%, sensitivity 46.9%, specificity 74.7%.

Post-hoc analyses

Decreasing GAF thresholds for impaired outcome (i.e. increasing the impairment) revealed improved balanced accuracies for MRI data after reducing the feature set to hippocampus and amygdala subregions (Table S4 in Multimedia Appendix 1). However, stacking amygdala and hippocampus features with clinical data revealed no improvement (73.0% versus 72.3% respectively). Data-driven clustering of GAF trajectories using 6 GAF timepoints into two clusters revealed a low- and high functioning cluster (Figure S3 in Multimedia Appendix 1). However, MRI data were not able to differentiate between the two clusters (Table 3)

Table 3. Balanced accuracies of the SVM classifiers for different GAF cut-off values as outcomes. The different balanced accuracy value for clinical features and GAF 60 to the main analysis are due to reduced sample size (N = 124, i.e. subjects with MRI). sMRI – structural MRI.

	Clinical	sMRI	Cortical thickness	Hippocampus and Amygdala subregions	Stacked Clinical + sMRI
GAF 55	78.0%	56.5%	52.7%	58.6%	69.4%
GAF 60	73.0%	51.9%	45.4%	57.1%	65.6%
GAF 65	58.2%	46.5%	47.2%	52.2%	51.3%
GAF clustered	66.6%	50.1%	52.4%	50.1%	64.8%

SVM regression to predict the maximum GAF values over the past year to follow-up using clinical features achieved a significant positive relationship between predicted and actual values ($R^2 = 0.12$, $p < 0.00001$) with a mean absolute error (MAE) = 10.0 GAF points for clinical data. MRI features revealed a weak relationship ($R^2 = 0.03$, $p < 0.04$, MAE = 11.3) and there was no relevant improvement using stacking ($R^2 = 0.14$, $p < 0.00001$, MAE = 9.8). Repeating the primary SVM analysis to predict impaired outcome using different machine learning methods did not improve the balanced accuracies for clinical and sMRI features (Table S5 in Multimedia Appendix 1).



Discussion

Machine learning predicted global functioning within the 2 years follow-up with good performance. A detailed analysis of most predictive features revealed baseline occupational functioning, interpersonal relationships, cognitive functioning, psychotic, affective as well as anxiety symptoms as most predictive. The feature set was further refined using a decision tree, providing an interpretable solution for possible clinical translation. Native, 'out-of-the-box' LLM achieved comparable performance using clinical data. Although MRI did not improve the prediction achieved by features based on clinical assessments only, amygdala and hippocampal subregions were more predictive than other MRI features, especially in individuals with lower functioning at follow-up.

SVM predicted the functional outcome with good accuracy using baseline clinical data. This is in alignment with Koutsouleris et al. 2018 [12], where machine learning predicted the 1-year functioning in individuals with clinical high risk for psychosis, although the accuracy in our sample was lower (69.2% versus 76.9%). This was most likely due to different populations between the cohorts, the latter containing mostly subjects with affective and anxiety symptoms and only a minority with psychotic symptoms (average PQ-16 was below 6, which denotes negative psychosis prodrome screening; less than 1% fulfilled the DSM-IV criteria for psychotic disorder). Individuals with predominantly affective and anxiety symptoms might display more heterogeneous baseline symptoms as well as trajectories than persons with clinical high risk for psychosis. This might lead to less accurate predictive models.

In contrast to the clinical high risk for psychosis sample [12], structural MRI in our sample did not improve the prediction beyond what was already achieved using clinical data. In models using structural neuroimaging only, Koutsouleris et al. [12] achieved balanced accuracy up to 76.2%. Whereas patients with schizophrenia display most prominent structural alterations among major psychiatric disorders (excluding dementia), persons with affective

anxiety disorders and ADHD tend to display less pronounced structural alterations than persons with psychosis [38]. The concept of clinical high risk for psychosis encloses persons with global structural deficits [39]. Interestingly, reducing the set of features to hippocampus and amygdala subregions improved the prediction accuracy of MRI data in our sample, especially by lower thresholds of GAF, i.e. while detecting individuals with poorer functional outcome. Indeed, hippocampus and amygdala tend to display structural alterations in affective disorders [40,41]. Focusing on these regions, even using multimodal or high-resolution MRI data [42], might further improve predictions.

A data driven selection of most predictive variables revealed a limited set of features related to occupational functioning, interpersonal relationships, cognitive functioning, psychotic, affective as well as anxiety symptoms as most predictive. For the SVM classifier, baseline GAF emerged as the most predictive feature, while the inclusion of additional items did not improve overall classification accuracy. Of note, in a clinical setting, prediction probability estimations, rather than categorical decisions are crucial. Multiple feature classifiers are necessary to produce smooth individualized probability ranges which are more suitable for clinical decision making (see Note S3 in Multimedia Appendix 1 for the comparison between single- and multiple-feature models). In this regard, the multi-feature SVM outperformed the single feature SVM. In contrast, the inclusion of additional features led to a modest improvement in performance for the decision tree classifier.

Interestingly, Llama-3 predicted functional outcome with a balanced accuracy comparable to trained SVM or decision-tree classifiers. While it did not outperform these traditional models, it is important to emphasize that we used a native, 'out-of-the-box' general-purpose LLM, not primarily trained for predicting clinical outcomes. Prompt-engineering [43] or fine-tuning might further improve predictive performance [7]. A significant disadvantage of LLMs is the lack of interpretability, bias risk due to non-disclosed training data [44] as well as of reliable probability estimates [45]. These issues need to be addressed in order to bring LLMs to

bedside. In order to explore the potential of LLMs for clinical predictions, unstructured text such as clinical notes or interview transcripts should become an integral part of longitudinal studies.

In this study, we mapped the trajectories of global functioning using GAF at follow-up as well as maximum GAF over the past year (assessed retrospectively). While this approach is common in predicting functional outcomes, it may overlook a more nuanced assessment of functioning over time. Although more suitable measures for tracking long-term global functioning have been developed, they have been rarely used. Examples include "days out of role" (i.e., the number of days an individual is unable to work or perform daily activities due to physical or mental health issues) [46] and "time use" (i.e., weekly engagement in structured activities) [47]. Future longitudinal studies should incorporate more differentiated long-term assessments.

Interestingly, the number of first-degree relatives of bipolar patients as well as transitions to bipolar disorder was higher in the non-impaired group. This might suggest, that persons with manic episodes might have better functioning in the long run, at least in the early stages. However, the number of transitions was too low to be statistically significant. Moreover, some individuals in the sample with depressive symptoms might have unrecognized bipolar disorder that would be revealed with a longer follow-up period.

We used a conservative leave-one-site-out validation without data harmonization. Although harmonization might improve the performance of structural MRI based models, it introduces a problem of data leakage between train and test sets. An external validation might provide better insights in generalizability of predictive models, however, as none of the models achieved >80% accuracy and therefore relevance for clinical use [48,49], the prediction models require further refinement prior to external validation. Multimodal data such as genetics, immunological markers [50], clinician prognostic estimates or natural language processing features [6,51] could improve classification. In addition, functional outcome

measures should map functional trajectories with sufficient temporal resolution.

In a transdiagnostic cohort of young individuals, machine learning and LLM can predict 2-year global functioning with good performance. Key predictors included occupational functioning, interpersonal relationships, cognitive functioning, psychotic and affective as well as anxiety symptoms. Although hippocampal and amygdala subregions showed potential as more predictive MRI features, particularly in those with poorer outcome, MRI data did not enhance overall performance. Further refinements of predictors as well as functional outcome measures are needed to better reflect clinical trajectories and enhance predictive accuracy.

Data availability

The analytic code of this study will be openly available in GitHub repository at https://github.com/pmiko/BL_GAF.

The data will be provided upon reasonable request.

Acknowledgments

Beyond analyses using Llama-3 described in the Methods section in detail, generative AI (ChatGPT, OpenAI) was used to assist with text editing and code development. All content was critically reviewed and verified by the authors.

Funding

Early-BipoLife was funded by the Federal Ministry of Education and Research (BMBF, grant numbers: 01EE1404A, 01EE1404E and 01EE1404F). This work was funded in part by the German Research Foundation (DFG) grant numbers 521379614 [SFB/TRR 393] and GRK2773/1- 454245598. MM was supported by the DFG grant numbers: 178833530 [SFB 940] and 402170461 [TRR 265]. UD was funded by the DFG grants FOR2107 DA1151/5-1, DA1151/5-2, DA1151/9-1, DA1151/10-1, DA1151/11-1 to UD; SFB/TRR 393, project grant no 521379614) and the Interdisciplinary Center for Clinical Research (IZKF) of the medical faculty of Münster (grant Dan3/022/22 to UD). JR was supported by the LOEWE program of the Hessian Ministry of Science and Arts (grant number: LOEWE1/16/519/03/09.001(0009)/98).

Conflict of Interest

K Leopold has been a consultant and/or advisor to or has received honoraria from: Janssen/J&J, Lundbeck, Otsuka, Recordati and ROVI. She has received grant support from Janssen/J&J and Otsuka. JR received speaker's honoraria from Janssen, Hexal, Neuraxpharm and Novartis. JNK declares consulting services for Owkin, France, DoMore Diagnostics, Norway, Panakeia, UK, Scailyte, Switzerland, Cancilico, Germany, Mindpeak, Germany, MultiplexDx, Slovakia, and Histofy, UK; furthermore, he holds shares in StratifAI GmbH, Germany, has received a research grant by GSK, and has received honoraria by AstraZeneca,

Bayer, Eisai, Janssen, MSD, BMS, Roche, Pfizer and Fresenius. ICW received honoraria from AstraZeneca All other authors report no biomedical financial interests or potential conflicts of interest. All other authors declared no conflict of interest.

Authors' Contributions

A.P., M.B., P.Ritter. and A.J. designed the study. K.B., C.B., J.M., B.M., A.J.F., T.E., A.Rau, T.K., U.D., I.F., M.L., G.L., C.M., V.K., K.L., A.B., A.Reif, S.M., T.S., F.B., J.F., G.J., V.F. and C.U.C. participated in the patient recruitment. F.H., P.M., J.R., E.M., P.H., F.G.V., I.W., J.K. and M.M. developed the machine learning analyses pipelines, performed data analyses and statistics. P.M., C.V. and F.H. performed the quality control. A.P., M.B., P.R., U.D. and T.K. obtained funding. P.M., M.M., F.H. and A.P. wrote the article. M.M., K.B., C.B., J.M., P.H., F.G.V., B.M., A.J.F., T.E., A.Rau, T.K., U.D., I.F., M.L., G.L., C.M., V.K., K.L., A.B., A.Reif, J.R., S.M., T.S., F.B., J.F., G.J., V.F., C.U.C., I.W., J.K., E.M., M.B. and P.Ritter. revised it critically for important intellectual content. All of the authors reviewed and approved the manuscript for publication.

Abbreviations

BARS - extended Bipolar-at-Risk criteria

BPSS-P - Bipolar Prodrome Symptom Scale

CTQ - Childhood Trauma Questionnaire

CV - cross-validation

FAST - Functioning Assessment Short Test

GAF - Global Assessment of Functioning

IDSC - depressive symptoms (Inventory for Depressive Symptomatology-Clinician

LLM - large language model

MAE - mean absolute error

PQ16 - Prodromal Questionnaire

SVM - support vector machine

SKID-I - Structured Clinical Interview for DSM-IV

References

1. Deutsche Rentenversicherung. Rentenversicherung in Zeitreihen. Deutsche Rentenversicherung. 2024. Available from: https://www.deutsche-rentenversicherung.de/SharedDocs/Downloads/DE/Statistiken-und-Berichte/statistikpublikationen/rv_in_zeitreihen.html [accessed Nov 27, 2024]
2. Arias D, Saxena S, Verguet S. Quantifying the global burden of mental disorders and their economic value. *eClinicalMedicine Elsevier*; 2022 Dec 1;54. PMID:36193171
3. Kjell ONE, Kjell K, Schwartz HA. Beyond rating scales: With targeted evaluation, large language models are poised for psychological assessment. *Psychiatry Research* 2024 Mar 1;333:115667. doi: 10.1016/j.psychres.2023.115667
4. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, Cool JA, Kanjee Z, Parsons AS, Ahuja N, Horvitz E, Yang D, Milstein A, Olson APJ, Rodman A, Chen JH. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Network Open* 2024 Oct 28;7(10):e2440969. doi: 10.1001/jamanetworkopen.2024.40969
5. Buckley TA, Crowe B, Abdunour R-EE, Rodman A, Manrai AK. Comparison of Frontier Open-Source and Proprietary Large Language Models for Complex Diagnoses. *JAMA Health Forum* 2025 Mar 14;6(3):e250040. doi: 10.1001/jamahealthforum.2025.0040
6. Wiest IC, Verhees FG, Ferber D, Zhu J, Bauer M, Lewitzka U, Pfennig A, Mikolas P, Kather JN. Detection of suicidality from medical text using privacy-preserving large language models. *The British Journal of Psychiatry* 2024 Dec;225(6):532–537. doi: 10.1192/bjp.2024.134
7. Ben Shoham O, Rappoport N. CPLLM: Clinical prediction with large language models. *PLOS Digit Health* 2024 Dec 6;3(12):e0000680. PMID:39642102

8. Brown KE, Yan C, Li Z, Zhang X, Collins BX, Chen Y, Clayton EW, Kantarcioglu M, Vorobeychik Y, Malin BA. Large language models are less effective at clinical prediction tasks than locally trained machine learning models. *Journal of the American Medical Informatics Association* 2025 Mar 8;ocaf038. doi: 10.1093/jamia/ocaf038
9. Yang Z, Mitra A, Liu W, Berlowitz D, Yu H. TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nat Commun Nature Publishing Group*; 2023 Nov 29;14(1):7857. doi: 10.1038/s41467-023-43715-z
10. Aas IM. Guidelines for rating Global Assessment of Functioning (GAF). *Annals of General Psychiatry United Kingdom: BioMed Central Limited*; 2011;10. doi: 10.1186/1744-859X-10-2
11. Cowman M, Godfrey E, Walsh T, Frawley E, Fowler D, Alvarez-Jimenez M, O'Connor K, Wykes T, Birchwood M, Donohoe G. Measures of Social and Occupational Function in Early Psychosis: A Systematic Review and Meta-analysis. *Schizophr Bull* 2023 May 13;50(2):266–285. PMID:37173277
12. Koutsouleris N, Kambeitz-Ilankovic L, Ruhrmann S, Rosen M, Ruef A, Dwyer DB, Paolini M, Chisholm K, Kambeitz J, Haidl T, Schmidt A, Gillam J, Schultze-Lutter F, Falkai P, Reiser M, Riecher-Rössler A, Upthegrove R, Hietala J, Salokangas RKR, Pantelis C, Meisenzahl E, Wood SJ, Beque D, Brambilla P, Borgwardt S, for the PRONIA Consortium. Prediction Models of Functional Outcomes for Individuals in the Clinical High-Risk State for Psychosis or With Recent-Onset Depression: A Multimodal, Multisite Machine Learning Analysis. *JAMA Psychiatry* 2018 Nov 1;75(11):1156. doi: 10.1001/jamapsychiatry.2018.2165
13. Slot MIE, Urquijo Castro MF, Winter-van Rossum I, van Hell HH, Dwyer D, Dazzan P, Maat A, De Haan L, Crespo-Facorro B, Glenthøj BY, Lawrie SM, McDonald C, Gruber O, van Amelsvoort T, Arango C, Kircher T, Nelson B, Galderisi S, Weiser M, Sachs G, Kirschner M, PSYSCAN Consortium, Fleischhacker WW, McGuire P, Koutsouleris N, Kahn RS. Multivariable prediction of functional outcome after first-episode psychosis: a crossover validation approach in EUFEST and PSYSCAN. *Schizophrenia (Heidelb)* 2024 Oct 7;10(1):89. PMID:39375356
14. de Nijs J, Burger TJ, Janssen RJ, Kia SM, van Opstal DPJ, de Koning MB, de Haan L, Cahn W, Schnack HG. Individualized prediction of three- and six-year outcomes of psychosis in a longitudinal multicenter study: a machine learning approach. *NPJ Schizophr* 2021 July 2;7:34. PMID:34215752
15. Oomen PP, Begemann MJH, Brand BA, de Haan L, Veling W, Koops S, van Os J, Smit F, Bakker PR, van Beveren N, Boonstra N, Gülöksüz S, Kikkert M, Lokkerbol J, Marcelis M, Rosema B-S, de Beer F, Gangadin SS, Geraets CNW, van 't Hag E, Haveman Y, van der Heijden I, Voppel AE, Willemse E, van Amelsvoort T, Bak M, Batalla A, Been A, van den Bosch M, van den Brink T, Faber G, Grootens KP, de Jonge M, Knegtering R, Kurkamp J, Mahabir A, Pijnenborg GHM, Staring T, Veen N, Veerman S, Wiersma S, Graveland E, Hoornaar J, Sommer IEC. Longitudinal clinical and functional outcome in distinct cognitive subgroups of first-episode psychosis: a cluster analysis. *Psychol Med* 2023 Apr;53(6):2317–2327. PMID:34664546
16. Arribas M, Oliver D, Patel R, Kornblum D, Shetty H, Damiani S, Krakowski K,

- Provenzani U, Stahl D, Koutsouleris N, McGuire P, Fusar-Poli P. A transdiagnostic prodrome for severe mental disorders: an electronic health record study. *Mol Psychiatry* Nature Publishing Group; 2024 Nov;29(11):3305–3315. doi: 10.1038/s41380-024-02533-5
17. Pfennig A, Leopold K, Martini J, Boehme A, Lambert M, Stamm T, BERPohl F, Reif A, Kittel-Schneider S, Juckel G, Fallgatter AJ, Kircher T, Jansen A, Pfeiffer S, Berndt C, Rottmann-Wolf M, Sauer C, Ritter P, Correll CU, Bechdolf A, Falkenberg I, Bauer M. Improving early recognition and intervention in people at increased risk for the development of bipolar disorder: study protocol of a prospective-longitudinal, naturalistic cohort study (Early-BipoLife). *Int J Bipolar Disord* 2020 Dec;8(1):22. doi: 10.1186/s40345-020-00183-4
 18. Leopold K, Ritter P, Correll CU, Marx C, Özgürdal S, Juckel G, Bauer M, Pfennig A. Risk constellations prior to the development of bipolar disorders: Rationale of a new risk assessment tool. *Journal of Affective Disorders* 2012 Feb;136(3):1000–1010. doi: 10.1016/j.jad.2011.06.043
 19. Martini J, Bröckel KL, Leopold K, Berndt C, Sauer C, Maicher B, Juckel G, Krüger-Özgürdal S, Fallgatter AJ, Lambert M, Bechdolf A, Reif A, Matura S, Biere S, Kittel-Schneider S, Stamm T, BERPohl F, Kircher T, Falkenberg I, Jansen A, Dannlowski U, Correll CU, Fusar-Poli P, Hempel LM, Mikolas P, Ritter P, Bauer M, Pfennig A. Young people at risk for developing bipolar disorder: Two-year findings from the multicenter prospective, naturalistic Early-BipoLife study. *European Neuropsychopharmacology* 2024 Jan 1;78:43–53. doi: 10.1016/j.euroneuro.2023.10.001
 20. Wittchen H-U, Zaudig M, Fydrich T, Wunderlich U, Gruschwitz S, Renneberg B, Schmitz B. SKID. Strukturiertes Klinisches Interview für DSM-IV. Achse I und II: Achse I: psychische Störungen; Achse II: Persönlichkeitstörungen. 1. Auflage. Göttingen: Hogrefe; 1997.
 21. Rush AJ, Giles DE, Schlesser MA, Fulton CL, Weissenburger J, Burns C. The Inventory for Depressive Symptomatology (IDS): preliminary findings. *Psychiatry Res* 1986 May;18(1):65–87. PMID:3737788
 22. Bernstein DP, Fink L, Handelsman L, Foote J. Childhood Trauma Questionnaire. 2011. doi: 10.1037/t02080-000
 23. Rosa AR, Sánchez-Moreno J, Martínez-Aran A, Salamero M, Torrent C, Reinares M, Comes M, Colom F, Van Riel W, Ayuso-Mateos JL, Kapczinski F, Vieta E. Validity and reliability of the Functioning Assessment Short Test (FAST) in bipolar disorder. *Clinical Practice and Epidemiology in Mental Health* 2007 June 7;3(1):5. doi: 10.1186/1745-0179-3-5
 24. Christensen MC, Schmidt SN, Grande I. The Functioning Assessment Short Test (FAST): Clinically meaningful response threshold in patients with major depressive disorder receiving antidepressant treatment. *Journal of Affective Disorders* 2024 Oct 15;363:634–642. doi: 10.1016/j.jad.2024.07.018
 25. Ising HK, Veling W, Loewy RL, Rietveld MW, Rietdijk J, Dragt S, Klaassen RMC, Nieman DH, Wunderink L, Linszen DH, van der Gaag M. The Validity of the 16-Item Version of the Prodromal Questionnaire (PQ-16) to Screen for Ultra High Risk of

- Developing Psychosis in the General Help-Seeking Population. *Schizophrenia Bulletin* 2012 Nov 1;38(6):1288–1296. doi: 10.1093/schbul/sbs068
26. Bechdolf A, Ratheesh A, Cotton SM, Nelson B, Chanen AM, Betts J, Bingmann T, Yung AR, Berk M, McGorry PD. The predictive validity of bipolar at-risk (prodromal) criteria in help-seeking adolescents and young adults: a prospective study. *Bipolar Disord* 2014 Aug;16(5):493–504. PMID:24797824
 27. Correll CU, Olvet DM, Auther AM, Hauser M, Kishimoto T, Carrión RE, Snyder S, Cornblatt BA. The Bipolar Prodrome Symptom Interview and Scale-Prospective (BPSS-P): description and validation in a psychiatric sample and healthy controls. *Bipolar Disord* 2014 Aug;16(5):505–522. doi: 10.1111/bdi.12209
 28. Vogelbacher C, Möbius TWD, Sommer J, Schuster V, Dannlowski U, Kircher T, Dempfle A, Jansen A, Bopp MHA. The Marburg-Münster Affective Disorders Cohort Study (MACS): A quality assurance protocol for MR neuroimaging data. *Neuroimage* 2018 May 15;172:450–460. PMID:29410079
 29. Esteban O, Birman D, Schaer M, Koyejo OO, Poldrack RA, Gorgolewski KJ. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. Bernhardt BC, editor. *PLoS ONE* 2017 Sept 25;12(9):e0184661. doi: 10.1371/journal.pone.0184661
 30. Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* 2006;31(3):968–980. doi: DOI:10.1016/j.neuroimage.2006.01.021
 31. Fischl B, van der Kouwe A, Destrieux C, Halgren E, Ségonne F, Salat DH, Busa E, Seidman LJ, Goldstein J, Kennedy D, Caviness V, Makris N, Rosen B, Dale AM. Automatically Parcellating the Human Cerebral Cortex. *Cerebral Cortex* 2004;14(1):11–22. doi: 10.1093/cercor/bhg087
 32. Iglesias JE, Augustinack JC, Nguyen K, Player CM, Player A, Wright M, Roy N, Frosch MP, McKee AC, Wald LL, Fischl B, Van Leemput K. A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: Application to adaptive segmentation of in vivo MRI. *NeuroImage* 2015 July;115:117–137. doi: 10.1016/j.neuroimage.2015.04.042
 33. Iglesias JE, Van Leemput K, Augustinack J, Insausti R, Fischl B, Reuter M, Alzheimer's Disease Neuroimaging Initiative. Bayesian longitudinal segmentation of hippocampal substructures in brain MRI using subject-specific atlases. *Neuroimage* 2016 Nov 1;141:542–555. PMID:27426838
 34. Mikolas P, Marxen M, Riedel P, Bröckel K, Martini J, Huth F, Berndt C, Vogelbacher C, Jansen A, Kircher T, Falkenberg I, Lambert M, Kraft V, Leicht G, Mulert C, Fallgatter AJ, Ethofer T, Rau A, Leopold K, Bechdolf A, Reif A, Matura S, Bermpohl F, Fiebig J, Stamm T, Correll CU, Juckel G, Flasbeck V, Ritter P, Bauer M, Pfennig A. Prediction of estimated risk for bipolar disorder using machine learning and structural MRI features. *Psychol Med* 2023 May 22;1–11. doi: 10.1017/S0033291723001319

35. Huth F, Tozzi L, Marxen M, Riedel P, Bröckel K, Martini J, Berndt C, Sauer C, Vogelbacher C, Jansen A, Kircher T, Falkenberg I, Thomas-Odenthal F, Lambert M, Kraft V, Leicht G, Mulert C, Fallgatter AJ, Ethofer T, Rau A, Leopold K, Bechdolf A, Reif A, Matura S, Biere S, BERPohl F, Fiebig J, Stamm T, Correll CU, Juckel G, Flasbeck V, Ritter P, Bauer M, Pfennig A, Mikolas P. Machine Learning Prediction of Estimated Risk for Bipolar Disorders Using Hippocampal Subfield and Amygdala Nuclei Volumes. *Brain Sciences Multidisciplinary Digital Publishing Institute*; 2023 June;13(6):870. doi: 10.3390/brainsci13060870
36. Koutsouleris N, Dwyer DB, Degenhardt F, Maj C, Urquijo-Castro MF, Sanfelici R, Popovic D, Oeztuerk O, Haas SS, Weiske J, Ruef A, Kambeitz-Ilankovic L, Antonucci LA, Neufang S, Schmidt-Kraepelin C, Ruhrmann S, Penzel N, Kambeitz J, Haidl TK, Rosen M, Chisholm K, Riecher-Rössler A, Egloff L, Schmidt A, Andreou C, Hietala J, Schirmer T, Romer G, Walger P, Franscini M, Traber-Walker N, Schimmelmann BG, Flückiger R, Michel C, Rössler W, Borisov O, Krawitz PM, Heekeren K, Buechler R, Pantelis C, Falkai P, Salokangas RKR, Lencer R, Bertolino A, Borgwardt S, Noethen M, Brambilla P, Wood SJ, Upthegrove R, Schultze-Lutter F, Theodoridou A, Meisenzahl E, PRONIA Consortium. Multimodal Machine Learning Workflows for Prediction of Psychosis in Patients With Clinical High-Risk Syndromes and Recent-Onset Depression. *JAMA Psychiatry* 2021 Feb 1;78(2):195–209. PMID:33263726
37. Kircher T, Wöhr M, Nenadic I, Schwarting R, Schratt G, Alferink J, Culmsee C, Garn H, Hahn T, Müller-Myhsok B, Dempfle A, Hahmann M, Jansen A, Pfefferle P, Renz H, Rietschel M, Witt SH, Nöthen M, Krug A, Dannlowski U. Neurobiology of the major psychoses: a translational perspective on brain structure and function—the FOR2107 consortium. *Eur Arch Psychiatry Clin Neurosci* 2019 Dec 1;269(8):949–962. doi: 10.1007/s00406-018-0943-x
38. Ching CRK, Hibar DP, Gurholt TP, Nunes A, Thomopoulos SI, Abé C, Agartz I, Brouwer RM, Cannon DM, Zwarte SMC, Eyler LT, Favre P, Hajek T, Haukvik UK, Houenou J, Landén M, Lett TA, McDonald C, Nabulsi L, Patel Y, Pauling ME, Paus T, Radua J, Soeiro-de-Souza MG, Tronchin G, Haren NEM, Vieta E, Walter H, Zeng L, Alda M, Almeida J, Alnæs D, Alonso-Lana S, Altimus C, Bauer M, Baune BT, Bearden CE, Bellani M, Benedetti F, Berk M, Bilderbeck AC, Blumberg HP, Bøen E, Bollettini I, Mar Bonnin C, Brambilla P, Canales-Rodríguez EJ, Caseras X, Dandash O, Dannlowski U, Delvecchio G, Díaz-Zuluaga AM, Dima D, Duchesnay É, Elvsåshagen T, Fears SC, Frangou S, Fullerton JM, Glahn DC, Goikolea JM, Green MJ, Grotegerd D, Gruber O, Haarman BCM, Henry C, Howells FM, Ives-Deliperi V, Jansen A, Kircher TTJ, Knöchel C, Kramer B, Lafer B, López-Jaramillo C, Machado-Vieira R, MacIntosh BJ, Melloni EMT, Mitchell PB, Nenadic I, Nery F, Nugent AC, Oertel V, Ophoff RA, Ota M, Overs BJ, Pham DL, Phillips ML, Pineda-Zapata JA, Poletti S, Polosan M, Pomarol-Clotet E, Pouchon A, Quidé Y, Rive MM, Roberts G, Ruhe HG, Salvador R, Sarró S, Satterthwaite TD, Schene AH, Sim K, Soares JC, Stäblein M, Stein DJ, Tamnes CK, Thomaidis GV, Upegui CV, Veltman DJ, Wessa M, Westlye LT, Whalley HC, Wolf DH, Wu M, Yatham LN, Zarate CA, Thompson PM, Andreassen OA, ENIGMA Bipolar Disorder Working Group. What we learn about bipolar disorder from large-scale neuroimaging: Findings and future directions from the ENIGMA Bipolar Disorder Working Group. *Hum Brain Mapp* 2020 July 29;43(1):hbm.25098. doi: 10.1002/hbm.25098
39. Vissink CE, Winter-van Rossum I, Cannon TD, Fusar-Poli P, Kahn RS, Bossong MG. Structural Brain Volumes of Individuals at Clinical High Risk for Psychosis: A Meta-analysis. *Biol Psychiatry Glob Open Sci* 2021 Sept 30;2(2):147–152. PMID:36325161

40. Schmaal L, Hibar DP, Sämann PG, Hall GB, Baune BT, Jahanshad N, Cheung JW, van Erp TGM, Bos D, Ikram MA, Vernooij MW, Niessen WJ, Tiemeier H, Hofman A, Wittfeld K, Grabe HJ, Janowitz D, Bülow R, Selonke M, Völzke H, Grotegerd D, Dannlowski U, Arolt V, Opel N, Heindel W, Kugel H, Hoehn D, Czych M, Couvy-Duchesne B, Rentería ME, Strike LT, Wright MJ, Mills NT, de Zubicaray GI, McMahon KL, Medland SE, Martin NG, Gillespie NA, Goya-Maldonado R, Gruber O, Krämer B, Hatton SN, Lagopoulos J, Hickie IB, Frodl T, Carballo A, Frey EM, van Velzen LS, Penninx BWJH, van Tol M-J, van der Wee NJ, Davey CG, Harrison BJ, Mwangi B, Cao B, Soares JC, Veer IM, Walter H, Schoepf D, Zuroski B, Konrad C, Schramm E, Normann C, Schnell K, Sacchet MD, Gotlib IH, MacQueen GM, Godlewska BR, Nickson T, McIntosh AM, Papmeyer M, Whalley HC, Hall J, Sussmann JE, Li M, Walter M, Aftanas L, Brack I, Bokhan NA, Thompson PM, Veltman DJ. Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the ENIGMA Major Depressive Disorder Working Group. *Mol Psychiatry* 2017 June;22(6):900–909. PMID:27137745
41. Hibar DP, Westlye LT, van Erp TGM, Rasmussen J, Leonardo CD, Faskowitz J, Haukvik UK, Hartberg CB, Doan NT, Agartz I, Dale AM, Gruber O, Krämer B, Trost S, Liberg B, Abé C, Ekman CJ, Ingvar M, Landén M, Fears SC, Freimer NB, Bearden CE, Sprooten E, Glahn DC, Pearlson GD, Emsell L, Kenney J, Scanlon C, McDonald C, Cannon DM, Almeida J, Versace A, Caseras X, Lawrence NS, Phillips ML, Dima D, Delvecchio G, Frangou S, Satterthwaite TD, Wolf D, Houenou J, Henry C, Malt UF, Bøen E, Elvsåshagen T, Young AH, Lloyd AJ, Goodwin GM, Mackay CE, Bourne C, Bilderbeck A, Abramovic L, Boks MP, van Haren NEM, Ophoff RA, Kahn RS, Bauer M, Pfennig A, Alda M, Hajek T, Mwangi B, Soares JC, Nickson T, Dimitrova R, Sussmann JE, Hagenaars S, Whalley HC, McIntosh AM, Thompson PM, Andreassen OA. Subcortical volumetric abnormalities in bipolar disorder. *Mol Psychiatry* 2016 Dec;21(12):1710–1716. doi: 10.1038/mp.2015.227
42. Ebneabbasi A, Mahdipour M, Nejati V, Li M, Liebe T, Colic L, Leutritz AL, Vogel M, Zarei M, Walter M, Tahmasian M. Emotion processing and regulation in major depressive disorder: A 7T resting-state fMRI study. *Human Brain Mapping* 2021;42(3):797–810. doi: 10.1002/hbm.25263
43. Verhees FG, Huth F, Meyer V, Wolf F, Bauer M, Pfennig A, Ritter P, Kather JN, Wiest IC, Mikolas P. clickBrick Prompt Engineering: Optimizing Large Language Model Performance in Clinical Psychiatry. Cold Spring Harbor Laboratory; 2025. doi: 10.1101/2025.06.28.25330267
44. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, Jurafsky D, Szolovits P, Bates DW, Abdunour R-EE, Butte AJ, Alsentzer E. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health* 2024 Jan 1;6(1):e12–e22. doi: 10.1016/S2589-7500(23)00225-X
45. Gu B, Desai RJ, Lin KJ, Yang J. Probabilistic medical predictions of large language models. *npj Digit Med Nature Publishing Group*; 2024 Dec 19;7(1):1–9. doi: 10.1038/s41746-024-01366-4
46. Alonso J, Petukhova M, Vilagut G, Chatterji S, Heeringa S, Üstün TB, Alhamzawi AO, Viana MC, Angermeyer M, Bromet E, Bruffaerts R, de Girolamo G, Florescu S, Gureje O, Haro JM, Hinkov H, Hu C -y, Karam EG, Kovess V, Levinson D, Medina-Mora ME, Nakamura Y, Ormel J, Posada-Villa J, Sagar R, Scott KM, Tsang A, Williams DR,

- Kessler RC. Days out of role due to common physical and mental conditions: results from the WHO World Mental Health surveys. *Mol Psychiatry* Nature Publishing Group; 2011 Dec;16(12):1234–1246. doi: 10.1038/mp.2010.101
47. Hodgekins J, French P, Birchwood M, Mugford M, Christopher R, Marshall M, Everard L, Lester H, Jones P, Amos T, Singh S, Sharma V, Morrison AP, Fowler D. Comparing time use in individuals at different stages of psychosis and a non-clinical comparison group. *Schizophrenia Research* 2015 Feb 1;161(2):188–193. doi: 10.1016/j.schres.2014.12.011
 48. Nunes, Nunes A, Schnack HG, Ching CRK, Agartz I, Akudjedu TN, Alda M, Alnæs D, Alonso-Lana S, Bauer J, Baune BT, Bøen E, Bonnin C del M, Busatto GF, Canales-Rodríguez EJ, Cannon DM, Caseras X, Chaim-Avancini TM, Dannlowski U, Díaz-Zuluaga AM, Dietsche B, Doan NT, Duchesnay E, Elvsåshagen T, Emden D, Eyler LT, Fatjó-Vilas M, Favre P, Foley SF, Fullerton JM, Glahn DC, Goikolea JM, Grotegerd D, Hahn T, Henry C, Hibar DP, Houenou J, Howells FM, Jahanshad N, Kaufmann T, Kenney J, Kircher TTJ, Krug A, Lagerberg TV, Lenroot RK, López-Jaramillo C, Machado-Vieira R, Malt UF, McDonald C, Mitchell PB, Mwangi B, Nabulsi L, Opel N, Overs BJ, Pineda-Zapata JA, Pomarol-Clotet E, Redlich R, Roberts G, Rosa PG, Salvador R, Satterthwaite TD, Soares JC, Stein DJ, Temmingh HS, Trappenberg T, Uhlmann A, van Haren NEM, Vieta E, Westlye LT, Wolf DH, Yüksel D, Zanetti MV, Andreassen OA, Thompson PM, Hajek T. Using structural MRI to identify bipolar disorders – 13 site machine learning study in 3020 individuals from the ENIGMA Bipolar Disorders Working Group. *Mol Psychiatry* 2020 Sept;25(9):2130–2143. doi: 10.1038/s41380-018-0228-9
 49. Radua J, Carvalho AF. Route map for machine learning in psychiatry: Absence of bias, reproducibility, and utility. *European Neuropsychopharmacology* 2021 Sept;50:115–117. doi: 10.1016/j.euroneuro.2021.05.006
 50. Lalousis PA, Schmaal L, Wood SJ, Reniers RLEP, Barnes NM, Chisholm K, Griffiths SL, Stainton A, Wen J, Hwang G, Davatzikos C, Wenzel J, Kambeitz-Ilankovic L, Andreou C, Bonivento C, Dannlowski U, Ferro A, Lichtenstein T, Riecher-Rössler A, Romer G, Rosen M, Bertolino A, Borgwardt S, Brambilla P, Kambeitz J, Lencer R, Pantelis C, Ruhrmann S, Salokangas RKR, Schultze-Lutter F, Schmidt A, Meisenzahl E, Koutsouleris N, Dwyer D, Uthegrove R. Neurobiologically Based Stratification of Recent-Onset Depression and Psychosis: Identification of Two Distinct Transdiagnostic Phenotypes. *Biological Psychiatry* 2022 Oct 1;92(7):552–562. doi: 10.1016/j.biopsych.2022.03.021
 51. Irving J, Patel R, Oliver D, Colling C, Pritchard M, Broadbent M, Baldwin H, Stahl D, Stewart R, Fusar-Poli P. Using Natural Language Processing on Electronic Health Records to Enhance Detection and Prediction of Psychosis Risk. *Schizophrenia Bulletin* 2021 Mar 16;47(2):405–414. doi: 10.1093/schbul/sbaa126

Figure 1: Graphical representation of functional outcome trajectories. The nodes of the Sankey plot represent the GAF values (lower value represents higher impairment). Defining the impaired outcome as $GAF \leq 60$ at year 1 and/or year 2 selected trajectories, that remained stable at a moderately impaired level of functioning, as well as trajectories that deteriorated to a moderate level during the follow-up.

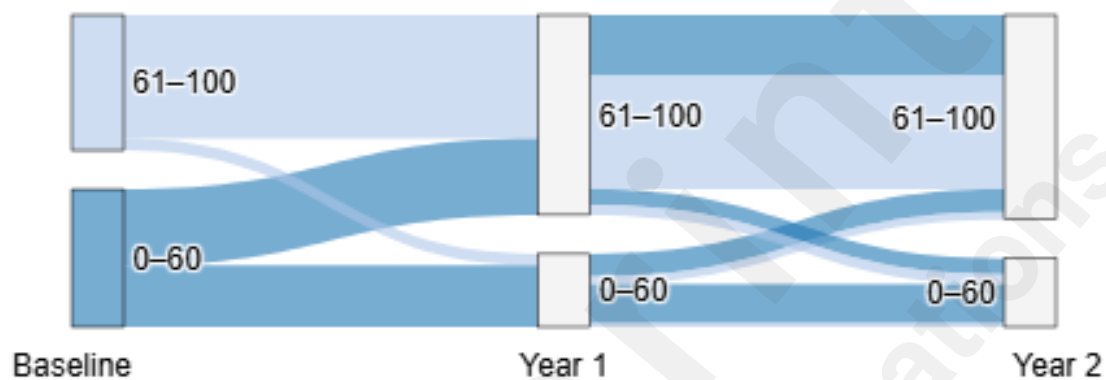
Figure 2: The ranking of clinical features as defined by the sign-based consistency. The red line indicates the significance level $p < 0.05$.

Figure 3. Decision tree predicting the impaired functional outcome within 2-year follow-up. The arrows pointing to the left indicate that the condition is fulfilled (true), the arrows to the right indicate not fulfilled (false). Colors indicate the decision being made by the tree when condition was fulfilled. The decision tree assigns the leaf label according to the corresponding majority class in that leaf. Number of samples and percentage of individuals with impaired outcome within each node / leaf is being displayed.

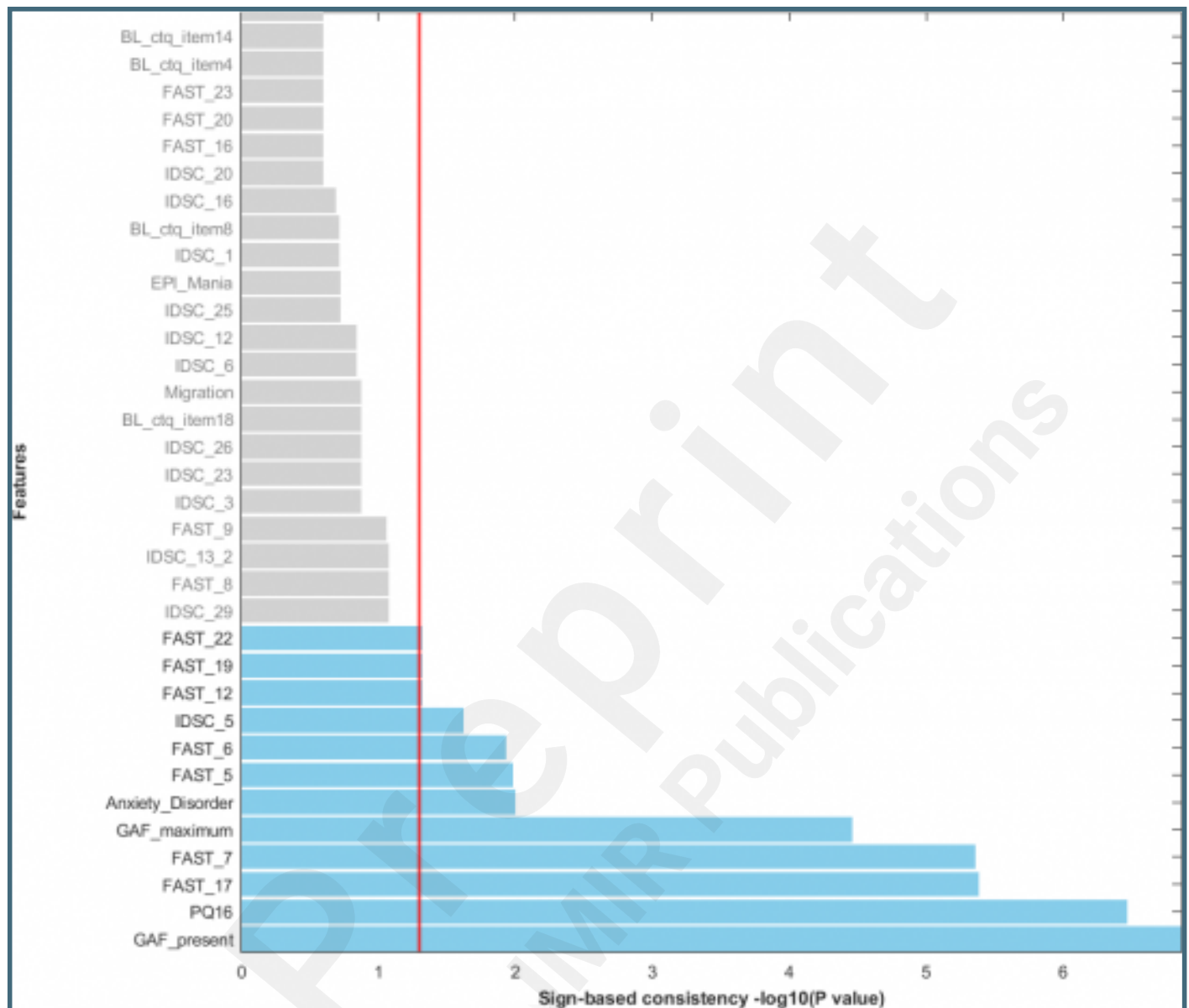
Supplementary Files

Figures

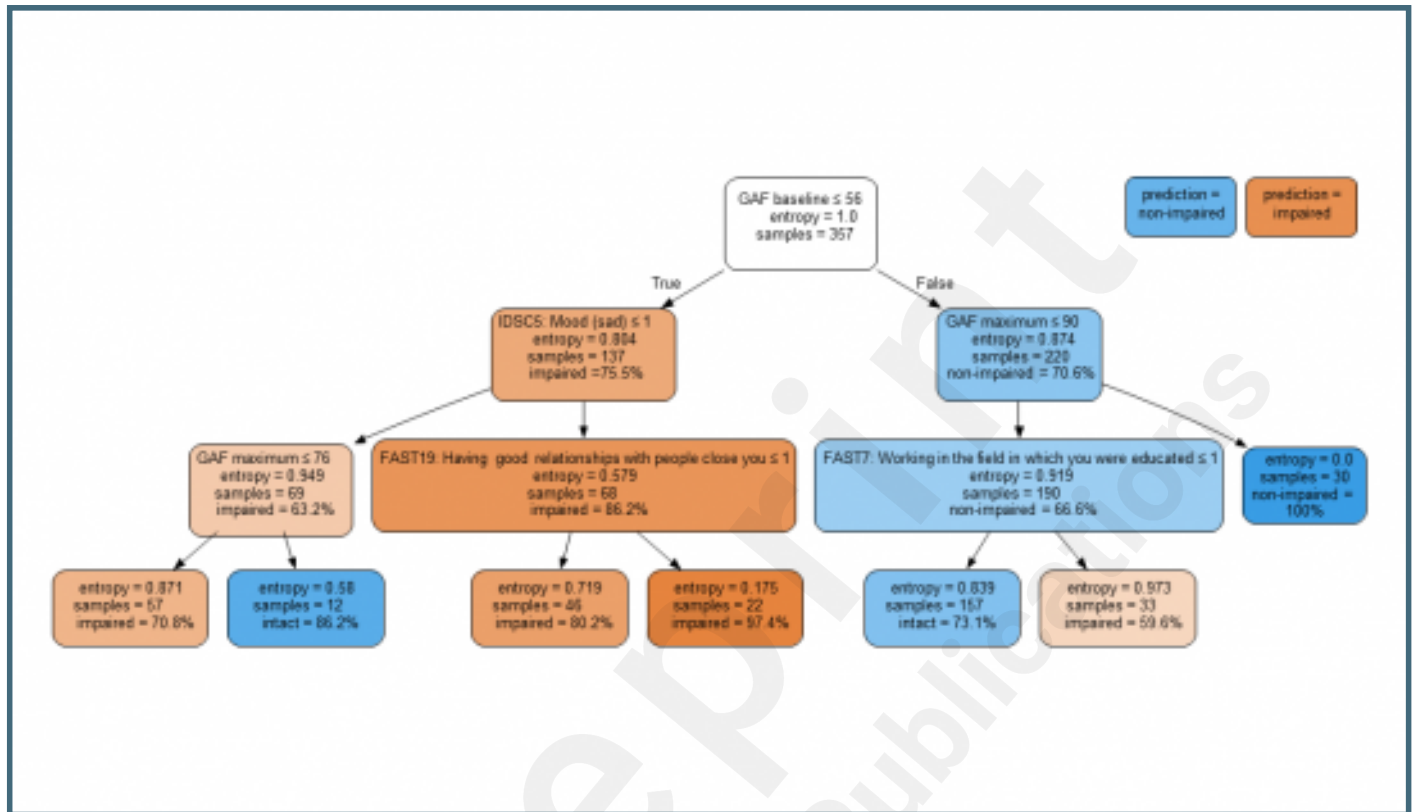
Graphical representation of functional outcome trajectories. The nodes of the Sankey plot represent the GAF values (lower value represents higher impairment). Defining the impaired outcome as GAF ≥ 60 at year 1 and/or year 2 selected trajectories, that remained stable at a moderately impaired level of functioning, as well as trajectories that deteriorated to a moderate level during the follow-up.



The ranking of clinical features as defined by the sign-based consistency. The red line indicates the significance level $p < 0.05$.



Decision tree predicting the impaired functional outcome within 2-year follow-up. The arrows pointing to the left indicate that the condition is fulfilled (true), the arrows to the right indicate not fulfilled (false). Colors indicate the decision being made by the tree when condition was fulfilled. The decision tree assigns the leaf label according to the corresponding majority class in that leaf. Number of samples and percentage of individuals with impaired outcome within each node / leaf is being displayed.



Multimedia Appendixes

Untitled.

URL: <http://asset.jmir.pub/assets/793bd01360963d7b3784d5ef309472b5.docx>

