Psyche and pixels: AI-generated images and their impact on stigma of mental illnesses in Germany

Janine Grimmer, Naiiri Khorikian-Ghazari, Alkomiet Hasan, Noa Lynn Hartmann, Lena Schoch, Sophie-Kathrin Greiner, Stefan Leucht, Irina Papazova

Angaben zur Veröffentlichung / Publication details:

Grimmer, Janine, Naiiri Khorikian-Ghazari, Alkomiet Hasan, Noa Lynn Hartmann, Lena Schoch, Sophie-Kathrin Greiner, Stefan Leucht, and Irina Papazova. 2026. "Psyche and pixels: AI-generated images and their impact on stigma of mental illnesses in Germany." *Journal of Psychiatric Research* 199: 276–83. https://doi.org/10.1016/j.jpsychires.2026.04.029.

Journal of Psychiatric Research 199 (2026) 276–283

Contents lists available at ScienceDirect

Journal of Psychiatric Research

journal homepage: www.elsevier.com/locate/jpsychires

Psyche and pixels: AI-generated images and their impact on stigma of mental illnesses in Germany

Janine Grimmer [a,*], Naiiri Khorikian-Ghazari [a], Alkomiet Hasan [a,b], Noa Lynn Hartmann [a], Lena Schoch [a], Sophie-Kathrin Greiner [a], Stefan Leucht [b,c], Irina Papazova [a]

[a] Department of Psychiatry, Psychotherapy and Psychosomatics, Medical Faculty, University of Augsburg, BKH Augsburg, Augsburg, Germany
[b] DZPG (German Center for Mental Health), partner site München/Augsburg, Germany
[c] Department of Psychiatry and Psychotherapy, TUM School of Medicine and Health, Technical University of Munich, Munich, Germany

ARTICLE INFO

Keywords:
AI-Generated images
Stigma
Severe mental illness
Psychiatric institution
Designer
DALL-E 3
Midjourney V6

ABSTRACT

Artificial intelligence (AI) has rapidly evolved in recent years. Image generating chatbots enjoy great popularity. But they bear the risk of maintaining or enhancing preexisting stereotypes and stigma against severe mental illnesses. The objective of the study is to investigate how AI-generated imagery on psychiatric diseases and institutions are perceived compared to (non-psychiatric) medical ones. Using the chatbots Designer, DALL-E 3 and Midjourney V6 we created images associated with psychiatric and medical contexts (disease, institution, incident). Participants (N = 239, 75.31% women) rated images generated by one of the chatbots upon a self-assessment manikin (SAM) rating, an adjective and emotion rating. Furthermore, participants were asked to title the images. Images containing psychiatric scenes were perceived as more negative and more arousing than other medical scenes (all $p < .001$). Further, they were often rated as having less control over the situation. Fear and anger were more often elicited by psychiatric than medical scenes. Psychiatric images were rated as more threatening and scarier (all $p \leq .001$). In sum, the perception of AI-generated images of psychiatric terms was aligned with pre-existing stigmatizing attitudes. This is the first study to systematically investigate affective perception and potential stigmatizing effects of AI-generated images in a psychiatric context. It raises important discussion about user information, usage guidelines and stricter regulations for generative AI.

1. Introduction

In recent years, artificial intelligence (AI) has been developed rapidly, and integrated into our daily life. Text-to-image generation has emerged as one of the most popular functions of generative AI with more than 34 million images being generated daily (Press, 2024). AI is stereotypically perceived as being fair and more objective than humans due to the machine-like appearance without subjective or emotional influences. However, previous studies have indicated bias (Rejmaniak, 2021), probably because AI models are at that stage often trained on data (mainly images and text) generated by humans. Therefore, AI outputs could reinforce pre-existing biases and stigmas against vulnerable groups. This trend is possibly more pronounced when the generated images become more realistic (Nightingale and Farid, 2022; Vázquez and Garrido-Merchán, 2024). Therefore, we aim to investigate how AI-generated imagery of severe mental illnesses (SMIs) and psychiatric institutions is perceived and its potential to exacerbate stigma.

People with a severe mental illness (SMI) suffer from stigmatization as they are often depicted as being aggressive, dangerous, unpredictable or incompetent in the public (Thornicroft et al., 2022). Stigmatization is evident in attitudes and languages (Howell et al., 2014; Rose et al., 2007) and can arise from stereotypes and prejudice against or negative attitudes towards a specific group (Thornicroft et al., 2022). These attitudes can lead to discrimination regarding employment, housing and health care (Rüsch et al., 2005). If internalized (self-stigma), stigma can reduce the person's self-esteem (Rüsch et al., 2005) and thus lead to decreased self-efficacy, avoidance of or delayed help-seeking behavior and treatment (Franz et al., 2010; Thornicroft, 2008).

(Social) Media content is a key factor for stigmatization as it is accessible easily and produces large amounts of data. Media can reinforce or decrease stigma against SMI depending on how they report about this topic (Thornicroft et al., 2022). AI algorithms use this large amount of media data to learn and generate output. However, even big data sets are shown to be incomplete and might not be representative for

* Corresponding author.
E-mail address: janine.grimmer@med.uni-augsburg.de (J. Grimmer).

https://doi.org/10.1016/j.jpsychires.2026.04.029
Received 17 October 2025; Received in revised form 1 April 2026; Accepted 22 April 2026
Available online 6 May 2026

people living with a SMI, thereby increasing the risk of being discriminated against in health care or employment (Monteith and Glenn, 2016).

As the use of AI increases, attitudes, prejudices, and stereotypes from (social) media are becoming more widespread, potentially intensifying stigmatization. In text-generating tasks, AIs maintained or even amplified stereotypes, gender and racial biases (Ananya, 2024; Caliskan et al., 2017; Lin et al., 2022; Mei et al., 2023; Nadeem et al., 2021; Obermeyer et al., 2019). Image-generating AI chatbots use related large language models as text-generating chatbots to learn text-image associations. Hence, it can be assumed that similar stigmatization will be present in the generated imagery (Saharia et al., 2022; van Kolfschooten and Pilottin, 2024). If AI-generated images become more prominent, this cannot only negatively influence help-seeking but also the interaction of health care providers with specific patient groups (van Kolfschooten and Pilottin, 2024).

In recent years, studies on stigmatization in AI-generated images have increased. For instance, King (2022) prompted the chatbot Midjourney with the word schizophrenia to create twelve images. They showed unnatural, partly horrifying figures that enhance existing stereotypes (King, 2022). A small exploratory study using the chatbots Midjourney and ChatGPT-4 demonstrated that a non-detailed and neutral formulation led to bias as patients who needed a psychiatric treatment were depicted as depressed only (van Kolfschooten and Pilottin, 2024). In another study, the chatbots Midjourney and DALL-E 3 were prompted with different mental disorders using lay terms, e.g. anxiety instead of agoraphobia. Stigmas concerning mental illness were reinforced, as images of people living with a narcissistic personality disorder were portrayed either with a mirror or a clown's mask, images concerning autism were colorful, whereas images portraying depression were gloomy (Flathers et al., 2024). Images of forensic patients living with an antisocial personality disorder created by the chatbot Creator in Bing were often depicted with weapons and a mask (Tortora, 2024). Further, a comparison of real-world epidemiology and generated images by the chatbots Bing Image Generator and Imagine showed inaccuracies concerning gender ratio for psychiatric diagnoses. Moreover, Bing revealed a racial bias depicting patients living with SMIs as mostly White, whereas Imagine revealed an age stereotype, portraying more older people than in the real-world (Wiegand et al., 2025).

Despite first qualitative reports on AI image generators reinforcing stereotypes and stigmas against people living with a SMI, systematic evidence is still scarce. The aim of the present study is to address this issue and to investigate quantitatively and qualitatively whether AI-generated images of psychiatric terms are perceived as more negative or stigmatizing compared to other hospital scenes as a control condition. We do not investigate personal stigma, i.e. participants' attitudes towards people living with a SMI or psychiatric institutions. Instead, we focus on perceived stigma in media i.e. the potential activation of stereotypes and prejudices through AI-generated images. Using three chatbots, we generated images of psychiatric and somatic contexts (e.g. diseases, ward). In an anonymous online-study in Germany, participants evaluated the evoked emotional state and interpreted the images. We hypothesized that images of psychiatric terms are perceived in alignment with pre-existing stereotypes and stigmas and evoke more negative emotions compared to images of medical terms.

2. Methods

2.1. Procedure

The anonymous online study was conducted on the online platform SoSci-Survey (version 3.6.12) between 13th November 2024 and 14th February 2025. An invitation to participate was sent via email distribution lists to psychiatric hospitals, mostly in Bavaria, to German psychiatric university hospitals, and to university lists. We further recruited participants through social networks. To avoid social desirability bias, we told participants that the study aimed to investigate how AI-generated images are perceived. The image prompts were not disclosed.

In a within-subject design (medical vs. psychiatric images), participants were randomly assigned to one of three groups. Each group viewed the generated images of one AI only. All participants were informed of the procedures and gave informed consent. After filling out demographic data, they were asked to rate the images on self-assessment manikin (SAM) rating scales. Then, they rated the images on adjective scales, selected evoked emotions, and provided a title for the image. Finally, they decided whether the images stigmatize specific groups.

Medical and psychiatric images were displayed alternately in a fixed order, except for Designer, where once two medical images were presented consequently. Study protocol was written according to the rules of the Declaration of Helsinki of 1975, revised in 2008, and approved by the local ethics committee (Medical Faculty of the Ludwig-Maximillian-University Munich; Number 25-0082-KB) and by the local data protection officer.

2.2. Sample

The sample consisted of 239 participants (75.31% women). Mean age was 38.50 years ($SD = 13.95$; Designer: $M = 37.52$, $SD = 14.47$; DALL-E 3: $M = 39.45$, $SD = 13.71$; Midjourney: $M = 38.45$, $SD = 13.76$). The majority (89.12%) works in the healthcare sector, and 62.34% work in a clinic for psychiatry or psychosomatics. Most participants (74.48%) had at least one relative living with a psychiatric disease, whereas 20.92% indicated having experienced a psychiatric disease. Table 1 shows the demographic characteristics of the sample, for further details see Table S1 in supplement A.

2.3. Material

Two researchers (IP, NKG) prompted the chatbots Designer (Microsoft, release date: 14.09.2022), DALL-E 3 (OpenAI, release date: 20./21.09.2023), and Midjourney (Discord, V6, release date: 21./20.12.2023) to generate various realistic medical scenes between 06.06.-03.07.2024. A detailed description can be found in the recently published work from our group (Papazova et al., 2025). Prompts were given only once and were structured as following: “Please generate a realistic image of a/an token.” Final used tokens for psychiatric and medical scenes per AI are shown in Table 2. If the chatbot produced more than one image per prompt, only the first four images were used to prevent the AI from generating user-adapted images. This resulted in 116 images (41 Designer, 12 DALL-E 3, 63 Midjourney) for the following

Table 1
Demographic characteristics of the sample.

Characteristic	%
Expertise	
Non-experts	7.53
Patient	3.35
Expert healthcare sector	19.67
Expert psychiatric or psychosomatic clinic	51.88
Patient and expert healthcare sector	7.11
Patient and expert psychiatric or psychosomatic clinic	10.46
Relationship to affected individuals	
Relation in the first degree	18.41
Relation in the second degree	27.61
Relation in the third degree	12.55
Spouse or life partner	11.30
Friends	32.21
Colleges	23.43
Neighbor	7.11
Others/not specified	14.23
Treatment experience	
Somatic inpatient	35.15
Psychiatric inpatient	5.86
Psychiatric or psychotherapeutic outpatient	16.32
No clinical treatment	53.97

Table 2
Tokens used per AI.

AI	Tokens
Designer	Incident in a hospital, incident in a mental health institution, severe illness, severe mental illness, hospital ward, CPR session, electroconvulsive therapy (ECT) session
DALL-E 3	Incident in a hospital, incident in a mental health institution, hospital ward, psychiatric ward, CPR session, ECT session
Midjourney	Incident in a hospital, incident in a mental health institution, severe illness, severe mental illness, hospital ward, psychiatric ward, CPR session, ECT session

selection process. For each chatbot, one image per prompt was selected randomly using a random samples algorithm from R (function sample from package base) for the online survey. This resulted in seven images for Designer, six images for DALL-E 3, and eight images for Midjourney. Fig. 1 depicts exemplary AI-generated images, all used images can be found in supplement A (Table S2 – Table S5).

2.3.1. Adjective rating

We developed a two-pole scale for this study. Individuals moved a slider across a visual analogue scale ranging from 1 to 101 with opposite adjectives at the extremes. The slider's position indicates the extent of agreement with the adjectives. Trying to illustrate common prejudices in psychiatric environments (e.g., as found in the media), the following pairs were built: safe (1) – threatening (101), inviting (1) – scary (101), realistic (1) – unrealistic (101), serious/sensible (1) – silly/childish (101), and neat (1) – rundown (101). Higher scores indicate a more negative attitude towards the image.

2.3.2. SAM rating

We used the self-assessment manikin (SAM) rating from Bradley and Lang (1994) as a non-verbal tool to assess valence (pleasantness), arousal (activation) and dominance (control) of an image. Each scale is rated on a five-point scale, having a neutral value for both valence and arousal in the middle of the scale. Valence ranges from negative (1) over neutral (3) to positive (5). Arousal ranges from relaxed, calm (1) over neutral/neither nor (3) to excited, nervous (5). Dominance ranges from no control (1) to being completely in control of the situation (5) (Bradley and Lang, 1994).

2.3.3. Emotion selection

To evaluate the emotions evoked by viewing the images, basic emotions identified by Ekman (1992) (anger, surprise, disgust, happiness, fear, and sadness) were presented, as well as feeling of shame often associated with SMI. The selection is designed as a forced-choice task in which participants must indicate at least one emotion, but have the opportunity to select more than one. (Non-)Selections of an emotion are counted and then compared between psychiatric and medical images.

2.3.4. Title

We asked participants to create a title using the questions "What is this image of? What title could this image have?" to assess both the subjective interpretation of the AI-generated images and possibly activated stigmatizing attitudes. The words used in the titles are grouped by context. The ten most frequently used words in psychiatric and medical contexts will be used as score. Furthermore, a three-part scale (0-2) was used to indicate the match of the title and the content displayed ranging from no (0%), over 50% (1) to 100% (2) match. The frequency of match grade is used to indicate the overall match for each context.

2.4. Data analysis

Data analysis was done using RStudio (2022.07.2 + 576, R 4.2.1, packages: readxl, dplyr, psych, gmodels, rcompanion, GGally, car, MVN, MultNonParam, mvnormtest, rstatix, effectsize, plotrix, confintr, akima,

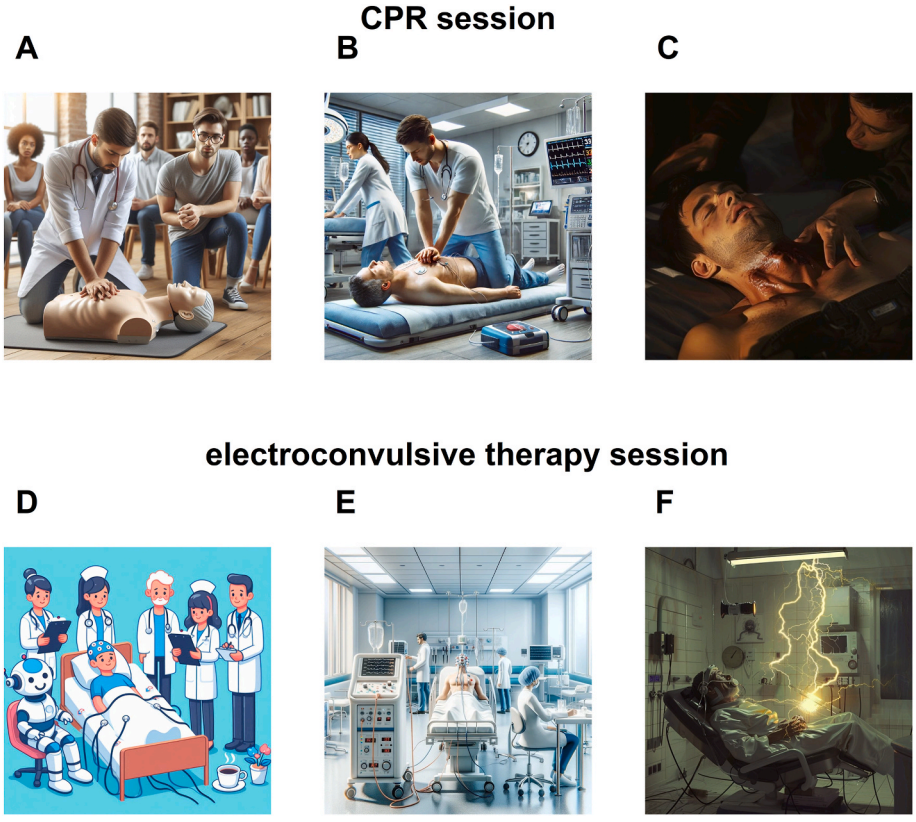

Fig. 1. Examples of AI-generated images
Note. Example of a medical (top) and a psychiatric image (bottom) for each chatbot (from left to right: Designer (A, D), DALL-E 3 (B, E), Midjourney (C, F)).

WRS repos = "http://R-Forge.R-project.org", reshape, TOSTER, ggpubr, ggplot2, ggsignif, plotrix, stats). The assumptions for homoskedasticity, multivariate normality, and homogeneity of variance-covariance matrices were not fulfilled. Therefore, a robust rank-based MANOVA using Munzel and Brunner's method (Munzel and Brunner, 2000) was employed to test the differences in adjective and SAM ratings between psychiatric and somatic prompts. The method is implemented in R as mulrank function (package WRS, Wilcox, 2012) and does not provide an effect size. Post-hoc analyses were conducted using a signed rank Wilcoxon test, since only two dependent groups were compared. Chi-squared tests were used to compute the relation between psychiatric and somatic prompts and evoked emotions and title matches. Post-hoc tests were conducted using standardized residuals for comparisons of more than two groups. To prevent a multiple comparisons problem, the alpha value for each of the 18 calculated tests (emotion rating (7), adjective rating (5), SAM-rating (3), title matching (3)) was set to $\alpha = .00278$ per AI.

3. Results

In the following sections, results will be presented in the chatbot order Designer, DALL-E 3, and Midjourney. Additional results to stigmatized groups can be found in the section Results: Stigmatized groups in supplement A.

The robust MANOVA showed a significant effect for *Designer* ($F = 51.08$, $p < .001$), for *DALLE-E 3* ($F = 9.22$, $p < .001$); and for *Midjourney* ($F = 115.38$, $p < .001$).

3.1. Adjectives

Fig. 2 shows the descriptive data of adjective ratings. For the data analysis of *Designer*, one image (prompt: hospital ward, as psychiatric ward could not be generated) was removed from all analyses as post-hoc paired Wilcoxon tests need the same number of observations per group.

Post hoc analyses for *Designer* revealed a significant difference between psychiatric and medical scenes for all adjectives, indicating higher ratings for images depicting psychiatric scenes. Table 3 shows the post-hoc test results of all AIs. The results of post hoc analyses for *DALL-E 3* demonstrated a significant difference between psychiatric and medical scenes for the adjectives threatening and scary, indicating higher ratings for psychiatric scenes. All other differences were not significant. Post hoc analyses revealed significant differences for all adjectives, indicating higher values for psychiatric scenes generated by *Midjourney*.

3.2. SAM rating

Post-hoc analyses showed a significant difference for valence, arousal, and dominance of images generated by *Designer* (see Table 4 for the results of the SAM rating for each AI and Fig. 3 for the descriptive data). Thus, psychiatric images are perceived as more negative, more arousing and having less control over the situation compared to medical scenes.

The post-hoc analyses of the SAM rating data from *DALL-E 3* showed a significant difference for valence and arousal but not for dominance. Psychiatric scenes are perceived as more negative and more arousing than medical scenes.

Furthermore, post-hoc analyses of images generated by *Midjourney* revealed a significant difference for valence, arousal, and dominance, showing that psychiatric images are perceived as more negative, more arousing and having less control over the situation.

3.3. Emotion selection

For the analysis of evoked emotions by the image, chi-squared tests were used for each AI. Table 5 shows the frequencies (for more details see Table S6 in supplement A) and statistical parameters.

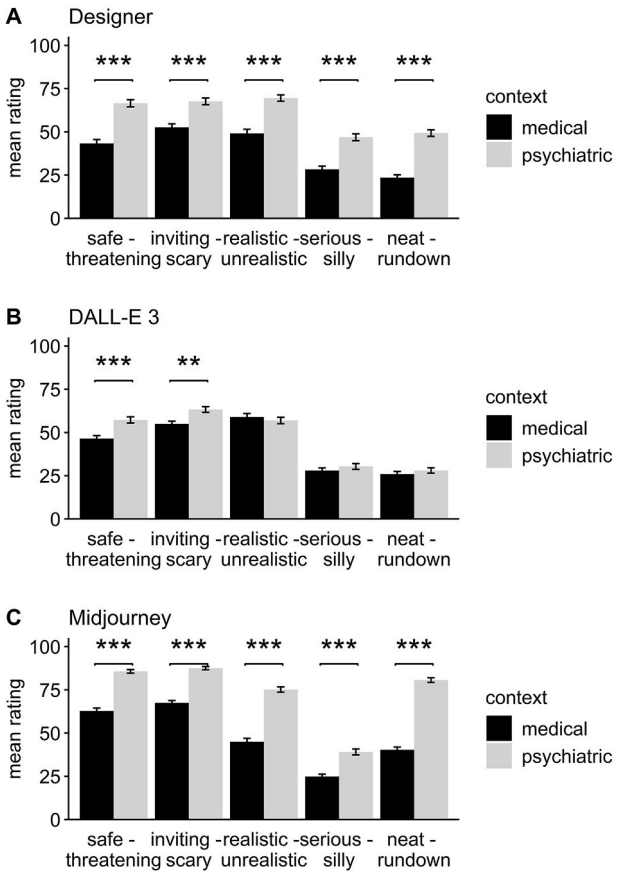

Fig. 2. Mean ratings of adjectives
Note. Mean ratings of adjectives describing the image per AI (scale 1-101). Higher values represent the second (negatively tuned) adjective and a higher stigmatization. Error bars indicate ±1 standard error. $^{**}p < .0027$, $^{***}p < .001$, $\alpha = .00278$.

For *Designer* a significant relation between the image context and evoked anger and fear was shown: psychiatric images evoke more fear and anger than medical images. There was no significant relation between other emotions and image context.

The analysis of evoked emotions by images generated by *DALL-E 3* revealed a significant relation between image context and evoked fear. Fear is slightly more often evoked in case of psychiatric images and less often in case of medical ones. No further relations reached significance.

The analysis for *Midjourney* yielded a significant relation between the image context and evoked anger, sadness, disgust, and happiness. Medical images evoke more often sadness and happiness and less often anger and disgust than psychiatric images. There was no significant relation between other emotions and image context.

3.4. Title analysis

Medical scenes were more often associated with specific rooms (e.g. patient room, multi-bed room, hospital ward) or actions in a hospital (e.g. care, everyday life in a hospital, medical examination), and crowding (e.g. overcrowded, mass care). In contrast, psychiatric scenes were described more negatively (e.g. run-down, catastrophe, without privacy, prison, desperate, attack, human trials). In general, qualitative analyses indicated that images generated by *DALL-E 3* were described using more neutral terms (e.g. resuscitation, hospital, intensive care unit, private clinic, or electroencephalography; for more details see Table S8 in supplement A).

As shown in Table 6, the degree of matching between the content

Table 3
Post hoc analyses for adjective ratings of psychiatric and medical scenes.

AI	Adjective[a]	Mdn_{med}	Mdn_{psy}	*V*	*p*	*r*
Designer	Threatening	40.00	75.00	3985.50	<.001***	.545
	Scary	50.00	75.00	5851.00	<.001***	.433
	Unrealistic	41.00	76.00	5871.50	<.001***	.409
	Silly/childish	20.00	43.00	6202.50	<.001***	.410
	Rundown	16.00	50.00	1722.50	<.001***	.728
DALLE-3	Threatening	46.00	60.00	11372.00	<.001***	.251
	Scary	56.00	67.00	5851.00	.001**	.200
	Unrealistic	65.00	62.00	16874.00	.639	.030
	Silly/childish	22.00	24.50	13824.00	.185	.084
	Rundown	20.00	23.00	13758.00	.448	.049
Midjourney	Threatening	67.50	91.00	4271.50	<.001***	.638
	Scary	69.00	94.00	2823.00	<.001***	.710
	Unrealistic	37.50	81.00	3685.50	<.001***	.668
	Silly/childish	19.00	36.00	6182.00	<.001***	.467
	Rundown	20.00	23.00	6184.00	<.001***	.784

[a] adjective pairs: safe - threatening, inviting - scary, realistic – unrealistic, serious/sensible - silly/childish, and neat – rundown, with higher values being in favor of the latter one. *Mdn* = Median, med = medical, psy = psychiatric. *V*= statistical parameter for the signed rank Wilcoxon test. **p< .0027, ***p < .001, α = .00278. *r* = effect size (correlation).

Table 4
Post hoc analyses for SAM rating of medical vs. psychiatric scenes.

AI	Dimension	Mdn_{med}	Mdn_{psy}	*V*	*p*	*r*
Designer	Valence	3.00	2.00	10446.00	<.001***	.492
	Arousal	2.00	4.00	1716.00	<.001***	.580
	Dominance	3.00	2.00	9985.50	<.001***	.443
DALL-E 3	Valence	3.00	2.00	11032.00	<.001***	.220
	Arousal	3.00	3.00	6513.50	<.001***	.231
	Dominance	3.00	3.00	11259.00	.100	.099
Midjourney	Valence	2.00	1.00	14592.00	<.001***	.569
	Arousal	3.00	4.00	3084.50	<.001***	.517
	Dominance	2.00	2.00	11966.00	<.001***	.196

Note. Mdn = Median, med = medical, psy = psychiatric. *V*= statistical parameter for the signed rank Wilcoxon test. ***p < .001, α = .00278. *r* = effect size (correlation).

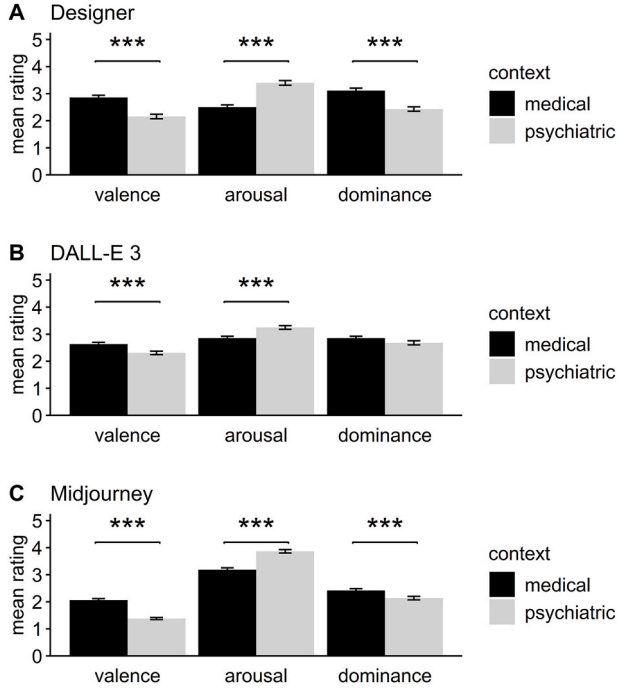

Fig. 3. Mean ratings of SAM rating for each AI
Note. Higher values represent a more positive (valence), more arousing (arousal) perception of the image, and being in control of the situation (dominance). Error bars represent ±1 standard error. ***p < .001, α = .00278.

Table 5
Absolute frequency of ticked emotion and chi-squared test for the relation between image context and emotion.

AI	emotion evoked	med:psy	X^2	*df*	*p*	*Cramer's V*
Designer	Fear	133:133	10.68	1	.001**	.139
(*N* = 553)	Anger	20:42	17.66	1	<.001***	.179
	Sadness	97:70	.09	1	.769	.013
	Disgust	32:36	3.22	1	.073	.076
	Surprise	72:54	.00	1	1.000	.000
	Happiness	76:43	2.80	1	.094	.071
	Shame	26:29	2.43	1	.119	.066
DALL-E 3	Fear	134: 168	9.23	1	.002**	.134
(*N* = 516)	Anger	33:22	2.46	1	.154	.069
	Sadness	84:58	6.57	1	.010	.113
	Disgust	21:20	.03	1	.871	.007
	Surprise	72:75	.09	1	.770	.013
	Happiness	35:22	3.33	1	.069	.080
	Shame	24:15	2.25	1	.133	.066
Midjourney	Fear	177:207	6.67	1	.010	.106
(*N* = 592)	Anger	23:58	17.52	1	<.001***	.172
	Sadness	131:83	16.86	1	<.001***	.169
	Disgust	71:153	48.29	1	<.001***	.286
	Surprise	53:57	.18	1	.673	.017
	Happiness	14:2	9.25	1	.002**	.125
	Shame	21:26	.58	1	.447	.031

Note. N= number of ratings (participants * number of images). med = medical. psy = psychiatric. med:psy: relation of ticked emotion depending on the context. **p< .0027, ***p < .001, α = .00278. *Cramer's V* = effect size.

displayed and the title given to the image is highest for *DALL-E 3*, followed by *Midjourney*, and lowest for *Designer*. For the analysis of title matching (fully, half, no) by context (medical, psychiatric), chi-squared

Table 6
Frequencies and chi-squared tests for title matching of medical vs. psychiatric scenes.

AI		*n*	med:psy	X^2	*df*	*p*	*Cramer's V*
Designer				34.57	2	<.001***	.250
(*N* = 553)	Full match	81	67:14				
	Half match	153	96:57				
	No match	319	153:166				
DALL-E 3				62.93	2	<.001***	.349
(*N* = 516)	Full match	246	168:78				
	Half match	116	39:77				
	No match	154	51:103				
Midjourney				107.52	2	<.001***	.426
(*N* = 592)	Full match	140	111:29				
	Half match	226	128:98				
	No match	226	57:169				

Note. N= number of ratings (participants * number of images). *n*= summed frequency; med = medical. psy = psychiatric. med:psy: relation of title match depending on the context. **p< .0027, ***p < .001, α = .00278. *Cramer's V* = effect size.

tests were used for each AI. For *Designer* a significant result was obtained (see Table 6 for all results). Post-hoc tests using standardized residuals showed less no matches and more 100% matches for medical scenes and more no matches and less 100 % matches for psychiatric scenes than expected.

Again, the chi-squared test for *DALL-E 3* was significant. Post hoc tests revealed less no matches and half-matches, and more 100% matches for medical images than expected. The opposite is true for psychiatric images: there were more no matches and half-matches and less 100% matches than expected.

Lastly, a significant result was shown for *Midjourney*. Post-hoc tests showed less no matching and more 100% matches for medical scenes and more no matches and less 100% matches for psychiatric scenes than expected. Overall, given titles often did not match fully when psychiatric scenes or SMIs were used as prompts across all AIs.

4. Discussion

To our knowledge, this is the first systematic study to investigate the perception of AI-generated images of psychiatric contexts compared to other non-psychiatric hospital scenes and to assess potential stigma. German health care providers, patients, and members of the general population evaluated images generated with the chatbots Designer, Midjourney, and DALL-E 3. Our findings indicate that AI-generated images of psychiatric contexts and SMI are perceived as more negative and arousing, having less control over the situation and more threatening. They were rated as more unrealistic and evoked more negative emotions such as anger. Our results are in line with previous research on stigma against SMI (Flathers et al., 2024; King, 2022). Unsurprisingly, AI-generated images potentially reinforce negative stereotypes and stigmas.

Images displaying a psychiatric context are rated as significantly more threatening and scarier than medical images across all AIs. Additionally, psychiatric images generated by Designer and Midjourney were perceived as more unrealistic, sillier and more rundown than medical ones. Generally, these two chatbots showed medium to large effects for all adjective comparisons, whereas DALL-E 3 reached small effects only for threatening and scary. Even though half of the sample were mental health experts, participants misinterpreted the content of psychiatric images: the generated titles rarely matched the prompts. Consequently, the chatbots were unable to generate images that clearly depicted psychiatric contexts.

Psychiatric images evoked more often anger or fear, our analyses indicate small effects for this result. Only for Midjourney further differences were found showing a small effect: images displaying psychiatric context were rated as evoking more frequently sadness and disgust and less often happiness than medical images. This emphasizes, that Midjourney probably makes greater differences between the contexts and could be more prone to stigmatizing psychiatric institutions and people living with a SMI. Similarly, Midjourney was shown to generate more images associated with negative moods when prompted with "people living with dementia" than when prompted with "elderly people without dementia". However, the total number of images depicting both negative and positive emotions was descriptively higher for DALL-E 3 than for Midjourney (Jintaganon et al., 2025).

Overall, the results demonstrate that people perceive AI-generated images as containing stereotypes about mental illness and psychiatric institutions. The results are not surprising since generative AI uses the same language models as chatbots (van Kolfschooten and Pilottin, 2024). Testing for differences in the stigmatization of chatbots is beyond the scope of this study. However, our descriptive data imply that the extent of stigmatization varies: it was the lowest for DALL-E 3 and the highest for Midjourney regarding adjective descriptions and evoked emotions. DALL-E 3 showed the smallest effect sizes and generated the most realistically perceived images from the given prompts.

Although, we strove for a diversified sample, it consisted mostly of experts, making a comparison between patient, expert and non-experts statistically not meaningful. Expertise probably distorted the title matching in favor of a higher overall matching grade, since health care professionals are more aware of the issue. Moreover, the majority of the sample (75%) were women. Thus, the composition of the sample limits the generalizability of our results. The evaluation of people without contact to a patient or health care professional is important, as stigma arises even more pronounced in this case (Angermeyer and Matschinger, 1997; Phelan et al., 2000).

The prompts used depicted comparable terms in medical or psychiatric care (e.g. hospital or psychiatric ward). To prevent priming, we did not pretest the prompts for emotional valence. We assumed that stigmatizing attitudes could affect the valence, thus invalidating the pretest and limiting its validity. We did not explore prejudice against specific mental disorders. Former studies indicate that, for example, alcohol abuse is more stigmatized than depression (Pescosolido et al., 2021). Investigating specific SMIs might yield more precise insights into the risk of reinforcing stigma using AI. AI is trained on large data sets which may include stereotypes that have been successfully fought against in the real world. It would be interesting to see whether AI-generated images reflect recent destigmatization trends relating to depression (Pescosolido et al., 2021). If not, it could lead to a regression in fighting stigmas and might even be able to reintroduce them in the real world.

To better understand these findings, future research should include larger representative samples allowing to analyze the effects of contact with SMI, gender and other vulnerability group characteristics. Moreover, a potential priming effect could be address by presenting the images in a randomized order. The use of standardized stigma scales before and after imagery presentation could asses the immediate effect of AI images on participants' stigma against SMIs. Future research could also include more chatbots (e.g. Stable Diffusion, Neuroflash, DeepSeek) and a continuous scale to assess basic emotions evoked. A comparison

between real photographs with AI-generated images of both contexts could reveal if AI amplifies existing stereotypes or simply reproduces stereotypes present in real-word imagery.

A major advantage of the present study is the systematic approach with the standardized SAM rating and basic emotions to assess AI-generated images increasing comparability. To our knowledge, this is the first study that attempted to go beyond the qualitive interpretation of AI-generated images by combining quantitative data with qualitative data. Another novel aspect is the inclusion of psychiatric institutions and contexts to further explore stereotypes. Moreover, we investigated different, in the operating system Windows integrated, very popular chatbots (Kumar, 2025; OpenAI, 2022; Similarweb LTD, 2025a; Similarweb LTD, 2025b), showing the varying extent of stigmatization.

AI develops rapidly and changes very fast so that our conclusions are limited to the current status. Our results stress the importance of informing users of the risk of misinformation and stigmatization of generative AI at this time. Alarmingly, even knowing about this problem and adding words to prevent or circumvent stigmas does not prevent the generation of stigmatizing images (Bianchi et al., 2023). However, raising awareness of this effect is crucial. It should be integrated into curricula as early as possible to increase AI usage literacy. Furthermore, recommendations on how to use AI should be easily accessible for users.

The European Union adopted an act which regulates the use of AI aiming to make it non-discriminatory. Generative AI is classified as limited risky and has to follow transparency requirements meaning that AI-generated content must be labeled (Regulation (EU) 2024/1689, 2024). However, this cannot prevent the generation of stigmas or misinformation. Respectively, AI chatbots cannot be recommended as source of information on SMIs and psychiatric institutions at the moment. Moreover, AIs need to be trained on data as stigma-free as possible that correspond to the real-world to generate more realistic and less stigmatizing images. However, measures to reduce bias and stereotypes such as enlarged training data sets or using filters to remove biased or stigmatizing content revealed mixed results. But the latter could be improved with human involvement (Ananya, 2024). This is crucial because people use AI to get medical information.

Designer and DALL-E 3 did not create all images as content policy guidelines were triggered. It remains unclear whether such restrictions reduce or increase stigma. Nevertheless, stricter ethical standards for creating images displaying mental health content are needed as well as more transparency about data processing. If a chatbot repeatedly generates highly stigmatizing content, another conceivable measure could be a restriction of usage to prevent the distribution of such content. Misinformation and amplified stigmas may result in self-stigma and delayed help-seeking behavior which causes suffering and sometimes higher healthcare costs as diseases become chronic. People living with a SMI are already a highly vulnerable group who have a higher risk of being disadvantaged. This might be increased through stigmatizing AI-generated content.

4.1. Conclusion

Our participants perceive AI-generated images depicting SMIs or psychiatric institutions as negative. The results demonstrate the need for a stricter control, a more transparent approach which and how data are used to generate images, and a dissemination of information that generated images might enhance negative perceptions, stereotypes and stigmas overall. AI-generated images can delay help-seeking and prevent the inclusion of and the reduction of existing stigmas of people living with a mental illness and psychiatric institutions. This has devastating consequences for the mental health of those affected.

CRediT authorship contribution statement

Janine Grimmer: Writing – review & editing, Writing – original draft, Visualization, Software, Investigation, Formal analysis, Data curation. **Naiiri Khorikian-Ghazari:** Writing – review & editing, Methodology, Conceptualization. **Alkomiet Hasan:** Writing – review & editing, Funding acquisition, Conceptualization. **Noa Lynn Hartmann:** Writing – review & editing, Methodology. **Lena Schoch:** Writing – review & editing, Methodology. **Sophie-Kathrin Greiner:** Writing – review & editing, Methodology. **Stefan Leucht:** Writing – review & editing, Methodology, Funding acquisition. **Irina Papazova:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Data statement

The data underlying this research paper are for data protection reasons not deposited in a data repository. Upon request the data can be made available.

Funding

This work was funded by the German Center for Mental Health (FKZ 01EE2503C, MUC7).

Declaration of competing interest

AH was or is a member of the advisory board and has received paid speakership from Boehringer Ingelheim, Lundbeck, Otsuka, Rovi, Teva (no speakership), AbbVie, and Recordati; is affiliated with AbbVie and Advanz; and is an editor of the German AWMF Guidelines for Schizophrenia. SKG is advisor to the GOLDKIND Foundation. SL has received honoraria for service as a consultant or adviser and/or for lectures from Angelini, Apsen, Boehringer Ingelheim, Eisai, Ekademia, Gedeon Richter, Janssen, Johnson & Johnson, Karuna, Kynexis, LB Pharma, LTS Lohmann, Lundbeck, MSD, Medichem, Medscape, Mitsubishi, Neurotorium, NovaNordisk, Otsuka, Recordati, Sandoz, Sanofi-Aventis, Sunovion, Rovi and TEVA.

All other authors have nothing to declare.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpsychires.2026.04.029.

References

Ananya, 2024. AI image generators often give racist and sexist results: can they be fixed? Nat 627, 722–725. https://doi.org/10.1038/d41586-024-00674-9.

Angermeyer, M.C., Matschinger, H., 1997. Social distance towards the mentally ill: results of representative surveys in the Federal Republic of Germany. Psychol. Med. 27, 131–141.

Bianchi, F., Ladhak, F., Hashimoto, T., Kalluri, P., Cheng, M., Jurafsky, D., Durmus, E., Nozza, D., Zou, J., Caliskan, A., 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In: Proc. of the 2023 ACM Conf. on Fairness, Accountability, and Transparency. Chicago, USA, pp. 1493–1504. https://doi.org/10.1145/3593013.3594095.

Bradley, M., Lang, P.J., 1994. Measuring emotion: the self-assessment manikin and the semantic differential. J. Behav. Ther. & Exp. Psychiatr. 25, 49–59.

Caliskan, A., Bryson, J.J., Narayanan, A., 2017. Semantics derived automatically from language corpora necessarily contain human biases. Sci. 356, 183–186. https://doi.org/10.1126/science.aal4230.

Ekman, P., 1992. An argument for basic emotions. Cognit. Emot. 6, 169–200.

Flathers, M., Smith, G., Wagner, E., Fisher, C.E., Torous, J., 2024. AI depictions of psychiatric diagnoses: a preliminary study of generative image outputs in midjourney V.6 and DALL-E 3. BMJ ment. Health 27, 1–6. https://doi.org/10.1136/bmjment-2024-301298.

Franz, L., Carter, T., Leiner, A.S., Bergner, E., Thompson, N.J., Compton, M.T., 2010. Stigma and treatment delay in first-episode psychosis: a grounded theory study. Early Interv. Psychiatr. 4, 47–56. https://doi.org/10.1111/j.1751-7893.2009.00155.x.

Howell, A.J., Ulan, J.A., Powell, R.A., 2014. Essentialist beliefs, stigmatizing attitudes, and low empathy predict greater endorsement of noun labels applied to people with mental disorders. Personality and Individ. Differences 66, 33–38. https://doi.org/10.1016/j.paid.2014.03.008.

Jintaganon, N., Osinga, C.J., Steijger, D., De Vugt, M., Neal, D., 2025. Biases in an Artificial Intelligence Image-generator's Depictions of Healthy Aging and Alzheimer's. https://doi.org/10.2139/ssrn.5177622.

King, M., 2022. Harmful biases in artificial intelligence. Lancet Psychiatry 9, e48.
Kumar, N., 2025. Midjourney statistics 2025 – users & revenue data. Midjourney Statistics in 2025. https://www.demandsage.com/midjourney-statistics/?utm_source=chatgpt.com. (Accessed 25 June 2025).
Lin, I.W., Njoo, L., Field, A., Sharma, A., Reinecke, K., Althoff, T., Tsvetkov, Y., 2022. Gendered mental health stigma in masked Language models. In: Proc. of the 2022 Conf. on Empir. Methods in Nat. Lang. Process. Abu Dhabi, United Arab Emirates, pp. 2152–2170.
Mei, K.X., Fereidooni, S., Caliskan, A., 2023. Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. In: Proc. of the 2023 ACM Conf. on Fairness, Accountability, and Transparency, pp. 1699–1710.
Monteith, S., Glenn, T., 2016. Automated decision-making and big data: concerns for people with mental illness. Curr. Psychiatry Rep. 18, 1–12. https://doi.org/10.1007/s11920-016-0746-6.
Munzel, U., Brunner, E., 2000. Nonparametric tests in the unbalanced multivariate one-way design. Biometrical J 42, 837–854.
Nadeem, M., Bethke, A., Reddy, S., 2021. StereoSet: measuring stereotypical bias in pretrained language models. In: Proc. of the 59th Annual Meet. of the Assoc. for Comput. Linguist. and the 11th Int. Joint Conf. on Nat. Lang. Process. Assoc. for. Comput. Linguist., pp. 5356–5371
Nightingale, S.J., Farid, H., 2022. AI-synthesized faces are indistinguishable from real faces and more trustworthy. Proc. Natl. Acad. Sci. 119, e2120481119. https://doi.org/10.1073/pnas.2120481119.
Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S., 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Sci 366, 447–453. https://doi.org/10.1126/science.aax2342.
OpenAI, 2022. DALL·E now available without waitlist. URL. https://openai.com/index/dall-e-now-available-without-waitlist/. (Accessed 25 June 2025).
Papazova, I., Hasan, A., Khorikian-Ghazari, N., 2025. Biased AI generated images of mental illness: does AI adopt our stigma? Eur. Arch. Psychiatr. Clin. Neurosci. 1–3. https://doi.org/10.1007/s00406-025-01998-x.
Pescosolido, B.A., Halpern-Manners, A., Luo, L., Perry, B., 2021. Trends in public stigma of mental illness in the US, 1996-2018. JAMA Netw. Open 4, e2140202. https://doi.org/10.1001/jamanetworkopen.2021.40202 (Reprinted).
Phelan, J.C., Link, B.G., Stueve, A., Pescosolido, B.A., 2000. Public conceptions of mental illness in 1950 and 1996: what is mental illness and is it to be feared? J. of Health and Soc. Beyond Behav. 41, 188–207. https://doi.org/10.2307/2676305.
Press, G., 2024. 34 million AI images created per day – AI Art Generator Stats 2024. https://whatsthebigdata.com/ai-art-generator-statistics/. (Accessed 14 May 2025).
Regulation (EU) 2024/1689, 2024. Regulation (EU) 2024/1689 of the European Parliament and of the council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending regulations (EC) no 300/2008, (EU) no 167/2013, (EU) no 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). http://data.europa.eu/eli/reg/2024/1689/oj. Pub. L. No. 2024/1689.
Rejmaniak, R., 2021. Bias in artificial intelligence systems. Białostockie Studia Prawnicze 26, 25–42. https://doi.org/10.15290/bsp.2021.26.03.02.
Rose, D., Thornicroft, G., Pinfold, V., Kassam, A., 2007. 250 labels used to stigmatise people with mental illness. BMC Health Services Res 7, 97. https://doi.org/10.1186/1472-6963-7-97.
Rüsch, N., Angermeyer, M.C., Corrigan, P.W., 2005. Mental illness stigma: concepts, consequences, and initiatives to reduce stigma. Eur. Psychiatry 20, 529–539. https://doi.org/10.1016/j.eurpsy.2005.04.004.
Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Gontijo-Lopes, R., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M., 2022. Photorealistic text-to-image diffusion models with deep language understanding. In: Adv. in Neural Inf. Process. Syst., pp. 36479–36494.
Similarweb LTD, 2025a. Microsoft Designer app analytics for June 22. https://www.similarweb.com/app/google/com.microsoft.designer/?utm_source=chatgpt.com#overview. (Accessed 25 June 2025).
Similarweb LTD, 2025b. Microsoft Designer app analytics for June 22. https://www.similarweb.com/app/google/com.microsoft.designer/germany/. (Accessed 25 June 2025).
Thornicroft, G., 2008. Stigma and discrimination limit access to mental health care. Epidemiol. Psichiatr. Soc. 17, 14–19.
Thornicroft, G., Sunkel, C., Aliev, A.A., Brohan, E., el Chammay, R., Davies, K., Demissie, M., Duncan, J., Fekadu, W., Gronholm, P.C., Guerrero, Z., Gurung, D., Habtamu, K., Hanlon, C., Heim, E., Henderson, C., Hijazi, Z., Hoffman, C., Hosny, N., Huang, F.-X., Kline, S., Kohrt, B.A., Lempp, H., Li, J., London, E., Ma, N., Mak, W.W. S., Makhmud, A., Maulik, P.K., Milenova, M., Cano, G.M., Ouali, U., Parry, S., Rangaswamy, T., Rüsch, N., Sabri, T., Sartorius, N., Schulze, M., Stuart, H., Salisbury, T.T., Juan, N.V.S., Votruba, N., Winkler, P., 2022. The Lancet Commission on ending stigma and discrimination in mental health. Lancet 400, 1438–1480. https://doi.org/10.1016/S0140-6736(22)01470-2.
Tortora, L., 2024. Stigmatised representations of forensic psychiatric patients in text-to-image (T2I) generative models. https://doi.org/10.2139/ssrn.5065579.
van Kolfschooten, H., Pilottin, A., 2024. Reinforcing stereotypes in health care through artificial intelligence-generated images: a call for regulation. Mayo Clin Proc Digit. Health 2, 335–341. https://doi.org/10.1016/j.mcpdig.2024.05.004.
Vázquez, A.F. de C., Garrido-Merchán, E.C., 2024. A taxonomy of the biases of the images created by generative artificial intelligence. arXiv preprint arXiv:2407.01556.
Wiegand, T.L.T., Jung, L.B., Gudera, J.A., Schuhmacher, L.S., Moehrle, P., Rischewski, J. F., Mehrzad, P., Jeong, S., Nguyen, L., Poeschla, M., Velezmoro, L.I., Kruk, L., Dimitriadis, K., Koerte, I.K., 2025. Demographic inaccuracies and biases in the depiction of patients by artificial intelligence text-to-image generators. npj Digit. Med. 8, 459. https://doi.org/10.1038/s41746-025-01817-6.
Wilcox, R., 2012. Introduction to Robust Estimation and Hypothesis Testing, third ed. Academic Press, Amsterdam.