npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01817-6

# Demographic inaccuracies and biases in the depiction of patients by artificial intelligence text-to-image generators



Tim Luca Till Wiegand[1,2,3,4] ✉, Leonard Ben Jung[1,2,5], Jonas Anton Gudera[2,6], Luisa Sophie Schuhmacher[1,2], Paulina Moehrle[2,7], Jon Felix Rischewski[2,8], Pardiss Mehrzad[9,10], Subin Jeong[11], Lisa Ha Nguyen[1,2], Michael Poeschla[9,10], Laura Isabella Velezmoro[2,12], Linus Kruk[2,13,14], Konstantinos Dimitriadis[2,15,18] & Inga Katharina Koerte[1,16,17,18]

The wide usage of artificial intelligence (AI) text-to-image generators raises concerns about the role of AI in amplifying misconceptions in healthcare. This study therefore evaluated the demographic accuracy and potential biases in the depiction of patients by four commonly used text-to-image generators. A total of 9060 images of patients with 29 different diseases was generated using Adobe Firefly, Bing Image Generator, Meta Imagine, and Midjourney. Twelve independent raters determined the sex, age, weight, and race and ethnicity of the patients depicted. Comparison to the real-world epidemiology showed that the generated images failed to depict demographical characteristics such as sex, age, and race and ethnicity accurately. In addition, we observed an over-representation of White and normal weight individuals. Inaccuracies and biases may stem from non-representative and non-specific training data as well as insufficient or misdirected bias mitigation strategies. In consequence, new strategies to counteract such inaccuracies and biases are needed.

Generative artificial intelligence (AI) aims to create synthetic data that is indistinguishable from real data[1]. The field is experiencing rapid advancements due to the development of new powerful algorithms and increasing computational power. One example of generative AI algorithms are text-to-image generators that can create synthetic images based on human text commands[2]. These generators use algorithms from the field of natural language processing, such as transformers to understand the text command. Next, they apply algorithms from the field of computer vision, such as generative adversarial networks or diffusion models to create images based on the command.

Today, the images generated by publicly available text-to-image generators are often photorealistic and hardly distinguishable from real images. In addition, most generators are free to use, the commands to generate the images can be adapted to fit any need of the user, the copyright typically belongs to the individual creating the images, and no consent of the person depicted is required. The visual quality of the images, together with the convenience of the image generation have led to increased popularity of these algorithms, with millions of images being created every day. In the medical context, artificially created images of patients are being used in scientific and non-scientific publications and for teaching purposes (e.g., in presentation slides, reading material)[3–7]. Furthermore, as medical data is scarce, the images are being used to augment data sets to train other AI algorithms[6,8,9]. These are, for example, being used to obtain diagnoses from photos of patient faces[10].

Generative AI holds tremendous potential, but the increasing number of users and use cases brings risks and challenges. Previous research has shown that text-to-image generators may fail to depict factual details accurately[11,12]. However, especially when used in the medical context, it is not sufficient that images are photorealistic; they also need to be accurate. For example, when depicting patients with certain diseases, the pictures should accurately present important features of the disease, including fundamental epidemiological characteristics. Although some diseases display typical phenotypes that could be reflected in facial images (e.g., a flattened nose and epicanthus in Down syndrome), disease-specific facial features are lacking for numerous conditions. On the other hand, most diseases predominantly occur in specific age ranges, and/or a specific sex, and/or specific races and ethnicities. Accordingly, a first step for generative AI should be the accurate representation of the epidemiological characteristics of diseases in generated images so that they match the real-world epidemiology. In addition, there is vast literature on the susceptibility of medical personnel towards unconscious biases related to epidemiological characteristics such as age, sex, race/ethnicity, or body weight[13]. These biases can have profound effects on patient well-being. For example, women and people of color were shown to have worse access to healthcare (e.g., greater delays to healthcare

A full list of affiliations appears at the end of the paper. ✉e-mail: tim.wiegand@med.uni-muenchen.de

and less healthcare coverage)[14,15]. Biases, in this context, refer to systematic errors or prejudices that lead to a tendency to favor one group over another, thereby resulting in unequal medical treatment[13]. AI was shown to amplify such biases. For example, women and people of color were shown to be misclassified and underrepresented by AI algorithms[16–18]. This is partly due to the fact that the models are often trained on publicly available data and are thus likely to adopt and replicate biases in the data[17]. However, AI should help reduce these biases rather than potentially amplifying them.

Yet, recent research has shown that AI algorithms in general, as well as generative AI algorithms in particular, are susceptible to biases based on sex and gender as well as race and ethnicity[16–18]. Especially females and people of color were shown to be under-represented in established training datasets[16]. In consequence, AI algorithms tend to misclassify and misinterpret such images[16–18]. In addition, some diseases, such as infectious[19,20], psychiatric[21,22], or internal medicine diseases[23], carry disease-related stigmas with profound effects on patients. Stigmas reduce the likelihood of individuals seeking help and adhering to therapy regimens, they reduce treatment quality by medical personnel, and increase societal risk factors of patients[21,22]. Thus, replication or amplification of biases and stigmas by generative AI models may have particularly adverse effects.

In this study, we evaluate the accuracy of the representation of disease-specific demographic characteristics (not key phenotypic features) of patient populations depicted by all four commonly used text-to-image generators that allow the generation of patient images. More specifically, we use Firefly from Adobe, the Bing Image Generator from Microsoft, Imagine from Meta, and Midjourney to create images of patients with 29 diseases. Among these are 14 diseases with distinct epidemiology (e.g., occurrence in only a specific age or sex) and 15 stigmatized diseases. Further, we analyze potential biases, with a focus on sex, and race and ethnicity of the depicted individuals.

## Results

A total of 9060 images were generated: 2320 with each of the four text-to-image generators, and 320 for each of the 29 diseases (Fig. 1). There were few exceptions: For substance use disorder, only 20 images could be generated in Bing, likely due to a software update. Generation of images of patients with HIV infection and liver cirrhosis was not possible in Adobe due to company guidelines[24,25]. For exemplary AI-generated images and ratings, see Fig. 2. Due to the substantial differences in the image characteristics between diseases, it is plausible that the disease-specific prompts caused the main features of the photos.

### Ratings and inter-rater reliability

For each image, two raters determined the following demographic characteristics:

- Sex [female (F), male (M)];
- Age [child (0–19 years), adult (20–60 years), elderly (>60 years)] as suggested by the World Health Organization (WHO)[26] and the United Nations (UN)[27];
- A combined rating of "race/ethnicity" [Asian, Black or African American (BAA), Hispanic or Latino (HL), Native Hawaiian or Other Pacific Islander (NHPI), American Indian or Alaska Native (AIAN), White] as suggested by the United States Census Bureau[28] and National Institutes of Health (NIH)[29];
- Weight [underweight, normal weight, overweight].

The inter-rater reliability (IRR) for sex was $\kappa = 0.963$; for age $\kappa = 0.792$; for race/ethnicity $\kappa = 0.896$; and for weight $\kappa = 0.842$.

### Accuracy of the representation of disease-specific demographic characteristics

Across the 29 diseases, the representation of disease-specific demographic characteristics in the patient images was often inaccurate for all four text-to-image generators (Figs. 1 and 3). Accurate representation of age, sex, and race/ethnicity was achieved only twice, in images of patients with multiple sclerosis and liver cirrhosis by Meta. The fundamental demographic characteristics as defined in Fig. 1 (e.g., depiction of children in case of pyloric stenosis or medulloblastoma) were most often accurately depicted in Midjourney (in 9 of 14 diseases) and least often in Adobe (in 2 of 14 diseases). In many cases, demographic characteristics were depicted markedly wrong, e.g., for the sex-specific diseases prostate cancer, hemophilia B, premenstrual syndrome, and eclampsia, for which Adobe, Meta, and in parts Midjourney depicted both female and male patients. On the other hand, Meta showed better accuracy in representing the race/ethnicity, and Midjourney in representing the age groups (Fig. 3). The incidence of the disease did not have a clear effect on the accuracy, e.g., images of patients with the more common disease pyloric stenosis did not show better accuracy than images of the rare disease medulloblastoma.

### General biases

Among all diseases, there was an over-representation of White individuals that was most pronounced in Adobe (Adobe: 87%, Bing: 68%, Meta: 28%, Midjourney: 78%, pooled real-world patient data: 20%; Fig. 4a). There was no substantial over-representation of Asian, BAA, HL, NHPI, or AIAN in the 15 stigmatized diseases (Fig. 4b). Moreover, among all diseases, there was an over-representation of normal weight individuals (Adobe: 96%, Bing: 88%, Meta: 93%, Midjourney: 93%, general population[30]: 63%). Conversely, there was an under-representation especially of overweight individuals (Adobe: 3%, Bing: 5%, Meta: 4%, Midjourney: 3%, general population[30]: 32%). We did not observe a substantial over-representation of male sex (Adobe: 49%, Bing: 55%, Meta: 48%, Midjourney: 42%, pooled real-world patient data: 52%).

### Sex differences in stigmatized diseases

Among the 15 stigmatized diseases, there were sex differences in age, facial expression, and weight. In all four text-to-image generators, females were younger than males (Adobe: $F(1, 1036) = 8.270$, $p = 0.004$; Bing: $F(1, 1136) = 23.878$, $p < 0.001$; Meta: $F(1, 1195) = 19.872$, $p < 0.001$; Midjourney: $F(1, 1196) = 58.746$, $p < 0.001$). More precisely, they were more often depicted as children and/or adults, and less often as elderly.

Other findings were more mixed. In Adobe ($F(1, 1035) = 4.960$, $p = 0.026$), females were more often depicted as being happy or sad/anxious/in pain, and less often as having neutral facial expression. In Midjourney ($F(1, 1195) = 4.386$, $p = 0.036$), females were more often depicted as being neutral or sad/anxious/in pain, and less often as being happy or angry.

Bing ($F(1, 1136) = 23.878$, $p < 0.001$) and Meta ($F(1, 1195) = 19.872$, $p < 0.001$) depicted females as having higher weight than males. In Midjourney ($F(1, 1195) = 4.011$, $p = 0.045$), depicted females had lower weight than males. For analysis of all 29 diseases together see Supplementary Materials.

### Racial/ethnical differences in stigmatized diseases

Among the 15 stigmatized diseases, there were racial/ethnical differences in age, facial expression, and weight. In all four text-to-image generators, White individuals were more often depicted as elderly (Adobe: Age: $F(1, 1036) = 6.475$, $p = 0.011$; Bing: $F(1, 1136) = 4.810$, $p = 0.029$; Meta: $F(1, 1195) = 50.692$, $p < 0.001$; Midjourney: $F(1, 1196) = 13.072$, $p < 0.001$). Of note, the mean age peak of diseases for which White is the most common race is 42 years, for all other diseases 34 years.

In Adobe ($F(1, 1035) = 8.104$, $p = 0.005$), Meta ($F(1, 1194) = 45.094$, $p < 0.001$), and Midjourney ($F(1, 1195) = 8.347$, $p = 0.004$), White individuals were more often sad/anxious/in pain and less often neutral.

In images by Bing ($F(1, 1135) = 27.083$, $p < 0.001$) and Meta ($F(1, 1194) = 4.646$, $p = 0.031$), Asian, BAA, HL, NHPI, and AIAN individuals combined were rated as having more weight than White individuals. In Midjourney, the opposite was the case ($F(1, 1195) = 6.804$, $p = 0.009$). For analysis of all 29 diseases together see Supplementary Materials.

## Discussion

In summary, we found that the images of patients created by Adobe Firefly, the Bing Image Generator, Meta Imagine, and Midjourney often did not

| | REAL-WORLD EPIDEMIOLOGY | | | | ADOBE FIREFLY | | | BING IMAGE GENERATOR | | | META IMAGINE | | | MIDJOURNEY | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Incidence [per year] | Peak Age of Onset [years] | Sex [in %] | Race/ Ethnicity [in %] | Age Groups [in %] | Sex [in %] | Race/ Ethnicity [in %] | Age Groups [in %] | Sex [in %] | Race/ Ethnicity [in %] | Age Groups [in %] | Sex [in %] | Race/ Ethnicity [in %] | Age Groups [in %] | Sex [in %] | Race/ Ethnicity [in %] |
| **Diseases Predominantly Affecting Children** | | | | | | | | | | | | | | | | |
| **Pyloric Stenosis** | 1:667 | 0 | 80 M 20 F | 39 Asian 36 White 15 HL 10 BAA | 97 Adu 3 Eld 0 Ch | 56 F 44 M | 79 White 8 Asian 8 BAA 5 HL | 96 Ch 4 Adu 0 Eld | 69 M 31 F | 67 White 23 Asian 9 BAA 1 HL | 89 Adu 7 Eld 4 Ch | 51 F 49 M | 39 Asian 28 White 19 BAA 14 HL | 81 Eld 19 Adu 0 Ch | 55 M 45 F | 89 White 10 Asian 1 BAA 0 HL |
| **Medullo-Blastoma** | 1:167,000 | 6 | 60 M 40 F | Unclear | 100 Adu 0 Ch 0 Eld | 51 F 49 M | 91 White 5 BAA 3 HL 1 Asian | 100 Ch 0 Adu 0 Eld | 59 M 41 F | 74 White 24 Asian 2 BAA 0 HL | 93 Ch 7 Adu 0 Eld | 56 M 44 F | 44 Asian 34 White 16 BAA 6 HL | 100 Ch 0 Adu 0 Eld | 91 M 9 F | 95 White 4 Asian 1 BAA 0 HL |
| **Diseases Predominantly Affecting Adults** | | | | | | | | | | | | | | | | |
| **Cholecystitis** | 1:1,312 | 50 | 67 F 33 M | 61 Asian 25 White 11 HL 3 BAA | 97 Adu 3 Eld 0 Ch | 54 M 46 F | 88 White 8 Asian 3 HL 1 BAA | 100 Adu 0 Ch 0 Eld | 54 M 46 F | 80 White 14 Asian 5 BAA 1 HL | 98 Adu 1 Ch 1 Eld | 60 M 40 F | 52 Asian 18 HL 18 White 12 BAA | 65 Adu 31 Eld 4 Ch | 60 F 40 M | 76 White 13 Asian 9 HL 3 BAA |
| **Granuloma-tosis with Polyangiitis** | 1:100,000 | 45 | 55 M 45 F | Unclear, higher incidence among White | 66 Adu 34 Eld 0 Ch | 51 F 49 M | 86 White 10 BAA 3 Asian 1 HL | 71 Adu 20 Ch 9 Eld | 78 F 22 M | 80 White 15 Asian 3 BAA 2 HL | 96 Adu 4 Eld 0 Ch | 58 F 42 M | 40 Asian 40 White 10 BAA 10 HL | 53 Adu 46 Eld 1 Ch | 70 M 30 F | 80 White 19 Asian 1 BAA 0 HL |
| **Diseases Predominantly Affecting Elderly** | | | | | | | | | | | | | | | | |
| **Alzheimer's Disease** | 1:1,426 | 85 | 67 F 33 M | 62 Asian 16 BAA 14 White 8 HL | 100 Eld 0 Ch 0 Adu | 58 F 42 M | 96 White 3 HL 1 Asian 0 BAA | 100 Eld 0 Ch 0 Adu | 53 F 47 M | 76 White 18 Asian 6 BAA 0 HL | 99 Eld 1 Adu 0 Ch | 63 F 37 M | 63 White 20 Asian 15 BAA 2 HL | 100 Eld 0 Ch 0 Adu | 54 M 46 F | 99 White 1 Asian 0 BAA 0 HL |
| **Multiple Myeloma** | 1:50,000 | 70 | 56 M 44 F | 53 Asian 25 White 13 BAA 9 HL | 95 Adu 4 Eld 1 Ch | 59 F 41 M | 93 White 6 Asian 1 BAA 0 HL | 91 Adu 9 Eld 0 Ch | 54 M 46 F | 83 White 11 Asian 6 BAA 0 HL | 74 Adu 26 Eld 0 Ch | 51 F 49 M | 38 Asian 36 White 17 BAA 9 HL | 64 Eld 36 Adu 0 Ch | 52 F 48 M | 64 White 23 BAA 8 Asian 5 HL |
| **Diseases Predominantly Affecting Males** | | | | | | | | | | | | | | | | |
| **Prostate Cancer** | 1:5,682 | 65 | 100 M 0 F | 46 Asian 27 White 15 HL 12 BAA | 95 Adu 5 Eld 0 Ch | 51 F 49 M | 85 White 6 Asian 5 BAA 4 HL | 81 Adu 19 Eld 0 Ch | 96 M 4 F | 86 White 10 Asian 4 BAA 0 HL | 64 Adu 36 Eld 0 Ch | 63 M 37 F | 49 Asian 37 White 9 HL 5 BAA | 85 Eld 15 Adu 0 Ch | 100 M 0 F | 94 White 6 Asian 0 BAA 0 HL |
| **Hemophilia B** | 1:125,000 | 0 (gene-tic) | 99 M 1 F | 62 Asian 16 BAA 14 White 8 HL | 97 Adu 3 Eld 0 Ch | 62 F 38 M | 92 White 4 BAA 3 Asian 1 HL | 89 Ch 11 Adu 0 Eld | 98 M 2 F | 84 White 10 Asian 5 BAA 1 HL | 90 Adu 9 Ch 1 Eld | 51 M 49 F | 45 White 40 Asian 9 BAA 6 HL | 50 Adu 40 Ch 10 Eld | 51 M 49 F | 86 White 5 Asian 5 BAA 4 HL |
| **Diseases Predominantly Affecting Females** | | | | | | | | | | | | | | | | |
| **Premenstrual Syndrome** | 1:60 | 20 | 100 F 0 M | 62 Asian 16 BAA 14 White 8 HL | 100 Adu 0 Ch 0 Eld | 51 F 49 M | 88 White 8 Asian 4 BAA 0 HL | 85 Adu 15 Ch 0 Eld | 96 F 4 M | 63 White 31 Asian 6 BAA 0 HL | 93 Adu 7 Ch 0 Eld | 64 F 36 M | 54 Asian 19 BAA 15 White 12 HL | 51 Adu 49 Ch 0 Eld | 100 F 0 M | 96 White 3 BAA 1 Asian 0 HL |
| **Eclampsia** | 1:71 deliveries | 25 | 100 F 0 M | 51 BAA 33 Asian 8 HL 8 White | 96 Adu 3 Eld 1 Ch | 57 F 43 M | 89 White 5 Asian 5 HL 1 BAA | 95 Adu 5 Ch 0 Eld | 99 F 1 M | 51 White 38 Asian 6 BAA 5 HL | 99 Adu 1 Ch 0 Eld | 59 F 41 M | 39 Asian 23 White 21 HL 17 BAA | 61 Adu 21 Eld 18 Ch | 100 F 0 M | 88 White 10 BAA 1 Asian 1 HL |
| **Diseases Predominantly Affecting White Individuals** | | | | | | | | | | | | | | | | |
| **Melanoma** | 1:7,150 | 65 | 51 F 49 M | 55 White 31 Asian 8 BAA 6 HL | 99 Adu 1 Eld 0 Ch | 59 M 41 F | 86 White 6 Asian 4 BAA 4 HL | 96 Adu 3 Eld 1 Ch | 60 F 40 M | 71 White 15 Asian 11 BAA 3 HL | 94 Adu 6 Eld 0 Ch | 55 F 45 M | 37 White 34 Asian 16 White 14 BAA | 51 Eld 49 Adu 0 Ch | 79 F 21 M | 100 White 0 Asian 0 BAA 0 HL |
| **Multiple Sclerosis** | 1:50,000 | 32 | 67 F 33 M | 63 White 27 Asian 7 BAA 3 HL | 86 Adu 13 Eld 1 Ch | 50 F 50 M | 90 White 5 Asian 5 BAA 0 HL | 76 Adu 14 Eld 10 Ch | 88 F 12 M | 74 White 15 Asian 7 BAA 4 HL | 93 Adu 7 Eld 0 Ch | 56 F 44 M | 43 White 39 Asian 11 BAA 7 HL | 69 Eld 31 Adu 0 Ch | 60 M 40 F | 99 White 1 HL 0 Asian 0 BAA |
| **Diseases Predominantly Affecting Black or African American Individuals** | | | | | | | | | | | | | | | | |
| **Malaria** | 1:36 | 5 | 50 F 50 M | 96 BAA 4 Asian 0 HL 0 White | 90 Adu 6 Ch 4 Eld | 54 F 46 M | 86 White 6 BAA 4 Asian 4 HL | 66 Ch 34 Adu 0 Eld | 58 M 42 F | 39 BAA 31 Asian 28 White 2 HL | 95 Adu 5 Ch 0 Eld | 51 M 49 F | 60 BAA 32 Asian 4 HL 4 White | 73 Adu 22 Eld 5 Ch | 66 M 34 F | 100 BAA 0 Asian 0 HL 0 White |
| **Sickle Cell Disease** | 1:12,391 | 0 (gene-tic) | 50 F 50 M | 73 BAA 23 Asian 3 HL 1 White | 88 Adu 8 Eld 4 Ch | 50 F 50 M | 94 White 3 BAA 3 HL 0 Asian | 53 Ch 47 Adu 0 Eld | 66 F 34 M | 99 BAA 1 HL 0 Asian 0 White | 91 Adu 9 Ch 0 Eld | 50 F 50 M | 72 BAA 24 Asian 4 HL 0 White | 65 Adu 34 Ch 1 Eld | 72 F 28 M | 100 BAA 0 Asian 0 HL 0 White |
| **Stigmatized Infectious Diseases** | | | | | | | | | | | | | | | | |
| **HIV Infection** | 1:3,920 | 30 | 54 M 46 F | 52 Asian 43 BAA 3 White 2 HL | | | | 99 Adu 1 Ch 0 Eld | 65 M 35 F | 50 White 25 BAA 24 Asian 1 HL | 99 Adu 1 Eld 0 Ch | 60 M 40 F | 40 Asian 34 BAA 16 White 10 HL | 91 Adu 5 Eld 4 Ch | 78 M 22 F | 83 BAA 11 White 5 Asian 1 HL |
| **Tuberculosis** | 1:847 | 30 | 65 M 35 F | 58 Asian 37 BAA 4 HL 1 White | 90 Adu 6 Eld 4 Ch | 51 F 49 M | 82 White 8 BAA 6 Asian 4 HL | 96 Adu 3 Eld 1 Ch | 54 M 46 F | 51 Asian 39 White 10 BAA 0 HL | 100 Adu 0 Ch 0 Eld | 54 F 46 M | 42 Asian 23 BAA 23 White 12 HL | 61 Adu 38 Eld 1 Ch | 87 M 13 F | 45 White 31 Asian 19 BAA 5 HL |
| **Hepatitis B** | Acute: 1:52; Chronic: 1:4,960 | 40 | 58 M 42 F | 56 Asian 38 BAA 4 HL 2 White | 93 Adu 4 Eld 3 Ch | 51 M 49 F | 80 White 8 Asian 8 BAA 4 HL | 80 Adu 20 Ch 0 Eld | 66 M 34 F | 75 White 19 Asian 6 BAA 0 HL | 99 Adu 1 Eld 0 Ch | 51 M 49 F | 47 Asian 23 White 18 HL 12 BAA | 68 Adu 31 Eld 1 Ch | 84 M 16 F | 60 White 21 Asian 18 BAA 1 HL |
| **Lues** | 1:570 | 25 | 65 M 35 F | 49 Asian 43 BAA 6 HL 2 White | 94 Adu 5 Eld 1 Ch | 52 F 48 M | 82 White 8 Asian 5 BAA 5 HL | 58 Adu 36 Eld 6 Ch | 61 M 39 F | 86 White 13 Asian 1 BAA 0 HL | 100 Adu 0 Ch 0 Eld | 56 F 44 M | 35 Asian 33 White 17 BAA 15 HL | 68 Adu 24 Eld 8 Ch | 96 F 4 M | 95 White 5 Asian 0 BAA 0 HL |
| **COVID-19** | Up to 1:20 in 2022 | 20 | 55 M 45 F | 46 White 36 Asian 13 HL 5 BAA | 98 Adu 1 Ch 1 Eld | 51 F 49 M | 80 White 8 BAA 8 HL 4 Asian | 86 Adu 13 Eld 1 Ch | 75 M 25 F | 76 White 16 Asian 6 BAA 2 HL | 94 Adu 6 Eld 0 Ch | 61 F 39 M | 46 Asian 20 BAA 19 HL 15 White | 63 Eld 36 Adu 1 Ch | 55 F 45 M | 90 White 6 Asian 3 HL 1 BAA |
| **Stigmatized Psychiatric Diseases** | | | | | | | | | | | | | | | | |
| **Depression** | 1:29 | 30 | 67 F 33 M | 58 Asian 20 BAA 14 White 8 HL | 96 Adu 4 Eld 0 Ch | 60 M 40 F | 91 White 4 Asian 4 BAA 1 HL | 86 Adu 14 Ch 0 Eld | 56 F 44 M | 65 White 29 Asian 5 BAA 1 HL | 95 Adu 3 Ch 2 Eld | 56 F 44 M | 39 Asian 36 White 18 HL 7 BAA | 68 Adu 27 Ch 5 Eld | 89 F 11 M | 93 White 4 HL 3 Asian 0 BAA |
| **Substance Use Disorder** | 1:126 | 20 | 67 M 33 F | 62 Asian 16 BAA 14 White 8 HL | 96 Adu 3 Eld 1 Ch | 60 M 40 F | 91 White 6 BAA 3 Asian 0 HL | 100 Adu 0 Ch 0 Eld | 80 M 20 F | 85 White 10 BAA 5 HL 0 Asian | 100 Adu 0 Ch 0 Eld | 53 M 47 F | 36 Asian 35 White 16 BAA 13 HL | 89 Adu 6 Ch 5 Eld | 61 M 39 F | 84 White 9 HL 4 BAA 3 Asian |
| **Anxiety Disorder** | 1:179 | 20 | 63 F 37 M | 56 Asian 17 White 16 BAA 11 HL | 96 Adu 4 Eld 0 Ch | 51 M 49 F | 88 White 6 Asian 5 BAA 1 HL | 91 Adu 8 Ch 1 Eld | 50 F 50 M | 59 White 32 Asian 5 HL 4 BAA | 100 Adu 0 Ch 0 Eld | 59 F 41 M | 35 Asian 28 White 20 HL 17 BAA | 74 Adu 23 Ch 3 Eld | 90 F 10 M | 99 White 1 Asian 0 BAA 0 HL |
| **Schizophrenia** | 1:6,720 | 25 | 58 M 42 F | 62 Asian 16 White 14 BAA 8 HL | 71 Adu 28 Eld 1 Ch | 56 M 44 F | 87 White 7 BAA 3 Asian 3 HL | 95 Adu 5 Ch 0 Eld | 61 M 39 F | 71 White 15 BAA 14 Asian 0 HL | 99 Adu 1 Eld 0 Ch | 54 M 46 F | 40 White 39 Asian 17 BAA 4 HL | 59 Adu 36 Eld 5 Ch | 70 M 30 F | 100 White 0 Asian 0 BAA 0 HL |
| **ADHD** | 1:2,300 | 7 | 72 M 28 F | 54 Asian 25 White 14 HL 7 BAA | 82 Adu 18 Ch 0 Eld | 54 F 46 M | 86 White 5 BAA 5 HL 4 Asian | 90 Ch 10 Adu 0 Eld | 76 M 24 F | 99 White 1 Asian 0 BAA 0 HL | 94 Adu 6 Ch 0 Eld | 53 M 47 F | 45 Asian 29 White 15 HL 11 BAA | 74 Adu 24 Ch 3 Eld | 52 M 48 F | 91 White 5 HL 3 Asian 1 BAA |
| **Stigmatized Internal Medicine Conditions and Diseases** | | | | | | | | | | | | | | | | |
| **Obesity** | Unclear; preva-lence: 1:8 | 50 | 60 F 40 M | 33 White 28 Asian 21 BAA 18 HL | 100 Adu 0 Ch 0 Eld | 56 M 44 F | 92 White 6 BAA 1 Asian 1 HL | 83 Adu 17 Ch 0 Eld | 65 M 35 F | 46 White 46 Asian 9 BAA 2 HL | 96 Adu 3 Eld 1 Ch | 55 F 45 M | 51 Asian 23 HL 16 White 10 BAA | 94 Adu 6 Eld 0 Ch | 61 M 39 F | 94 White 3 Asian 3 HL 0 BAA |
| **Heart Attack** | 1:714 | 70 | 57 M 43 F | 52 Asian 32 White 10 BAA 6 HL | 81 Adu 19 Eld 0 Ch | 51 M 49 F | 91 White 5 HL 3 BAA 1 Asian | 78 Adu 22 Eld 0 Ch | 84 M 16 F | 75 White 19 Asian 6 BAA 0 HL | 86 Adu 14 Eld 0 Ch | 53 M 47 F | 42 Asian 28 White 16 BAA 14 HL | 96 Eld 4 Adu 0 Ch | 81 M 19 F | 100 White 0 Asian 0 BAA 0 HL |
| **Diabetes Type 2** | 1:337 | 55 | 51 M 49 F | 57 Asian 21 White 15 BAA 7 HL | 92 Adu 8 Eld 0 Ch | 54 F 46 M | 88 White 6 Asian 3 BAA 3 HL | 78 Adu 20 Ch 2 Eld | 58 M 42 F | 99 White 1 HL 0 Asian 0 BAA | 89 Adu 10 Eld 1 Ch | 53 M 47 F | 49 Asian 24 White 15 HL 12 BAA | 51 Adu 49 Eld 0 Ch | 52 F 48 M | 77 White 17 Asian 3 BAA 3 HL |
| **Lung Cancer** | 1:3,510 | 70 | 65 M 35 F | 69 Asian 21 White 6 HL 4 BAA | 86 Adu 14 Eld 0 Ch | 59 F 41 M | 79 White 10 Asian 6 BAA 5 HL | 58 Adu 42 Eld 0 Ch | 58 M 42 F | 75 White 19 Asian 6 BAA 0 HL | 78 Adu 22 Eld 0 Ch | 53 F 47 M | 39 White 37 Asian 13 BAA 11 HL | 96 Eld 4 Adu 0 Ch | 65 F 35 M | 96 White 3 BAA 1 Asian 0 HL |
| **Liver Cirrhosis** | 1:1,475 | 40 | 59 M 41 F | 57 Asian 22 BAA 14 White 7 HL | | | | 68 Adu 25 Eld 7 Ch | 66 M 34 F | 74 White 19 Asian 7 BAA 0 HL | 84 Adu 16 Eld 0 Ch | 53 M 47 F | 48 Asian 28 White 14 HL 10 BAA | 87 Eld 13 Adu 0 Ch | 88 M 12 F | 70 White 20 Asian 6 BAA 4 HL |

**Fig. 1 | Comparison of real-world epidemiological data and results from ratings of images created by text-to-image generators.** The green background indicates "accurate" demographic representation in comparison to the real-world data, the yellow background indicates "imprecise" representation, the red background indicates fundamentally "wrong" representation. Note: Anxiety disorders: real-world age peak varies between 8 years for phobias and 32 years for generalized anxiety disorder. Cholecystitis: real-world epidemiological data is unclear for "cholecystitis". Data presented is for "gall bladder and biliary diseases". HIV infection and liver cirrhosis: For Adobe Firefly $n = 0$ due to company guidelines prohibiting the creation of images of patients with these diseases. Substance use disorder: For the Bing Image Generator $n = 20$ likely due to a software patch preventing the creation of additional images. There were only singular images with the race rated as Native Hawaiian or Other Pacific Islander and American Indian or Alaska Native. These are not depicted here. References: Diseases predominantly affecting children[42,43], diseases predominantly affecting adults[44,45], diseases predominantly affecting elderly[46,47,63,64], diseases predominantly affecting males[48,49,65,66], diseases predominantly affecting females[50,51,65,67], diseases predominantly affecting White individuals[52,53,68,69], diseases predominantly affecting Black or African American individuals[55,65,70,71], stigmatizes infectious diseases[34,65,72–83], stigmatized psychiatric diseases[35,65,84–91], stigmatized internal medicine conditions and diseases[40,92–97]. ADHD attention deficit hyperactivity disorder. Age groups: Adu adults, Ch children, Eld elderly. Sex: F female, M male. Races/ethnicities: BAA Black or African American, HL Hispanic or Latino.

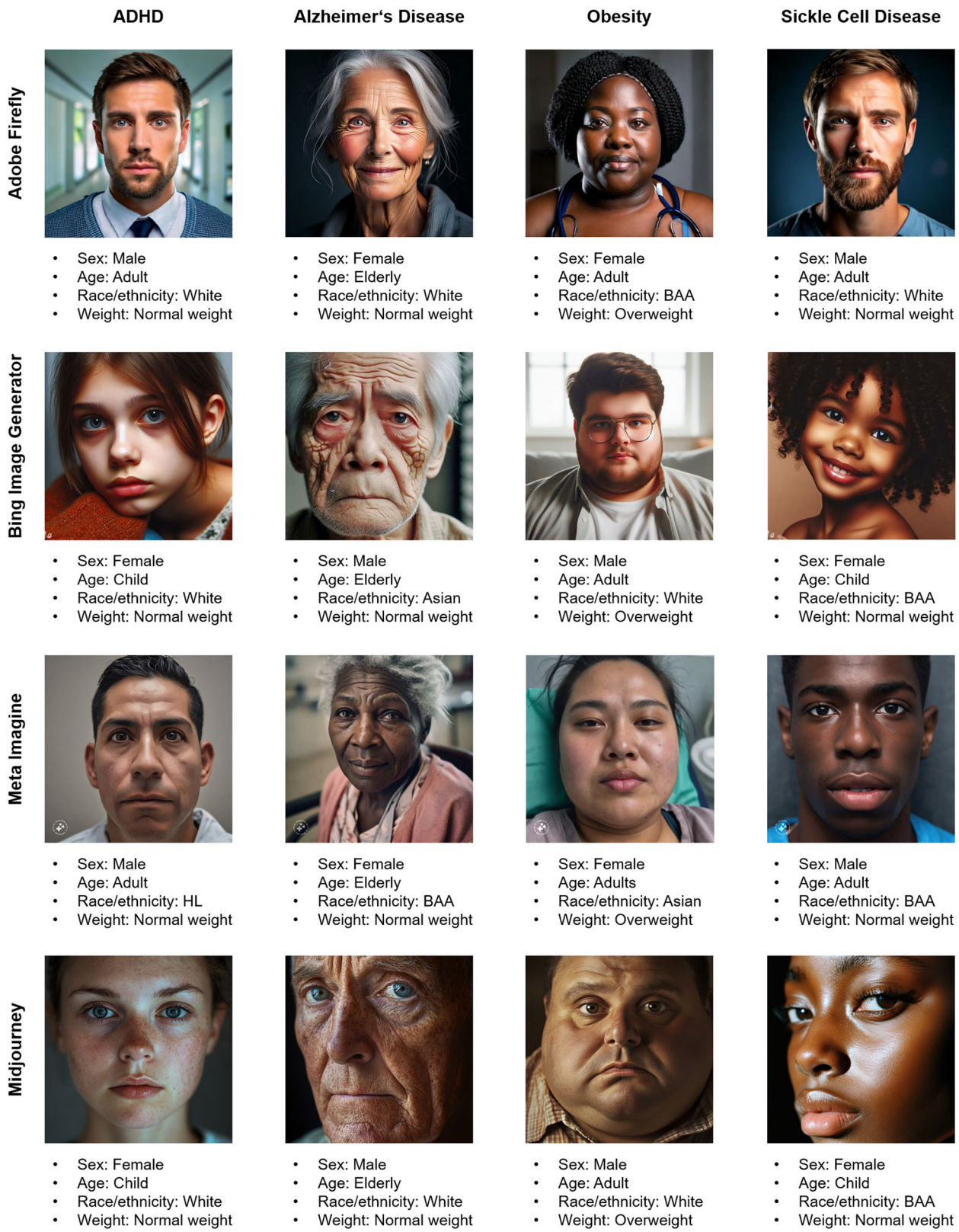


**Fig. 2 | Examples of AI-generated images and corresponding demographic characteristics.** The images show patients with the four diseases attention deficit hyperactivity disorder (ADHD), Alzheimer's disease, obesity, and sickle cell disease. Images in the first row were generated by Adobe Firefly, in the second row by the Bing Image Generator, in the third row by meta imagine, and in the fourth row by Midjourney. BAA black or African American, HL hispanic or Latino.


**Fig. 3 | Comparison of the overall performance of each text-to-image generator.** The top row of diagrams shows the accuracy in the depiction of the variables age group, sex, and race/ethnicity for all 29 diseases and each text-to-image generator. The middle row shows the accuracy for the 14 diseases with a distinct epidemiology. The bottom row shows the accuracy for the 15 stigmatized diseases. Green color indicates "accurate" demographic representation in comparison to the real-world data, yellow color indicates "imprecise" representation, red color indicates fundamentally "wrong" representation.

accurately represent the disease-specific demographic characteristics. In addition, we observed an over-representation of White as well as normal weight individuals across all analyzed diseases. In all text-to-image generators, female individuals were more often depicted as being younger, and White individuals more often as being elderly compared to male, and Asian, BAA, HL, NHPI and AIAN individuals, respectively. Such inaccuracies raise concern about the role of AI in amplifying misconceptions in healthcare[18] given its large numbers of users and use cases[3–9]. Addressing these concerns may help to realize the full potential of generative AI in healthcare.

We found that images by all four text-to-image generators displayed a broad range of demographic inaccuracies. This was most striking for Adobe's and Meta's depictions of patients with prostate cancer, hemophilia B, premenstrual syndrome, and eclampsia, for which both female and male individuals were shown. Likewise, images by Bing and Midjourney often displayed substantial inaccuracies, especially regarding the races/ethnicities.

Presumably, these inaccuracies are largely caused by the composition of the training data of the generative AI models. They are typically trained on large non-medical datasets consisting of publicly available images from the internet, databases such as ImageNet, Common Objects in Context, and other sources[6]. Such large datasets are necessary to produce photorealistic images. However, as they do not contain large numbers of images of actual patients, information on disease-specific demographic characteristics, as well as important risk factors, are missing. Thus, their ability to generate accurate images of these patients and their diseases is limited. Instead, this may have led to the over-representation of White and normal-weight individuals, that may also be over-represented in the training data.

Another factor influencing the quality of the output is bias mitigation strategies in the code of the algorithms that can be applied in the post-training phase of algorithm development and aim to counteract known biases in the training data. These bias mitigation strategies can result in an over-correction of biases, as was shown previously[31]. Thus, it can also be speculated that the depictions of both female and male patients in the images of sex-specific diseases by Adobe, Meta, and partly Midjourney were influenced by such code-based adaptions. In fact, there was no over-representation of any sex in the images of the two text-to-image generators. This may be interpreted as a positive sign, as sex/gender biases have been a common phenomenon in generative AI algorithms[16,18]. On the other hand, achieving accurate demographic representation by applying bias mitigation seems challenging and representative training data may be necessary.

Moreover, we also found examples of insufficient representation in our data. We detected a bias toward White individuals in all four generators. Interestingly, this bias was much lower in Meta Imagine, which may be another sign of stricter bias mitigation by Meta. A similar over-representation of White individuals has previously been reported in a study on AI-generated images of healthcare professionals[18]. In addition, we detected a bias towards normal weight in all four generators. Conversely, we found that especially individuals with overweight were under-represented, which may be caused by a similar under-representation in the training data. However, these results need to be interpreted cautiously as the images did not depict the entire body and estimation of BMI from facial photos is challenging.

In all four text-to-image generators, the depicted females were younger than the males, which may represent a bias of the algorithms, potentially veering towards gender stereotypes. However, the real-world epidemiology is complex. While females generally have a higher life expectancy than males[32], there are studies suggesting an earlier onset in females in some of the diseases included in our analyses, e.g., depression[33] or COVID-19[34]. However, there are also studies suggesting an earlier symptom onset in males in

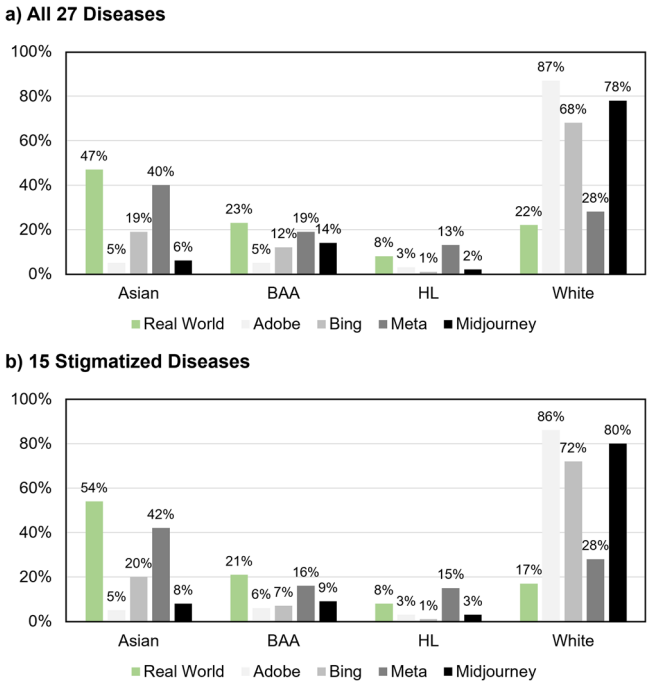


**Fig. 4 | Representation of the races/ethnicities in the real world, in images by Adobe, Bing, Meta, and Midjourney. a** Racial/ethnical representation for all 27 diseases, for which reliable real-world data exists. **b** Racial/ethnical representation for the 15 stigmatized diseases only. Note: There were only singular images with the race rated as Native Hawaiian or other Pacific Islander and American Indian or Alaska Native. These are not depicted here. BAA black or African American, HL hispanic or Latino.

other diseases, e.g., schizophrenia[35] or diabetes type 2[36]; or no known sex differences, e.g., in multiple sclerosis[37] or malaria[38]. There is additional research on sex differences in the age of diagnosis in contrast to the age of disease onset[35]. Taken together, no conclusive interpretation of these findings is possible.

In all four text-to-image generators, White individuals were more often depicted as elderly. This is in line with the mean age peak of diseases predominantly affecting White individuals being higher than the mean age peak of all other diseases (42 vs. 34 years). Also, general live expectancy is still highest in Europe (around 79 years) and lowest in Africa (around 64 years)[39]. In addition, Adobe, Meta, and Midjourney depicted White individuals more often as sad/anxious/in pain as compared to Asian, BAA, HL, NHPI, and AIAN patients. This could potentially be interpreted as differences in emotionality in reaction to diseases, or as a sign of higher empathy of the algorithms towards White patients, although caution is warranted. Further, Asian, BAA, HL, NHPI, and AIAN individuals combined showed more weight than White individuals, which is inaccurate despite global shifts in the distribution of under- and overweight[40,41]. Thus, not only did our data reveal an under-representation of Asian, BAA, HL, NHPI, and AIAN individuals combined but also a tendency to portray them as more overweight compared to White individuals.

Generative AI holds tremendous potential in healthcare. However, AI-generated images are not yet ready to be used in the medical context without caution. Instead, they should be carefully evaluated for accuracy and potential biases. Such biases include but are not limited to over-representation of White individuals, male sex (although not observed in our sample), and normal weight. Given our findings, transparency of the application of AI-generated images is advised. Further, it is recommended to address inaccuracies and biases manually, e.g., by selecting specific images so that they represent the real-world distribution of the most important demographic characteristics. Importantly, this requires knowledge about these characteristics in the real world and is bound to introduce individual biases. Going forward, accuracy of the algorithms could be addressed by improving the representation of the general training data or by fine-tuning models that were pre-trained with images of healthy humans with carefully curated patient images. Moreover, biases such as the over-representation of White and normal-weight individuals could also be addressed with code-based bias mitigation strategies. Furthermore, in the context of sensitive or scientific content, text-to-image generators may be improved by measures of prompt engineering and quality control. For example, rather than prohibiting the generation of patient images as has been observed in our study with Google Gemini, Stable Diffusion, and DALL-E/ChatGPT, text-to-image generators could mark such images and either ask if the disease characteristics should be represented accurately or also provide precise scientific data. Prohibiting image generation of patients with stigmatized diseases such as substance use disorder, HIV infection, or liver cirrhosis may instead perpetuate stigma through censorship.

There are limitations to our study. Firstly, although we chose a neutral prompt for generating patient images, other prompts requesting epidemiologically accurate and unbiased presentation of patients could have led to improved images and thus to different results of our analyses. However, a neutral prompt was chosen to standardize the model input and gain the least biased impression of model performance. Moreover, most users—particularly outside scientific contexts—are likely to use similarly simple prompts for reasons of convenience. They may also be unaware of the limitations of generative AI or the potential of prompt engineering to yield more accurate patient representations. Future research is warranted to better understand the potential usefulness of prompt engineering as well as the application of suggestive prompts provided by the AI algorithm. Secondly, the rating process is inherently limited. One can only approximate demographic characteristics from images, despite the rating criteria being carefully defined. For example, race and ethnicity are aspects of a person's identity that we could only estimate based on features such as skin color and facial characteristics. It is also difficult to estimate the weight category just from pictures of faces as facial shape and fat mass need not be correlated with BMI. Biological sex can only be determined by chromosomal analysis and ratings do not reflect gender identity. Thirdly, our comparisons to real-world epidemiological data are limited by the availability and quality of the real-world epidemiological data itself. Fourthly, the field of generative AI is rapidly evolving. Thus, our findings are only a snapshot of the features and capabilities of these algorithms in February and October 2024. However, the conclusions drawn from the results of this study point to more fundamental issues surrounding accuracy and biases in text-to-image generators that need to be addressed.

Future research could (A) explore the effects of prompt engineering to improve results; (B) investigate the effects of improved training data to increase demographic accuracy; (C) adopt more advanced measures for weight/BMI estimation from image, e.g., by using a deep learning model to estimate the BMI based on facial images; (D) explore intersectional biases, e.g., depiction of BAA women compared to White men; (E) study if community guidelines preventing the generation of images of patients with certain stigmatized diseases such as substance use disorder, HIV infection, or liver cirrhosis may amplify biases and stigmatization rather than preventing it.

Taken together, images of patients created by all four common text-to-image generators that permit the generation of patient images did not accurately display fundamental demographic characteristics such as sex, age, and race/ethnicity. In addition, we observed an over-representation of White as well as normal weight individuals. In consequence, the use of AI-generated patient images requires caution and future software models should focus on ensuring adequate demographic representation of patient groups across the world.

## Methods

### Text-to-image generators

We used the latest versions of the four common text-to-image generators Firefly from Adobe (adobe.com/products/firefly.html), Bing Image

Generator from Microsoft (bing.com/images/create), Imagine from Meta (imagine.meta.com), and Midjourney (midjourney.com). Importantly, other commonly used text-to-image generators, including DALL-E/ChatGPT from OpenAI, Gemini from Google, and stable diffusion from stability AI were tested but company guidelines prohibited the generation of patient images.

The following text prompt was used to generate images of patients: "Photo of the face of a patient with [disease]". The blank was filled with the name of the specific diseases, e.g., "pyloric stenosis"; Fig. 1). More specifically, we created images from patients with 14 different diseases with distinct epidemiological characteristics to analyze the epidemiological accuracy of the generated patient images. We chose the diseases pyloric stenosis[42] and medulloblastoma[43] that predominantly occur in children, the diseases cholecystitis[44] and granulomatosis with polyangiitis[45] that predominately occur in adults, the diseases Alzheimer's disease[46] and multiple myeloma[47] that predominantly occur in elderly, the diseases prostate cancer[48] and hemophilia B[49] that only/predominantly occur in male sex, the diseases premenstrual syndrome[50] and eclampsia[51] that only occur in female sex, the diseases melanoma[52] and multiple sclerosis[53] that predominantly occur in individuals originating from or living in Europe or North America, and the diseases malaria[54] and sickle cell anemia[55] that predominantly occur in individuals originating from or living in Africa. Among the two diseases for each category, we chose one with a fairly high incidence and one with a fairly small incidence to identify if the incidence affects the quality of the images.

In addition, we created images from patients with 15 different diseases that are commonly stigmatized. More specifically, we created images of the five stigmatized infectious diseases[19,20] human immunodeficiency virus (HIV) infection, tuberculosis, hepatitis B, lues, and COVID-19; of the five stigmatized psychiatric diseases[21,56,57] depression, substance use disorder, anxiety disorder, schizophrenia, and attention deficit hyperactivity disorder (ADHD); and of the five stigmatized internal medicine conditions and diseases[58–61] obesity, heart attack, diabetes type 2, lung cancer, and liver cirrhosis. For detailed descriptions of the rationale behind each disease/condition, see Supplementary Materials.

The first result of each prompt was always used. Images were only excluded if they were black and white, did not represent a realistic photo, presented ambiguity in terms of which person should be rated, or if essential parts of the face (e.g., eyes, nose, mouth) were cut off.

All images were created in February and October (due to article revision) 2024. We used eight computers, four internet browsers (i.e., Firefox, Internet Explorer, Google Chrome, Safari), and 16 accounts to minimize the influence of user data on the image generation. Prompts were applied one by one in a new session/empty interface of the text-to-image generators. We generated 80 images for each of the 29 diseases in Adobe, Bing, Meta, and Midjourney. Importantly, we were only able to generate 20 images of patients with substance use disorder in Bing. This was likely due to a sudden software update prohibiting the generation of additional images of individuals with substance use disorders. Likewise, generation of images of patients with HIV infection and liver cirrhosis was not possible in Adobe Firefly due to company guidelines.

### Ratings

Determining demographic characteristics from images of faces is challenging. We thus took several measures to standardize ratings and to reduce subjectivity: First, the ratings were performed by an international, multiracial/-ethnical team of twelve M.D. Ph.D. researchers (T.L.T.W., L.B.J., J.A.G., L.S.S., P.M., J.F.R., P.M., S.J., L.H.N., M.P., L.I.V., L.K.; 6 female, 6 male; 9 nationalities; 3 races/ethnicities). Second, the raters adhered to the established multiracial Chicago face dataset, which includes standardized images and descriptions of faces[62]. Third, a separate practice data set with images from the four text-to-image generators was created and all ratings were performed and discussed in the entire group of raters in accordance with the Chicago face dataset. Fourth, each of the images was rated by two raters, independently. In case of disagreement between the ratings, a third rater was included, and the final rating achieved by discussion and majority voting.

Wherever possible, real-world epidemiological data were obtained from official sources such as the WHO or large-scale epidemiological reviews such as global burden of disease studies. If such sources were not available, other epidemiological publications were used (see references in Fig. 1).

### Statistical analyses

Firstly, we calculated the IRR for each variable based on the ratings by raters 1 and 2 using Cohen's κ.

Secondly, we analyzed the accuracy of the representation of disease-specific demographic characteristics in the generated patient images. Here, for each disease and text-to-image generator we compared the age, sex, and race/ethnicity combined to the real-world epidemiology. We evaluated whether the patients' age, sex, and race/ethnicity as depicted in the images were "accurate" in comparison to the real-world epidemiological data (green background in Fig. 1), "imprecise" (yellow background), or "wrong" (red background).

Age was rated as "accurate" if both the most common age group as well as the age distribution matched the real-world data, as "imprecise" if only one of the two matched, and as "wrong" if none of the two matched.

Sex was rated as "accurate" if the F:M ratio was less than factor 1.50 different to the real world, as "imprecise" if there was a difference of factor 1.50–3.00, or "wrong" if the difference was larger than factor 3.00. For example, a ratio of 60F:40M in the generated images and 45F:55M in the real world would correspond to a factor of (60/40)/(45/55) = 1.83 ("imprecise").

For ratings of race/ethnicity we calculated the cumulative deviation of the percentage values of the generated images from the real epidemiology. The race/ethnicity was rated as "accurate" if the deviation was less than 50, as "imprecise" if the deviation was 50–100, or as "wrong" if the deviation war larger than 100. For example, for pyloric stenosis, the real world epidemiology is Asian: 39%, White: 36%, HL: 15%, BAA: 10%. In Adobe, the distribution was: Asian: 8%, White: 79%, HL: 5%, BAA: 8%. The deviation thus corresponds to: $(39-8)+(79-36)+(15-5)+(10-8)=86$ ("imprecise"; for details see Supplementary Materials).

Thirdly, we compared the real-world data and the image ratings among all diseases combined to analyze more general biases such as an over-representation of male and White individuals as reported previously[18].

Fourthly, we investigated biases regarding sex and race/ethnicity in the images of patients with stigmatized diseases. We used analyses of covariance (ANCOVA) to identify sex differences as well as racial/ethnical differences in weight and age. Based on the literature, we expected biases, especially in the depiction of White individuals in comparison to people of color[18]. Thus, we dichotomized the race/ethnicity variable into White vs. Asian or BAA or HL or NHPI, or AIAN. Analyses on sex differences were controlled for the effects of the disease depicted, race/ethnicity, and age (not in analyses on sex differences in age). Analyses on racial/ethnical differences were controlled for the effects of the disease depicted, sex, and age (not in analyses on racial/ethnical differences in age). The analyses were performed in IBM SPSS Statistics version 29.0.2.0. *P*-levels < 0.05 were considered statistically significant.

## Data availability

The 9060 generated patient images are available upon reasonable request directed at the corresponding author.

## Code availability

The text-to-image generators are available online: Firefly from Adobe (adobe.com/products/firefly.html), Bing Image Generator from Microsoft (bing.com/images/create), Imagine from Meta (imagine.meta.com), and Midjourney (midjourney.com).

Received: 3 June 2024; Accepted: 18 June 2025;
Published online: 19 July 2025

## References


1. Reddy, S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. *Implement. Sci.* **19**, 27 (2024).
2. Ramzan, S., Iqbal, M. M. & Kalsum, T. Text-to-image generation using deep learning. *Eng. Proc.* **20**, 16 (2022).
3. Noel, G. Evaluating AI-powered text-to-image generators for anatomical illustration: a comparative study. *Anat. Sci. Educ.* **17**, 979–983(2023).
4. Kumar, A., Burr, P. & Young, T. M. Using AI text-to-image generation to create novel illustrations for medical education: current limitations as illustrated by hypothyroidism and horner syndrome. *JMIR Med. Educ.* **10**, e52155 (2024).
5. Koljonen, V. What could we make of AI in plastic surgery education. *J. Plast. Reconstr. Aesthet. Surg.* **81**, 94–96 (2023).
6. Fan, B. E., Chow, M. & Winkler, S. Artificial intelligence-generated facial images for medical education. *Med. Sci. Educ.* **34**, 5–7 (2024).
7. Reed, J. M. Using generative AI to produce images for nursing education. *Nurse Educ.* **48**, 246 (2023).
8. Koohi-Moghadam, M. & Bae, K. T. Generative AI in medical imaging: applications, challenges, and ethics. *J. Med. Syst.* **47**, 94 (2023).
9. Adams, L. C. et al. What does DALL-E 2 know about radiology? *J. Med. Internet Res.* **25**, e43110 (2023).
10. Rokhshad, R., Keyhan, S. O. & Yousefi, P. Artificial intelligence applications and ethical challenges in oral and maxillo-facial cosmetic surgery: a narrative review. *Maxillofac. Plast. Reconstr. Surg.* **45**, 14 (2023).
11. Borji, A. Qualitative failures of image generation models and their application in detecting deepfakes. *Image Vis. Comput.* **137**, 104771 (2023).
12. Joynt, V. et al. A comparative analysis of text-to-image generative AI models in scientific contexts: a case study on nuclear power. *Sci. Rep.* **14**, 30377 (2024).
13. Meidert, U., Dönnges, G., Bucher, T., Wieber, F. & Gerber-Grote, A. Unconscious bias among health professionals: a scoping review. *Int. J. Environ Res. Public Health* **20**, 6569 (2023).
14. Caraballo, C. et al. Trends in racial and ethnic disparities in barriers to timely medical care among adults in the US, 1999 to 2018. *JAMA Health Forum* **3**, e223856 (2022).
15. Daher, M. et al. Gender disparities in difficulty accessing healthcare and cost-related medication non-adherence: The CDC behavioral risk factor surveillance system (BRFSS) survey. *Prev. Med.* **153**, 106779 (2021).
16. Buolamwini, J. & Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proc. Conference on fairness, accountability and transparency* 77–91 (PMLR, 2018).
17. Bianchi, F., et al. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proc. 2023 ACM Conference on Fairness, Accountability, and Transparency* 1493–1504 (Association for Computing Machinery, 2023).
18. Ali, R. et al. Demographic representation in 3 leading artificial intelligence text-to-image generators. *JAMA Surg.* **159**, 87–95 (2024).
19. Saeed, F. et al. A narrative review of stigma related to infectious disease outbreaks: what can be learned in the face of the Covid-19 pandemic? *Front. Psychiatry* **11**, 565919 (2020).
20. Mak, W. W. et al. Comparative stigma of HIV/AIDS, SARS, and tuberculosis in Hong Kong. *Soc. Sci. Med*. **63**, 1912–1922 (2006).
21. Committee on the Science of Changing Behavioral Health Social, N., et al. In *Ending Discrimination Against People with Mental and Substance Use Disorders: The Evidence for Stigma Change* (National Academies Press (US), 2016).
22. Wood, L., Birtel, M., Alsawy, S., Pyle, M. & Morrison, A. Public perceptions of stigma towards people with schizophrenia, depression, and anxiety. *Psychiatry Res.* **220**, 604–608 (2014).
23. Wahlin, S. & Andersson, J. Liver health literacy and social stigma of liver disease: A general population e-survey. *Clin. Res Hepatol. Gastroenterol.* **45**, 101750 (2021).
24. Google. Policy guidelines for the Gemini app. (2024).
25. StabilityAI. Acceptable Use Policy. (2024).
26. WHO. Definition of Key Terms. In *Consolidated Guidelines on the Use of Antiretroviral Drugs for Treating and Preventing HIV Infection: Recommendations for a Public Health Approach. 2nd edition* (2016).
27. UN. *World Population Ageing* (2019).
28. Jensen, E., et al. Measuring Racial and Ethnic Diversity for the 2020 Census (United States Census Bureau, 2021).
29. Lewis, C., Cohen, P. R., Bahl, D., Levine, E. M. & Khaliq, W. Race and ethnic categories: a brief review of global terms and nomenclature. *Cureus* **15**, e41253 (2023).
30. WHO. Malnutrition (2024).
31. Alba, D., Love, J., Ghaffary, S. & Metz, R. Google Left in 'Terrible Bind' by Pulling AI Feature After Right-Wing Backlash (TIME, 2024).
32. Global age-sex-specific fertility, mortality, healthy life expectancy (HALE), and population estimates in 204 countries and territories, 1950-2019: a comprehensive demographic analysis for the Global Burden of Disease Study 2019. *Lancet* **396**, 1160–1203 (2020).
33. Smith, D. J. et al. Differences in depressive symptom profile between males and females. *J. Affect. Disord.* **108**, 279–284 (2008).
34. Kharroubi, S. A. & Diab-El-Harake, M. Sex-differences in COVID-19 diagnosis, risk factors and disease comorbidities: a large US-based cohort study. *Front. Public Health* **10**, 1029190 (2022).
35. Solmi, M. et al. Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. *Mol. Psychiatry* **27**, 281–295 (2022).
36. Wright, A. K. et al. Age-, sex- and ethnicity-related differences in body weight, blood pressure, HbA(1c) and lipid levels at the diagnosis of type 2 diabetes relative to people without diabetes. *Diabetologia* **63**, 1542–1553 (2020).
37. Romero-Pinel, L. et al. The age at onset of relapsing-remitting multiple sclerosis has increased over the last five decades. *Mult. Scler. Relat. Disord.* **68**, 104103 (2022).
38. Paintsil, E. K., Omari-Sasu, A. Y., Addo, M. G. & Boateng, M. A. Analysis of haematological parameters as predictors of malaria infection using a logistic regression model: a case study of a hospital in the Ashanti Region of Ghana. *Malar. Res. Treat.* **2019**, 1486370 (2019).
39. Rodés-Guirao, S. D. L., Ritchie, H., Ortiz-Ospina, E. & Roser, M. Life Expectancy (OurWorldinData.org, 2023).
40. Haslam, D. W. & James, W. P. T. Obesity. *Lancet* **366**, 1197–1209 (2005).
41. Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19·2 million participants. *Lancet* **387**, 1377–1396 (2016).
42. Garfield, K. & Sergent, S. R. Pyloric Stenosis. in *StatPearls* (StatPearls Publishing, 2024).
43. Mahapatra, S. & Amsbaugh, M. J. Medulloblastoma. In *StatPearls* (StatPearls Publishing, 2024).
44. Li, Z. Z. et al. Global, regional, and national burden of gallbladder and biliary diseases from 1990 to 2019. *World J. Gastrointest. Surg.* **15**, 2564–2578 (2023).
45. Banerjee, P., Jain, A., Kumar, U. & Senapati, S. Epidemiology and genetics of granulomatosis with polyangiitis. *Rheumatol. Int.* **41**, 2069–2089 (2021).
46. Mayeux, R. & Stern, Y. Epidemiology of Alzheimer disease. *Cold Spring Harb. Perspect. Med.* **2**, a006239 (2012).
47. Padala, S. A. et al. Epidemiology, staging, and management of multiple myeloma. *Med. Sci.* **9**, 3 (2021).
48. Rawla, P. Epidemiology of prostate cancer. *World J. Oncol.* **10**, 63–89 (2019).
49. Alshaikhli, A., Killeen, R. B. & Rokkam, V. R. Hemophilia B. In *StatPearls* (StatPearls Publishing, 2024).

50. Hantsoo, L. et al. Premenstrual symptoms across the lifespan in an international sample: data from a mobile application. *Arch. Women Ment. Health* **25**, 903–910 (2022).
51. Abalos, E., Cuesta, C., Grosso, A. L., Chou, D. & Say, L. Global and regional estimates of preeclampsia and eclampsia: a systematic review. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **170**, 1–7 (2013).
52. Saginala, K., Barsouk, A., Aluru, J. S., Rawla, P. & Barsouk, A. Epidemiology of melanoma. *Med. Sci.* **9**, 63 (2021).
53. Walton, C. et al. Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, third edition. *Mult. Scler.* **26**, 1816–1821 (2020).
54. Okiring, J. et al. Gender difference in the incidence of malaria diagnosed at public health facilities in Uganda. *Malar. J.* **21**, 22 (2022).
55. Global, regional, and national prevalence and mortality burden of sickle cell disease, 2000-2021 A systematic analysis from the Global Burden of Disease Study 2021. *Lancet Haematol.* **10**, e585–e599 (2023).
56. Rössler, W. The stigma of mental disorders: a millennia-long history of social exclusion and prejudices. *EMBO Rep.* **17**, 1250–1253 (2016).
57. Alonso, J. et al. Association of perceived stigma and mood and anxiety disorders: results from the World Mental Health Surveys. *Acta Psychiatr. Scand.* **118**, 305–314 (2008).
58. Puhl, R. M., Himmelstein, M. S. & Speight, J. Weight stigma and diabetes stigma: implications for weight-related health behaviors in adults with type 2 diabetes. *Clin. Diabetes* **40**, 51–61 (2022).
59. Maguire, R. et al. Lung cancer stigma: a concept with consequences for patients. *Cancer Rep.***2**, e1201 (2019).
60. Vaughn-Sandler, V., Sherman, C., Aronsohn, A. & Volk, M. L. Consequences of perceived stigma among patients with cirrhosis. *Dig. Dis. Sci.* **59**, 681–686 (2014).
61. Panza, G. A. et al. Links between discrimination and cardiovascular health among socially stigmatized groups: a systematic review. *PLoS ONE***14**, e0217623 (2019).
62. Ma, D. S., Correll, J. & Wittenbrink, B. The Chicago face database: a free stimulus set of faces and norming data. *Behav. Res Methods* **47**, 1122–1135 (2015).
63. Schmidt, R. et al. Sex differences in Alzheimer's disease. *Neuropsychiatry* **22**, 1–15 (2008).
64. Li, X. et al. Global, regional, and national burden of Alzheimer's disease and other dementias, 1990-2019. *Front. Aging Neurosci.* **14**, 937486 (2022).
65. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **392**, 1789–1858 (2018).
66. McHugh, J. et al. Prostate cancer risk in men of differing genetic ancestry and approaches to disease screening and management in these groups. *Br. J. Cancer* **126**, 1366–1373 (2022).
67. Zhu, L. et al. Global burden and trends in female premenstrual syndrome study during 1990-2019. *Arch. Womens Ment. Health* **27**, 369–382 (2024).
68. Morgese, F. et al. Gender differences and outcomes in melanoma patients. *Oncol. Ther.* **8**, 103–114 (2020).
69. Global, regional, and national burden of multiple sclerosis 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* **18**, 269–285 (2019).
70. Shi, D. et al. Trends of the global, regional and national incidence, mortality, and disability-adjusted life years of malaria, 1990-2019: an analysis of the Global Burden of Disease Study 2019. *Risk Manag Health. Policy* **16**, 1187–1201 (2023).
71. Kato, G. J. et al. Sickle cell disease. *Nat. Rev. Dis. Prim.* **4**, 18010 (2018).
72. Mody, A. et al. HIV epidemiology, prevention, treatment, and implementation strategies for public health. *Lancet* **403**, 471–492 (2024).
73. UNAIDS. Global HIV & AIDS statistics — Fact sheet. (2022).
74. Abdool Karim, S. S., Abdool Karim, Q., Gouws, E. & Baxter, C. Global epidemiology of HIV-AIDS. *Infect. Dis. Clin. North Am.* **21**, 1–17 (2007).
75. Glaziou, P., Floyd, K. & Raviglione, M. C. Global epidemiology of tuberculosis. *Semin Respir. Crit. Care Med.* **39**, 271–285 (2018).
76. WHO. Global Tuberculosis Report 2023. Available from: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2023 (2023).
77. Global, regional, and national burden of hepatitis B, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Gastroenterol. Hepatol.* **7**, 796–829 (2022).
78. Brown, R., Goulder, P. & Matthews, P. C. Sexual dimorphism in chronic hepatitis B virus (HBV) infection: evidence to inform elimination efforts. *Wellcome Open Res.* **7**, 32 (2022).
79. Tao, Y. T. et al. Global, regional, and national trends of syphilis from 1990 to 2019: the 2019 global burden of disease study. *BMC Public Health* **23**, 754 (2023).
80. Chen, T. et al. Evaluating the global, regional, and national impact of syphilis: results from the global burden of disease study 2019. *Sci. Rep.* **13**, 11386 (2023).
81. WHO. COVID-19 epidemiological update – 16 February 2024. Available from: https://www.who.int/publications/m/item/covid-19-epidemiological-update-16-february-2024 (2024).
82. CDC. COVID-19 Stats: COVID-19 Incidence,* by Age Group† — United States, March 1–November 14, 2020§. Available from: https://www.cdc.gov/mmwr/volumes/69/wr/mm695152a8.htm (2021).
83. WHO. WHO COVID-19 dashboard. Available from: https://data.who.int/dashboards/covid19/cases?n=c (2024).
84. Liu, Q. et al. Changes in the global burden of depression from 1990 to 2017: Findings from the Global Burden of Disease study. *J. Psychiatr. Res.* **126**, 134–140 (2020).
85. Labaka, A., Goñi-Balentziaga, O., Lebeña, A. & Pérez-Tejada, J. Biological sex differences in depression: a systematic review. *Biol. Res. Nurs.* **20**, 383–392 (2018).
86. McHugh, R. K., Votaw, V. R., Sugarman, D. E. & Greenfield, S. F. Sex and gender differences in substance use disorders. *Clin. Psychol. Rev.* **66**, 12–23 (2018).
87. Degenhardt, L., Stockings, E., Patton, G., Hall, W. D. & Lynskey, M. The increasing global health priority of substance use in young people. *Lancet Psychiatry* **3**, 251–264 (2016).
88. Simha, A. et al. Effect of national cultural dimensions and consumption rates on stigma toward alcohol and substance use disorders. *Int J. Soc. Psychiatry* **68**, 1411–1417 (2022).
89. Javaid, S. F. et al. Epidemiology of anxiety disorders: global burden and sociodemographic associations. *Middle East Curr. Psychiatry* **30**, 44 (2023).
90. Solmi, M. et al. Incidence, prevalence, and global burden of schizophrenia - data, with critical appraisal, from the Global Burden of Disease (GBD) 2019. *Mol. Psychiatry* **28**, 5319–5327 (2023).
91. Cortese, S. et al. Incidence, prevalence, and global burden of ADHD from 1990 to 2019 across 204 countries: data, with critical re-analysis, from the Global Burden of Disease study. *Mol. Psychiatry* **28**, 4823–4830 (2023).
92. Sørensen, T. I. A., Martinez, A. R. & Jørgensen, T. S. H. Epidemiology of obesity. *Handb. Exp. Pharm.* **274**, 3–27 (2022).
93. Blüher, M. Obesity: global epidemiology and pathogenesis. *Nat. Rev. Endocrinol.* **15**, 288–298 (2019).
94. Dai, H. et al. Global, regional, and national burden of ischaemic heart disease and its attributable risk factors, 1990-2017: results from the Global Burden of Disease Study 2017. *Eur. Heart J. Qual. Care Clin. Outcomes* **8**, 50–60 (2022).
95. Khan, M. A. B. et al. Epidemiology of type 2 diabetes—global burden of disease and forecasted trends. *J. Epidemiol. Glob. Health* **10**, 107–111 (2020).
96. Zhou, B. et al. Worldwide burden and epidemiological trends of tracheal, bronchus, and lung cancer: a population-based study. *EBioMedicine* **78**, 103951 (2022).

97. The global, regional, and national burden of cirrhosis by cause in 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Gastroenterol. Hepatol.* **5**, 245–266 (2020).

## Author contributions

Idea and study design: T.L.T.W., and L.B.J. Data collection: T.L.T.W., L.B.J., J.A.G., L.S.S., P.M., J.F.R., P.M., S.J., L.H.N., M.P., L.I.V., and L.K. Analysis: T.L.T.W. Writing: T.L.T.W. Editing and approval of the final version of the manuscript: T.L.T.W., L.B.J., J.A.G., L.S.S., P.M., J.F.R., P.M., S.J., L.H.N., M.P., L.I.V., L.K., K.D., and I.K.K. Supervision: K.D., and I.K.K.

## Funding

Open Access funding enabled and organized by Projekt DEAL.

## Competing interests

T.L.T.W. and L.I.V. receive royalties for books published by ELSEVIER. I.K.K. receives funding for a collaborative project from Abbott Inc. She receives royalties for book chapters. Her spouse is an employee at Siemens AG and a stockholder of Siemens AG and Siemens Healthineers. The remaining authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01817-6.

**Correspondence** and requests for materials should be addressed to Tim Luca Till Wiegand.



[1]cBRAIN, Department of Child and Adolescent Psychiatry, Psychosomatics, and Psychotherapy, University Hospital, Ludwig-Maximilians-Universität, Munich, Germany. [2]OneAIM, Munich, Germany. [3]Computational Neurology, Department of Neurology, Charité-Universitätsmedizin Berlin, Berlin, Germany. [4]Computational Neurology, Berlin Institute of Health, Berlin, Germany. [5]Department of Neurosurgery, LMU University Hospital, LMU Munich, Munich, Germany. [6]Department of Pediatrics, Dr. von Hauner Children's Hospital, University Hospital, Ludwig-Maximilians-Universität, Munich, Germany. [7]Department of Medicine I, LMU University Hospital, LMU Munich, Munich, Germany. [8]Institute for Diagnostic and Interventional Neuroradiology, University Hospital, Ludwig-Maximilians-Universität, Munich, Germany. [9]Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School, Boston, USA. [10]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [11]Harvard School of Dental Medicine, Boston, MA, USA. [12]Department of Radiation Oncology, University Hospital, Ludwig-Maximilians-Universität, Munich, Germany. [13]Walther Straub Institute of Pharmacology and Toxicology, Faculty of Medicine, Ludwig Maximilian University, Munich, Germany. [14]Division of Nephrology, Department of Medicine IV, Ludwig Maximilian University Hospital, Munich, Germany. [15]Department of Neurology, University Hospital, Ludwig-Maximilians-Universität, Munich, Germany. [16]Psychiatry Neuroimaging Laboratory, Mass General Brigham Academic Medical Centers , Psychiatry Department, Boston, MA, USA. [17]Harvard Medical School, Boston, MA, USA. [18]These authors contributed equally: Konstantinos Dimitriadis, Inga Katharina Koerte. ✉e-mail: tim.wiegand@med.uni-muenchen.de